biomolecules

MDPI

Article

G2PDeep-v2: A Web-Based Deep-Learning Framework for Phenotype Prediction and Biomarker Discovery for All Organisms Using Multi-Omics Data

Shuai Zeng [1,2], Trinath Adusumilli [1], Sania Zafar Awan [3], Manish Sridhar Immadi [1], Dong Xu [1,2,3] and Trupti Joshi [1,2,3,4,5,*]

[1] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA; zengs@missouri.edu (S.Z.); tafbr@missouri.edu (T.A.); mizy9@missouri.edu (M.S.I.); xudong@missouri.edu (D.X.)
[2] Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
[3] Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA; sah2p@missouri.edu
[4] Department of Biomedical Informatics, Biostatistics and Medical Epidemiology, University of Missouri, Columbia, MO 65211, USA
[5] Department of Biomedical Sciences, Joan C. Edwards School of Medicine, Marshall University, Huntington, WV 25703, USA
* Correspondence: joshitr@marshall.edu; Tel.: +1-(573)-884-5963

Abstract

Multi-omics data offers rich insights into complex traits across organisms, yet integrating and analyzing these datasets for phenotype prediction and marker discovery remains challenging. Researchers need accessible tools that combine deep learning, hyperparameter optimization, visualization, and downstream analysis in a unified web platform. To address this, we developed G2PDeep-v2, a web-based platform powered by deep learning for phenotype prediction and marker discovery from multi-omics data across a wide range of organisms, including humans and plants. The server provides multiple services for researchers to create deep-learning models through an interactive interface and train these models using an automated hyperparameter tuning algorithm on high-performance computing resources. Users can visualize the results of phenotype and markers predictions and perform Gene Set Enrichment Analysis for the significant markers to provide insights into the molecular mechanisms underlying complex diseases, conditions and other biological phenotypes being studied.

Keywords: multi-omics; biomarker; phenotype prediction; deep learning; automated hyperparameters tunning; reproducibility; web-platform

Academic Editors: Mohsin Saleet Jafri and Frank Krause

Received: 9 October 2025
Revised: 13 November 2025
Accepted: 28 November 2025
Published: 1 December 2025

Citation: Zeng, S.; Adusumilli, T.; Awan, S.Z.; Immadi, M.S.; Xu, D.; Joshi, T. G2PDeep-v2: A Web-Based Deep-Learning Framework for Phenotype Prediction and Biomarker Discovery for All Organisms Using Multi-Omics Data. *Biomolecules* **2025**, *15*, 1673. https://doi.org/10.3390/biom15121673

1. Introduction

With the advances in molecular profiling technologies, the ability to observe large-scale multi-omics data from patients or other biological organisms has grown remarkably over the past decade. Genome-wide data encompassing various molecular processes, such as gene expression, microRNA (miRNA) expression, protein expression, DNA methylation, single nucleotide polymorphisms (SNP), and copy number variations (CNV), can be obtained for the same set of samples, resulting in multi-omics data for numerous disease and crop studies. Although each type of multi-omics data captures a portion of the biological information, integrating multi-omics data helps researchers comprehensively

understand biological systems from different perspectives [1,2]. Researchers have utilized multi-omics data to address many significant breeding and biomedical problems, including plant breeding [3], drug target discovery [4], disease therapy [5,6], and survival analysis. Specifically, muti-omics data allows researchers to predict the phenotypes and identify biomarkers that affect the diversity of phenotypes. To effectively take advantage of complementary information in multi-omics data, it is important to have a one-stop-shop platform for researchers to integrate multi-omics data, train customized deep-learning models for predicting phenotypes using high-performance computing resources and discover potential biomarkers along with their biological relevance.

Many approaches have been proposed over the past decade to utilize one type of omics data analysis for various bioinformatics problems. Early attempts have employed supervised learning methods for biomedical classification tasks. For example, DeepGS [7] applies a deep convolutional neural network combined with a fully connected neural network to predict phenotype based on SNP. Blaise et al. [8] proposed an approach for the biological interpretation of deep learning models for phenotype prediction from gene expression data. However, these methods only consider one of the multi-omics data types and fail to utilize useful biological information from other types of multi-omics data. Recently, more supervised methods focused on exploiting the interactions across different omics data types for better prediction. MOGONET [9] integrates multi-omics data using graph convolutional networks for biomedical classification tasks such as Alzheimer's disease patient classification and kidney cancer type classification. Sammut et al. [10] introduced an ensemble-based machine learning framework to integrate representations from different multi-omics data types for breast cancer therapy response. Some efforts focus on biologically informed deep learning models with multi-omics data to enhance the interpretability of models [11–13].

Although these methods have shown some good performance, there are still challenges in adopting such models in different types of studies. The models used in these methods are typically designed for a specific study with a particular set of data, which means that researchers must invest considerable effort to adapt the model for other studies, as they are not generalizable. Inappropriate hyperparameter optimization is a common issue, which often negatively affects the performance of model and analytical outcomes. In other words, manually tuning the optimal hyperparameters is challenging due to the vast number of possible combinations. These methods have steep learning curves and often require complicated installation steps. Furthermore, training models with large-scale multi-omics data requires computing resources and storage exceeding the capacities of most potential non-computer savvy users. Moreover, few existing methods integrate functionalities to identify significant multi-omics signatures and biomarkers related to biomedical and biological studies, resulting in researchers spending additional time on confirming evidence for the findings.

Along this line of research, we have been developing the deep learning method G2PDeep. The first original v1 model was made available in 2019 [14], followed by the web server published in 2021 [15]. In its first version, G2PDeep enabled the quantitative phenotype prediction and marker discovery by using a dual-CNN model trained from scratch using only SNP. This work has gained a lot of interest from researchers worldwide, with more than 500 submissions for model training conducted via web-based access. To address the limitations discussed above, we have further expanded it to G2PDeep-v2, a comprehensive web-based platform for phenotype prediction using multi-omics data and biomarkers discovery for all organisms. Unlike the previous version of G2PDeep, the new version, G2PDeep-v2, now supports multiple inputs for multi-omics data, offers a broader array of model selection options, advanced settings for tuning model hyperparameters, and

includes comprehensive Gene Set Enrichment Analysis (GSEA) functionalities. The difference between the previous and the new version of G2PDeep is clearly depicted in Table 1. Precisely, compared with other available applications, G2PDeep-v2 provides end-to-end management of machine learning projects from multi-omics dataset creation through to model interpretation, which also supports individual omics or any combination of up to three multi-omics data for the predictions. It is equipped with a fully automated pipeline to process and organize multi-omics data such as gene expression, miRNA expression, DNA methylation, protein expression SNP, and CNV. It provides an interactive web interface enabling machine learning and deep learning models to be created, and customized predictions according to different research tasks. It also provides automated hyperparameters search with Bayesian optimization algorithm, discovering a top-performing model configuration from a huge number of combinations of hyperparameters, without any manual effort necessary beyond just the initial set-up. It supports real time monitoring for ongoing model training and optimization history through a real-time web dashboard. The G2PDeep-v2 server is publicly available at https://g2pdeep.org/ and can be utilized for all organisms.

Table 1. Comparison of functionalities between the previous and latest versions of G2PDeep.

Categories	Functionality	G2PDeep-v1	G2PDeep-v2
Dataset creation	single nucleotide polymorphisms (SNP)/Zygosity	✔	✔
	gene expression		✔
	copy number variation (CNV)		✔
	Protein expression		✔
	microRNA (miRNA) expression		✔
	DNA Methylation		✔
Custom models	dual-CNN/multi-CNN	✔	✔
	Support Vector Machine (SVM)		✔
	Logistic Regression (LR)		✔
	Random Forest (RF)		✔
	Decision Tree (DT)		✔
	Multiple inputs		✔
Task	Regression	✔	✔
	Classification		✔
Model training	Online training	✔	✔
	Training monitoring	✔	✔
	Automate hyperparameter tunning		✔
	Hyperparameter tunning monitoring		✔
Online prediction	Prediction with test dataset	✔	✔
Marker discovery	Identifying significant markers	✔	✔
	GSEA with KEGG/Reactome		✔
	Studies related to significant markers		✔

The datasets and well-trained models are serialized and stored in user accounts to protect the privacy of research information from unauthorized parties. The well-trained models can be retrieved from a pool of models to predict the phenotype and discover the significant biomarkers associated with the phenotype, making the models reusable

and reproducible. The predicted results of phenotype are summarized in an interactive figure, and its raw results can be downloaded as a comma-separated values (CSV) file. The GSEA can be performed using significant biomarkers, Kyoto Encyclopedia of Genes and Genomes (KEGG) [16] and Reactome [17] pathway information, providing insights into pathways underlying the phenotype. The publications strongly associated with significant biomarkers in phenotype of user's interest are listed in a table along with their abstracts and URL links, identifying the newest evidence from relevant research for the researchers.

Here, we present our multi-omics datasets exemplar studies for 23 different cancer with long-term-survival labels, originally provided by The Cancer Genome Atlas (TCGA) project [18] for biomedical applications and Soybean Cyst Nematode (SCN) resistance prediction in soybean for agribiotech application. We have utilized G2PDeep-v2 to train models with automating hyperparameters search on different combinations of multi-omics data and identified multiple sets of significant biomarkers. All these datasets, models, biomarkers with GSEA results are retrievable for all users and visitors. We demonstrated that G2PDeep-v2 can identify significant biomarkers associated with patient outcomes, as well as markers linked to resistance traits. To the best of our knowledge, G2PDeep-v2 is the only end-to-end, web-based deep-learning framework that supports phenotype prediction, biomarker discovery, and functional annotation from multi-omics data across diverse organisms, lowering the barriers for utilization and application of deep learning techniques, especially for non-informatics and computer savvy users, new to such techniques. Users can apply G2PDeep-v2 not only to human disease studies but also to other organisms including research in plants. The G2PDeep-v2 server is publicly available at https://g2pdeep.org (accessed on 1 August 2023).

