Computational and Structural Biotechnology Journal 19 (2021) 2719–2725

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

DeepOmix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis

Lianhe Zhao [a,b,1], Qiongye Dong [a,1], Chunlong Luo [a,b], Yang Wu [a], Dechao Bu [a], Xiaoning Qi [a,b], Yufan Luo [a,b], Yi Zhao [a,c,*]

[a] Key Laboratory of Intelligent Information Processing, Advanced Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Hwa Mei Hospital, University of Chinese Academy of Sciences, Ningbo 315000, China

ARTICLE INFO

Article history:
Received 1 March 2021
Received in revised form 26 April 2021
Accepted 27 April 2021
Available online 1 May 2021

Keywords:
Multi-omics
Deep learning
Survival analysis
Prognosis prediction
Interpretable model

ABSTRACT

Integrative analysis of multi-omics data can elucidate valuable insights into complex molecular mechanisms for various diseases. However, due to their different modalities and high dimension, utilizing and integrating different types of omics data suffers from great challenges. There is an urgent need to develop a powerful method to improve survival prediction and detect functional gene modules from multi-omics data. To deal with these problems, we present DeepOmix (a scalable and interpretable multi-Omics Deep learning framework and application in cancer survival analysis), a flexible, scalable, and interpretable method for extracting relationships between the clinical survival time and multi-omics data based on a deep learning framework. DeepOmix enables the non-linear combination of variables from different omics datasets and incorporates prior biological information defined by users (such as signaling pathways and tissue networks). Benchmark experiments demonstrate that DeepOmix outperforms the other five cutting-edge prediction methods. Besides, Lower Grade Glioma (LGG) is taken as the case study to perform the prognosis prediction and illustrate the functional module nodes which are associated with the prognostic result in the prediction model.

1. Introduction

The advent of high-throughput omics technologies, such as the next-generation sequencing [1], DNA microarrays [2], and DNA methylation arrays [3] enable the measurement of thousands of molecules at the same time from a biological sample comprehensively. Each type of omics data provides unbiased characterization on one aspect of genome, transcriptome, and epigenome, which raises opportunities for biological and medical research explorations [4,5]. However, analysis of single omics data is limited to exploring the underlying biological mechanisms and capturing intricacy for various complex diseases, which only can explain its molecular field respectively [6]. Therefore, integrating multi-omics data at different levels yields a better understanding of overall disease alterations, has an enormous impact in cancer profiling, diagnosis, and treatment, and elucidates the relationships among different types of omics data for one specific disease [7].

Developing accurate survival prediction models of cancer benefits the identification of effective prognostic biomarkers, improvement of risk stratification, and personalized treatment. With the accumulation of a tremendous number of multiple omics data in the past decades, it brings opportunities to build the prediction model and infer the systematic underlying biological mechanisms through making an integrative analysis. However, it raises new computational challenges in the data integration due to the heterogeneous characteristics and distribution of different types of data, the high dimensionality of each level of a molecular dataset, and a limited number of observations [8–10]. To address these issues, a variety of regression methods have been proposed to build the prognostic model through integrating multi-omics data [11].

There are mainly four kinds of approaches to predict the survival time, namely penalized regression, boosting, random forest, and deep learning-based methods [12]. Integrative LASSO with Penalty Factors (IPF-LASSO) [13], an extension of LASSO method, is designed to make the L1-penalized regression analysis by using different penalty weights for each type of omics data to train the

* Corresponding author at: Key Laboratory of Intelligent Information Processing, Advanced Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: biozy@ict.ac.cn (Y. Zhao).
[1] Lianhe Zhao, Qiongye Dong have contributed equally to this work.

https://doi.org/10.1016/j.csbj.2021.04.067

model. Component-wise gradient boosting with a generalized linear model (glmboost) is proposed for regression, time-to-event, classification analysis with variable selection [14]. Both glmboost and IPF-LASSO can be applied to multi-omics data with high dimension, which perform the feature selection and reduce the complexity of the model. However, they could only detect the linear relationships between features and outcome variables through fitting linear models. Block forest, a variant from random forest, incorporates the block structure of multi-omics data and makes the non-linear prediction of the clinical outcomes [15]. The limitation of block forest is that they can't extrapolate the data.

Some deep learning-based methods were also applied to train the survival model. DeepSurv was a method of the deep feed-forward neural network to perform a prediction of time-to-event to make personalized treatment recommendations. It estimated each sample's effect on their hazard rates concerning parametrized weights of the network, which was configurable with multiple numbers of hidden layers [16]. DeepHIT was a deep neural network architecture to learn the joint distribution of survival time and event directly without making assumptions about the underlying stochastic process [17]. The parameters of the model and the form of the stochastic process depended on the features of the input dataset used for survival analysis. However, these two methods were not designed for integrating multi-omics data and lacked interpretability. It was urgent to propose an interpretable nonlinear model for multi-omics data integration and survival prediction.

