


## Article

# Integrating Multi-Omics Using Bayesian Ridge Regression with Iterative Similarity Bagging

Talal Morizig Almutiri <sup>1,\*</sup>, Khalid Hamad Alomar <sup>1</sup> and Nofe Ateq Alganmi <sup>2</sup> 

<sup>1</sup> Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; khalomar@kau.edu.sa

<sup>2</sup> Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; nalghanimi@kau.edu.sa

\* Correspondence: tnajialmutiri@stu.kau.edu.sa

**Abstract:** Cancer research has increasingly utilized multi-omics analysis in recent decades to obtain biomolecular information from multiple layers, thereby gaining a better understanding of complex biological systems. However, the curse of dimensionality is one of the most significant challenges when handling omics or biological data. Additionally, integrating multi-omics by transforming different omics types into a new representation can reduce a model's interpretability, as the extracted features may lose the biological context. This paper proposes Iterative Similarity Bagging (ISB), assisted by Bayesian Ridge Regression (BRR). BRR serves as a domain-oriented supervised feature selection method, choosing essential features by calculating the coefficients for each feature. Despite this, the BRR output datasets contain many features, leading to complexity and high dimensionality. To address this, ISB was introduced to dynamically reduce dimensionality and complexity without losing the biological integrity of the omics data, which often occurs with transformation-based integration approaches. The evaluation measures employed were Root Mean Square Error (RMSE), the Pearson Correlation Coefficient (PCC), and the coefficient of determination ( $R^2$ ). The results demonstrate that the proposed method outperforms some current models in terms of regression performance, achieving an RMSE of 0.12, a PCC of 0.879, and an  $R^2$  of 0.77 for the CCLE. For the GDSC, it achieved an RMSE of 0.029, a PCC of 0.90, and an  $R^2$  of 0.80.

**Keywords:** anti-cancer; Bayesian Ridge Regression; deep learning; drug response prediction; multi-omics integration



**Citation:** Almutiri, T.M.; Alomar, K.H.; Alganmi, N.A. Integrating Multi-Omics Using Bayesian Ridge Regression with Iterative Similarity Bagging. *Appl. Sci.* **2024**, *14*, 5660. <https://doi.org/10.3390/app14135660>

Academic Editor: Andrea Prati

Received: 6 May 2024

Revised: 20 June 2024

Accepted: 25 June 2024

Published: 28 June 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In recent years, the discipline of systems biology has seen a significant increase in the application of multi-omics techniques to the study of complex biological systems [1]. By combining data from genomics, transcriptomics, proteomics, metabolomics, and other omics technologies, researchers can better understand biological processes and disease mechanisms [2,3]. This integrative approach enables the discovery of critical molecular players and pathways that may be overlooked when evaluating distinct omics datasets separately [3,4]. Omics is a field of molecular biology that aims to comprehensively analyze and measure the genome, transcriptome, and proteome to understand and influence a biological entity's function, structure, and dynamics [5,6]. Each category of omics data corresponds to a distinct layer of biological information, including genomics, epigenomics, transcriptomics, proteomics, and metabolomics [5]. These various data types offer a complementary medical perspective on a biological system or an individual [7]. Advancements in biotechnology have enabled scientists to compile extensive molecular databases and conduct investigations that are either independent or integrative in nature, spanning several domains like genomics, transcriptomics, and proteomics [6].

Drug response prediction is central to the vision of personalized medicine, which aims to individualize treatments based on an individual's unique biological constitution. These

predictions are significantly enhanced by multi-omics integration methods, which pool information across genomics, transcriptomics, proteomics, metabolomics, and other omics layers [8]. Utilizing diverse molecular data provides a robust basis for associating drug efficacy and toxicity with various biological processes and pathways. This comprehensive approach can identify new biomarkers that more accurately predict patient responses compared to single-omics analysis [1]. For example, multi-omics analysis has revealed complex gene–environment interactions and epigenetic alterations that affect drug metabolism and action. Consequently, multi-omics integration has become fundamental in developing personalized therapeutic strategies, minimizing unnecessary healthcare expenses, and improving patient outcomes [1].

Progress in cancer treatment using single-omics datasets, such as those produced by the Human Genome Project and early genomic profiling from the Cancer Genome Atlas (TCGA) studies [9], has been insufficient [6]. Multi-omics analysis, which has become increasingly important in cancer research in recent decades, is the only method to understand cancer behavior thoroughly and uncover previously unknown treatment vulnerabilities [1]. Integrating multi-omics data, which provide information on biomolecules from several layers, holds great promise for systematically and comprehensively understanding complicated biology [1]. Integrated techniques combine individual omics data either sequentially or simultaneously to elucidate the interactions between molecules [8]. The integration of multi-omics can be classified into three main approaches [6,7]: early integration (also called concatenation-based), mixed or middle integration (transformation-based), and late integration (model-based) [6,7]. Early integration involves combining all datasets into one large table or matrix, which leads to increased complexity, noise, and high dimensionality, making the learning process more difficult [7]. Additionally, varying sizes of omics datasets can lead to an imbalance in learning, as the algorithm may focus more on the omics with a greater number of variables and overlook the others [10]. The concatenation-based method is simple, easy to implement, and is performed early, before learning [7]. Most studies utilized feature selection methods to reduce the dimensionality and complexity associated with the early integration approach [11,12].

The middle integration approach reduces the complexity of omics by transforming them into a simple or low-dimensional representation [6]. Each dataset is independently transformed into a new representation and then combined with others before the learning process [7]. This approach employs various techniques, such as graph-embedding [13,14], network-based methods [15–17], and kernel-based methods [18–20]. Graph embedding focuses on modeling each dataset using a graph by learning a low-dimensional representation of the nodes and their relationships [7]. The network-based approach transforms each dataset and fuses them into a homogeneous network [12]. The kernel-based approach involves creating a kernel to integrate multiple omics blocks for each omics dataset and then combining them in a single matrix [7,21]. Late integration consists of multiple layers, with each layer in the model representing a single omics dataset [6]. Each layer analyzes its data independently without combining all other types. Integration occurs at the results level or by using fully connected layers, as in neural networks or deep learning models [6].

Consequently, the “curse of dimensionality”—having many features ( $p$ ) and few available data samples ( $n$ )—is among the most challenging when dealing with omics or biological data [12]. Therefore, the limited number of data instances ( $n$ ) and the large number of features ( $p$ ) present significant hurdles to adopting early concatenation in multi-omics integration [6]. Additionally, most studies introduce unsupervised methods that integrate multi-omics without considering domain contexts such as drug response and cancer classification [22].

In addition, transforming omics data types into a new representation reduces the interpretability of a model because the extracted features may lose the biological context [7]. Transformation-based methods frequently include complex mathematical transformations or fusion processes, which can lead to less interpretability of the models. Consequently, understanding the precise biological measurements of features or relationships within

the integrated data might be challenging [7,23]. The reduction in interpretability from transformation-based multi-omics integration refers to the loss of understanding of biological values when raw biological data are converted into abstract features using elaborate transformations [7]. This reduction occurs because the produced features do not directly relate to the original biological variables (e.g., genes or proteins). For example, methods like variational autoencoders and Similarity Network Fusion (SNF) [24] generate new features from the original ones, with new values that do not reflect the original biological variables. This loss of interpretability makes it challenging to map model predictions back to precise biological processes or mechanisms, thus obscuring direct biological relevance and hindering meaningful inferences from the model output. Consequently, transformation methods tend to mask intrinsic relationships within raw data—relationships that are often pivotal in uncovering valuable insights for biomedical research.

This paper introduces a new method called Iterative Similarity Bagging (ISB), supported by Bayesian Ridge Regression (BRR) as a domain-oriented feature selection method. Drug response is assigned to each omics type as an input for BRR. Subsequently, BRR works as a domain-oriented supervised feature selection tool to choose essential features by calculating the coefficient for each feature. Bayesian regression estimates linear models using probability distributions rather than singular point values [25]. The posterior probability of the model parameters depends on training inputs and outcomes [25]. The Ridge approach is used with Bayesian regression to reduce model complexity and multicollinearity through coefficient reduction [26]. Given that most omics data have a small number of samples, which are considered small-scale, Bayesian analysis is appropriate for such scenarios [27].

However, the BRR output datasets contain many features, leading to complexity and dimensionality issues. More details will be elaborated on in the Results and Discussion section. Therefore, Iterative Similarity Bagging (ISB) was introduced as the next step in solving these problems. Iterative Similarity Bagging was inspired by the bagging ensemble technique, utilizing column-based grouping to set features in bags without replacement [28]. The similarity is then computed based on a Euclidean distance for each group's features or genes. The nearest features inside each bag are considered redundant and are removed according to a dynamic threshold. ISB is an unsupervised method that dynamically reduces dimensionality without losing the original biological values. By preserving interpretability and retaining the original values, ISB helps to measure the relationships between drug response and omics features, as well as between the omics features themselves. Therefore, features are studied and selected according to their original values without transforming them into new features [28].

Furthermore, there is a strong correlation between features represented by probes with either similar molecular activities (as shown in gene expression analysis) or genomic positions (as observed in DNA copy number analysis) [29]. Traditional machine learning and feature selection techniques exhibit instability when confronted with strong correlations across features [29,30]. Darst et al. [31] examined the influence of correlated features in large-scale omics datasets by employing a random forest model with a recursive feature elimination technique. The researchers determined that RF-RFE may not be suitable for dealing with high-dimensional omics data containing numerous highly correlated features [31]. The contribution of this paper can be summarized in the following points:

1. Bayesian Ridge Regression (BRR) is introduced as a supervised domain-oriented feature selection method to reduce omics complexity and dimensionality. Features are selected based on domain contexts, such as drug response and cancer classification.
2. A new method named Iterative Similarity Bagging (ISB) is presented to perform a dynamic reduction of dimensionality and complexity without losing the biological measurements of omics data, which is a common issue with some transformation-based integration approaches.

This study presents BRR-ISB, a method that discovers informative features by utilizing two relationships. The first relationship focuses on drug response and multi-omics types, exploring how drug response can be predicted using multi-omics features. Bayesian Ridge

Regression (BRR) is used to study this relationship. BRR is a supervised feature selection method that ranks features according to their relevance to drug response. By estimating coefficients for all its features, BRR can identify influential ones on drug response, ensuring these selected features are contextually meaningful.

The following relationship studies the connections among omics features themselves. It is assessed via Iterative Similarity Bagging (ISB), an unsupervised approach that removes similar genes based on their distance metrics. ISB works by iteratively pruning the feature set, selecting the most similar ones, and leaving others that are more distinctively informative. This paper combines these two approaches to produce a more robust model for predicting drug response, thereby balancing domain-specific relevance and internal feature variety.

The rest of this manuscript is structured as follows: Section 2 introduces the related studies. Section 3 details the methods and materials utilized in this investigation, including the datasets, BRR, ISB, and evaluation metrics. Section 4 elaborates on the results and discussion. Finally, Section 5 presents the conclusion, limitations, and future work.

## 2. Related Work

Several recent reviews have presented methods for integrating multi-omics data using statistical models, machine learning, or deep learning, either supervised or unsupervised [6,7,10,12,22,32,33]. This brief discussion will introduce some recent methods, focusing on integration strategies: early integration, mixed or middle integration, and late integration.

