

G2PDeep-v2: a web-based deep-learning framework for phenotype prediction and biomarker discovery using multi-omics data

Shuai Zeng^{1,2}, Trinath Adusumilli¹, Sania Zafar Awan³, Manish Sridhar Immadi¹, Dong Xu^{1,2,3} and Trupti Joshi^{1,2,3,4,*}

¹ Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, 65211, USA

² Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, 65211, USA

³ MU Institute for Data Science and Informatics, University of Missouri-Columbia, Columbia, MO, 65211, USA

⁴ Department of Biomedical Informatics, Biostatistics and Medical Epidemiology, University of Missouri, Columbia, MO, 65211, USA

* To whom correspondence should be addressed. Tel: +1 (573) 884-5963; Email: joshitr@health.missouri.edu

Email:

Shuai Zeng, zengs@missouri.edu

Trinath Adusumilli, tafbr@missouri.edu

Sania Zafar Awan, sah2p@missouri.edu

Manish Sridhar Immadi, mizy9@missouri.edu

Dong Xu, xudong@missouri.edu

Trupti Joshi, joshitr@health.missouri.edu

Abstract

The G2PDeep-v2 server is a web-based platform powered by deep learning, for phenotype prediction and markers discovery from multi-omics data in any organisms including humans, plants, animals, and viruses. 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 and other biological processes. The G2PDeep-v2 server is publicly available at <https://g2pdeep.org/>.

Keywords

Multi-omics, Biomarker, Phenotype prediction, Deep learning, Automated hyperparameters tuning, Reproducibility, Web-platform

Background

With the advances in molecular profiling technologies, the ability to observe large-scale multi-omics data from patients or other biological samples 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, multi-omics data allows researchers to predict the phenotypes and identify biomarkers that affect the diversities 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 the potential biomarkers along with their biological relevance.

Many approaches have been proposed over the past decade to perform 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 failed 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. 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 installations. Furthermore, training models with large-scale multi-omics data requires computing resources and storage exceeding the capacities of most potential users. Moreover, few of existing methods integrate functionalities to identify significant multi-omics signatures and biomarkers related to the biomedical and biological studies, resulting in researchers spending additional time on confirming evidence for the findings.

Introduction

Along this line of research, we have been developing the deep learning method G2PDeep. The first original 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 the web-based access. To address issues discussed above, we have further developed G2PDeep-v2, a comprehensive web-based platform for phenotype prediction using multi-omics data and biomarkers discovery. 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 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 3 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 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 datasets and well-trained models are serialized and stored in user accounts to protect 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 to 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 studies for 23 different cancer with long-term-survival labels, originally provided by The Cancer Genome Atlas (TCGA) project [18]. 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. To the best of our knowledge, G2PDeep-v2 is the first web-based deep-learning framework available for phenotype prediction, biomarker discovery and annotation for multi-omics data. Users can apply G2PDeep-v2 not only to human disease studies but also to other organisms including research in plants, animals, bacteria, and viruses. The G2PDeep-v2 server is publicly available at <https://g2pdeep.org>. The Python-based deep-learning model is available at https://github.com/shuaizengMU/G2PDeep_model.

Results

Overview of the web server

The overview of G2PDeep-v2 is depicted in Fig. 1. Starting from a multi-omics dataset, G2PDeep-v2 integrates samples from each type of multi-omics and splits merged samples into 5 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 these 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.

Dataset creation

Initiating 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 Fig. 2A). 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 [19,20], 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, it requires data containing 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 the dataset is from human, animal, plants and other. The G2PDeep-v2 validates uploaded files to guarantee the data can be used in model creation. For any invalid format or unsupported data, it has a function to stop data creation and show a corresponding error message. It also shows a progress bar with duration and percentage, 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 sharing data with the community in 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 [21] and Bandillo's SNP datasets [22] (see Fig. 2B). Comprehensive details for each dataset, including links to data, type of data, number of samples, and features, are retrievable from the website. Once the datasets are created, users can build their models for the datasets.

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 Fig. 3). 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, 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 [23], 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.

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 hours to train a model using automated hyperparameter tuning settings, involving 400 training samples across three types of multi-omics data and CPU resources.

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 Fig. 4. 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.

Additionally, the optimization history plot for automated hyperparameter tuning provides valuable insights into the efficacy of different hyperparameters.

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 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^2). All predicted results and interactive plots are downloadable as CSV files and PNG images.

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 an average, less than 30 seconds to predict phenotype and marker significance for 1,000 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 messages. 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.