2. Materials and Methods

2.1. Data Pre-Processing

To enhance the scalability of the dataset, G2PDeep-v2 employs one-hot encoding and normalization individually on six different types of omics data: gene expression, miRNA expression, DNA methylation, protein expression, SNP, and CNV. Regarding features in expression data, such as gene expression, miRNA expression, DNA methylation, and protein expression, the values in each sample undergo normalization through z-score normalization. Focusing on DNA methylation data, only CpG islands occurring in promoter regions or genes are included. For SNP data, the four genotypes (adenine (A), thymine (T), cytosine (C), and guanine (G)) and missing data undergo encoding through one-hot binary encoding. In the case of gene-level CNV data, the encoding includes homozygous deletion, single copy deletion, diploid normal copy, low-level copy number amplification, and high-level copy number amplification, utilizing one-hot binary encoding. Notably, missing values for expression data are set to 0, while none of the SNP and CNV datasets undergo any imputation process. For each input modality, G2PDeep-v2 automatically verifies the data format to ensure compatibility with the platform. This includes checking feature types, value ranges, and dimensional consistency. The system provides informative feedback if any discrepancies are detected, preventing misclassification of data types or incorrect combination of different omics modalities, and ensuring that integration and model training are performed accurately.

2.2. Modeling in G2PDeep

2.2.1. Multi-CNN

Our proposed multi-CNN is an extended version of the dual-CNN reported in our previous work [14,15]. The multi-CNN model (as shown in Figure 1) takes up to three types of omics data combinations as input. The model consists of multiple parallel CNN

layers and a fully connected neural network. The encoded genotypes for each type of omics are individually passed into multiple parallel CNN layers. These layers generate representations for each type of omics data to discover patterns and provide a better understanding of the biomarkers. The representations for each type of omics are concatenated, integrating the information of biomarkers from different perspectives. The concatenated representations are then passed into the fully connected neural network with an output layer for phenotype prediction. To prevent the model overfitting, a Batch Normalization [19] layer is added at the end of representation and Dropout [20] layers are added in each layer of fully connected neural network. The Leaky Rectified Linear Unit (Leaky-ReLU) [21] activation function is added to each layer of model. The loss function of the model is cross-entropy and mean squared error for categorical phenotype and quantitative phenotype prediction, respectively. The model is optimized by Adam [22], an adaptive learning rate optimization algorithm. In TCGA cancer studies, the output of the model is a vector of probabilities converted by the Softmax function, representing the probability to LTS or non-LTS.

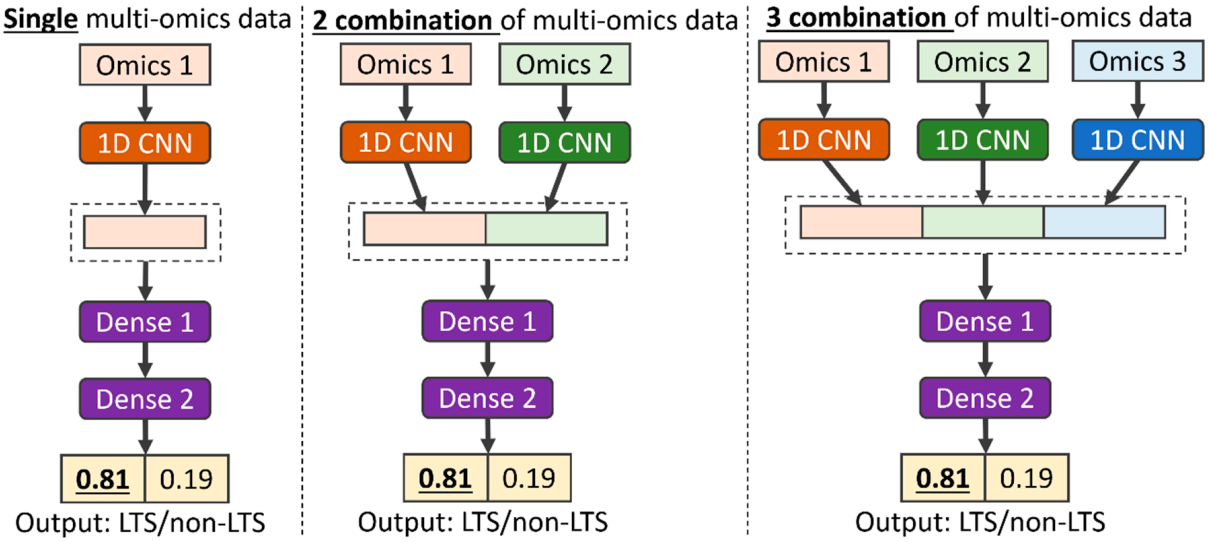

Figure 1. An example architecture of the multi-CNN model designed for long-term survival prediction using input data with single, two combinations, and three combinations of multi-omics data.

2.2.2. Traditional Machine Learning Models

G2PDeep-v2 integrates various traditional machine learning methods, such as Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) for providing easy access to the commonly used tools and to serve as benchmark models for comparison with deep-learning approaches. The input for these models is a vector of values concatenated from each type of omics. For logistic regression, it uses an L2 penalty term to deal with multicollinearity problems and penalize insignificant biomarkers. The SVM model uses the radial basis function (RBF) kernel, which makes the data separable using a hyperplane by projecting non-linearly separable data into higher-dimensional space. The decision tree, a nonparametric machine learning algorithm, facilitates training the data without strong assumptions or prior knowledge. The random forest, an ensemble learning method, can handle both linear and non-linear types of data.

2.3. Biomarkers Discovery and Annotation

The significant biomarkers associated with phenotypes of interest to researchers are estimated using models in G2PDeep-v2. Based on our previous work [14], the saliency map algorithm is applied to the multi-CNN to identify SNPs highly associated with the phenotype, while the coefficients of traditional ML models are also utilized to pinpoint

significant biomarkers. Biomarkers with higher estimated values are considered significant. To facilitate the functional annotation of these identified significant biomarkers, the Gene Set Enrichment Analysis (GSEA) function of GSEApy [23] (version 1.1.11), a Python library, is employed.

2.4. Web Server Implementation

G2PDeep-v2 is developed using Model-View-Controller (MVC) architectural pattern and deployed in Docker. This containerized deployment is hosted on a server equipped with an Intel(R) Xeon(R) Gold 6248 CPU and 384 GB of memory, signifying a robust computing environment capable of efficiently handling the computational demands of G2PDeep-v2. G2PDeep-v2 is designed to provide users with a clean and orderly appearance of interface components, reducing the chances of faulty operations and improving user experience. It utilizes high-performance computing resources to guarantee efficient, sustainable, and reliable services with a high volume of tasks. The architectural framework of G2PDeep comprises four modules, complemented by a security policy as illustrated in Figure 2.

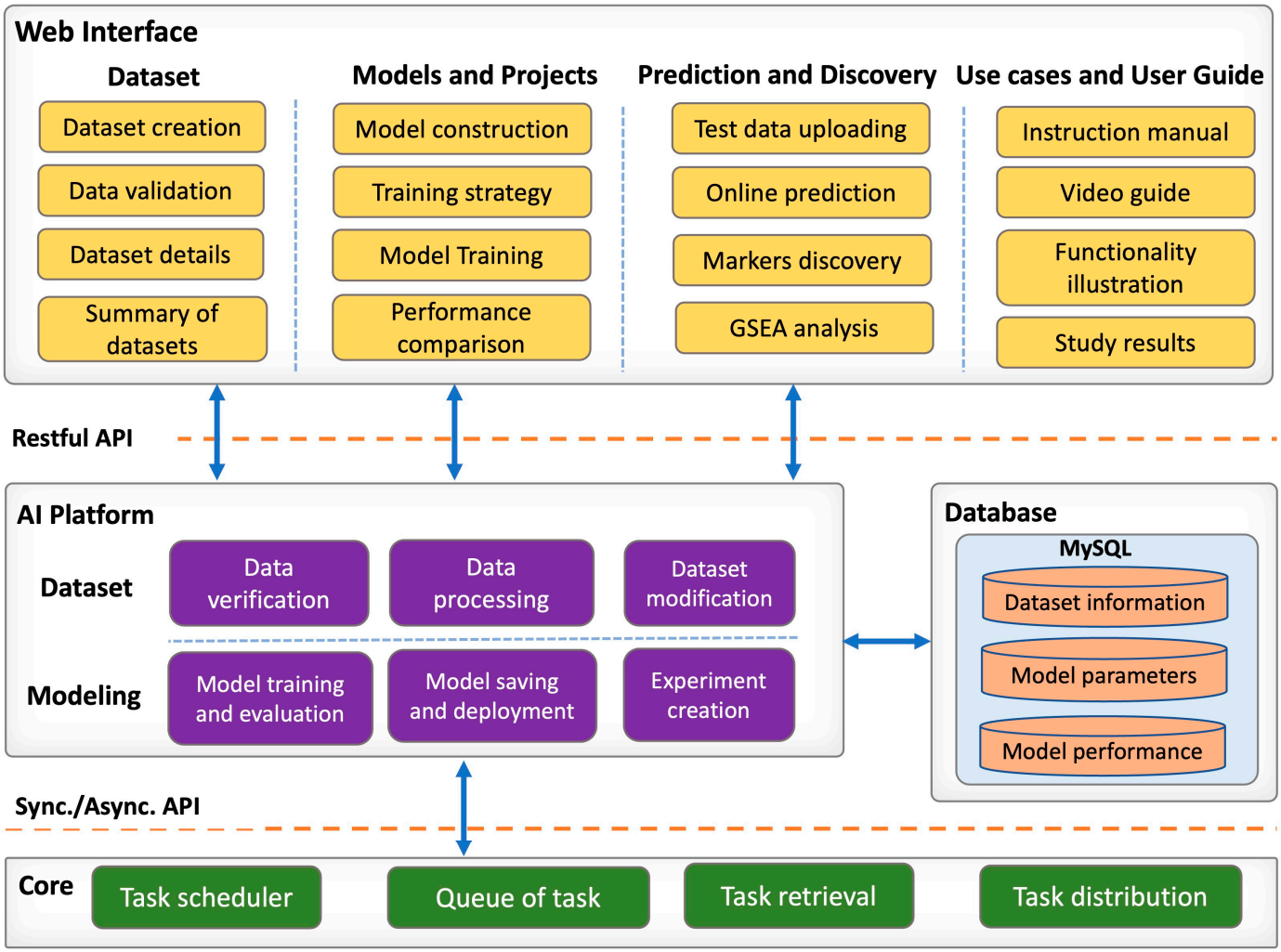

Figure 2. The architecture of G2PDeep. The architecture consists of four modules, and these modules communicate with each other via appropriate APIs, as indicated by the blue arrows.