For early integration, Park et al. [34] introduced iMO-BSPC (Integrative Analysis of Multi-omics Data Based on Blockwise Sparse Principal Components). The algorithm conducted variable clustering for each omics dataset. Their approach identified the initial sparse principal components (sPCs) as surrogate variables for dimensionality reduction for each cluster or block. The sPCs obtained from each omics data source were then combined to create a unified multi-omics dataset.

Xie et al. [35] presented a novel approach called GDP (Group Lasso Regularized Deep Learning for Cancer Prognosis) to analyze survival. This method leverages gene-level group prior knowledge. GDP was employed to integrate clinical data with other types of molecular data, including RNA-seq, copy number variation (CNV), normalized RPPA protein expression data, and DNA somatic mutation data. The GDP technique was utilized to regularize the neural network's input layer coefficients. In their study, "group prior knowledge" refers to the ability to combine many characteristics of the same gene, such as copy number variation, gene expression level, protein expression level, and single-nucleotide polymorphism, during regularization.

For graph-based methods, a mixed integration approach has been considered. Xie et al. [36] developed Learning Graph Representation for Drug Response Prediction (LGRDRP) to predict cell line drug responses. LGRDRP first builds a heterogeneous network that integrates cell line miRNA expression profiles, drug chemical structural similarity, gene–gene interaction, known cell line drug responses, and cell line–gene interaction. The learning graph representation and Laplacian feature selection are then combined for each cell line to obtain network topology features related to the cell line. The learning graph representation method learns network topology structural features, and the Laplacian feature selection works to choose the most informative features.

Wen et al. [17] presented a multi-omics data integration approach that utilizes random walk with restart (RWR) over a multiplex network. The resulting process is a Random Walk with Restart for Multi-dimensional Data Fusion (RWRF). RWRF utilizes the similarity network of samples as the basis framework for integration. The process involves creating a similarity network for each data type and linking the appropriate samples from numerous similarity networks to form a multiplex sample network. RWRF uses the stationary probability distribution to combine similarity networks by implementing RWR on the multiplex network.

Graph-embedding approaches were introduced to integrate multi-omics by extracting new features representing each “omic”. Xuan et al. [13] proposed DTIGBDT, a gradient boosting decision tree-based drug–target interaction prediction approach. They created a drug–target heterogeneous network with drug similarities based on chemical structures, target sequence similarities, and known drug–target interactions. Random walks updated drug or target similarities by capturing network topology. Multiple groups of drug–target paths were identified, and their features were extracted.

To integrate multi-omics using kernel-based approaches for predicting breast cancer survival, He et al. [19] utilized multiple kernel learning (MKL) to effectively integrate somatic mutation with contemporary molecular data such as methylation, copy number variation (CNV), gene expression, and protein expression. Maximum relevance minimum redundancy (mRMR) was implemented as a feature selection method to select informative features for each data type.

For late integration, Chu et al. [37] introduced Graph Transformer for Drug Response Prediction (GrapTransDRP). This innovative neural network structure can extract a more refined drug representation from molecular graphs to forecast drug responses on cell lines. The Graph Transformer was combined with the Graph Attention Network (GAT) and Graph Convolutional Networks (GCNs) to acquire knowledge about the features of drugs. Subsequently, 1D convolutional neural network (CNN) layers were employed for each omics data type to acquire latent gene expression, mutation, copy number aberration, and methylation features.

Malik et al. [38] proposed a late multi-omics integration framework for robustly quantifying survival and drug response in breast cancer patients. Each omics dataset was individually processed using a neighborhood component analysis (NCA)-based feature selection algorithm. The most important features identified were then utilized in classifier and regressor models based on neural networks.

A summary of the shortcomings in the related studies reveals that the curse of dimensionality inherent in biological data affects most proposed models. Recent studies have independently processed each “omic” to reduce dimensionality and data complexity [34,35], often exploring unsupervised methods, which may lose domain-specific context. Domains such as drug response or cancer classification are crucial in identifying informative features. Supervised feature selection approaches, such as those considering drug responses [22], help in selecting significant features based on domain relevance. Transforming omics datasets into new representations, like graph embeddings or unsupervised dimension reduction methods, may reduce model interpretability as the biological context of extracted features could be lost [13,17,24].

Overfitting occurs when a model fits training data perfectly but requires improvement in testing or unseen data. Overfitting is one of the challenges of using machine learning in drug response prediction [39]. Recent studies have employed various strategies to combat overfitting. Regularization techniques include Lasso (Least Absolute Shrinkage and Selection Operator) and Ridge Regression. For example, Lasso regression adds an  $L_1$  penalty to the loss function, leading to some coefficients becoming zero, which performs as a feature selection [40]. On the other hand, Ridge Regression shrinks the fitting coefficients using an  $L_2$  penalty so that they have less variance and control the overfitting of the data. Nevertheless, it does not push the coefficients of each feature to be zero [41].

Ensemble methods, such as Gradient Boosting Machines (GBMs) and random forests, have proven effective in controlling overfitting and improving prediction accuracy [42]. GBMs build models sequentially to correct errors made by previous models, while random forests aggregate predictions of several weak learners, such as decision trees, and learn from various data samples, thereby averting overfitting and enhancing prediction [42].

In deep learning, dropout is a critical technique to prevent overfitting [43]. This method randomly sets a fraction of the input neurons to zero during training, effectively “dropping out” some neurons. Thus, any contribution they would have produced in a backward or forward pass while training is prohibited. Therefore, the model training occurs with

noise, enabling the model to learn more robust and generalizable features, thereby reducing overfitting. Additionally, cross-validation is an essential stage often employed to assess how well a model conducts with different datasets, assuring that the model is generalized [39].

In the context of drug response prediction, Partin et al. [44] and Chang et al. utilized the early stopping technique to prevent overfitting. In this technique, model training halts when performance improvements in specific iterations of model validation are not observed [45]. They also employed various approaches to mitigate overfitting. They utilized dropout layers, max-pooling layers for dimensionality reduction, and five-fold cross-validation. Zhu et al. [46] utilized LightGBM as an ensemble method along with early stopping and regularization to mitigate overfitting. Sotudian et al. [47] introduced iterative thresholding for complexity reduction and utilized Lasso regularization to induce sparsity in coefficient vectors, alongside cross-validation used to prevent overfitting. Roder et al. [48] combined k-nearest neighbor (kNN) with logistic regression and employed strong regularization through a dropout strategy to mitigate overfitting risks. This technique is iterated into training and test sets for numerous random sample divisions. The continuous variable outputs of these numerous classifiers are averaged together using an ensemble method known as bagging. To find potential drugs in DTI datasets, Xiaolin et al. [49] presented OverfitDTI, a straightforward but efficient method for predicting drug–target interactions (DTIs). When utilizing OverfitDTI, a DNN model is trained using all the data. After being overfitted, a DNN model can “remember” the dataset’s features and use them to recreate the dataset. The features of the unseen drug and targets were obtained using a variational autoencoder (VAE) model. This allowed them to employ all the data for overfitting training, even if the drugs and targets were not labeled and the binding affinities between them were not given. VAE’s reconstruction function can build new data and extract features from existing datasets.

### 3. Materials and Methods

This section describes the dataset and the two main steps of the proposed method. In the first step, Bayesian Ridge Regression (BRR) is presented as a supervised domain-oriented feature selection approach to reduce omics complexity and dimensionality. In the second step, a new method called Iterative Similarity Bagging (ISB) is introduced to dynamically reduce dimensionality and complexity without losing biological measurements of omics data. In this research, drug response prediction serves as the application to test the effectiveness of BRR-ISB.

#### 3.1. Datasets

The Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE) are comprehensive drug response datasets used to validate the proposed method, Bayesian Ridge Regression with Iterative Similarity Bagging (BRR-ISB).

The (GDSC) project [50] tested 1000 cancer cell lines against 250 different chemicals for treatment sensitivity and identified genetic changes associated with drug efficacy. The TCGA’s genomic data on 11,289 tumors were compared to cancer cell lines, highlighting “cancer functional events” (CFEs), key molecular abnormalities. Data from cell lines sharing similar CFEs are widely used to predict responsiveness to a specific pharmaceutical treatment. Mutation and copy number information predict drug responses in specific tissues, while gene expression predicts pan-cancer response [51].

The Cancer Cell Line Encyclopedia (CCLE) project is a collaborative effort to analyze the genetic and pharmacological characteristics of a vast collection of human cancer cell lines [52]. The main objective of the CCLE is to provide researchers with a comprehensive repository for understanding the molecular and genetic characteristics of various cancer types and their responses to distinct medications and treatments. By analyzing diverse cancer cell lines, the CCLE enables investigation into cancer heterogeneity and the discovery of possible targets for therapy and biomarkers [53].

The (CCLE) project characterized over 1000 human cancer cell lines at the molecular level [52]. Analysis identified 24 susceptibility profiles to anti-cancer drugs among a sample of 363 cell lines. The CCLE and GDSC [52,54] datasets were utilized in this study. The IC50, representing drug responses for cell lines for different medications, was denoted as ( $y_{res,c}$ ) for a specific cell line  $c$ . Three omics were employed: single-nucleotide mutation (represented by  $x_{snv,g}$ ) for gene  $g$ , gene expression (represented by  $x_{exp,g}$ ), and copy number alteration/variation (represented by  $x_{cnv,g}$ ). Gene expression and copy number alteration are continuous variables, while single-nucleotide mutation is binary. In the case of mutation, “1” is used to indicate the presence of mutation, while “0” represents the absence of mutation (wild type) [54]. The gene expression data do not contain any missing values. Rows containing more than 50% missing values were excluded from copy number alteration and single-nucleotide mutation analysis. The mean weight method was employed to compensate for the missing values in the remaining cell lines [54].

The distance was computed to get the closest  $k$ , which was then utilized to fill in missing values for gene expression, as specified below:

$$distance(c,k) = \left\| x_{exp,c} - x_{exp,k} \right\|_2^2 \quad (1)$$

where  $c$  represents the cell line,  $k$  represents the nearest cell line, and  $x$  is the gene expression value for each cell line.

The missing value of cell line  $c$  in the copy number alteration of gene  $g$  was imputed using the mean value of the nearest cell lines.

$$missingCNV(c,g) = \sum_{k=1}^K \frac{distance(c,c_k)}{\sum_{k=1}^K distance(c,c_k)} missingCNV(c_k,g) \quad (2)$$

The single-nucleotide mutation features are represented by binary values, where a value of 1 indicates a mutation and a value of 0 indicates the wild type. The average feature value of cell line  $c$ , calculated among the  $k$ -nearest cell lines, was utilized to adjust for the absence of the SNV (single-nucleotide mutation or variation) value of gene  $g$  in the following manner [54]:

$$missingSNV(c,g) = \begin{cases} 1 \text{ if } \left( \sum_{k=1}^K missingSNV(c_k,g) > \sum_{k=1}^K (1 - missingSNV(c_k,g)) \right) \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

Similarly, the missing IC50 value was imputed using the same method as the copy number change. The missing value of cell line  $c$  was imputed using the average value of the closest cell lines.

$$missingIC50(c,g) = \sum_{k=1}^K \frac{distance(c,c_k)}{\sum_{k=1}^K distance(c,c_k)} missingIC50(c_k,g) \quad (4)$$

A 2D chemical structure of the drugs was retrieved from PubChem [55] in SMILES [56] format. Simplified Molecular Input Line Entry System (SMILES) strings can be converted into graphical representations of molecular structures, with atoms as nodes and chemical bonds as edges [57,58]. This graph-based representation enables the implementation of graph algorithms and methodologies for analyzing chemical structures and predicting drug responses [58]. SMILES provides a versatile means of describing complex molecular structures, encompassing cyclic compounds, stereochemistry, and functional groups. Graph-based representations derived from SMILES strings facilitate the extraction of diverse features that capture significant structural attributes of drugs. These features include properties associated with nodes (such as atom types and charges) and attributes associated

with edges (such as bond types and distances). These attributes serve as input features for machine learning models used in predicting drug responses [57,59].