Upon submission, G2PDeep-v2 generates a bar chart illustrating predicted values and a plot highlighting significant biomarkers (shown in Fig. 5A). 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 analysis based on KEGG [16] and Reactome [17] pathway databases (shown in Fig. 5B), 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 genes 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.

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.

As of now, G2PDeep-v2 provides comprehensive studies on cancer, with our investigation delving into 23 TCGA cancer studies 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 [24]. 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 ≤ 3 years. Individuals who survived with the last follow-up of ≤ 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 6 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 Table 2 and 3 respectively.

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 [20], SVM [21], DT [22], and RF [23], 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. The Fig. 6 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 S3 respectively, G2PDeep-v2 using our proposed multi-CNN also outperforms or competes effectively with other ML models across the majority of 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 expand upon the study results by incorporating significant biomarkers and conducting corresponding GSEA analysis.

Application of G2PDeep-v2

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 [25]. We found that 6 out of the 100 genes are oncogenes (see Supplementary Fig. S1A). We then performed GSEA analysis 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 S4 Fig. 1B). Todd et al. [26] 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. [27] 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 [28]. These evidences support the relevance of the identified biomarkers and their contribution towards these predictions.

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 with Copy Number Variation (CNV) data gathered from WGRS datasets publicly available for 1066 Soybean lines. There was a total of 228 samples of SCN with two classes Susceptible (S) and Resistant (R). We trained and evaluated multi-CNN model, and the performance of the model was evaluated on 5-fold cross-validation using area under the ROC curve. We performed the saliency map for genomic selection and selected the top 100 most significant biomarkers for SCN phenotype. The results of the saliency map suggested a novel gene Glyma.13g030200 ranked tenth in saliency list, protein from the same family was previously published as a candidate for nematode resistance in rice [29]. Upon investigation using SNPviz [30,31] we found some big indels in the promotor region, these indels can potentially regulate the function of this gene. We performed enrichment analysis with GO and KEGG pathway on these biomarkers and found evidence related to abiotic stresses and defense response.

Discussion

G2PDeep 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. 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. Performance, compatibility, usability, and interpretability are 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. It has applications for predicting phenotypes in a wide range of research domains, including human diseases, agriculture, animal, and viral studies. 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 can be challenging. In the future, we plan to employ meta-learning techniques to enable models to learn from small amounts of data by leveraging prior knowledge learned from other tasks or experiences. To reduce the batch effect in multi-omics datasets, we also plan to utilize contrastive learning to learn feature representations that are invariant to batch effects. By comparing different representations of data from different batches, our models can identify common patterns that are independent of the batch effect. We are also planning to enhance G2PDeep-v2 by enabling models to cater to multi-class prediction scenarios. We will also deploy G2PDeep-v2 on a server equipped with both CPU and GPU resources to expedite model training and inference processes. 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. G2PDeep-v2 will continue to expand and develop in response to the evolving needs of the research community.

Conclusions

G2PDeep-v2 is a novel and comprehensive web-platform that enables researchers to perform phenotype prediction, biomarker discovery, and GSEA analysis for a range of applications in research in human disease and plant breeding. With its user-friendly interface, advanced machine learning algorithms and automated hyperparameter tuning, G2PDeep-v2 allows for easy customization and optimization of models without the experience of machine learning required. 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.

Methods

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.

Modeling in G2PDeep

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 Fig. 7) 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 [32] layer is added at the end of representation and Dropout [33] layers are added in each layer of fully connected neural network. The Leaky Rectified Linear Unit (Leaky-ReLU) [34] 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 [35], 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. In Juan's rice study, the output of model is a single value representing the quantity of phenotype.

Traditional machine learning models

G2PDeep-v2 integrates various traditional machine learning methods, such as LR, SVM, DT, and RF for comparisons. 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.

Biomarkers discovery and annotation

The significant biomarkers associated with phenotypes of interest to researchers are estimated using models in G2PDeep-v2. The saliency map algorithm is applied to the multi-CNN to estimate these significant biomarkers, and the coefficients of traditional ML models are utilized to identify them. 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 [39], a Python library, is employed.

Web server implementation

G2PDeep-v2 is developed using Model-View-Controller (MVC) architectural pattern and deployed in Docker. This Dockerized 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 Fig. 8.

Web interface module

G2PDeep-v2 provides user-friendly web interface developed using ReactJS [36] and Material UI [37], 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. The Plotly [38], 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.

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 [39], a Python-based powerful and flexible server-side web framework, for managing 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 [40] and SciPy [41], to perform a wide variety of mathematical operations on high-dimensional input data and results. The Celery [42], a Python-based extension of Django, schedules model training tasks in a queue and completes expensive operations of training asynchronously.