2.4.1. Web Interface Module

G2PDeep-v2 provides user-friendly web interface developed using ReactJS [24] and Material UI [25], enterprise-level user interface (UI) libraries. It is designed to be responsive

and to render content freely across all screen resolutions on computer and tablet. Plotly [26], a Python graphing library, is used for publication-quality graphs on cross-platform web browsers including Google Chrome, Firefox, Microsoft Edge, and Safari. High-quality interactive charts help users not only summarize the most interesting results easily but also understand the omics-based finding comprehensively.

2.4.2. Core Backend Module

The core backend of G2PDeep-v2 is a middle platform connecting to web interface, database, and the AI platform. It is developed based on the Django REST framework [27], a Python-based powerful and flexible server-side web framework, for managing a high volume of requests and tasks robustly. The Hypertext Transfer Protocol (HTTP) is used to communicate between web interface and backend. The backend integrates different pipelines for dataset creation, models training, and results summarization. It uses Python-based libraries, such as Pandas, NumPy [28] and SciPy [29], to perform a wide variety of mathematical operations on high-dimensional input data and results. The Celery [30], a Python-based extension of Django, schedules model training tasks in a queue and completes expensive operations of training asynchronously.

2.4.3. AI Platform Module

The AI platform is designed for construction, modification, training, and inference of deep learning neural networks and machine learning-based models. The deep learning models and their mathematical optimization are developed based on TensorFlow [31] and Keras [32], high-level deep learning frameworks. The machine learning based models are implemented by scikit-learn [33], free software machine learning library for the Python programming language. Optuna [34], an automatic hyperparameter optimization software framework, provides black box and hyperparameter optimization to maximize the performance of the deep learning and machine learning models.

2.4.4. Database Module

MySQL [35] and Redis [36] databases are used in G2PDeep-v2. MySQL, a relational database, enables meaningful information by joining various organized tables. It manages various multi-omics data, project information, modeling information, training information, and user information. Redis is a NoSQL database and in-memory database, extremely fast in reading and writing the data in random access memory. Redis stores the model training information and details of scheduler, bring the reliability of data storage and transactions during multiple tasks processing.

2.4.5. Security Policy

The G2PDeep-v2 leverages JSON Web Token (JWT) token [37] to control the access to private datasets and models. The JWT is a protocol providing authentication, authorization, and other security features for enterprise applications. Users can create an account by filling out a registration form on the sign-up page with the required information. The activation link for the new account is then sent to users. Users can log into G2PDeep using their registered username and password. The login credential remains valid for 12 h, providing access without having to prompt the user to log in again.

3. Results

3.1. Overview of the Web Server

The overview of G2PDeep-v2 is depicted in Figure 3. Starting from a multi-omics dataset, G2PDeep-v2 integrates samples from each type of multi-omics and splits merged

samples into five equally sized sets with 5-fold cross-validation. G2PDeep-v2 provides a variety of machine learning and deep learning models, including our proposed multi-CNN, Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). The platform also features a web-based interactive interface that allows users to create, train, and monitor the performance of these models, which is a unique aspect of bioinformatics. All models are trained using our high-performance computing resources and stored in the database for future inference. G2PDeep-v2 provides prediction for large-scale datasets, and visualization for predicted results and biomarkers associated with corresponding phenotypes. The results of Gene Set Enrichment Analysis (GSEA) for these biomarkers are generated automatically. It also provides complete documentation on the website, including a user guide describing all tools, examples, and frequently asked questions. To accelerate scientific research for survival analysis in cancer studies, we utilized G2PDeep-v2 and established biomarkers associated with long-term survival for 23 cancer studies.

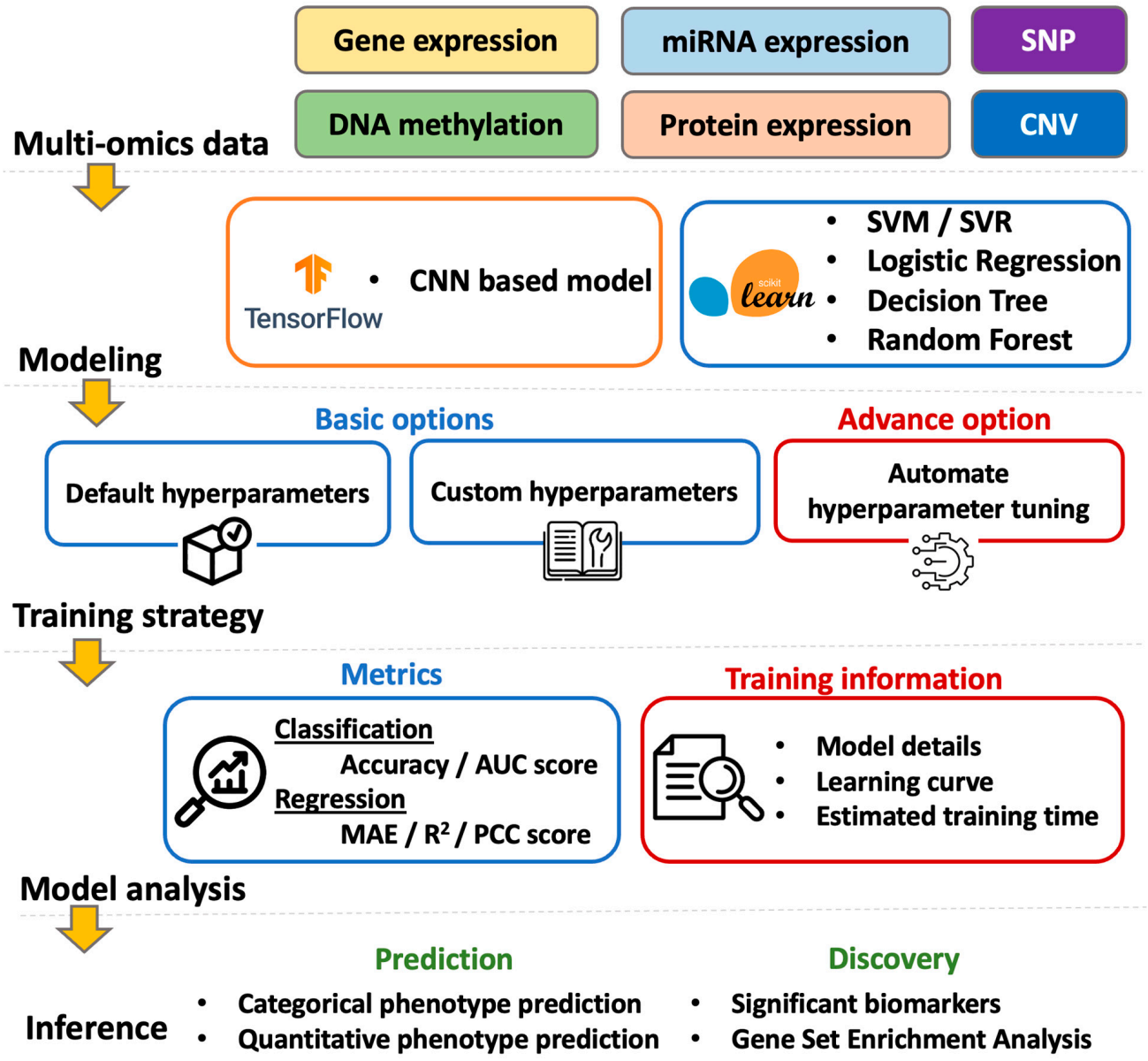

Figure 3. Overview of G2PDeep-v2.

3.1.1. Dataset Creation

To initiate the use of G2PDeep-v2, the pivotal first step involves creating datasets. G2PDeep-v2 allows users to create datasets with two options: uploading a CSV file or transferring data from a link (see Figure 4A). For a small dataset (up to 50 MB), users can create a dataset by uploading their own data from their local machine. For a large dataset (up to 10 GB), users can enter a shared link of data from Google Drive, OneDrive, CyVerse Data Store [38,39], or other public repositories. Users can upload multi-omics data, including gene expression, miRNA expression, DNA methylation, protein expression SNP, and CNV. Once the files are uploaded, G2PDeep-v2 performs z-score normalization for each expression sample and imputes missing values automatically. To merge multi-omics data from various sources, the datasets must share a column with unique IDs for each sample. By combining data from multiple sources, users can create more comprehensive datasets that may be better suited to their research questions. Users can also enter the type of data source to indicate whether the dataset is from humans or plants. The G2PDeep-v2 validates uploaded files to guarantee the data can be used in model creation. For any invalid format or unsupported data type, it has a function to stop data creation and notify users about the corresponding error message. It also shows a progress bar with duration remaining, allowing users to monitor the status of the dataset creation. The created datasets are private and only retrievable by the owners of the datasets. G2PDeep-v2 supports user's needs of sharing data with the community after anonymization by removing identifiable information for samples, making it available to other researchers to work on same data and share insights while protecting dataset privacy. G2PDeep-v2 also integrates the publicly available datasets, such as 23 TCGA cancer datasets, SoyNAM datasets [40] and Bandillo's SNP datasets [41] (see Figure 4B). Comprehensive details for each dataset, including links to data, type of data, number of samples, and features, are directly retrievable from the website. Once the datasets are created, users can build their models for the datasets.

3.1.2. Model Creation

Transitioning to model creation, G2PDeep-v2 emphasizes customization as a key feature. Hyperparameters, critical components influencing machine learning model performance, can be tailored by users on the Model Creation page (See Figure 5). The range of suggested hyperparameters and training parameters for models in G2PDeep-v2 are shown in Supplementary Table S1. Users can also select up to three different types of data as input and determine whether the model is designed for quantitative phenotype prediction or categorical phenotype prediction.

To strike a balance between training speed and model performance, G2PDeep-v2 provides three strategic options for setting hyperparameters. The first involves using default pre-tuned hyperparameters based on models created using data from 23 different TCGA studies and WGRS dataset for SCN resistance, enabling users to quickly generate models without additional tuning. Alternatively, users can opt for the second strategy, customizing hyperparameters through an interactive interface, aligning their models with specific datasets and research questions. The third strategy employs an automated hyperparameter search using a Bayesian optimization algorithm [42], efficiently exploring a large search space to identify optimal hyperparameters challenging to pinpoint through manual tuning.