Subsequently, RDKit [60], an open-source chemical informatics software tool, generated a molecular graph representing the interactions among atoms within the drug. The atom feature design from DeepChem [61] was utilized to characterize the nodes within the graph. Each node encompasses five distinct atom characteristics: the symbol of the atom, the atom's degree, determined by the number of linked neighbors, including hydrogen atoms, the overall count of hydrogen atoms, the implicit value of the atom, and whether the atom is aromatic [37]. These atom features form a binary feature vector with multiple dimensions [62]. An edge is established between two atoms when a link exists. Consequently, a binary graph with attributed nodes was indirectly constructed for each input SMILES string [62].

The atom symbol represents chemical elements such as carbon (C), nitrogen (N), and oxygen (O). Each element is encoded using one-hot encoding, where a vector of zeros represents every potential atom type (every unique atom form). The value one is placed in the corresponding cell dimension for the atom's element, and the zeros are assigned to all other dimensions. This encoding enables the model to accurately differentiate between atom types in molecules [63]. The atom degree is defined by the number of adjacent atoms, including hydrogen atoms. This function detects the atom's connection inside the molecule, revealing critical information about the local chemical structure, which is crucial for determining each atom's immediate surroundings within the molecular graph [61]. For instance, a carbon atom in methane (CH<sub>4</sub>) has a degree of four due to its bonds with four hydrogen atoms. Another crucial characteristic is the total number of hydrogen atoms bound to one atom. This count offers valuable information about the hydrogenation state of the atom, which in turn affects the reactivity and interactions of the molecule [63]. The implicit valence of an atom is the number of bonds it forms with its current bonding state/type. The inferred valence is important when interpreting the chemical stability and likely reactivity of the atom in the molecule. This feature helps predict how the molecule may interact with other molecules in various chemical reactions [61]. Aromaticity is a binary attribute indicating whether the element is an atom of an aromatic system (e.g., a benzene ring). These aromatic atoms are typically found in delocalized  $\pi$ -electron systems, which impart special electronic properties to the molecule. Aromaticity is a crucial aspect of the model, as aromaticity has a significant influence on the stability and reactivity of each molecule [64].

The IC<sub>50</sub> matrix was transformed into a tabular format consisting of 8712 rows, each containing the cell line name, drug name, and response value. The total number of samples for the CCLE data is shown in Table 1.

**Table 1.** The number of samples and features of genomics data in the used CCLE dataset.

Type	Raw Data	Processed
Drugs	24	24
Cell lines	1061	363
Gene expression	20,049	19,389
Copy number alteration	24,960	24,960
Single-nucleotide mutation	1667	1667

In GDSC, the IC<sub>50</sub> matrix was converted into a tabular format consisting of 54,390 rows, each including the cell line name, drug name, and response value. Table 2 presents the total number of samples for the GDSC data.



**Table 2.** The number of samples and features of genomics data in the used GDSC dataset.

Type	Raw Data	Processed
Drugs	98	98
Cell lines	1124	555
Gene expression	11,833	11,712
Copy number alteration	24,960	24,959
Single-nucleotide mutation	70	54

### 3.2. Bayesian Ridge Regression

Bayesian Ridge Regression (BRR) was employed as a feature selection strategy. This approach involves calculating coefficients to estimate the relevance score of each feature. Bayesian regression focuses on utilizing the Bayesian approach, wherein the estimation of linear models is conducted by considering probability distributions rather than singular point values [25]. The posterior probability of model parameters is influenced by training inputs and outputs. Additionally, the ridge method is integrated into Bayesian regression to mitigate the issues of model complexity and multicollinearity through coefficient shrinkage. Most omics data are typically characterized as small-scale, making Bayesian analysis appropriate for such scenarios. The process combines the existing information about the parameter, known as the prior parameter distribution, with the observed data [27].

Bayesian Ridge Regression (BRR) differs from traditional Ridge Regression by utilizing a probabilistic framework and including prior distributions for the regression coefficients [65]. Traditional Ridge Regression aims to minimize the sum of squared errors while imposing a penalty term that restricts the regression coefficients'  $L^2$  norm (Euclidean norm) [65]. Including this penalty term alleviates the problem of multicollinearity and enhances the stability of the regression coefficient estimation, particularly in the context of high-dimensional data. In contrast, BRR treats regression coefficients as stochastic variables and assigns a prior distribution to them, typically Gaussian [27]. This Bayesian approach facilitates the integration of prior knowledge or beliefs about coefficient distributions into the modeling process. During parameter estimation, Bayesian Ridge Regression (BRR) not only maximizes data likelihood but also updates with observed data. This results in a posterior distribution representing the regression coefficients' uncertainty. The final approximations are typically derived by computing posterior mean or mode, striking a balance between prior information and observed data [65–67].

The introduction of uninformative priors on hyperparameters allows regularization akin to  $L_2$  regularization in Ridge Regression and classification. This regularization seeks maximum a posteriori estimation under a Gaussian prior over the coefficients  $w$  with precision  $\lambda^{-1}$ . Instead of setting lambda manually, it is possible to treat it as a random variable to be estimated from the data [66]. The output  $y$  is assumed to be Gaussian-distributed around  $Xw$  to create a fully probabilistic model [68].

$$p(y|X, w, \alpha) = \mathcal{N}(y|Xw, \alpha) \quad (5)$$

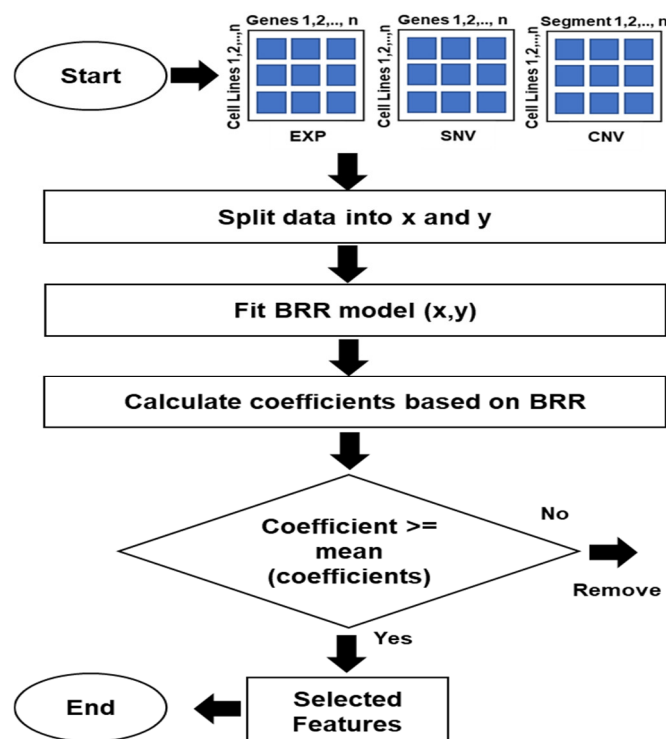
where  $\alpha$  is once more considered a random variable that needs to be estimated from the data.

As previously described, Bayesian Ridge Regression estimates a probabilistic model for the regression problem, employing a spherical Gaussian as the prior for the coefficient  $w$  [65].

$$p(w|\lambda) = \mathcal{N}(w|0, \lambda^{-1}I_p) \quad (6)$$

The priors over  $\alpha$  and  $\lambda$  are selected as gamma distributions, which are conjugated prior for Gaussian precision. The resulting model, similar to the classical Ridge Regression, is known as Bayesian Ridge Regression. During model fitting, the parameters  $w$ ,  $\alpha$ , and  $\lambda$  are estimated jointly, with the regularization parameters  $\alpha$  and  $\lambda$  estimated by maximizing the log marginal likelihood [65,69].

This study utilized BRR to select informative features based on drug response. For each omics dataset, a drug response value was assigned to each sample; then, data were split into X as the multiple independent variables and Y as the dependent variable. The BRR model was fitted to predict the response variable using a weighted sum of the independent variables. This process involved calculating coefficients that were used as a weighted total to make a prediction. The coefficients were utilized as feature relevance ratings to choose the optimal features from each omics dataset. In the final stage, the coefficients greater than or equal to the calculated mean were chosen as significant features. Figure 1 shows the flowchart of BRR for selecting informative features for each omics type.



**Figure 1.** The flowchart of BRR to select essential features for each omics type.

### 3.3. Iterative Similarity Bagging

A new method called Iterative Similarity Bagging (ISB) was introduced as the next step to address the issue of dimensionality resulting from the many features selected in the BRR step. ISB is an unsupervised method that eliminates redundant features based on their distance. If neighboring features are deemed redundant, one of them will be removed. ISB draws inspiration from the bagging ensemble technique, utilizing columns-based grouping to partition features into bags without replacement. Subsequently, similarities are computed based on the Euclidean distance for each group of features or genes. Redundant features within each bag are identified and removed using a dynamic threshold. Figure 2 shows the architecture of ISB.

The architecture shown in Figure 2 consists of several steps. (A) ISB receives all omics datasets (EXP, SNV, and CNV) produced from BRR. The number of iterations is a parameter that needs to be set, and steps B–F are repeated independently for each omics dataset according to the iterations count. (B) Column-based bagging: each set of features collected in the bag is parameter  $k$ . Each number of features  $k$  will be grouped into the bag without replacement and with the same number of samples. (C) The bag will be received as a matrix, and then the matrix will be transposed, which interchanges its columns into rows to be prepared for similarity calculation. In step (D), the similarity between genes/features will be computed based on Euclidean distance, as shown in Equation (7). (E) Correlated features are removed based on a threshold. Generally, if the distance between neighboring

points is close to zero, one of them is considered a redundant feature. ISB supports various thresholds such as mean, median, half-mean, and first percentile (Q1). According to the literature review, the mean value is commonly used as a threshold to filter features [70]. This means features with scores higher or lower than the calculated mean of all feature scores can be selected or eliminated, depending on the experiment's scope and purpose. Using a mean-level threshold or dynamic threshold overcomes the challenges associated with fixed threshold techniques, such as the uncertainty problem [71]. In this experiment, the half-mean threshold was selected because the method aims to eliminate the neighboring points, thus favoring the smallest value. The half-mean threshold was chosen after several experiments to determine the best threshold for this study. The final step in each iteration is F, wherein selected features from the current bag are added to a list. All lists generated from all iterations are concatenated to select unique features of each omics dataset. Algorithm 1 shows the input, output, and main steps that represent how ISB works.