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 [43] and Keras [44], high-level deep learning frameworks. The machine learning based models are implemented by scikit-learn [45], free software machine learning library for the Python programming language. The Optuna [46], an automatic hyperparameter optimization software framework, provides black box and hyperparameter optimization to maximize the performance of the deep learning and machine learning models.

Database module

MySQL [47] and Redis [48] 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.

Security policy

The G2PDeep-v2 leverages JSON Web Token (JWT) token [49] to control the access to private datasets and models. The JWT token 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 hours, providing access without having to prompt the user to log in again.

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

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not Applicable.

Availability of data and materials

The G2PDeep-v2 server is publicly available at <https://g2pdeep.org>. The Python-based deep-learning model is available at https://github.com/shuaizengMU/G2PDeep_model.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by funding from Missouri Department of Health and Senior Services (MDHSS) - Contract #AOC23380006, National Science Foundation (NSF) Cybersecurity Innovation OAC-2232889; National Institutes of Health (R35-GM126985) and U.S. Department of Energy under Award DE-SC0023142.

Authors' contributions

SZ, DX, TJ conceived the research. SZ, TA and MI wrote the software. SZ and SA conducted the deep learning and machine learning experiments. SZ wrote the manuscript with suggestions from DX and TJ. DX and TJ provided valuable input and advice for the project. All authors read and approved the final manuscript.

Acknowledgements

Our appreciation goes out to Ajay Kumar for their efforts in assisting with functional testing of G2PDeep-v2.

Author information

Authors and Affiliations

Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, 65211, USA

Shuai Zeng, Trinath Adusumilli, Manish Sridhar Immadi, Dong Xu, Trupti Joshi

Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, 65211, USA

Shuai Zeng, Dong Xu, Trupti Joshi

MU Institute for Data Science and Informatics, University of Missouri-Columbia, Columbia, MO, 65211, USA

Ajay Kumar, Dong Xu, Trupti Joshi

Department of Biomedical Informatics, Biostatistics and Medical Epidemiology, University of Missouri, Columbia, MO, 65211, USA

Trupti Joshi

Corresponding author

Correspondence to Trupti Joshi.

References

1. Menyhárt O, Györfy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput Struct Biotechnol J*. 2021;19:949.
2. Hasin Y, Seldin M, Lusk A. Multi-omics approaches to disease. *Genome Biol*. 2017;18:1–15.
3. Sandhu KS, Lozada DN, Zhang Z, Pumphrey MO, Carter AH. Deep Learning for Predicting Complex Traits in Spring Wheat Breeding Program. *Front Plant Sci*. 2021;11.
4. Dimitrakopoulos C, Hindupur SK, Colombi M, Liko D, Ng CK, Piscuoglio S, et al. Multi-omics data integration reveals novel drug targets in hepatocellular carcinoma. *BMC Genomics*. 2021;22:1–26.
5. Wang C, Lye X, Kaalia R, Kumar P, Rajapakse JC. Deep learning and multi-omics approach to predict drug responses in cancer. *BMC Bioinformatics*. 2021;22:1–15.

6. Leo IR, Aswad L, Stahl M, Kunold E, Post F, Erkers T, et al. Integrative multi-omics and drug response profiling of childhood acute lymphoblastic leukemia cell lines. *Nat Commun.* 2022;13:1691.
7. Ma W, Qiu Z, Song J, Li J, Cheng Q, Zhai J, et al. A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta.* 2018;248:1307–18.
8. Hanczar B, Zehraoui F, Issa T, Arles M. Biological interpretation of deep neural network for phenotype prediction based on gene expression. *BMC Bioinformatics.* 2020;21:501.
9. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun.* 2021;12:3445.
10. Sammut S-J, Crispin-Ortuzar M, Chin S-F, Provenzano E, Bardwell HA, Ma W, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature.* 2022;601:623–9.
11. Elmarakeby HA, Hwang J, Arafeh R, Crowdis J, Gang S, Liu D, et al. Biologically informed deep neural network for prostate cancer discovery. *Nature.* 2021;598:348–52.
12. Oh JH, Choi W, Ko E, Kang M, Tannenbaum A, Deasy JO. PathCNN: interpretable convolutional neural networks for survival prediction and pathway analysis applied to glioblastoma. *Bioinformatics.* 2021;37:i443–50.
13. Poirion OB, Jing Z, Chaudhary K, Huang S, Garmire LX. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med.* 2021;13:1–15.
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.
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:W228–36.
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:D353–61.
17. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* 2020;48:D498–503.
18. Institute NHGR. The Cancer Genome Atlas (TCGA) portal [Internet]. 2022. Available from: <https://www.cancer.gov/tcga>
19. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, et al. The iPlant collaborative: cyberinfrastructure for plant biology. *Front Plant Sci.* 2011;2:34.
20. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, et al. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol.* 2016;14:e1002342.