Once users complete model creation, G2PDeep-v2 automatically saves the model as a private entry in the database. Users can conveniently access and manage their private and public models, along with corresponding configurations. Additionally, the platform supports model sharing within the community, fostering collaboration and knowledge exchange.

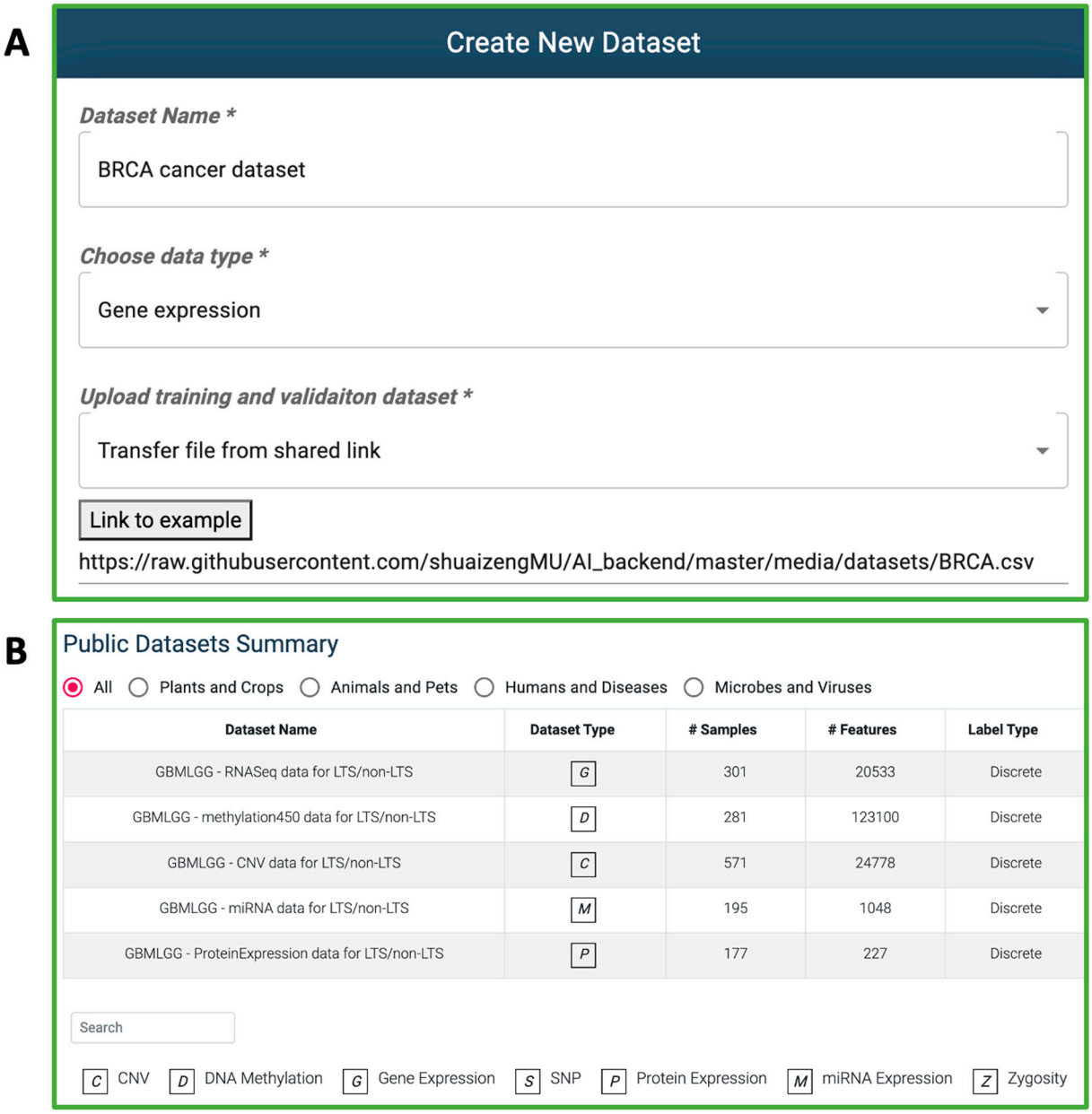

Figure 4. Dataset creation and retrieval in G2PDeep-v2. (**A**) Example of dataset creation by a shared link to the required data, indicated with a * symbol. (**B**) Publicly available datasets are shown with structured information.

3.1.3. Project for Model Training and Evaluation

Once the dataset and model are prepared, users can seamlessly leverage G2PDeep-v2 to train models using the uploaded datasets. On the Project Creation page, users can conveniently access all publicly available models as well as their private models, categorized based on the type of multi-omics data they are interested in. To initiate a new project of models training, users are prompted to select a dataset for each type of multi-omics data to serve as input for the model. After dataset selection, users have the flexibility to experiment with different hyperparameter-setting strategies to identify the optimal configuration for their specific data. Upon submission of the project, it enters a task queue, awaiting allocation of computing resources. The project settings and model configurations are securely stored in the database. Notably, for cancer data, the server typically takes around 2 h to train a model using automated hyperparameter tuning settings, involving 400 training samples across three types of multi-omics data and only CPU resources.

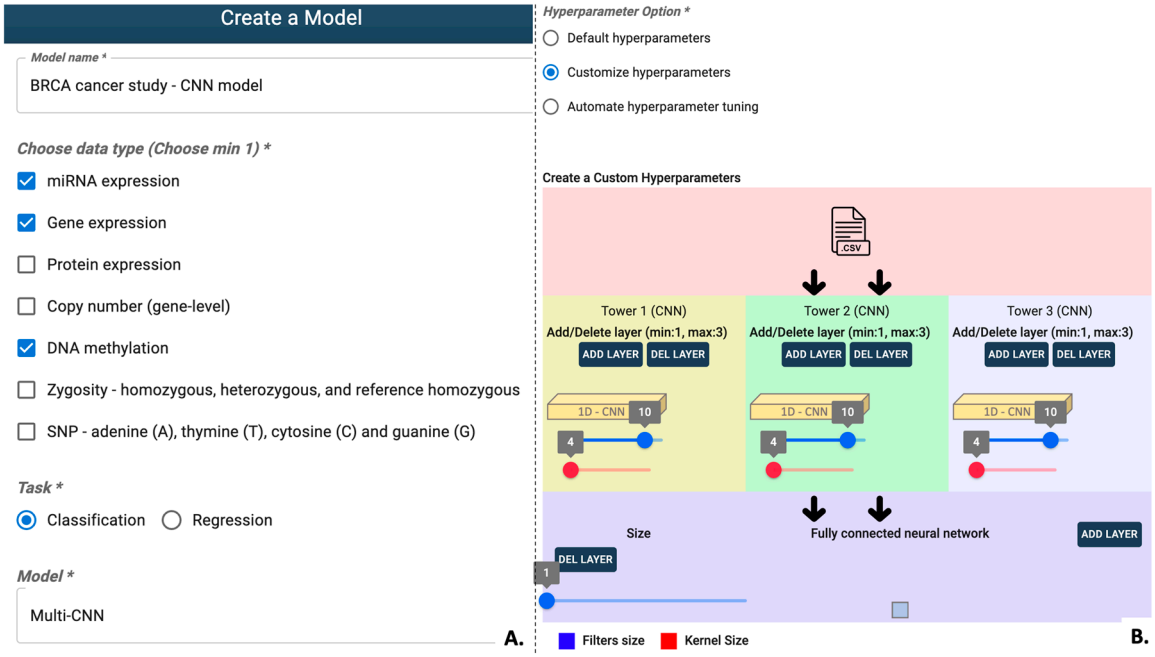

Figure 5. Interactive chart to configure the deep-learning model in G2PDeep-v2. (**A**) Required options (indicated with a * symbol) for inputting details such as the model, task, and input data. (**B**) Hyperparameters tuning options.

Users can track progress via a detailed summary page throughout the model training process. A progress bar with duration and percentage is displayed on the summary page, along with the estimated time to completion and model information. Further insights into the model, dataset, and training information are accessible on the Detail page, as illustrated in Figure 6. Dataset details include names, omics types, number of samples, and features, presented in a clear tabular format. Model information encompasses the model type and a diagram illustrating the kernel size and number of filters for each layer. The learning curve graphically portrays the performance of model on both training and validation datasets, aiding in assessing overfitting or underfitting.

Once the model reaches optimal training, G2PDeep-v2 provides interactive plots illustrating predicted results and model performance on both training and validation datasets. For categorical phenotype prediction tasks, a bar chart depicts the frequency of predicted labels alongside ground truth. Receiver Operating Characteristic (ROC) curves and Precision-Recall curves offer a visual representation of the diagnostic capabilities of the model. In cases of quantitative phenotype prediction tasks, a scatter plot compares predicted values with ground truth, accompanied by metrics like the Pearson correlation coefficient (PCC) and coefficient of determination (R squared). All predicted results and interactive plots are downloadable as CSV files and PNG images.

3.1.4. Prediction and Significant Biomarkers Discovery

Users can utilize G2PDeep-v2 to make predictions and visualize results using multi-omics data and a well-trained model. The predictions take, on average, less than 30 s to predict phenotype and marker significance for 1000 samples. Precisely, users can effortlessly input data by uploading a CSV file directly to the server for each type of multi-omics data. The system performs thorough validation, ensuring adherence to the required format, and promptly notifies users of any invalid input data through error notification. Notably, the system accommodates up to 10,000 samples, and a user-friendly progress bar allows for real-time monitoring of prediction status. All predicted results are securely stored in the database, readily retrievable for future analysis and comparison.