---

**Algorithm 1:** Iterative Similarity Bagging method

---

**Input:**

- Iterations count:  $i$
- Bag size (Columns in the bag):  $k$
- Single-omics dataset:  $data$

**Output:**

- Selected features list after all iterations:  $selected\_features$

**Begin**

**Declare** List  $selected\_features = []$

// Increment value after each iteration

**Declare** integer increment value:  $c$

**Set**  $index = 0$  // Starting column to select genes in the bag.

**Set**  $c = k$

**For**  $index = 0$ :  $i$

**For**  $j = 0$ :  $range(0, len(data.columns))$

**IF**  $index < len(data.columns)$ :

$df\_bag = SELECT\_COLUMNS(data, index : k$

$df = TRANSPOSE(df\_bag)$

$df\_Sim = compute\_similarity(df)$  // Euclidean distance

$threshold = get\_threshold(df\_Sim)$  // Half-mean threshold

$iteration\_selected\_features = df\_Sim[col] > threshold$

$selected\_features += list(unique(iteration\_selected\_features))$

$index = index + c$

$k = k + c$

**End IF**

**End For**

**End For**

**End**

---

The similarity between genes was calculated using Euclidean distance, represented as follows:

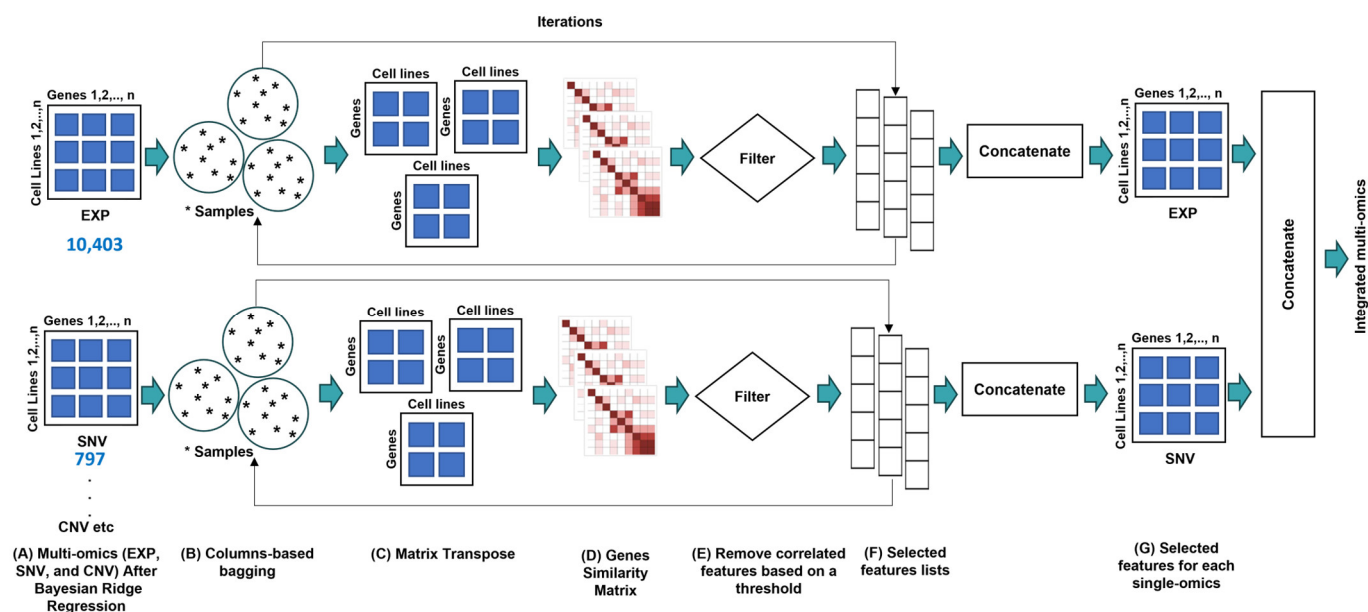
$$Distance(X, Y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2} \quad (7)$$

where  $X$  and  $Y$  are arrays of gene values, with  $n$  representing the number of genes.

The half-mean threshold was employed in this study to filter features and eliminate redundant ones. The formula is represented as follows:

$$Half\_Mean = \frac{\sum_{i=1}^n x_i}{n} / 2 \quad (8)$$

where  $x$  represents the distance value between genes, and  $n$  denotes the number of genes.



**Figure 2.** The architecture of the Iterative Similarity Bagging method.

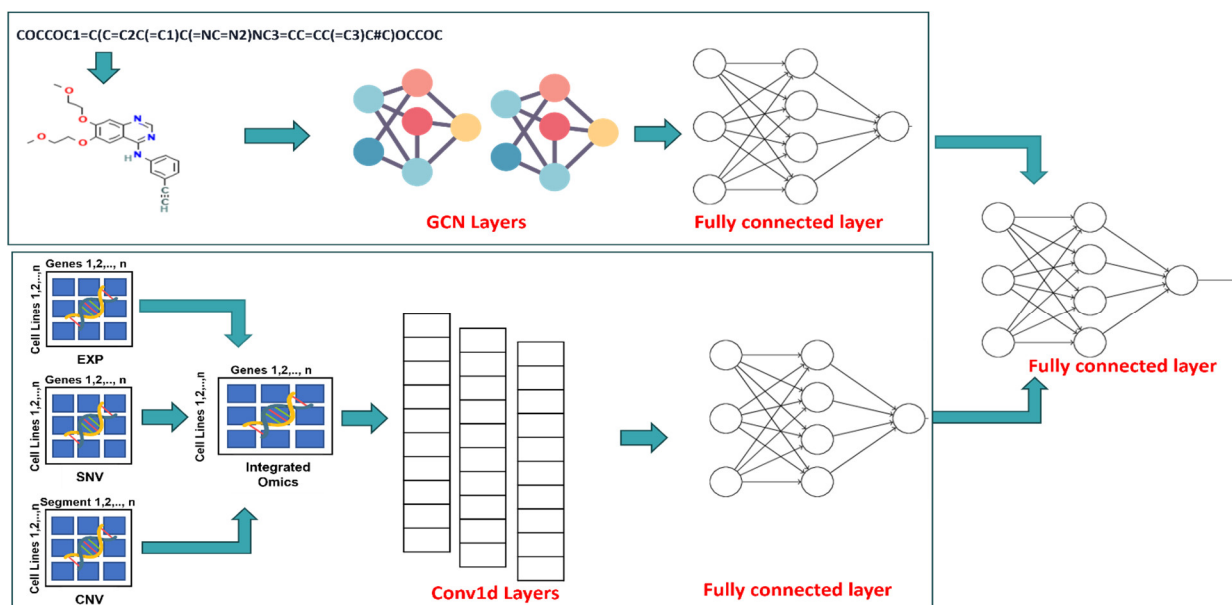
### 3.4. Drug Response Prediction Using Graph Convolutional Network and Convolutional Neural Network

Personalized oncology or medicine is a cancer therapy method aimed at identifying the most effective therapeutic solutions for each patient. This discovery was facilitated by combining genetic and drug sensitivity data, and the subsequent creation of drug response associations allowed this discovery. While personalized medicine is not yet utilized as a regular treatment, it is possible for most cancer patients due to the progress in multi-omics features and drug sensitivity testing [72]. Personalized treatment regimens based on genetics are a primary goal of systems medicine. For the development of individualized cancer therapy treatments with a projected efficacy much above existing standard-of-care methods, inferred models' ability to correctly forecast a tumor's responsiveness to a medicine or drug combination might benefit that process. Various methods for predicting drug response rely on genetic profiling data [72]. If a training set has experimentally measured genomic characteristics (such as protein expression, RNA expression, single-nucleotide polymorphisms (SNPs), DNA methylation, or other types) and responses to different drugs, then supervised or unsupervised approaches can be developed for each drug, considering one or more omics data features based on these assumptions [51].

Accurately predicting how individual patients will respond to drugs is a critical focus within the realm of personalized medicine [73]. By exploring the intricate interplay among multi-omics data derived from cancer cell lines or patient tumors, novel opportunities arise for developing individualized therapeutics tailored to each cancer patient [72]. Thus, drug response prediction was employed to evaluate the proposed model. BRR-ISB can also be applied across different domains. BRR-ISB integrates three omics: gene expression, single-nucleotide mutations, and copy number variations. This integrated multi-omics approach, combined with the drug's chemical structure, served as inputs to the drug response model shown in Figure 3.

Graph Convolutional Networks (GCNs) have increasingly been employed to handle drug chemical structures due to the capability of the GCNs to model intricate relationships in molecular graphs. In this context, GCNs represent each molecule as a graph, where atoms are nodes and chemical bonds are edges. GCNs perform convolutional operations on these graphs, pooling features of each atom with those of its neighboring atoms to reproduce a specific representation of the atom within the molecule. This approach enables

the model to consider both local atomic features and the global molecular structure, thereby making accurate predictions of molecular properties and activities. GCNs have demonstrated success in various tasks, such as predicting molecular properties, bioactivity, and pharmacokinetic properties, outperforming traditional machine learning methods that rely on hand-crafted features [63,74]. For instance, Wu et al. [63] showcased the effectiveness of GCNs on MoleculeNet, a benchmark dataset for molecular machine learning, reporting significant improvements in predictive performance across multiple datasets. GCNs facilitate efficient learning from graph representations of chemical structures, proving to be a powerful tool for drug discovery and development.



**Figure 3.** The architecture of the GCN and 1D-CNN model for drug response prediction.

The architecture of the proposed model was inspired by [37,62] and customized to fit the CCLE dataset and integrate multi-omics. The Graph Convolutional Network (GCN) model is specifically designed to combine and analyze graph-structured and sequential data. This allows the model to predict intricate biological events such as drug response prediction. The model combines graph convolutional layers with 1D convolutional neural network layers. The graph component of the model processes drug features, represented as nodes in a graph. Each node is defined by a feature vector of 75 elements representing different molecular properties. Connections between nodes describe the connections and interactions between distinct features of drugs. These connections are represented by the adjacency matrix (edge index) and the properties of the connections. The model employs two Graph Convolutional Network (GCN) layers to process the graph-structured data. The initial GCN layer reduces node features from their original size to an embedded size of 512 and applies a ReLU activation function. Subsequently, node embeddings undergo Top-K Pooling to retain 80% of the most significant nodes. The aggregated representation then enters a subsequent Graph Convolutional Network (GCN) layer, followed by another Top-K Pooling layer, preserving 50% of the nodes. Global Max Pooling (GMP) and Global Average Pooling (GAP) combine the node features into a single graphical representation after each pooling layer.

The omics data are processed using 1D convolutional layers [62]. Initially, the data are reshaped with a channel dimension and processed through a convolutional layer with 40 channels and a kernel size of 8, followed by ReLU activation. Subsequently, a max-pooling layer with a pool size of 3 reduces dimensionality. The output then passes through a second convolutional layer with 80 channels, and a similar ReLU activation function is

used, followed by another max-pooling layer. The resulting feature maps are flattened and converted into a 128-dimensional vector through fully connected layers.

### 3.5. Evaluation Metrics

The data were split based on the Pareto Principle [75,76], also known as the 80:20 rule, wherein 80% was allocated for training, 10% for validation, and 10% for testing. Additionally, this study utilized cross-validation [77,78] to validate model prediction and BRR feature selection. The model underwent training and validation in each epoch, ensuring that every iteration yielded the best model compared to its predecessor, and was subsequently exported. Finally, the best model was employed to predict training, validation, and testing splits within a 10-fold to guarantee that the results did not happen randomly. The final result is the mean of the 10-fold. Bayesian Ridge Regression (BRR) feature selection was utilized with GridSearchCV with a specified cross-validation strategy (in this case, 3-fold); this searches for the optimal hyperparameters that maximize model effectiveness. The feature coefficients were obtained from the best estimator [78,79]. This study used three evaluation metrics: Root Mean Square Error (RMSE), the Pearson Correlation Coefficient (PCC), and the coefficient of determination ( $R^2$ ) [37,62,79].