21. Song Q, Yan L, Quigley C, Jordan BD, Fickus E, Schroeder S, et al. Genetic characterization of the soybean nested association mapping population. *Plant Genome* 10 (2). 2017.
22. Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J, et al. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome*. 2015;8:plantgenome2015.04.0024.
23. Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. *Adv Neural Inf Process Syst*. 2012;25.
24. FireBrowse Team. FireBrowse [Internet]. 2016. Available from: <http://firebrowse.org/>
25. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: a precision oncology knowledge base. *JCO Precis Oncol*. 2017;1:1–16.
26. Miller TW, Rexer BN, Garrett JT, Arteaga CL. Mutations in the phosphatidylinositol 3-kinase pathway: role in tumor progression and therapeutic implications in breast cancer. *Breast Cancer Res*. 2011;13:1–12.
27. Ščupáková K, Adelaja OT, Balluff B, Ayyappan V, Tressler CM, Jenkinson NM, et al. Clinical importance of high-mannose, fucosylated, and complex N-glycans in breast cancer metastasis. *Jci Insight*. 2021;6.
28. Farnie G, Clarke RB, Spence K, Pinnock N, Brennan K, Anderson NG, et al. 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;99:616–27.
29. Li Z, Huang Q, Lin B, Guo B, Wang J, Huang C, et al. 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.
30. Zeng S, Škrabišová M, Lyu Z, Chan YO, 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. 2020 IEEE Int Conf Bioinforma Biomed BIBM. 2020. p. 1408–15.
31. Zeng S, Škrabišová M, Lyu Z, Chan YO, Dietz N, Bilyeu K, et al. 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 Bioinforma*. 2021;25:65–85.
32. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Int Conf Mach Learn*. pmlr; 2015. p. 448–56.
33. 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;15:1929–58.
34. Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. *Proc Icml*. Atlanta, Georgia, USA; 2013. p. 3.
35. Kingma DP, Ba J. Adam: A method for stochastic optimization. *ArXiv Prepr ArXiv14126980*. 2014;

36. Facebook. React [Internet]. 2022. Available from: <https://reactjs.org/>
37. Google. Material-UI [Internet]. 2023. Available from: <https://material-ui.com/>
38. Inc PT. Collaborative data science [Internet]. Montreal, QC: Plotly Technologies Inc.; 2015. Available from: <https://plot.ly>
39. Django Software Foundation. Django [Internet]. 2019. Available from: <https://djangoproject.com>
40. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature*. 2020;585:357–62.
41. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17:261–72.
42. Celery Team. Celery [Internet]. 2021. Available from: <https://github.com/celery/celery>
43. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems [Internet]. 2015. Available from: <https://www.tensorflow.org/>
44. Chollet F, others. Keras [Internet]. 2015. Available from: <https://keras.io>
45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
46. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. *Proc 25th ACM SIGKDD Int Conf Knowl Discov Data Min*. 2019. p. 2623–31.
47. Oracle Corporation. MySQL [Internet]. 2021. Available from: <https://www.mysql.com/>
48. Sanfilippo S, Labs R. Redis [Internet]. 2022. Available from: <https://redis.io/>
49. JWT Team. JSON Web Token [Internet]. 2015. Available from: <https://jwt.io>