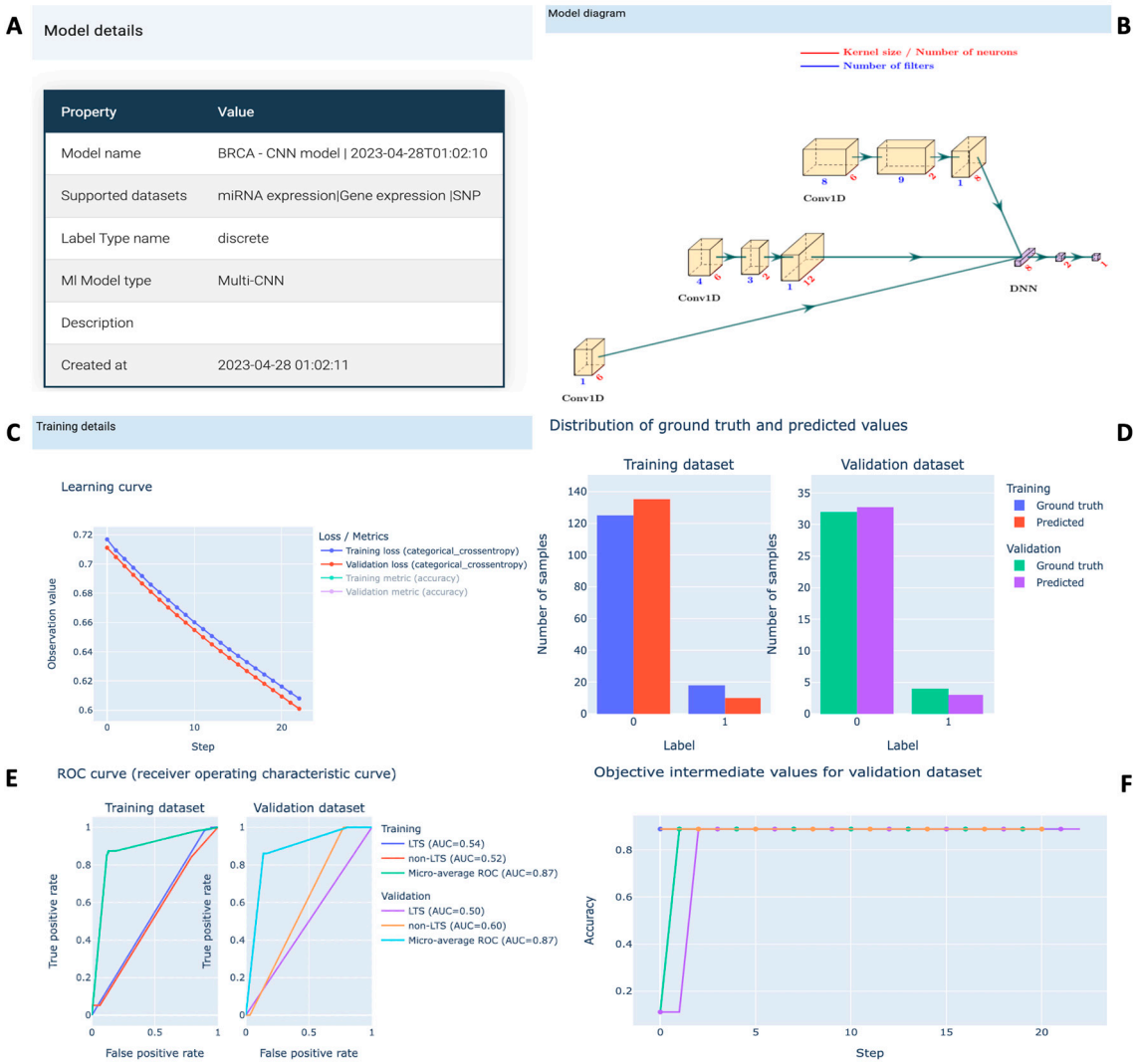

Figure 6. Project page in G2PDeep-v2. (**A**) Model summary showing the type of model and corresponding training dataset; (**B**) visualization of the multi-CNN architecture, illustrating the convolutional and fully connected layers used for multi-omics feature extraction and integration; (**C**) learning curves showing training and validation loss across epochs, allowing users to monitor model convergence and potential overfitting; (**D**) distribution of ground and predicted values for training and validation datasets; (**E**) ROC for phenotype prediction, demonstrating the discriminative ability of model on training and validation datasets; (**F**) optimization history from the Bayesian hyperparameter tuning process, highlighting how model performance improves over successive iterations of parameter adjustment. Each line represents a single trial.

Upon completion, G2PDeep-v2 generates a bar chart illustrating predicted values and a plot highlighting significant biomarkers (shown in Figure 7A). Users retain the flexibility to adjust the number of displayed biomarkers by setting a threshold based on the highest saliency values, focusing on the most relevant biomarkers for their specific research requirements. The plot presents significant biomarkers sorted by decreasing saliency values, and this information can be conveniently saved as a CSV file. G2PDeep-v2 also provides GSEA for significant biomarkers. It performs GSEA based on KEGG [16] and Reactome [17] pathway databases (shown in Figure 7B), which are widely used and comprehensive resources for pathway information. In cases where the biomarkers are not genes, such as CpG islands identified from methylation data, G2PDeep-v2 converts these markers to the corresponding neighboring gene that they regulate to fetch significance. It also provides users with a scatterplot for top 10 enriched pathways from KEGG and

Reactome for the gene sets, making it easy to gain insights into the molecular mechanisms underlying complex diseases and other biological phenomena. Detailed information on enriched pathways is presented in tabular form, including corresponding *p*-values, adjusted *p*-values, and gene sets. Additionally, a table listing literature evidence associated with significant biomarkers and relevant cancer or other studies enhances the interpretability of the results.

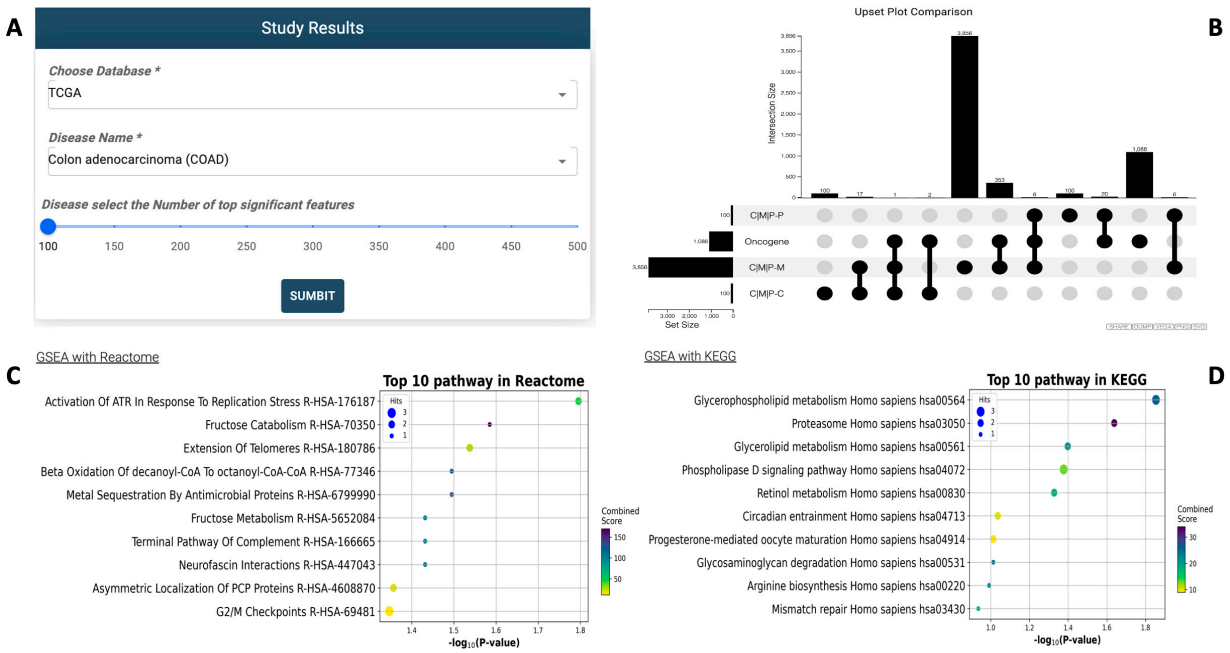

Figure 7. Study Results page in G2PDeep-v2. (**A**) Panel to select study with required options, as indicated by a * symbol; (**B**) Upset plot shows overlapping significant biomarkers; (**C**) GSEA with Reactome for significant biomarkers; (**D**) GSEA with KEGG for significant biomarkers.

3.1.5. Study Results in G2PDeep-v2

We regularly update and share the outcomes of cancer studies on the Study Results Page within G2PDeep-v2. Users can effortlessly access and retrieve results tailored to their specific interests, thereby facilitating enhanced accessibility for subsequent analysis and exploration.

Currently in G2PDeep-v2, we conducted several comprehensive studies using the 23 TCGA cancer studies dataset encompassing six distinct types of multi-omics data independently. The diverse array of multi-omics data, including gene expression, miRNA expression, DNA methylation, protein expression SNP, and CNV, was downloaded from the Broad Institute Fire Browse portal [43]. To ensure a robust analysis, we systematically created 41 datasets for each cancer study. These datasets include individual types of omics (6 datasets), combinations of two omics (15 datasets), and combinations of three omics (20 datasets). The phenotypes of these studies are long-term survival (LTS) and non-long-term survival (non-LTS) groups. The LTS is defined as survival > 3 years after diagnosis, and the non-LTS is defined as survival $\leq$ 3 years. Individuals who survived with the last follow-up of $\leq$3 years are excluded from further analysis.

To make 23 TCGA studies applicable to both ideal scenarios and real-world conditions, we categorized them into two types: studies with uniform multi-omics data and those with non-uniform multi-omics data. In the context of ideal scenarios, uniform data denotes that patient cohorts in these studies encompass all six types of multi-omics data, while non-uniform data for real-world conditions indicates that cohorts may lack some types of multi-omics data. Precisely, the uniform data can be considered a subset of the non-uniform

data. The studies with uniform omics data are tailored to investigate the significance of multi-omics data combinations. Due to limitations in the cohort of patients, we specifically designated six out of the total 23 studies as studies with uniform omics data. On the other hand, studies with non-uniform data are designed to explore biomarkers under scenarios that more closely mirror the complexities of real-world conditions. We finally made a total of 23 studies specifically with non-uniform data. The specifics of uniform and non-uniform multi-omics data for each cancer study, including information such as sequencing platforms, the number of features, and samples, are comprehensively listed in Tables 2 and 3, respectively.