The Root Mean Square Error (RMSE) [45,80] was used to quantify the discrepancy between the observed and expected drug response values. It is mathematically defined as

$$RMSE = \frac{\sqrt{\sum (y_i - \tilde{y}_i)^2}}{N} \quad (9)$$

where  $N$  is the sample size,  $y$  is the actual value of drug response,  $\tilde{y}_i$  is the predicted value of drug response.

The PCC [81] value was utilized to quantify the extent of association or correlation between the drug response and predictors generated by multi-omics integration, which may be defined as

$$PCC = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \quad (10)$$

where  $y_i$  is the drug response value,  $x_i$  is the value of predictors,  $\bar{x}$  and  $\bar{y}$  indicate the means of values.

The coefficient of determination ( $R^2$ ) [82] was employed to quantify the extent to which the variability in drug response can be explained by its relationship to other independent variables. This might be formulated as

$$R^2 = 1 - \frac{\sum (y_i - \tilde{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} \quad (11)$$

where  $y$  is the actual value of the drug response,  $\tilde{y}_i$  is the predicted value of the drug response.

### 3.6. Experimental Setup

The experiments were implemented using Python PyTorch (2.3.0) and scikit-learn (1.2.2) to develop the proposed solution. The models were executed on Google Colab with T4 GPU and 50 GB RAM. The baseline scenario was implemented using a different GPU (TPU V2) because using a T4 GPU causes memory crashing.

## 4. Results and Discussion

This section presents the results and effectiveness of the proposed solution. BRR-ISB was implemented for drug response prediction in three steps. BRR demonstrated informative features selected by BRR as a supervised domain-oriented feature selection method. ISB presented the final features chosen by ISB as an approach for the dynamic

reduction of dimensionality and complexity without losing biological measurements of omics data. Finally, the results of the predictive model for drug response are presented and discussed.

#### 4.1. Genomic Features Selected by BRR

Drug response is assigned to each omics type as an input for BRR. Thereafter, BRR functions as a domain-oriented supervised feature selection to determine essential features based on the calculating coefficients for each feature. Figure 4 shows the selected features of each omics type after implementing BRR.



**Figure 4.** Architecture of BRR and the selected features of each omics type after implementing BRR.

However, the selected features generated by BRR are still considered numerous and, thus, potentially time-consuming. Therefore, Iterative Similarity Bagging (ISB) was employed as an additional step to dynamically reduce dimensionality and complexity. Table 3 shows the top ten features for each multi-omics produced by BRR.

**Table 3.** The top ten features for each multi-omics after utilizing BRR.

Type	Genes
Gene expression	TFPI2, SGCE, PPIC, ATP1B1, DSP, PEG10, MAGEA4, C1S, CPVL, GATA6.
Copy number alteration	RASSF8AS1, MIR4302, CCNE1, RASSF8, LMNTD1, LOC102724958, STARD3, LINC00906, KRAS, LYRM5
Single-nucleotide mutation	AKAP12, TP53, NLRP3, ATRX, OBSCN, CARD10, KRAS, ATR, FZD1, GPR112

Expression values of certain genes play a crucial role in various types of cancers. The loss of sarcoglycan, epsilon gene (SGCE) enhances the responsiveness of breast cancer to

chemotherapy by targeting breast cancer stem cells (BCSCs) [83]. ATP1B1, a pivotal gene involved in copy number variations, has been identified as a significant driver in diffuse large B-cell lymphoma. This gene holds potential as a target for therapeutic development [84]. An exploration of paternally expressed gene 10 (PEG10) as a novel therapeutic strategy to overcome resistance to cyclin-dependent kinases 4 and 6 (CDK4/6) inhibitors in breast cancer has also been reported [85]. In addition, the role of A-kinase anchor protein 12 (AKAP12) as a tumor suppressor in various malignancies has been extensively investigated and confirmed.

However, the role of the immune system in stomach adenocarcinoma (STAD) remains unclear [86]. Furthermore, one in eight breast cancers is “HER2-positive,” which is more aggressive and is treated with surgery, chemotherapy, a targeted drug against HER2, radiation, and endocrine therapy. HER2 is strongly linked to STARD3, a protein discovered by Lodi et al. [87]. STARD3 predicts a pathological complete response and shows significant correlation with prognosis and cancer recurrence.

#### 4.2. Genomic Features Selected by ISB

The extensive number of features selected by BRR introduced complexity and dimensionality. Therefore, ISB was implemented to dynamically reduce the number of features without compromising biological context. Table 4 presents the top ten features of each omics type based on ten iterations, with a bag size of 200. The following features were selected across all iterations.

**Table 4.** The top ten features for each multi-omics after implementation of ISB.

Type	Genes
Gene expression	TFPI2, SGCE, ATP1B1, DSP, PEG10, MAGEA4, C1S, CPVL, GATA6, RP11-490M8.1
Copy number alteration	RASSF8-AS1, STARD3, PPP1R1B, SLC35E3, ZNF536, SOX5, TRIT1, TMEM75, ZNF879, ST8SIA1
Single-nucleotide mutation	AKAP12, TP53, NLRP3, ATRX, OBSCN, CARD10, KRAS, ATR, FZD1, GPR112

The features presented in Table 3 closely resemble those in Table 4. In gene expression, one gene (PPIC) was removed by ISB, which was considered correlated with other features, with one of them being selected and considered sufficient for the model. In copy number alteration, RASSF8-AS1 and STARD3 were the only genes selected by ISB after BRR. Finally, the genes in single-nucleotide mutation were identical in both tables.

#### 4.3. Effectiveness of BRR-ISB in Drug Response Prediction

Different scenarios were implemented to assess the effectiveness of BRR-ISB in drug response prediction. (1) Baseline: the first scenario implemented the model with all 46,016 features of the three omics. (2) The model relied on the features produced for BRR after utilizing BRR as a feature selection method. (3) Similarity Network Fusion (SNF) [24] is one of the state-of-the-art methods for integrating genomic data. SNF works to create patient clustering for each available data type and then efficiently fuses these into one network that represents the full spectrum of underlying data. SNF can often achieve promising results compared with other methods, such as iCluster [88] or KMeans. SNF is a graph-embedding method in which transforming high-dimensional to new, low-dimensional representation may reduce the interpretability of a model as the extracted features are no longer biological measurements. Therefore, BRR-ISB was compared with SNF to evaluate how preserving biological context affects the model results. (4) BRR was combined with SNF. (5) Finally, different parameters were used to test BRR-ISB, as shown in Tables 5 and 6. All scenarios were implemented to the CCLE and GDSC dataset and drug chemical structure data, utilizing 100 epochs and 512 batches.



**Table 5.** The scenarios were implemented using the same CCLE dataset and drug chemical structure for drug response prediction.

Method	Input Features	Model Features	Training			Validation			Testing			Time
			RMSE	PCC	R <sup>2</sup>	RMSE	PCC	R <sup>2</sup>	RMSE	PCC	R <sup>2</sup>	
Baseline	46,016	46,016	0.088	0.935	0.873	0.12	0.864	0.744	0.13	0.13	0.737	4:03:11
Iterative Similarity Bagging (ISB)												
ISB bag size = 50, iterations = 5	46,016	13,844	0.087	0.935	0.875	0.118	0.871	0.755	0.126	0.87	0.754	10:17
ISB bag size = 50, iterations = 10	46,016	12,270	0.103	0.909	0.824	0.127	0.847	0.716	0.13	0.858	0.736	9:16
ISB bag size = 100, iterations = 10	46,016	5926	0.08	0.946	0.894	0.117	0.87	0.755	0.129	0.862	0.74	5:50
ISB bag size = 200, iterations = 10	46,016	2390	0.091	0.929	0.863	0.116	0.875	0.764	0.124	0.873	0.76	3:51
ISB bag size = 300, iterations = 10	46,016	2261	0.091	0.929	0.863	0.119	0.866	0.747	0.126	0.868	0.75	3:48
ISB bag size = 400, iterations = 10	46,016	2119	0.1	0.917	0.837	0.119	0.866	0.75	0.127	0.865	0.747	3:42
Bayesian Ridge Regression with Iterative Similarity Bagging (BRR-ISB)												
BRR	23,683	23,683	0.087	0.937	0.877	0.117	0.872	0.758	0.125	0.87	0.754	15:43
BRR-ISB bag size = 50, iterations = 5	23,683	5740	0.093	0.926	0.857	0.115	0.876	0.766	0.124	0.872	0.759	5:48
BRR-ISB bag size = 50, iterations = 10	23,683	4822	0.094	0.924	0.854	0.121	0.86	0.739	0.127	0.865	0.748	5:20
BRR-ISB bag size = 100, iterations = 10	23,683	2133	0.099	0.917	0.84	0.117	0.871	0.758	0.125	0.869	0.754	3:57
BRR-ISB bag size = 200, iterations = 10	23,683	1273	0.097	0.919	0.845	0.114	0.879	0.771	0.121	0.879	0.77	3:50
BRR-ISB bag size = 300, iterations = 10	23,683	1245	0.099	0.918	0.84	0.116	0.872	0.76	0.122	0.877	0.768	3:47
BRR-ISB bag size = 400, iterations = 10	23,683	1260	0.097	0.921	0.846	0.116	0.874	0.763	0.127	0.867	0.749	3:50
Similarity Network Fusion												
SNF	46,016	363	0.13	0.854	0.721	0.13	0.841	0.699	0.134	0.848	0.716	3:02
BRR-SNF	23,683	363	0.126	0.859	0.738	0.127	0.844	0.713	0.138	0.847	0.7	3:02

The results from ISB pipeline experiments and BRR-ISB are nearly similar. However, the primary objective of the proposed method is to select informative features based on two relationships. Firstly, the relationship between drug response and multi-omics features is assessed using Bayesian Ridge Regression (BRR), a supervised feature selection method focused on domain-oriented contexts such as drug response. Secondly, the relationship among omics features themselves is measured using Iterative Similarity Bagging (ISB), an unsupervised method that eliminates similar genes according to distance.

In general, features from the CCLE and GDSC were manually selected using the BRR-ISB or the ISB alone. Therefore, the determination of the bag size and the number of iterations requires that it is set automatically via a method such as GridSearchCV in scikit-learn. In addition, considering alternatives to BRR for feature selection methods may be considered to improve the proposed method.

Combining all features in the baseline scenario caused memory crushing; therefore, a different GPU was utilized to solve the memory issue. BRR-ISB, employing a bag size of 200 features and ten iterations, demonstrated superior results regarding RMSE, PCC, and R<sup>2</sup> compared to other scenarios. It also significantly reduced the execution time by 77% compared to the BRR scenario. Similarity Network Fusion (SNF), a transformation-based

integration method, may decrease the interpretability of a model due to the potential loss of biological context in the extracted features. In contrast, BRR-ISB, with its dynamic reduction in dimensionality and complexity without compromising the biological measurements of omics data, outperformed SNF regarding RMSE, PCC,  $R^2$ , and execution time.

**Table 6.** The scenarios were implemented using the same GDSC dataset and drug chemical structure for drug response prediction.