To fill the gaps of the algorithm in this field, we presented DeepOmix, a scalable and interpretable deep learning framework for multi-omics data integration and survival prediction. DeepOmix learned meaningful information by incorporating prior biological knowledge of gene functional module networks as the function module layer since genes perform functions in cells in the form of a synergistic and regulatory system [18]. Getting the low-dimensional representations in the functional module layer facilitates extracting significant modules corresponding to the prognostic prediction result. The functional modules can be defined by users, including tissue networks [19], gene co-expression networks [20], or prior biological signaling pathways [21]. We performed experiments of benchmark comparison and elucidated that the performance of DeepOmix outperformed other existing state of art prediction models. Then, Low-Grade Glioma (LGG) was taken as the case study. Patients were grouped into two subtypes with significant differences in survival time based on the output layer of the prediction result. The difference of functional nodes on the module layers in these subtypes was tested and top-ranked functional modules were detected. DeepOmix can integrate multi-omics data by incorporating prior biological knowledge to conduct prognosis prediction and learn the low representations on the module layers to understand the underlying mechanisms for further study.

2. Results

2.1. DeepOmix: A multi-omics scalable and interpretable prognosis prediction framework

DeepOmix efficiently implemented a non-linear combination of variables from different omics datasets (Fig. 1A) and incorporated prior biological information defined by users (Fig. 1B) (such as signaling pathways and tissue networks). The deep learning framework of DeepOmix was designed as in Fig. 1C, the detail of which was in the method section.

DeepOmix integrated different omics data as input gene layer and the gene layer nodes were connected with functional layer according to the prior information from pathways or functional modules as input defined. The basic idea behind this model was that in different biological processes, genes exercised their functions in the form of functional modules instead of working alone. The functional module layer was the low-dimensional representations, and each node was a non-linear function of the values at the different molecular levels (mutations, copy number alterations, gene expressions, and DNA methylation) of the genes it contained. Meanwhile, the relationships were various due to different diseases or biological processes. Users can define their own functional modules according to their clinical needs or trail experiments and signaling pathway gene sets were used in our analysis.

After training the model, samples would be classified into two groups, namely high-risk and low risk, according to the values in the output layer (Fig. 1D). Then, the nodes in the functional module layer were treated as the low-dimensional representations, namely new features to elucidate the underlying mechanisms among these two groups with different prognoses. To adumbrate the prognosis result and interpretability of DeepOmix, LGG was taken as the case study.

2.2. Performance comparison with other methods

We compared DeepOmix with other five state-of-the-art methods on eight different cancer datasets. According to the evaluation review published recently [12], top-ranked methods for multi-omics data integration and survival time prediction were selected in our comparison, including IPF-LASSO [13], glmboost [14], and block forest [15]. Besides, two widely used deep learning-based survival time prediction methods, DeepSurv [16] and DeepHIT [17] were also included. Eight different cancer types from The Cancer Genome Atlas (TCGA) project included bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRAC), head-and-neck squamous cell carcinoma (HNSCC), lower grade glioma (LGG), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), ovarian serous cystadenocarcinoma (OV) and stomach adenocarcinoma (STAD).

Concordance index (C-index) [22], a widely used performance metric of the survival prognosis prediction model was applied in the evaluation of the experiments. The performance results for six algorithms on eight datasets mentioned above were summarized in Supplementary Table 2. Each method was conducted on eight datasets, and the eight calculated metric values of the C-index were used to evaluate its performance. C-index differences between the other five methods and DeepOmix were shown in Fig. 2A, and statistical method of one-tailed *t*-test was carried out to elucidate that the mean differences were significantly smaller than zero within all the comparisons. On six out of eight datasets, DeepOmix performed best with the highest average C-index (Fig. 2B). The comparison result showed that DeepOmix was robust and outperformed other methods significantly.