Table 2. Uniform dataset for 6 different TCGA cancer studies.

Study	Number of Samples (LTS/Non-LTS)	Number of Features					
		Gene Expression	miRNA Expression	DNA Methylation	Protein Expression	SNP	CNV
BLCA	42 (15/27)	20,533	1048	300,869	225	18,634	24,778
HNSC	39 (14/25)	20,533	1048	300,973	239	17,796	24,778
LUAD	33 (16/17)	20,533	1048	300,822	239	18,950	24,778
LUSC	28 (15/13)	20,533	1048	300,970	239	18,822	24,778
SARC	26 (15/11)	20,533	1048	299,776	219	12,422	24,778
SKCM	41 (29/12)	20,533	1048	300,455	225	19,488	24,778

Table 3. Dataset for 23 different TCGA cancer studies.

Study	Number of Samples (LTS/Non-LTS)				
	Gene Expression	miRNA Expression	DNA Methylation	Protein Expression	SNP
ACC	62 (44/18)	63 (44/19)	63 (44/19)	36 (28/8)	73 (50/23)
BLCA	248 (87/161)	250 (89/161)	252 (89/163)	215 (76/139)	252 (89/163)
BRCA	506 (437/69)	344 (296/48)	364 (314/50)	410 (351/59)	455 (395/60)
CESC	146 (91/55)	146 (91/55)	146 (91/55)	65 (44/21)	138 (86/52)
CHOL	26 (11/15)	26 (11/15)	26 (11/15)	22 (9/13)	26 (11/15)
COAD	126 (78/48)	91 (56/35)	130 (81/49)	133 (79/54)	172 (101/71)
ESCA	86 (17/69)	87 (18/69)	87 (18/69)	51 (12/39)	86 (18/68)
HNSC	327 (144/183)	298 (128/170)	331 (145/186)	230 (89/141)	318 (135/183)
KICH	53 (47/6)	53 (47/6)	53 (47/6)	51 (45/6)	53 (47/6)
KIRC	404 (293/111)	177 (132/45)	228 (157/71)	246 (173/73)	226 (173/53)
KIRP	127 (100/27)	127 (100/27)	120 (94/26)	95 (75/20)	120 (94/26)
LIHC	195 (91/104)	195 (92/103)	199 (94/105)	109 (35/74)	189 (89/100)
LUAD	270 (133/137)	223 (109/114)	230 (112/118)	204 (102/102)	269 (133/136)
LUSC	305 (149/156)	196 (95/101)	222 (111/111)	204 (106/98)	302 (146/156)
MESO	80 (14/66)	80 (14/66)	80 (14/66)	58 (8/50)	76 (14/62)
PAAD	108 (20/88)	108 (20/88)	114 (21/93)	70 (11/59)	112 (21/91)
READ	38 (27/11)	33 (23/10)	40 (29/11)	46 (30/16)	49 (36/13)
SARC	177 (108/69)	177 (108/69)	179 (109/70)	150 (87/63)	159 (96/63)
SKCM	335 (227/108)	322 (219/103)	336 (227/109)	236 (152/84)	334 (226/108)
STAD	196 (48/148)	184 (47/137)	189 (49/140)	170 (38/132)	208 (49/159)
THCA	208 (199/9)	209 (200/9)	210 (201/9)	169 (160/9)	205 (198/7)
UCEC	69 (44/25)	183 (127/56)	193 (137/56)	217 (163/54)	273 (208/65)
UCS	42 (12/30)	41 (12/29)	42 (12/30)	36 (8/28)	42 (12/30)

The G2PDeep-v2 conducted a thorough analysis of phenotype prediction using both studies with uniform and non-uniform multi-omics data. Various models, including our proposed multi-CNN, LR [44], SVM [45], DT [22,46], and RF [47], were employed for predictions. To ensure reproducibility, the data for each cancer study underwent a systematic division into a training dataset (60% of the entire data) for model training, a validation dataset (20% of the entire data) for hyper-parameter tuning, and a test dataset (20% of the entire data) to evaluate model performance. The model was constructed in each cross-validation iteration and rigorously evaluated on the designated test set. Quantification of predictive performance was achieved by calculating the mean area under the curve (AUC) over a 5-fold cross-validation framework. Figure 8 illustrates that G2PDeep-v2 using our proposed multi-CNN outperforms other ML models in predicting phenotypes for the Skin Cutaneous Melanoma (SKCM) study with uniform multi-omics data. Based on the metrics recorded for models applied to both studies with uniform and non-uniform multi-omics, as depicted in Supplementary Table S2 and Table S3, respectively, G2PDeep-v2 using our proposed multi-CNN also outperforms or competes effectively with other ML models across most of the cancer studies. All performance details are conveniently accessible on the Study Result Page, providing a consolidated view of the effectiveness of models across various multi-omics data scenarios for user convenience. Furthermore, we expanded upon the study results by incorporating significant biomarkers and conducting corresponding GSEA.

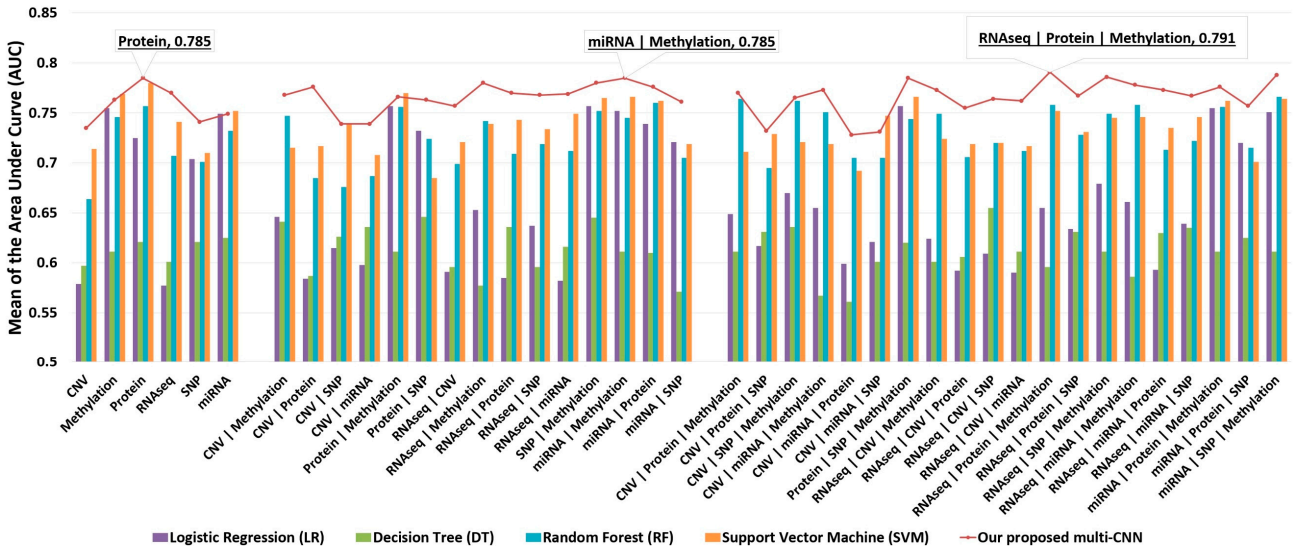

Figure 8. Comparison of model performance on 41 datasets from the Skin Cutaneous Melanoma study. Each model was trained and evaluated separately on individual datasets, and the mean area under the AUC was calculated. The results show that the proposed multi-CNN model (dotted line) consistently outperforms traditional machine learning methods (bars), including logistic regression, support vector machines, decision trees, and random forests.

3.2. Application of G2PDeep-v2

3.2.1. Use Case #1: Long-Term-Survival Prediction and Markers Discovery for Cancer

The motivation for this use case is to highlight the advantages of G2PDeep-v2 for long-term survival prediction and biomarker discovery in Breast Invasive Carcinoma (BRCA) cancer. We used G2PDeep-v2 to predict the phenotype of BRCA patients based on their multi-omics data, including gene expression, miRNA expression, DNA methylation, protein expression, SNP, and CNV data. We created and trained deep learning models to accurately predict the long-term survival of BRCA patients. According to our results (See Supplementary Table S2), the best model trained on three combinations of omics is the

CNN model, which achieved a mean AUC score of 0.907. The three combinations of omics are gene expression, miRNA expression, and SNP. We generated significant biomarkers and sorted them by saliency values. We selected the biomarkers with the top 100 highest saliency values and compared these biomarkers with oncogenes from the OncoKB database [48]. We found that 6 out of the 100 genes are oncogenes (see Supplementary Figure S1A). We then performed GSEA on these 100 genes and found seven pathways with *p*-values lower than 0.05. We noticed that most of the enriched pathways are related to breast cancer development (see Supplementary Table S4 and Figure S1B). Todd et al. [49] have reported that breast cancer with aberrant activation of the PI3K pathway can be identified by somatic mutations, suggesting potential dependence on the phosphatidylinositol signaling system pathway. Klara et al. [50] reported that N-glycosylation of breast cancer cells during metastasis is observed in a site-specific manner, highlighting the significance of high-mannose, fucosylated, and complex N-glycans as potential diagnostic markers and therapeutic targets in metastatic breast cancer. The Notch signaling pathway promotes tumor progression and survival and induces a breast cancer stem cell (CSC) phenotype [51]. This evidence supports the relevance of the identified biomarkers and their contribution towards these predictions.