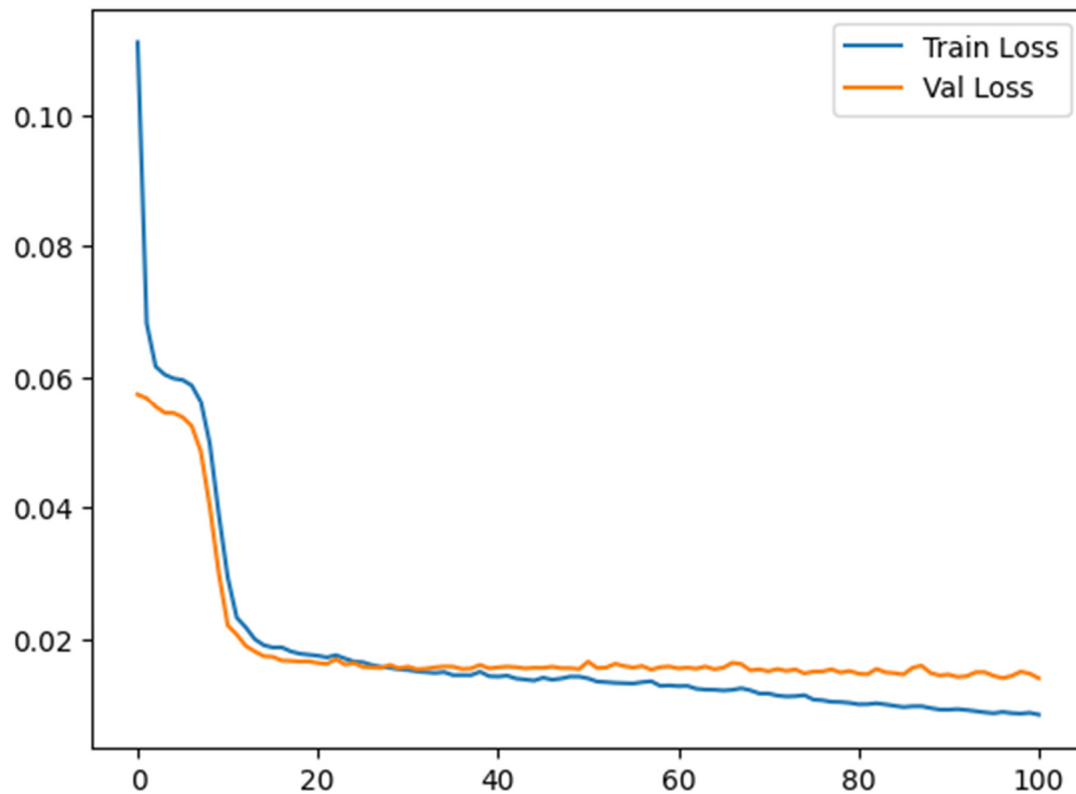
Method	Input Features	Model Features	Training			Validation			Testing			Time
			RMSE	PCC	$R^2$	RMSE	PCC	$R^2$	RMSE	PCC	$R^2$	
Baseline	36,725	36,725	0.023	0.934	0.872	0.032	0.879	0.771	0.03	0.89	0.791	19:38:19
Iterative Similarity Bagging (ISB)												
ISB bag size = 50, iterations = 5	36,725	11,987	0.022	0.943	0.889	0.031	0.881	0.776	0.03	0.895	0.8	59:48
ISB bag size = 50, iterations = 10	36,725	10,367	0.023	0.935	0.874	0.031	0.88	0.774	0.03	0.895	0.799	51:44
ISB bag size = 100, iterations = 10	36,725	4211	0.024	0.933	0.867	0.031	0.887	0.783	0.029	0.9	0.808	32:52
ISB bag size = 200, iterations = 10	36,725	1174	0.026	0.919	0.843	0.031	0.881	0.775	0.03	0.894	0.798	26:07
ISB bag size = 300, iterations = 10	36,725	974	0.025	0.92	0.846	0.031	0.883	0.78	0.029	0.896	0.803	24:31
ISB bag size = 400, iterations = 10	36,725	956	0.025	0.92	0.846	0.031	0.882	0.777	0.029	0.896	0.801	24:43
Bayesian Ridge Regression with Iterative Similarity Bagging (BRR-ISB)												
BRR	36,725	18,392	0.024	0.93	0.866	0.032	0.878	0.771	0.03	0.894	0.798	1:23:11
BRR-ISB bag size = 50, iterations = 5	18,392	5369	0.023	0.938	0.879	0.031	0.885	0.783	0.029	0.899	0.807	36:19
BRR-ISB bag size = 50, iterations = 10	18,392	4509	0.024	0.928	0.859	0.031	0.882	0.775	0.03	0.894	0.797	32:07
BRR-ISB bag size = 100, iterations = 10	18,392	1681	0.026	0.915	0.835	0.031	0.881	0.774	0.03	0.892	0.794	23:28
BRR-ISB bag size = 200, iterations = 10	18,392	606	0.026	0.916	0.838	0.031	0.883	0.777	0.029	0.896	0.801	19:58
BRR-ISB bag size = 300, iterations = 10	18,392	549	0.028	0.904	0.817	0.032	0.878	0.771	0.03	0.892	0.796	21:08
BRR-ISB bag size = 400, iterations = 10	18,392	566	0.028	0.903	0.815	0.032	0.877	0.769	0.03	0.892	0.796	21:33
Similarity Network Fusion												
SNF	36,725	555	0.029	0.896	0.802	0.033	0.87	0.756	0.031	0.884	0.782	20:23
BRR-SNF	18,392	555	0.029	0.894	0.799	0.033	0.869	0.756	0.031	0.884	0.781	20:19

In addition to a reduction in computational cost and a substantial dimension reduction, the feature selection methods offer several benefits. These benefits include enhanced model interpretability, reduced overfitting risk, and potentially improved generalization to new data. Studying multi-omics with a few informative features can facilitate interpretability and discovering relationships between variables.

Furthermore, BRR-ISB demonstrated no overfitting between training and validation, as depicted in Figure 5, which shows the loss values between training and validation in the BRR-ISB scenario with a bag size of 200 and ten iterations.

This study addresses overfitting through several approaches. Regularization is a feature embedded within BRR that prevents overfitting and enhances the model's ability to generalize, particularly when multicollinearity or high-dimensional data are present [89]. In addition, ISB works iteratively to select essential features by removing highly correlated

features, which prevents overfitting and leads to better generalization [90,91]. Furthermore, dropout is a technique designed to combat the overfitting problem [43]; dropout was utilized in the GCN-1CNN model.



**Figure 5.** The loss value between training and validation.

#### 4.4. Comparison with Related Works

The proposed method was benchmarked against four models that introduced notable approaches to predicting drug response:

1. Researchers utilized Weighted Graph Regularized Matrix Factorization (WGRMF) [92] to predict the responses of cell lines to anti-cancer drugs. This model used the CCLE, which has 491 cell lines and 23 drugs with 10,870 known responses. WGRMF employed gene expression and drug fingerprints as inputs for the model.
2. EBSRMF [81]: Researchers proposed Ensemble-based Similarity-Regularized Matrix Factorization, a bagging-based technique to enhance drug response prediction accuracy on the CCLE dataset. The dataset comprises 24 drugs and 363 types of cell lines. It utilized gene expression profiles and chemical structure.
3. DeepDSC [80]: Gene expression data were utilized to extract features of cell lines by a stacked deep autoencoder. Subsequently, the gene expression data were combined with chemical structure information to forecast drug response. DeepDSC utilized the Cancer Cell Line Encyclopedia (CCLE), which has 491 cell lines and 23 drugs, along with 10,870 documented responses.
4. SRMF [93]: Drug response prediction was accomplished by combining gene expression data with chemical structures using a Similarity-Regularized Matrix Factorization model. The CCLE dataset has 10,870 known responses, encompassing 491 distinct cell lines and 23 drugs.

Table 7 compares the BRR-ISB using 200 features as a bag size and ten iterations with the four models using the CCLE dataset.

**Table 7.** Comparison of performances with other related studies.

Model	RMSE	PCC	R <sup>2</sup>
WGRMF	0.56	0.72	-
EBSRMF	0.21	0.86	-
DeepDSC	0.23	-	78
SRMF	0.57	0.71	-
BRR-ISB (Proposed)	0.12	0.879	77

The proposed model demonstrated improvements by achieving the lowest RMSE compared with the other models. Additionally, BRR-ISB achieved the highest PCC compared with the other methods. DeepDSC showed the highest R<sup>2</sup> score compared with BRR-ISB.

## 5. Conclusions

This paper introduced BRR-ISB as a multi-omics integration method aimed at overcoming challenges such as dimensionality and the loss of biological context inherent in transformation-based methods. Bayesian Ridge Regression (BRR) was proposed as a supervised approach for feature selection in omics to reduce complexity and dimensionality. Features were selected based on domain-specific contexts, such as drug response and cancer classification. The Iterative Similarity Bagging (ISB) method was introduced to further reduce omics data's dimensionality and complexity without compromising the biological measurements. This methodology addresses the limitations associated with transformation-based integration methods. Various scenarios were employed to evaluate the efficacy of BRR-ISB in drug response prediction. The scenario used for BRR-ISB, utilizing a bag size of 200 features and conducting ten iterations, had superior performance in terms of RMSE, PCC, and R<sup>2</sup> when compared to other scenarios.

Furthermore, the execution time was reduced by 77% compared to the BRR scenario. SNF is a transformation-based integration method that converts omics types into a new representation. However, this transformation might decrease a model's interpretability, as the extracted features may lose their biological context.

Moreover, BRR-ISB, a method that effectively reduces the dimensionality and complexity of omics data while preserving biological measurements, shows improvement compared to SNF regarding RMSE, PCC, R<sup>2</sup>, and execution time. Furthermore, the BRR-ISB model demonstrated no overfitting between the training and validation datasets. This was observed by comparing the loss values in the BRR-ISB scenario, which involved 200 bag sizes and ten iterations.

However, there are some limitations of the BRR-ISB. BRR is a wrapper feature selection method that is considered time-consuming. It can be replaced by other filter feature selection methods suitable for small-scale datasets. Wrapper feature selection approaches are perceived as time-consuming since they include iterations, necessitate model training, and explore a vast search space [94]. These techniques entail the training and assessment of several models, each utilizing a distinct set of features. Cross-validation is typically employed to ensure the reliability and stability of the results. Exploring the extensive range of potential feature combinations can be computationally demanding, particularly for datasets with several features, resulting in a longer computation time. In addition, wrapper approaches can utilize iterative algorithms to continuously modify the feature subsets based on their impact on model performance [94,95].

Furthermore, features from the CCLE and GDSC were selected manually using either BRR-ISB or ISB alone. Therefore, the process of selecting the optimal number that represented informative features was not precise.

In future work, different filter feature selection methods need to be investigated as an alternative to Bayesian Ridge Regression that may reduce the time requirement and complexity of BRR. In addition, more data, such as methylation and metabolomics data, may need to be investigated to integrate them using ISB to improve drug response prediction. Furthermore, the current study utilized the half-mean as a threshold of the

ISB; different thresholds, such as mean, median, and first percentile Q1, will be tested and compared. Other distance metrics, such as the Manhattan distance and Pearson correlation, will also be studied to improve the effectiveness of ISB.

**Author Contributions:** Conceptualization, K.H.A. and N.A.A.; Methodology, T.M.A. and N.A.A.; Software, T.M.A.; Validation, T.M.A. and N.A.A.; Formal analysis, T.M.A.; Investigation, K.H.A.; Data curation, T.M.A.; Writing—original draft, T.M.A.; Writing—review & editing, K.H.A. and N.A.A.; Visualization, T.M.A.; Supervision, K.H.A. and N.A.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are openly available at <https://doi.org/10.3389/fgene.2019.00233> (accessed on 24 June 2024), reference number [54].