Figures

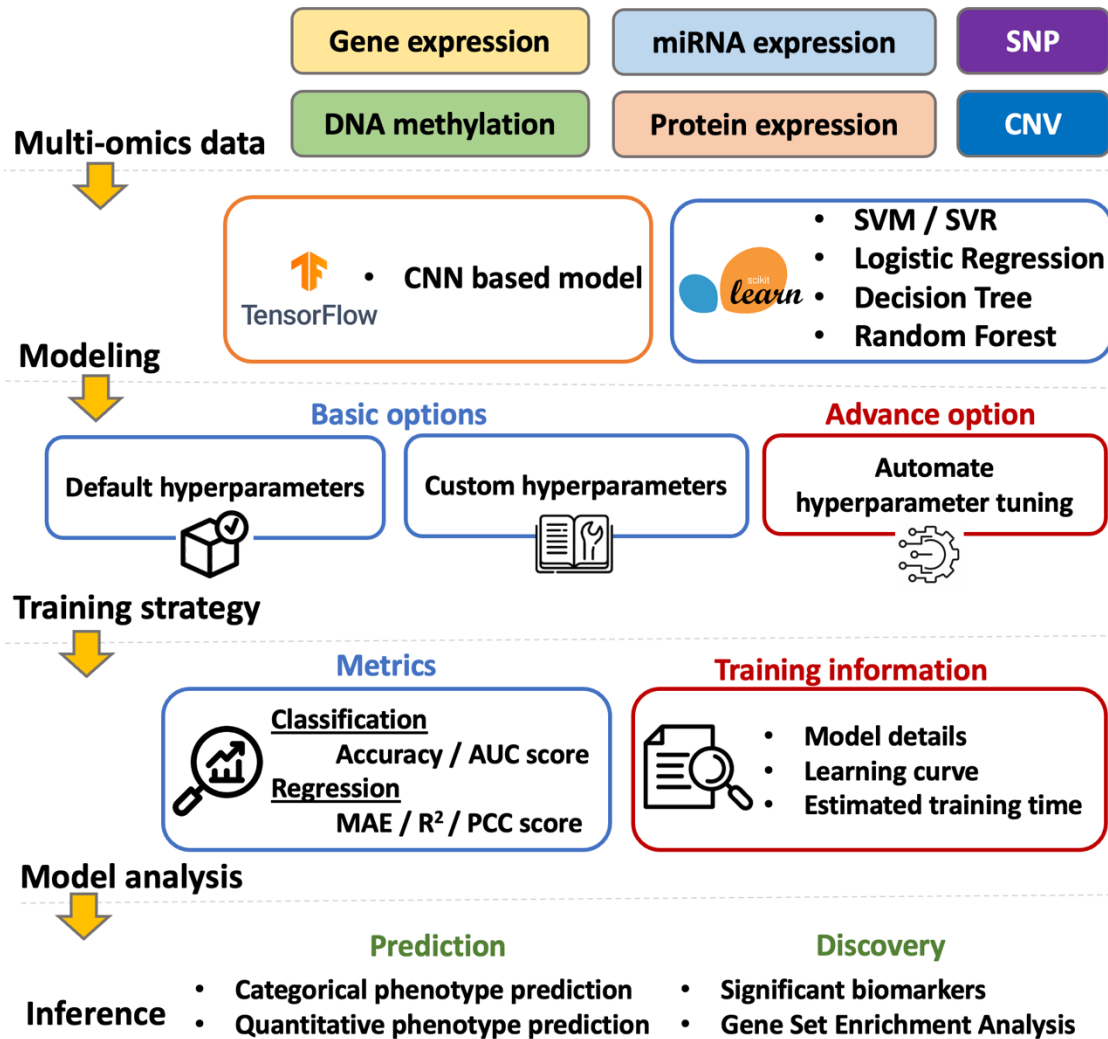


Fig. 1. Overview of G2PDeep-v2.

A

Create New Dataset

Dataset Name *

Choose data type *

Gene expression▼

Upload training and validation dataset *

Transfer file from shared link▼

[Link to example](#)

https://raw.githubusercontent.com/shuaizengMU/AI_backend/master/media/datasets/BRCA.csv

B

Public Datasets Summary

All Plants and Crops Animals and Pets Humans and Diseases Microbes and Viruses

Dataset Name	Dataset Type	# Samples	# Features	Label Type
GBMLGG - RNASeq data for LTS/non-LTS	G	301	20533	Discrete
GBMLGG - methylation450 data for LTS/non-LTS	D	281	123100	Discrete
GBMLGG - CNV data for LTS/non-LTS	C	571	24778	Discrete
GBMLGG - miRNA data for LTS/non-LTS	M	195	1048	Discrete
GBMLGG - ProteinExpression data for LTS/non-LTS	P	177	227	Discrete

C CNV D DNA Methylation G Gene Expression S SNP P Protein Expression M miRNA Expression Z Zygosity

Fig. 2. Dataset creation and retrieval in G2PDeep-v2. (A) Example of dataset creation by a shared link to data. (B) Publicly available datasets are shown with structured information.