3.2.2. Use Case #2: Disease Resistance Prediction for Soybean Cyst Nematode (SCN) in Soybean 1066 Lines

In this use case, we tested G2PDeep-v2 for Soybean Cyst Nematode (SCN) using Copy Number Variation (CNV) data, extracted from publicly available Whole-Genome Resequencing (WGRS) datasets for 1066 Soybean accessions [52]. The dataset consisted of multiple phenotypes, 228 samples from this dataset had readings for SCN phenotype, with class categories, Susceptible (S) and Resistant (R). G2PDeep's multi-CNN model was trained on 80 percent of this dataset, and its performance was evaluated on 5-fold cross-validation, using the AUC. The model performed consistently well on all 5-folds. To interpret the model's predictions and identify main genomic regions responsible for prediction of SCN resistance, we implemented a saliency map approach. This approach ranked the resultant SNP list based on the saliency values. In a further step to simplify rankings, saliency value was converted to dense rank; the higher the saliency value, lower it is rank would be. SNPs are mapped to genes based on their chromosomal positions using the soybean genome to generate a corresponding gene list. Based on the ranked gene list, the model identified a novel gene *Glyma.13g030200* (as shown in Supplementary Table S4), which ranked tenth in the saliency list. Interestingly, protein from the same family was previously published as a candidate for nematode resistance in rice [53]. To validate these results further, we looked at the regulatory aspects, to explore the Transcription Factors (TF) binding to *Glyma.13g030200* promoter region. GenVarX tool [54] in SoyKB [52,55], identified 81 TF binding sites within a 2 kb upstream region of the new candidate gene. Notably, 37 of these sites were particularly found to contain variantsTo explore Indels in the identified promotor regions, SNPViz tool [56–58] in SoyKB was utilized. This identified large insertions within the promotor region (as shown in Supplementary Table S5), which can potentially regulate the function of this gene affecting its role in SCN resistance. Further functional enrichment was performed on the resulting gene list, using GProfiler [59], to analyze Gene Ontology (GO) and KEGG pathway enrichment, where results revealed GO terms associated with defense response and stress response (as shown in Supplementary Table S6). The overall findings suggest *Glyma.13g030200* as a promising candidate which can be further investigated for SCN resistance phenotype. Further studies may be required to experimentally validate its precise function in SCN resistance.

4. Discussion

G2PDeep-v2 webserver is developed as a one-stop shop platform that addresses the need for efficient and accurate phenotype predictions from multi-omics data with customizable deep learning and machine learning models for any organisms. G2PDeep-v2 is the first web server that allows models to be created, trained with automated hyperparameter tuning, and used for inference on multi-omics data uploaded by researchers. We have deployed G2PDeep-v2 on a server equipped with both CPU and GPU resources to expedite model training and inference processes. Performance, compatibility, usability, and interpretability are all central principles of G2PDeep-v2. G2PDeep-v2 integrates numerous deep learning and machine learning models that are well-trained on 23 different TCGA cancer studies, SoyNAM, and Bandillo's SNP datasets, allowing researchers to reuse these models to predict phenotypes and identify significant biomarkers for biomedical and agribiotech purposes. In the context of G2PDeep-v2, significant refers to features identified by the CNN as important contributors to model predictions, rather than statistically significant features in the classical sense, which limits the interpretability of the results in traditional statistical terms. It has applications for predicting phenotypes in a wide range of research domains, including human and agriculture. It can also further help uncover the specific multi-omics data types that may be best suited for respective phenotype predictions.

In many real-world scenarios, such as medical research and rare disease studies, obtaining sufficient labeled data remains a major challenge. To address this issue, we plan to incorporate meta-learning techniques that enable models to learn effectively from limited data by leveraging prior knowledge from related tasks or experiences. To mitigate batch effects in multi-omics datasets, we also plan to employ contrastive learning to derive feature representations that are invariant to batch effects and robust to missing values. By comparing representations from different batches, the model can identify shared biological patterns independent of technical variations. Currently, G2PDeep-V2 supports a maximum combination of three omics data types, as samples containing four or more omics types are extremely limited in most available datasets. We aim to expand the framework to accommodate higher-order omics integration as more comprehensive datasets become available. Furthermore, we plan to enhance G2PDeep-V2 to support multi-class prediction tasks. As the current version does not yet include cross-species analysis, we plan to implement a new cross-organism validation module and perform comprehensive evaluations to rigorously assess its performance. We also plan to incorporate advanced deep learning interpretability methods, such as Grad-CAM, Layer-wise Relevance Propagation (LRP), and SHAP-based analyses. In future versions, we plan to integrate more sophisticated imputation techniques, such as KNN or mean imputation, to further enhance flexibility and performance. Additionally, future versions will include survival models and classical evaluation metrics, such as the concordance index (C-index) and Kaplan–Meier curves, as well as functionality to handle data heterogeneity, skewed distributions, and outliers in phenotypic data with the idea from Wu et al. [60]. Currently, we are working on combining scRNA-seq with bulk RNA-seq to improve the accuracy and resolution of transcriptomic analysis. By integrating scRNA-seq and bulk RNA-seq data, we can identify cell-type-specific gene expression patterns in complex tissues, enabling a deeper understanding of cellular heterogeneity and the identification of new biomarkers, than can be achieved by bulk transcriptomics alone. G2PDeep-v2 features will continue to expand and develop in response to the evolving needs of the research community.

5. Conclusions

G2PDeep-v2 is a novel and comprehensive web-platform that enables researchers to perform phenotype prediction, biomarker discovery, and GSEA for a range of applications

in research in human disease and plant breeding. G2PDeep-v2 allows for easy customization and optimization of models without the need for extensive experience in machine learning. By integrating various multi-omics datasets and pre-trained models, G2PDeep-v2 enables the creation of robust and reproducible predictions and biomarkers, while also providing access to a wealth of downstream analysis tools and results from multiple studies. Overall, G2PDeep-v2 represents a single one-stop-shop solution for phenotype predictions, with potential applications in precision medicine, drug discovery, precision agriculture, genomic epidemiology and other areas of research that rely on complex omics data.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/biom15121673/s1, Figure S1: Plots for significant biomarkers; Table S1: Suggested hyperparameters; Table S2: Performance on 6 uniform datasets; Table S3: Performance on 23 nonuniform datasets; Table S4: Significant promoter regions; Table S5: Promoter results based on SNPViz; Table S6: Functional Enrichement shows pathways related to defense and stress response.

Author Contributions: S.Z., D.X. and T.J. conceived the research. S.Z., T.A. and M.S.I. wrote the software. S.Z. and S.Z.A. conducted the deep learning and machine learning experiments. S.Z. wrote the manuscript with suggestions from D.X. and T.J.; D.X. and T.J. provided valuable input and advice for the project. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by funding from Missouri Department of Health and Senior Services (MDHSS)—Contract #AOC23380006, National Science Foundation (NSF) IOS-2343815, National Science Foundation (NSF) Cybersecurity Innovation OAC-2232889; National Institutes of Health (R35-GM126985) and U.S. Department of Energy under Award DE-SC0023142. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM103434. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The G2PDeep-v2 server is publicly available at https://g2pdeep.org (accessed on 1 August 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

AUC	area under the curve
BRCA	Breast Invasive Carcinoma
CNV	copy number variations
CSC	cancer stem cell
CSV	comma-separated values
DT	Decision Tree
GSEA	Gene Set Enrichment Analysis
HTTP	Hypertext Transfer Protocol
JWT	JSON Web Token
KEGG	Kyoto Encyclopedia of Genes and Genomes
LR	Logistic Regression
LTS	long-term survival
miRNA	microRNA
MVC	Model-View-Controller
non-LTS	non-long-term survival
PCC	Pearson correlation coefficient

RF	Random Forest
ROC	Receiver Operating Characteristic
SKCM	Skin Cutaneous Melanoma
SNP	single nucleotide polymorphisms
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
UI	user interface

References

1. Menyhárt, O.; Győrffy, B. Multi-Omics Approaches in Cancer Research with Applications in Tumor Subtyping, Prognosis, and Diagnosis. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 949. [CrossRef]
2. Hasin, Y.; Seldin, M.; Lusis, A. Multi-Omics Approaches to Disease. *Genome Biol.* **2017**, *18*, 83. [CrossRef]
3. Sandhu, K.S.; Lozada, D.N.; Zhang, Z.; Pumphrey, M.O.; Carter, A.H. Deep Learning for Predicting Complex Traits in Spring Wheat Breeding Program. *Front. Plant Sci.* **2021**, *11*, 613325. [CrossRef]
4. Dimitrakopoulos, C.; Hindupur, S.K.; Colombi, M.; Liko, D.; Ng, C.K.; Piscuoglio, S.; Behr, J.; Moore, A.L.; Singer, J.; Ruscheweyh, H.J. Multi-Omics Data Integration Reveals Novel Drug Targets in Hepatocellular Carcinoma. *BMC Genom.* **2021**, *22*, 592. [CrossRef]
5. Wang, C.; Lye, X.; Kaalia, R.; Kumar, P.; Rajapakse, J.C. Deep Learning and Multi-Omics Approach to Predict Drug Responses in Cancer. *BMC Bioinform.* **2021**, *22*, 632. [CrossRef]
6. Leo, I.R.; Aswad, L.; Stahl, M.; Kunold, E.; Post, F.; Erkers, T.; Struyf, N.; Mermelekas, G.; Joshi, R.N.; Gracia-Villacampa, E. Integrative Multi-Omics and Drug Response Profiling of Childhood Acute Lymphoblastic Leukemia Cell Lines. *Nat. Commun.* **2022**, *13*, 1691. [CrossRef] [PubMed]
7. Ma, W.; Qiu, Z.; Song, J.; Li, J.; Cheng, Q.; Zhai, J.; Ma, C. A Deep Convolutional Neural Network Approach for Predicting Phenotypes from Genotypes. *Planta* **2018**, *248*, 1307–1318. [CrossRef]
8. Hanczar, B.; Zehraoui, F.; Issa, T.; Arles, M. Biological Interpretation of Deep Neural Network for Phenotype Prediction Based on Gene Expression. *BMC Bioinform.* **2020**, *21*, 501. [CrossRef]
9. Wang, T.; Shao, W.; Huang, Z.; Tang, H.; Zhang, J.; Ding, Z.; Huang, K. MOGONET Integrates Multi-Omics Data Using Graph Convolutional Networks Allowing Patient Classification and Biomarker Identification. *Nat. Commun.* **2021**, *12*, 3445. [CrossRef] [PubMed]
10. Sammut, S.-J.; Crispin-Ortuzar, M.; Chin, S.-F.; Provenzano, E.; Bardwell, H.A.; Ma, W.; Cope, W.; Dariush, A.; Dawson, S.-J.; Abraham, J.E.; et al. Multi-Omic Machine Learning Predictor of Breast Cancer Therapy Response. *Nature* **2022**, *601*, 623–629. [CrossRef] [PubMed]
11. Elmarakeby, H.A.; Hwang, J.; Arafeh, R.; Crowdis, J.; Gang, S.; Liu, D.; AlDubayan, S.H.; Salari, K.; Kregel, S.; Richter, C.; et al. Biologically Informed Deep Neural Network for Prostate Cancer Discovery. *Nature* **2021**, *598*, 348–352. [CrossRef]
12. Oh, J.H.; Choi, W.; Ko, E.; Kang, M.; Tannenbaum, A.; Deasy, J.O. PathCNN: Interpretable Convolutional Neural Networks for Survival Prediction and Pathway Analysis Applied to Glioblastoma. *Bioinformatics* **2021**, *37*, i443–i450. [CrossRef] [PubMed]
13. Poirion, O.B.; Jing, Z.; Chaudhary, K.; Huang, S.; Garmire, L.X. DeepProg: An Ensemble of Deep-Learning and Machine-Learning Models for Prognosis Prediction Using Multi-Omics Data. *Genome Med.* **2021**, *13*, 112. [CrossRef] [PubMed]
14. Liu, Y.; Wang, D.; He, F.; Wang, J.; Joshi, T.; Xu, D. Phenotype Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean. *Front. Genet.* **2019**, *10*, 1091. [CrossRef]
15. Zeng, S.; Mao, Z.; Ren, Y.; Wang, D.; Xu, D.; Joshi, T. G2PDeep: A Web-Based Deep-Learning Framework for Quantitative Phenotype Prediction and Discovery of Genomic Markers. *Nucleic Acids Res.* **2021**, *49*, W228–W236. [CrossRef]
16. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* **2017**, *45*, D353–D361. [CrossRef] [PubMed]
17. Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **2020**, *48*, D498–D503. [CrossRef]
18. National Human Genome Research Institute (NHGRI). *The Cancer Genome Atlas (TCGA) Portal 2022*; National Human Genome Research Institute (NHGRI): Bethesda, MD, USA, 2022.
19. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 448–456.
20. 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.
21. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In Proceedings of the International Conference on Machine Learning 2013, Atlanta, GA, USA, 16–21 June 2013; Volume 30, p. 3.

22. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
23. Fang, Z.; Liu, X.; Peltz, G. GSEApy: A Comprehensive Package for Performing Gene Set Enrichment Analysis in Python. *Bioinformatics* **2023**, *39*, btac757. [CrossRef]
24. *Facebook React*; Meta Platforms: Menlo Park, CA, USA, 2022.
25. *Google Material-UI*; Google LLC: Mountain View, CA, USA, 2023.
26. Plotly Technologies Inc. P.T. Collaborative Data Science. Available online: https://plot.ly (accessed on 1 January 2012).
27. Django Software Foundation. *Django*; Django Software Foundation: Lawrence, KS, USA, 2019.
28. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array Programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef]
29. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
30. Celery Team. Celery. 2021. Available online: https://github.com/celery/celery (accessed on 27 November 2025).
31. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, S.G.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. *arXiv* **2015**, arXiv:1603.04467.
32. Chollet, F. Keras. 2015. Available online: https://keras.io/ (accessed on 27 November 2025).
33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
34. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-Generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.
35. Oracle Corporation. *MySQL*; Oracle Corporation: Austin, TX, USA, 2021.
36. Sanfilippo, S. *Redis Labs. Redis*; Redis Labs: Mountain View, CA, USA, 2022.
37. JWT Team. *JSON Web Token*; Auth0: Bellevue, WA, USA, 2015.
38. Goff, S.A.; Vaughn, M.; McKay, S.; Lyons, E.; Stapleton, A.E.; Gessler, D.; Matasci, N.; Wang, L.; Hanlon, M.; Lenards, A. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front. Plant Sci.* **2011**, *2*, 34. [CrossRef]
39. Merchant, N.; Lyons, E.; Goff, S.; Vaughn, M.; Ware, D.; Micklos, D.; Antin, P. The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol.* **2016**, *14*, e1002342. [CrossRef] [PubMed]
40. Song, Q.; Yan, L.; Quigley, C.; Jordan, B.D.; Fickus, E.; Schroeder, S.; Song, B.H.; Charles An, Y.Q.; Hyten, D.; Nelson, R. Genetic Characterization of the Soybean Nested Association Mapping Population. *Plant Genome* **2017**, *10*, plantgenome2016-1. [CrossRef] [PubMed]
41. Bandillo, N.; Jarquin, D.; Song, Q.; Nelson, R.; Cregan, P.; Specht, J.; Lorenz, A. A Population Structure and Genome-wide Association Analysis on the USDA Soybean Germplasm Collection. *Plant Genome* **2015**, *8*, plantgenome2015-04. [CrossRef]
42. Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian Optimization of Machine Learning Algorithms. *Adv. Neural Inf. Process. Syst.* **2012**, *4*, 2951–2959.
43. FireBrowse Team. FireBrowse. 2016. Available online: http://firebrowse.org/ (accessed on 27 November 2025).
44. Kleinbaum, D.G.; Dietz, K.; Gail, M.; Klein, M.; Klein, M. *Logistic Regression*; Springer: Berlin/Heidelberg, Germany, 2002; ISBN 0-387-95397-3.
45. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support Vector Machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18–28. [CrossRef]
46. Song, Y.Y.; Ying, L.U. Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130.
47. Biau, G.; Scornet, E. A Random Forest Guided Tour. *Test* **2016**, *25*, 197–227. [CrossRef]
48. Chakravarty, D.; Gao, J.; Phillips, S.; Kundra, R.; Zhang, H.; Wang, J.; Rudolph, J.E.; Yaeger, R.; Soumerai, T.; Nissan, M.H. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* **2017**, *1*, 1–16. [CrossRef] [PubMed]
49. Miller, T.W.; Rexer, B.N.; Garrett, J.T.; Arteaga, C.L. Mutations in the Phosphatidylinositol 3-Kinase Pathway: Role in Tumor Progression and Therapeutic Implications in Breast Cancer. *Breast Cancer Res.* **2011**, *13*, 224. [CrossRef] [PubMed]
50. Ščupáková, K.; Adelaja, O.T.; Balluff, B.; Ayyappan, V.; Tressler, C.M.; Jenkinson, N.M.; Claes, B.S.; Bowman, A.P.; Cimino-Mathews, A.M.; White, M.J. Clinical Importance of High-Mannose, Fucosylated, and Complex N-Glycans in Breast Cancer Metastasis. *Jci Insight* **2021**, *6*, e146945. [CrossRef] [PubMed]
51. Farnie, G.; Clarke, R.B.; Spence, K.; Pinnock, N.; Brennan, K.; Anderson, N.G.; Bundred, N.J. Novel Cell Culture Technique for Primary Ductal Carcinoma in Situ: Role of Notch and Epidermal Growth Factor Receptor Signaling Pathways. *J. Natl. Cancer Inst.* **2007**, *99*, 616–627. [CrossRef]
52. Joshi, T.; Wang, J.; Zhang, H.; Chen, S.; Zeng, S.; Xu, B.; Xu, D. The Evolution of Soybean Knowledge Base (SoyKB). In *Plant Genomics Databases: Methods and Protocols*; Humana Press: New York, NY, USA, 2017; pp. 149–159.

53. Li, Z.; Huang, Q.; Lin, B.; Guo, B.; Wang, J.; Huang, C.; Liao, J.; Zhuo, K. CRISPR/Cas9-Targeted Mutagenesis of a Representative Member of a Novel PR10/Bet v1-like Protein Subfamily Significantly Reduces Rice Plant Height and Defense against Meloidogyne Graminicola. *Phytopathol. Res.* **2022**, *4*, 38. [CrossRef]
54. Chan, Y.O.; Biová, J.; Mahmood, A.; Dietz, N.; Bilyeu, K.; Škrabišová, M.; Joshi, T. Genomic Variations Explorer (GenVarX): A Toolset for Annotating Promoter and CNV Regions Using Genotypic and Phenotypic Differences. *Front. Genet.* **2023**, *14*, 1251382. [CrossRef]
55. Joshi, T.; Patil, K.; Fitzpatrick, M.R.; Franklin, L.D.; Yao, Q.; Cook, J.R.; Wang, Z.; Libault, M.; Brechenmacher, L.; Valliyodan, B. Soybean Knowledge Base (SoyKB): A Web Resource for Soybean Translational Genomics. *BMC Genom.* **2012**, *13*, S15. [CrossRef]
56. Zeng, S.; Škrabišová, M.; Lyu, Z.; Chan, Y.O.; Dietz, N.; Bilyeu, K.; Joshi, T. Application of SNPViz v2.0 Using next-Generation Sequencing Data Sets in the Discovery of Potential Causative Mutations in Candidate Genes Associated with Phenotypes. *Int. J. Data Min. Bioinform.* **2021**, *25*, 65–85. [CrossRef]
57. Zeng, S.; Škrabišová, M.; Lyu, Z.; Chan, Y.O.; Bilyeu, K.; Joshi, T. SNPViz v2.0: A Web-Based Tool for Enhanced Haplotype Analysis Using Large Scale Resequencing Datasets and Discovery of Phenotypes Causative Gene Using Allelic Variations. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Republic of Korea, 16–19 December 2020; pp. 1408–1415.
58. Langewisch, T.; Zhang, H.; Vincent, R.; Joshi, T.; Xu, D.; Bilyeu, K. Major Soybean Maturity Gene Haplotypes Revealed by SNPViz Analysis of 72 Sequenced Soybean Genomes. *PLoS ONE* **2014**, *9*, e94150. Available online: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0094150 (accessed on 14 November 2024). [CrossRef] [PubMed]
59. Reimand, J.; Kull, M.; Peterson, H.; Hansen, J.; Vilo, J. G:Profiler—A Web-Based Toolset for Functional Profiling of Gene Lists from Large-Scale Experiments. *Nucleic Acids Res.* **2007**, *35*, W193–W200. Available online: https://academic.oup.com/nar/article/35/suppl_2/W193/2920757 (accessed on 14 November 2024). [CrossRef] [PubMed]
60. Wu, C.; Zhou, F.; Ren, J.; Li, X.; Jiang, Y.; Ma, S. A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. *High Throughput* **2019**, *8*, 4. [CrossRef] [PubMed]