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; Anamika, K. Multi-Omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* **2020**, *14*, 117793221989905. [[CrossRef](#)] [[PubMed](#)]
- Chen, C.; Wang, J.; Pan, D.; Wang, X.; Xu, Y.; Yan, J.; Wang, L.; Yang, X.; Yang, M.; Liu, G. Applications of Multi-omics Analysis in Human Diseases. *MedComm* **2023**, *4*, e315. [[CrossRef](#)] [[PubMed](#)]
- Kreitmaier, P.; Katsoula, G.; Zeggini, E. Insights from Multi-Omics Integration in Complex Disease Primary Tissues. *Trends Genet.* **2023**, *39*, 46–58. [[CrossRef](#)] [[PubMed](#)]
- Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D.S.; Xia, J. MetaboAnalyst 4.0: Towards More Transparent and Integrative Metabolomics Analysis. *Nucleic Acids Res.* **2018**, *46*, W486–W494. [[CrossRef](#)] [[PubMed](#)]
- López de Maturana, E.; Alonso, L.; Alarcón, P.; Martín-Antoniano, I.A.; Pineda, S.; Piorno, L.; Calle, M.L.; Malats, N. Challenges in the Integration of Omics and Non-Omics Data. *Genes* **2019**, *10*, 238. [[CrossRef](#)] [[PubMed](#)]
- Cai, Z.; Poulos, R.C.; Liu, J.; Zhong, Q. Machine Learning for Multi-Omics Data Integration in Cancer. *iScience* **2022**, *25*, 103798. [[CrossRef](#)] [[PubMed](#)]
- Picard, M.; Scott-Boyer, M.-P.; Bodein, A.; Périn, O.; Droit, A. Integration Strategies of Multi-Omics Data for Machine Learning Analysis. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3735–3746. [[CrossRef](#)]
- Hasin, Y.; Seldin, M.; Lusis, A. Multi-Omics Approaches to Disease. *Genome Biol.* **2017**, *18*, 83. [[CrossRef](#)] [[PubMed](#)]
- Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial Sequencing and Analysis of the Human Genome. *Nature* **2001**, *409*, 860–921. [[CrossRef](#)]
- Reel, P.S.; Reel, S.; Pearson, E.; Trucco, E.; Jefferson, E. Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review. *Biotechnol. Adv.* **2021**, *49*, 107739. [[CrossRef](#)]
- Almutiri, T.; Alomar, K.; Alganmi, N. Predicting Drug Response on Multi-Omics Data Using a Hybrid of Bayesian Ridge Regression with Deep Forest. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 470–482. [[CrossRef](#)]
- Nicora, G.; Vitali, F.; Dagliati, A.; Geifman, N.; Bellazzi, R. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. *Front. Oncol.* **2020**, *10*, 1030. [[CrossRef](#)] [[PubMed](#)]
- Xuan, P.; Sun, C.; Zhang, T.; Ye, Y.; Shen, T.; Dong, Y. Gradient Boosting Decision Tree-Based Method for Predicting Interactions Between Target Genes and Drugs. *Front. Genet.* **2019**, *10*, 459. [[CrossRef](#)] [[PubMed](#)]
- Yue, X.; Wang, Z.; Huang, J.; Parthasarathy, S.; Moosavinasab, S.; Huang, Y.; Lin, S.M.; Zhang, W.; Zhang, P.; Sun, H. Graph Embedding on Biomedical Networks: Methods, Applications and Evaluations. *Bioinformatics* **2020**, *36*, 1241–1251. [[CrossRef](#)]
- Ma, T.; Zhang, A. Affinity Network Fusion and Semi-Supervised Learning for Cancer Patient Clustering. *Methods* **2018**, *145*, 16–24. [[CrossRef](#)] [[PubMed](#)]
- Gligorijević, V.; Barot, M.; Bonneau, R. DeepNF: Deep Network Fusion for Protein Function Prediction. *Bioinformatics* **2018**, *34*, 3873–3881. [[CrossRef](#)] [[PubMed](#)]
- Wen, Y.; Song, X.; Yan, B.; Yang, X.; Wu, L.; Leng, D.; He, S.; Bo, X. Multi-Dimensional Data Integration Algorithm Based on Random Walk with Restart. *BMC Bioinform.* **2021**, *22*, 97. [[CrossRef](#)] [[PubMed](#)]
- Zhang, Y.; Li, A.; Peng, C.; Wang, M. Improve Glioblastoma Multifforme Prognosis Prediction by Using Feature Selection and Multiple Kernel Learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2016**, *13*, 825–835. [[CrossRef](#)]
- He, Z.; Zhang, J.; Yuan, X.; Zhang, Y. Integrating Somatic Mutations for Breast Cancer Survival Prediction Using Machine Learning Methods. *Front. Genet.* **2021**, *11*, 632901. [[CrossRef](#)]

20. Ammad-ud-din, M.; Khan, S.A.; Malani, D.; Murumägi, A.; Kallioniemi, O.; Aittokallio, T.; Kaski, S. Drug Response Prediction by Inferring Pathway-Response Associations with Kernelized Bayesian Matrix Factorization. *Bioinformatics* **2016**, *32*, i455–i463. [[CrossRef](#)]
21. Costello, J.C.; Heiser, L.M.; Georgii, E.; Gönen, M.; Menden, M.P.; Wang, N.J.; Bansal, M.; Ammad-ud-din, M.; Hintsanen, P.; Khan, S.A.; et al. A Community Effort to Assess and Improve Drug Sensitivity Prediction Algorithms. *Nat. Biotechnol.* **2014**, *32*, 1202–1212. [[CrossRef](#)] [[PubMed](#)]
22. Vahabi, N.; Michailidis, G. Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review. *Front. Genet.* **2022**, *13*, 854752. [[CrossRef](#)] [[PubMed](#)]
23. Gligorijević, V.; Pržulj, N. Methods for Biological Data Integration: Perspectives and Challenges. *J. R. Soc. Interface* **2015**, *12*, 20150571. [[CrossRef](#)] [[PubMed](#)]
24. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nat. Methods* **2014**, *11*, 333–337. [[CrossRef](#)] [[PubMed](#)]
25. Efendi, A.; Effrihan, E. A Simulation Study on Bayesian Ridge Regression Models for Several Collinearity Levels. *AIP Conf. Proc.* **2017**, *1913*, 020031.
26. Yassen, M.F.; Al-Duais, F.S.; Almazah, M. Ridge Regression Method and Bayesian Estimators under Composite LINEX Loss Function to Estimate the Shape Parameter in Lomax Distribution. *Comput. Intell. Neurosci.* **2022**, *2022*, 1200611. [[CrossRef](#)] [[PubMed](#)]
27. Flavin, T.; Steiner, T.; Mitra, B.; Nagaraju, V. Bayesian Ridge Regression Based Model to Predict Fault Location in HVdc Network. In Proceedings of the 2022 IEEE Power & Energy Society General Meeting (PESGM), Denver, CO, USA, 17–21 July 2022; pp. 1–5.
28. Ngo, G.; Beard, R.; Chandra, R. Evolutionary Bagging for Ensemble Learning. *Neurocomputing* **2022**, *510*, 1–14. [[CrossRef](#)]
29. Tološi, L.; Lengauer, T. Classification with Correlated Features: Unreliability of Feature Ranking and Solutions. *Bioinformatics* **2011**, *27*, 1986–1994. [[CrossRef](#)]
30. Jain, I.; Jain, V.K.; Jain, R. Correlation Feature Selection Based Improved-Binary Particle Swarm Optimization for Gene Selection and Cancer Classification. *Appl. Soft Comput.* **2018**, *62*, 203–215. [[CrossRef](#)]
31. Darst, B.F.; Malecki, K.C.; Engelman, C.D. Using Recursive Feature Elimination in Random Forest to Account for Correlated Variables in High Dimensional Data. *BMC Genet.* **2018**, *19*, 65. [[CrossRef](#)]
32. Misra, B.B.; Langefeld, C.; Olivier, M.; Cox, L.A. Integrated Omics: Tools, Advances and Future Approaches. *J. Mol. Endocrinol.* **2019**, *62*, R21–R45. [[CrossRef](#)] [[PubMed](#)]
33. Wörheide, M.A.; Krumsiek, J.; Kastenmüller, G.; Arnold, M. Multi-Omics Integration in Biomedical Research—A Metabolomics-Centric Review. *Anal. Chim. Acta* **2021**, *1141*, 144–162. [[CrossRef](#)]
34. Park, M.; Kim, D.; Moon, K.; Park, T. Integrative Analysis of Multi-Omics Data Based on Blockwise Sparse Principal Components. *Int. J. Mol. Sci.* **2020**, *21*, 8202. [[CrossRef](#)]
35. Xie, G.; Dong, C.; Kong, Y.; Zhong, J.; Li, M.; Wang, K. Group Lasso Regularized Deep Learning for Cancer Prognosis from Multi-Omics and Clinical Features. *Genes* **2019**, *10*, 240. [[CrossRef](#)] [[PubMed](#)]
36. Xie, M.; Lei, X.; Zhong, J.; Ouyang, J.; Li, G. Drug Response Prediction Using Graph Representation Learning and Laplacian Feature Selection. *BMC Bioinform.* **2022**, *23*, 532. [[CrossRef](#)] [[PubMed](#)]
37. Chu, T.; Nguyen, T.T.; Hai, B.D.; Nguyen, Q.H.; Nguyen, T. Graph Transformer for Drug Response Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 1065–1072. [[CrossRef](#)] [[PubMed](#)]
38. Malik, V.; Kalakoti, Y.; Sundar, D. Deep Learning Assisted Multi-Omics Integration for Survival and Drug-Response Prediction in Breast Cancer. *BMC Genom.* **2021**, *22*, 214. [[CrossRef](#)]
39. Wang, Z.; Li, H.; Carpenter, C.; Guan, Y. Challenge-Enabled Machine Learning to Drug-Response Prediction. *AAPS J.* **2020**, *22*, 106. [[CrossRef](#)]
40. Bühlmann, P.; Van De Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011; ISBN 364220192X.
41. Bøvelstad, H.M.; Nygård, S.; Størvold, H.L.; Aldrin, M.; Borgan, Ø.; Frigessi, A.; Lingjærde, O.C. Predicting Survival from Microarray Data—A Comparative Study. *Bioinformatics* **2007**, *23*, 2080–2087. [[CrossRef](#)]
42. Natekin, A.; Knoll, A. Gradient Boosting Machines, a Tutorial. *Front. Neurobot.* **2013**, *7*, 21. [[CrossRef](#)]
43. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
44. Partin, A.; Brettin, T.; Evrard, Y.A.; Zhu, Y.; Yoo, H.; Xia, F.; Jiang, S.; Clyde, A.; Shukla, M.; Fonstein, M. Learning Curves for Drug Response Prediction in Cancer Cell Lines. *BMC Bioinform.* **2021**, *22*, 252. [[CrossRef](#)] [[PubMed](#)]
45. Chang, Y.; Park, H.; Yang, H.-J.; Lee, S.; Lee, K.-Y.; Kim, T.S.; Jung, J.; Shin, J.-M. Cancer Drug Response Profile Scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. *Sci. Rep.* **2018**, *8*, 8857. [[CrossRef](#)] [[PubMed](#)]
46. Zhu, Y.; Brettin, T.; Evrard, Y.A.; Partin, A.; Xia, F.; Shukla, M.; Yoo, H.; Doroshow, J.H.; Stevens, R.L. Ensemble Transfer Learning for the Prediction of Anti-Cancer Drug Response. *Sci. Rep.* **2020**, *10*, 18040. [[CrossRef](#)] [[PubMed](#)]
47. Sotudian, S.; Paschalidis, I.C. Machine Learning for Pharmacogenomics and Personalized Medicine: A Ranking Model for Drug Sensitivity Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *19*, 2324–2333. [[CrossRef](#)] [[PubMed](#)]