Create a Model

Model name *
BRCA cancer study - CNN model

Choose data type (Choose min 1) *

- miRNA expression
- Gene expression
- Protein expression
- Copy number (gene-level)
- DNA methylation
- Zygosity - homozygous, heterozygous, and reference homozygous
- SNP - adenine (A), thymine (T), cytosine (C) and guanine (G)

Task *

- Classification Regression

Model *

Multi-CNN

Hyperparameter Option *

- Default hyperparameters
- Customize hyperparameters
- Automate hyperparameter tuning

Create a Custom Hyperparameters

CSV

Tower 1 (CNN) Tower 2 (CNN) Tower 3 (CNN)

Add/Delete layer (min:1, max:3) Add/Delete layer (min:1, max:3) Add/Delete layer (min:1, max:3)

1D - CNN 10 4

Size Fully connected neural network

DEL LAYER ADD LAYER

Filters size Kernel Size

A. **B.**

Fig. 3. Interactive chart to configure the deep-learning model in G2PDeep-v2. (A) Options for inputting details such as the model, task, and input data. (B) Hyperparameters tuning options.



Fig. 4. Project page in G2PDeep-v2. (A) Model details show the type of model and corresponding training dataset; (B) Figure of multi-CNN model; (C) Learning curve for training and validation datasets; (D) Distribution of ground truth and predicted values for training and validation datasets; (E) ROC curve for phenotype prediction; (F) Optimization history shows improvement of the model during the automate hyperparameter tuning.

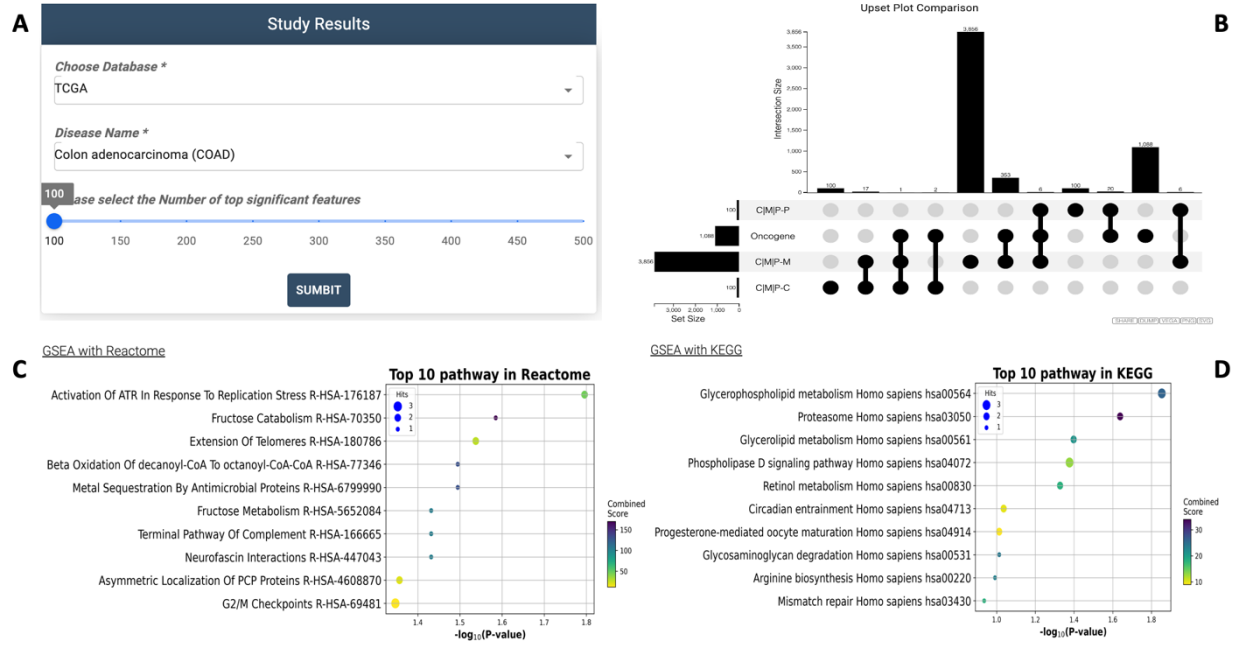


Fig. 5. Study Results page in G2PDeep-v2. (A) Panel to select study; (B) Upset plot shows overlapping significant biomarkers; (C) GSEA analysis with Reactome for significant biomarkers; (D) GSEA analysis with KEGG for significant biomarkers.