2.3. Case study

2.3.1. Prognosis prediction result of DeepOmix on LGG (Lower Grade Glioma)

Dataset of LGG was taken as an example to perform the case study. First, DeepOmix was applied to the LGG dataset to make the prediction. Then, patient samples were classified into two subtypes with two different prognostic status, namely high-risk and low-risk, according to the values of the output layer. Fig. 3A showed that the difference of survival time was significantly different between these two subgroups (Kaplan-Meier curve) through cox-PH model. Multi-Dimensional Scaling (MDS) was conducted on the second hidden layer nodes for further visualization on the lower dimensional space (Fig. 3B). The features were scaled into

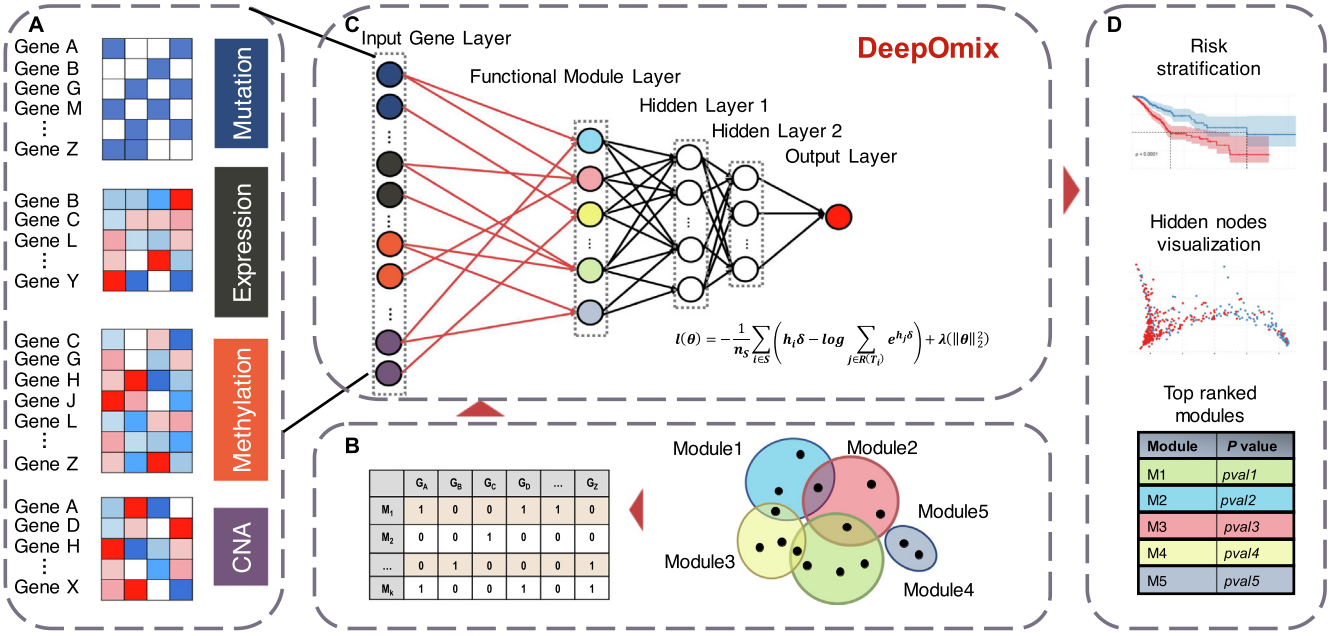

Fig. 1. Workflow of DeepOmix. A. Multi-omics data at the gene-level was used as the input data on the input gene layer. B. Functional module gene sets defined by users determines the number of nodes in the functional module layer and the edges between this layer and the input gene layer. C. The framework in DeepOmix. It includes five layers, namely the input gene layer, functional module layer, two hidden layers, and the output layer. The functional module layer is the low-dimensional representations of the gene layer, which is a non-linear function of the gene nodes. D. Samples were classified into high and low risk subgroups according to the output layer, treated as the prognostic values. *Kolmogorov-Smirnov* test was performed to rank the meaningful pathways and visualization of nodes on hidden layers among two groups.