48. Roder, J.; Oliveira, C.; Net, L.; Tsy-pin, M.; Linstid, B.; Roder, H. A Dropout-Regularized Classifier Development Approach Optimized for Precision Medicine Test Discovery from Omics Data. *BMC Bioinform.* **2019**, *20*, 325. [[CrossRef](#)] [[PubMed](#)]
49. Xiaolin, X.; Xiaozhi, L.; Guoping, H.; Hongwei, L.; Jinkuo, G.; Xiyun, B.; Zhen, T.; Xiaofang, M.; Yanxia, L.; Na, X. Overfit Deep Neural Network for Predicting Drug-Target Interactions. *iScience* **2023**, *26*, 107646. [[CrossRef](#)]
50. Iorio, F.; Knijnenburg, T.A.; Vis, D.J.; Bignell, G.R.; Menden, M.P.; Schubert, M.; Aben, N.; Gonçalves, E.; Barthorpe, S.; Lightfoot, H.; et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **2016**, *166*, 740–754. [[CrossRef](#)] [[PubMed](#)]
51. Kurilov, R.; Haibe-Kains, B.; Brors, B. Assessment of Modelling Strategies for Drug Response Prediction in Cell Lines and Xenografts. *Sci. Rep.* **2020**, *10*, 2849. [[CrossRef](#)]
52. Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A.A.; Kim, S.; Wilson, C.J.; Lehár, J.; Kryukov, G.V.; Sonkin, D.; et al. The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature* **2012**, *483*, 603–607. [[CrossRef](#)]
53. Yang, W.; Soares, J.; Greninger, P.; Edelman, E.J.; Lightfoot, H.; Forbes, S.; Bindal, N.; Beare, D.; Smith, J.A.; Thompson, I.R. Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells. *Nucleic Acids Res.* **2012**, *41*, D955–D961. [[CrossRef](#)] [[PubMed](#)]
54. Xu, X.; Gu, H.; Wang, Y.; Wang, J.; Qin, P. Autoencoder Based Feature Selection Method for Classification of Anticancer Drug Response. *Front. Genet.* **2019**, *10*, 233. [[CrossRef](#)] [[PubMed](#)]
55. Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202–D1213. [[CrossRef](#)] [[PubMed](#)]
56. O’Boyle, N.M. Towards a Universal SMILES Representation—A Standard Method to Generate Canonical SMILES Based on the InChI. *J. Cheminform.* **2012**, *4*, 22. [[CrossRef](#)]
57. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [[CrossRef](#)]
58. Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput. Aided Mol. Des.* **2016**, *30*, 595–608. [[CrossRef](#)] [[PubMed](#)]
59. Goh, G.B.; Siegel, C.; Vishnu, A.; Hodas, N. Using Rule-Based Labels for Weak Supervised Learning: A ChemNet for Transferable Chemical Property Prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 302–310.
60. Landrum, G. Rdkit: Open-Source Cheminformatics Software. 2016. Volume 149. p. 650. Available online: <http://www.rdkit.org/> (accessed on 24 June 2024).
61. Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*; O’Reilly Media, Inc.: Sebastopol, CA, USA, 2019; ISBN 1492039780.
62. Nguyen, T.; Nguyen, G.T.T.; Nguyen, T.; Le, D.-H. Graph Convolutional Networks for Drug Response Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *19*, 146–154. [[CrossRef](#)]
63. Wu, Z.; Ramsundar, B.; Feinberg, E.N.; Gomes, J.; Geniesse, C.; Pappu, A.S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530. [[CrossRef](#)] [[PubMed](#)]
64. Fernández, I.; Frenking, G.; Merino, G. Aromaticity of Metallabenzenes and Related Compounds. *Chem. Soc. Rev.* **2015**, *44*, 6452–6463. [[CrossRef](#)]
65. Tipping, M.E. Sparse Bayesian Learning and the Relevance Vector Machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
66. Bishop, C.M.; Nasrabadi, N.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4.
67. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67. [[CrossRef](#)]
68. Neal, R.M. *Bayesian Learning for Neural Networks*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 118, ISBN 1461207452.
69. MacKay, D.J.C. Bayesian Interpolation. *Neural Comput.* **1992**, *4*, 415–447. [[CrossRef](#)]
70. Ozdemir, S.; Susarla, D. *Feature Engineering Made Easy: Identify Unique Features from Your Dataset in Order to Build Powerful Machine Learning Systems*; Packt Publishing Ltd.: Birmingham, UK, 2018; ISBN 1787286479.
71. Tancredi, A.; Anderson, C.; O’Hagan, A. Accounting for Threshold Uncertainty in Extreme Value Estimation. *Extremes* **2006**, *9*, 87–106. [[CrossRef](#)]
72. Goodspeed, A.; Heiser, L.M.; Gray, J.W.; Costello, J.C. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. *Mol. Cancer Res.* **2016**, *14*, 3–13. [[CrossRef](#)] [[PubMed](#)]
73. Gambardella, V.; Tarazona, N.; Cejalvo, J.M.; Lombardi, P.; Huerta, M.; Roselló, S.; Fleitas, T.; Roda, D.; Cervantes, A. Personalized Medicine: Recent Progress in Cancer Therapy. *Cancers* **2020**, *12*, 1009. [[CrossRef](#)] [[PubMed](#)]
74. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
75. Joseph, V.R. Optimal Ratio for Data Splitting. *Stat. Anal. Data Min. ASA Data Sci. J.* **2022**, *15*, 531–538. [[CrossRef](#)]
76. Dunford, R.; Su, Q.; Tamang, E. The Pareto Principle. 2014. Available online: <https://core.ac.uk/download/pdf/200202097.pdf> (accessed on 24 June 2024).
77. Nti, I.K.; Nyarko-Boateng, O.; Aning, J. Performance of Machine Learning Algorithms with Different K Values in K-Fold Cross-Validation. *Int. J. Inf. Technol. Comput. Sci.* **2021**, *13*, 61–71.

78. Wong, T.-T.; Yeh, P.-Y. Reliable Accuracy Estimates from K-Fold Cross Validation. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 1586–1594. [[CrossRef](#)]
79. Liu, Q.; Hu, Z.; Jiang, R.; Zhou, M. DeepCDR: A Hybrid Graph Convolutional Network for Predicting Cancer Drug Response. *Bioinformatics* **2020**, *36*, i911–i918. [[CrossRef](#)]
80. Li, M.; Wang, Y.; Zheng, R.; Shi, X.; Li, Y.; Wu, F.-X.; Wang, J. DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 575–582. [[CrossRef](#)] [[PubMed](#)]
81. Shahzad, M.; Tahir, M.A.; Khan, M.A.; Jiang, R.; Malick, R.A.S. EBSRMF: Ensemble Based Similarity-Regularized Matrix Factorization to Predict Anticancer Drug Responses. *J. Intell. Fuzzy Syst.* **2022**, *43*, 3443–3452. [[CrossRef](#)]
82. Golbraikh, A.; Tropsha, A. Beware of Q2! *J. Mol. Graph. Model.* **2002**, *20*, 269–276. [[CrossRef](#)] [[PubMed](#)]
83. Zhao, L.; Qiu, T.; Jiang, D.; Xu, H.; Zou, L.; Yang, Q.; Chen, C.; Jiao, B. SGCE Promotes Breast Cancer Stem Cells by Stabilizing EGFR. *Adv. Sci.* **2020**, *7*, 1903700. [[CrossRef](#)] [[PubMed](#)]
84. Zhang, S.; Wang, H.; Liu, A. Identification of ATP1B1, a Key Copy Number Driver Gene in Diffuse Large B-Cell Lymphoma and Potential Target for Drugs. *Ann. Transl. Med.* **2022**, *10*, 1136. [[CrossRef](#)] [[PubMed](#)]
85. Katuwal, N.B.; Kang, M.S.; Ghosh, M.; Hong, S.D.; Jeong, Y.G.; Park, S.M.; Kim, S.-G.; Sohn, J.; Kim, T.H.; Moon, Y.W. Targeting PEG10 as a Novel Therapeutic Approach to Overcome CDK4/6 Inhibitor Resistance in Breast Cancer. *J. Exp. Clin. Cancer Res.* **2023**, *42*, 325. [[CrossRef](#)] [[PubMed](#)]
86. Xu, Z.; Xiang, L.; Peng, L.; Gu, H.; Wang, Y. Comprehensive Analysis of the Immune Implication of AKAP12 in Stomach Adenocarcinoma. *Comput. Math. Methods Med.* **2022**, *2022*, 3445230. [[CrossRef](#)]
87. Lodi, M.; Voilquin, L.; Alpy, F.; Molière, S.; Reix, N.; Mathelin, C.; Chenard, M.-P.; Tomasetto, C.-L. STARD3: A New Biomarker in HER2-Positive Breast Cancer. *Cancers* **2023**, *15*, 362. [[CrossRef](#)] [[PubMed](#)]
88. Shen, R.; Mo, Q.; Schultz, N.; Seshan, V.E.; Olshen, A.B.; Huse, J.; Ladanyi, M.; Sander, C. Integrative Subtype Discovery in Glioblastoma Using ICluster. *PLoS ONE* **2012**, *7*, e35236. [[CrossRef](#)]
89. Bishop, C.M.; Tipping, M.E. Bayesian Regression and Classification. *Nato Sci. Ser. Sub Ser. III Comput. Syst. Sci.* **2003**, *190*, 267–288.
90. Ying, X. An Overview of Overfitting and Its Solutions. *J. Phys. Conf. Ser.* **2019**, *1168*, 022022. [[CrossRef](#)]
91. Zhang, Z.; Zhang, Y.; Li, Z. Removing the Feature Correlation Effect of Multiplicative Noise. *Adv. Neural Inf. Process. Syst.* **2018**, *31*. Available online: [https://papers.nips.cc/paper\\_files/paper/2018/hash/e7b24b112a44fdd9ee93bdf998c6ca0e-Abstract.html](https://papers.nips.cc/paper_files/paper/2018/hash/e7b24b112a44fdd9ee93bdf998c6ca0e-Abstract.html) (accessed on 24 June 2024).
92. Guan, N.-N.; Zhao, Y.; Wang, C.-C.; Li, J.-Q.; Chen, X.; Piao, X. Anticancer Drug Response Prediction in Cell Lines Using Weighted Graph Regularized Matrix Factorization. *Mol. Ther. Nucleic Acids* **2019**, *17*, 164–174. [[CrossRef](#)] [[PubMed](#)]
93. Wang, L.; Li, X.; Zhang, L.; Gao, Q. Improved Anticancer Drug Response Prediction in Cell Lines Using Matrix Factorization with Similarity Regularization. *BMC Cancer* **2017**, *17*, 513. [[CrossRef](#)] [[PubMed](#)]
94. Kohavi, R.; John, G.H. Wrappers for Feature Subset Selection. *Artif. Intell.* **1997**, *97*, 273–324. [[CrossRef](#)]
95. Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.