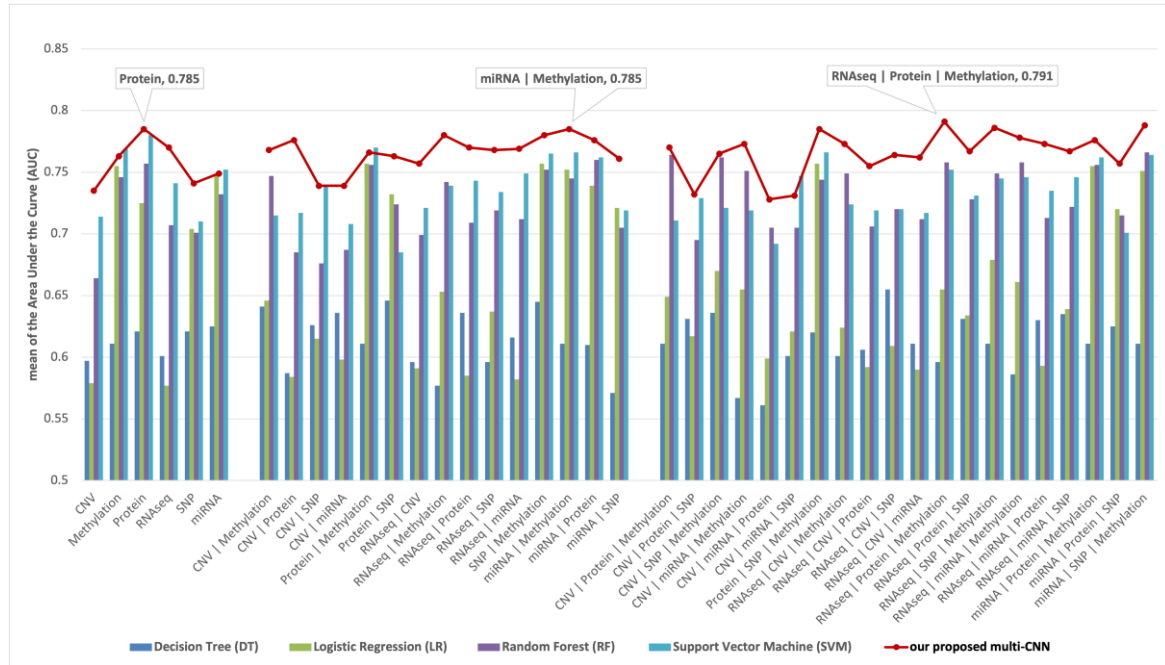


Fig. 6. Mean AUC of models on 41 datasets from the Skin Cutaneous Melanoma (SKCM) study. Models are trained on each dataset individually. The result indicates our proposed multi-CNN model outperforms other traditional machine learning models.

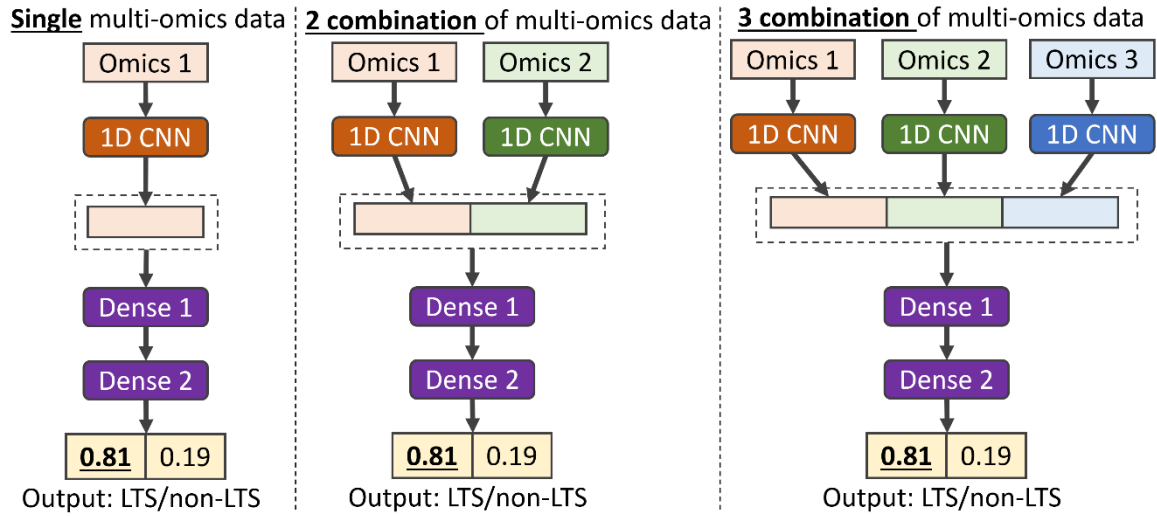


Fig. 7. An example architecture of the multi-CNN model designed for long-term survival prediction using input data with single, two combinations, and three combinations of multi-omics data.