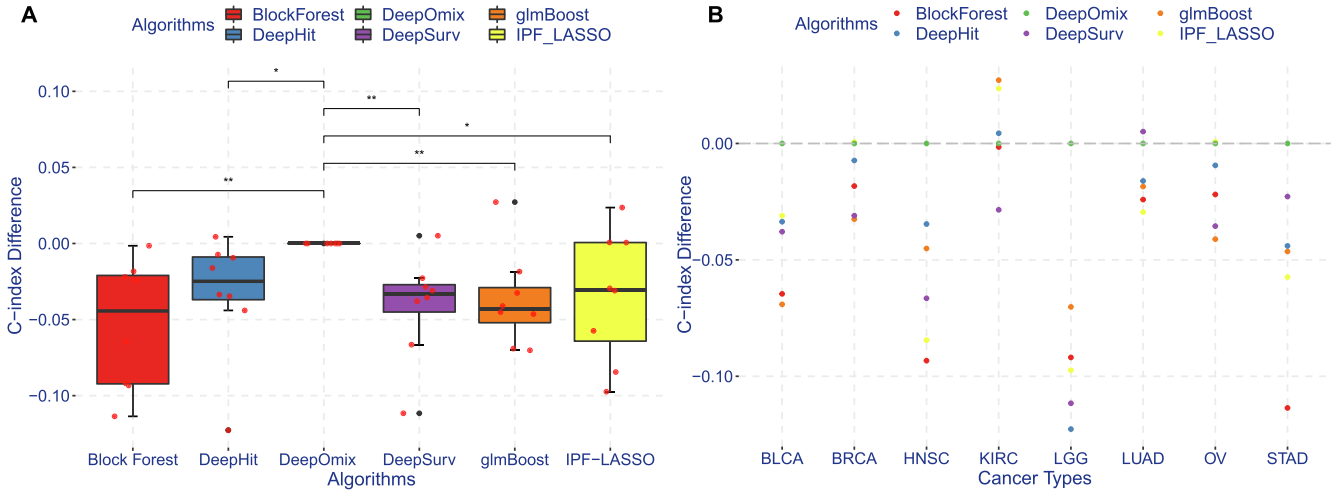

Fig. 2. Performance comparison between DeepOmix and five other methods (A) Boxplot of Differences between the average C-index of other methods and DeepMusicsBoxplot of C-index difference between other methods and DeepOmix. One-tailed *t*-test was applied to test the difference of the means of the differential C-index for each comparison. (B) Differences between the average C-index of other methods and DeepOmix on each cancer dataset.

two dimensions, and the scatter plot shown that samples in the two groups of different prognosis prediction results were split.

2.3.2. Detection of the functional modules significantly different among two prognostic groups for LGG

DeepOmix was biologically interpretable since the functional module information was incorporated as the functional layer. It was able to capture the nonlinear and hierarchical effects of biological pathways associated with survival time. The nodes in the functional layer were used as the lower representations for the patients. For each node of the functional module, the *Kolmogorov-Smirnov* test [23] was performed to test whether the distribution was significantly different between the previously defined groups of samples. The top pathways were listed in Table 1 and Supplementary Table 3.

The top one pathway, formation of incision complex in GG-NER participates the process of DNA repair, participates the process of DNA repair and affects the sensitivity to cancer therapeutics [24]. A recent study elucidated that DDR-related cytokines had prognostic implications on glioma patients [25] and it affected the malignant progression of LGG after temozolomide treatment [26].

Interestingly, Advanced Glycosylation End-product Receptor (AGER) signaling pathway, helped amyloid-beta peptide (Aβ) and mediated Aβ neurotoxicity, and promoted Aβ influx into the brain [27]. Aβ was accumulated naturally in glioma tumors and nearby blood vessels in a mouse model of glioma [28]. The Folate

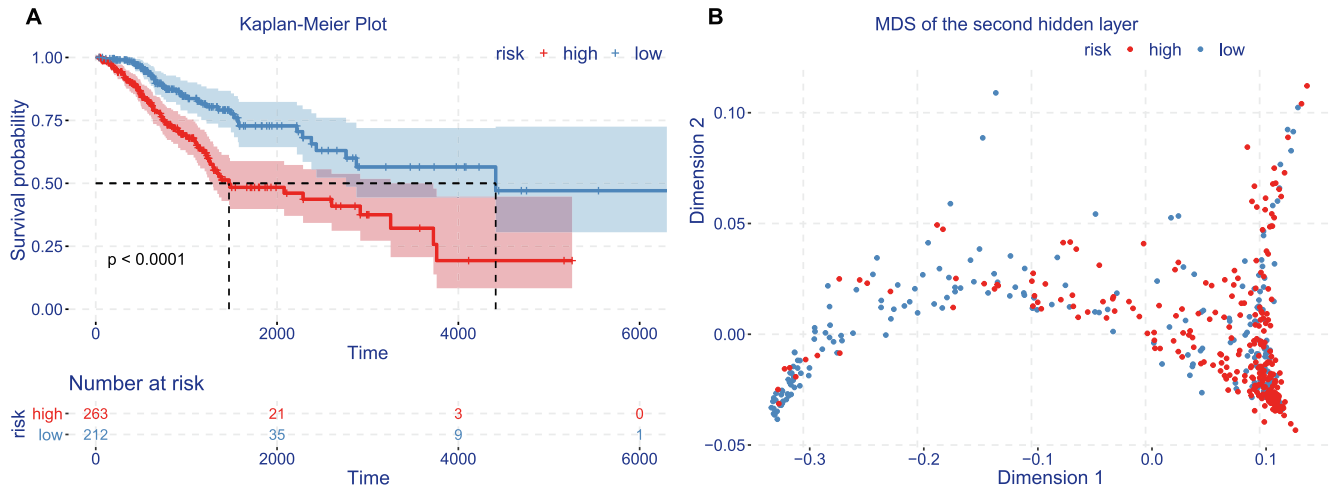

Fig. 3. Prognosis prediction result of DeepOmix on LGG. A. Kaplan-Meier plot for two different survival risk groups. B. Visualization of Multi-Dimensional Scaling (MDS) result of the second hidden layer in the two different prognostic groups.