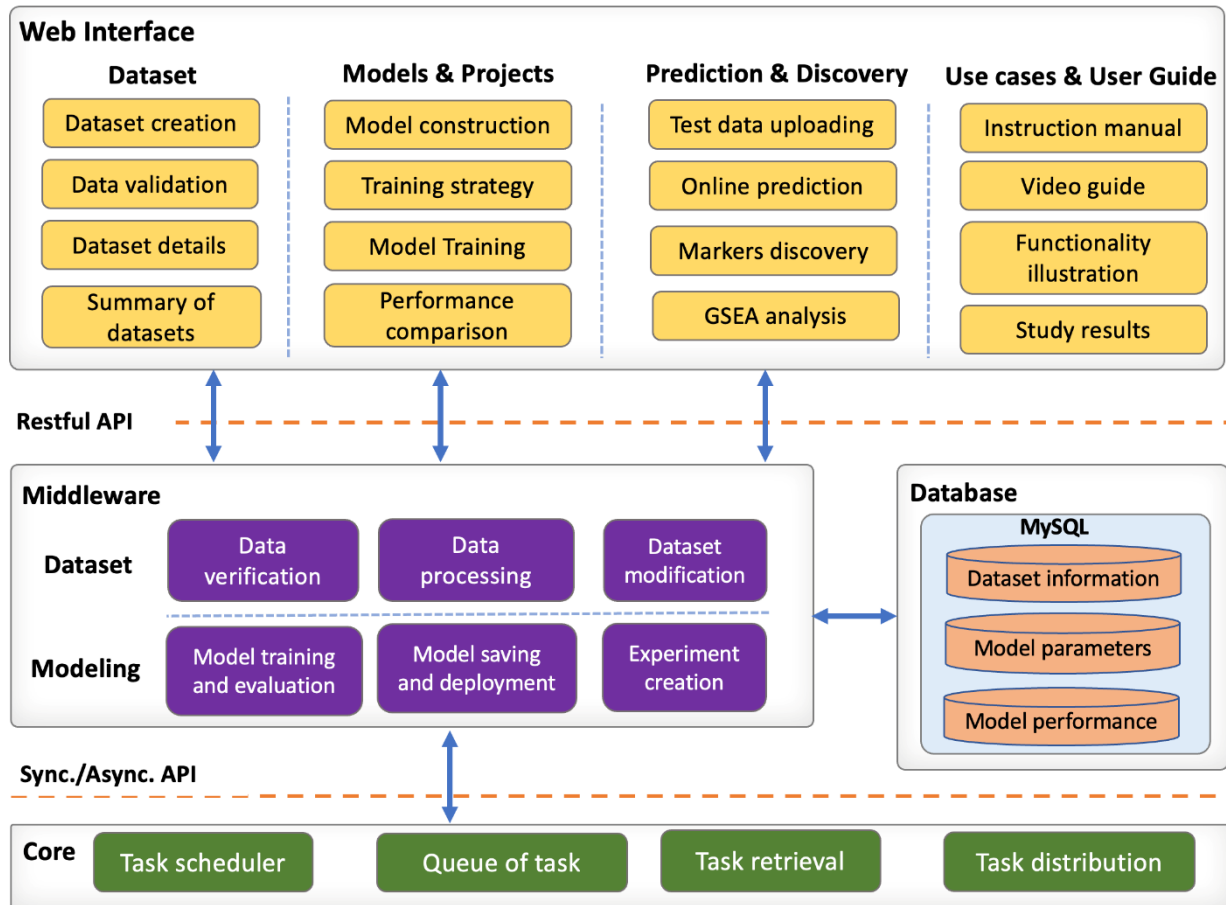


Fig. 8. The architecture of G2PDeep. The architecture consists of four modules and these modules communicate with each other via appropriate APIs.

Tables

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 tuning		✓
	Hyperparameter tuning monitoring		✓
Online prediction	Prediction with test dataset	✓	✓
Marker discovery	Identifying significant markers	✓	✓
	GSEA with KEGG/Reactome		✓
	Literatures related to significant markers		✓

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

Study	# of samples (LTS/Non- LTS)	Number of features					
		Gene expression	miRNA expression	DNA methylation	protein expression	SNP	CNV
BLCA	42 (15/27)	20,533	1,048	300,869	225	18,634	24,778
HNSC	39 (14/25)	20,533	1,048	300,973	239	17,796	24,778
LUAD	33 (16/17)	20,533	1,048	300,822	239	18,950	24,778
LUSC	28 (15/13)	20,533	1,048	300,970	239	18,822	24,778
SARC	26 (15/11)	20,533	1,048	299,776	219	12,422	24,778
SKCM	41 (29/12)	20,533	1,048	300,455	225	19,488	24,778

Table 2. Uniform dataset for 6 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)