Table 1
Ten top-ranked pathways in LGG.

Pathway name	*p*-value	Bonferroni adjusted *p*-value
Formation of Incision Complex In GG-NER	1.90E-14	1.63E-11
Glycoprotein Hormones	5.73E-14	4.93E-11
KEGG of Colorectal Cancer	2.69E-12	2.32E-09
Adenylate Cyclase Inhibitory Pathway	3.94E-11	3.39E-08
Advanced Glycosylation End-product Receptor Signalling	1.03E-10	8.89E-08
KEGG of Acute Myeloid Leukaemia	5.61E-09	4.83E-06
RNA Pol III Chain Elongation	7.42E-09	6.38E-06
Folate Biosynthesis	6.31E-08	5.43E-05
KEGG of Alpha Linolenic Acid Metabolism	6.36E-08	5.47E-05
HS GAG Biosynthesis	9.85E-08	8.47E-05

Biosynthesis pathway was reported to be at risk of childhood brain tumors [29]. Regulation of insulin secretion by acetylcholine was reported that insulin-mediated signaling facilitates resistance to PDGFR inhibition in pro-neural hPDGFB-driven gliomas [30]. SNPs in gene ADCY, the key gene in the Adenylate Cyclase (ADCY) inhibitory pathway, affected glioma risk in a sex-specific fashion, elevating the risk for females while protecting males [31].

3. Methods

3.1. Overview

The core principle behind DeepOmix is to learn the representations of the modules by integrating multi-omics data and user-defined functional modules. Each module is represented by a non-linear function of the multiple omics' values of genes it contains. Embedding is conducted to learning each module representation on the lower dimension for each input sample.

3.2. Data acquisition and pre-processing

Three kinds of data were used to train the model, namely patients' multi-omics data, their clinical survival time data, and functional module data defined by users. The top eight cancer datasets by the sample size available of multi-omics and survival time data were obtained from LinkedOmics [32], which contains multi-omics data and clinical data of patients from The Cancer Genome Atlas (TCGA) project. For one certain cancer, only the samples with four types of omics data, including somatic mutation data, copy number alteration data, gene expression data, and DNA methylation data, were used in the following analysis. Prior biological knowledge, signaling pathways were taken as the functional module input in our research. KEGG and Reactome [33,34] pathway gene sets were obtained from Molecular Signatures Database (MSigDB) [35]. Pathways with over 200 genes or less than 20 genes were excluded since small pathways might be redundant with other larger pathways and large pathways might be related to the general biological pathways, rather than specific to a certain disease [36].

For one certain cancer, the processing of the datasets consists of two steps. First, four types of omics data at the gene-level were collected (details in Fig. 4), and variables in each type of omics data were filtered by overlapping with the genes in the functional modules respectively. For mRNA data, protein-coding genes were selected from the raw counts of mRNA expression data, and genes with zero read counts in more than 20% samples were filtered out. Then, the original raw read counts of RNAseq data were normalized through conducting the R package *DESeq2* [37], and read counts were transferred into the logarithmic space through "*log2(counts + 1)*". Second, since variables in the gene expression value of mRNA data, beta-value of methylation data, and log2 ratio of GISTIC2 [38] of CNA data were continuous, we processed these three types of omics data in the same way. We kept the top 5000 variables according to their standard deviations among the patients. Each variable was normalized into a standard normal distribution through the R function *scale*.

3.3. Model construction

DeepOmix was designed as a feed-forward neural network, built with five layers (Fig. 1), including the input layer, one functional module layer, two hidden layers, and the output layer of survival time. The input layer was composed of the normalized four different omics data. For the i-th sample, x_{igk} represent the k-th omics data of gene g (k = 4 in our analysis). The second layer represented the gene functional modules, the number of nodes in which was the number of functional modules (signaling pathways in our analysis). The edges between the gene layer and functional layer were constructed based on the prior knowledge of pathway gene sets. If the gene belongs to the pathway, an edge was added

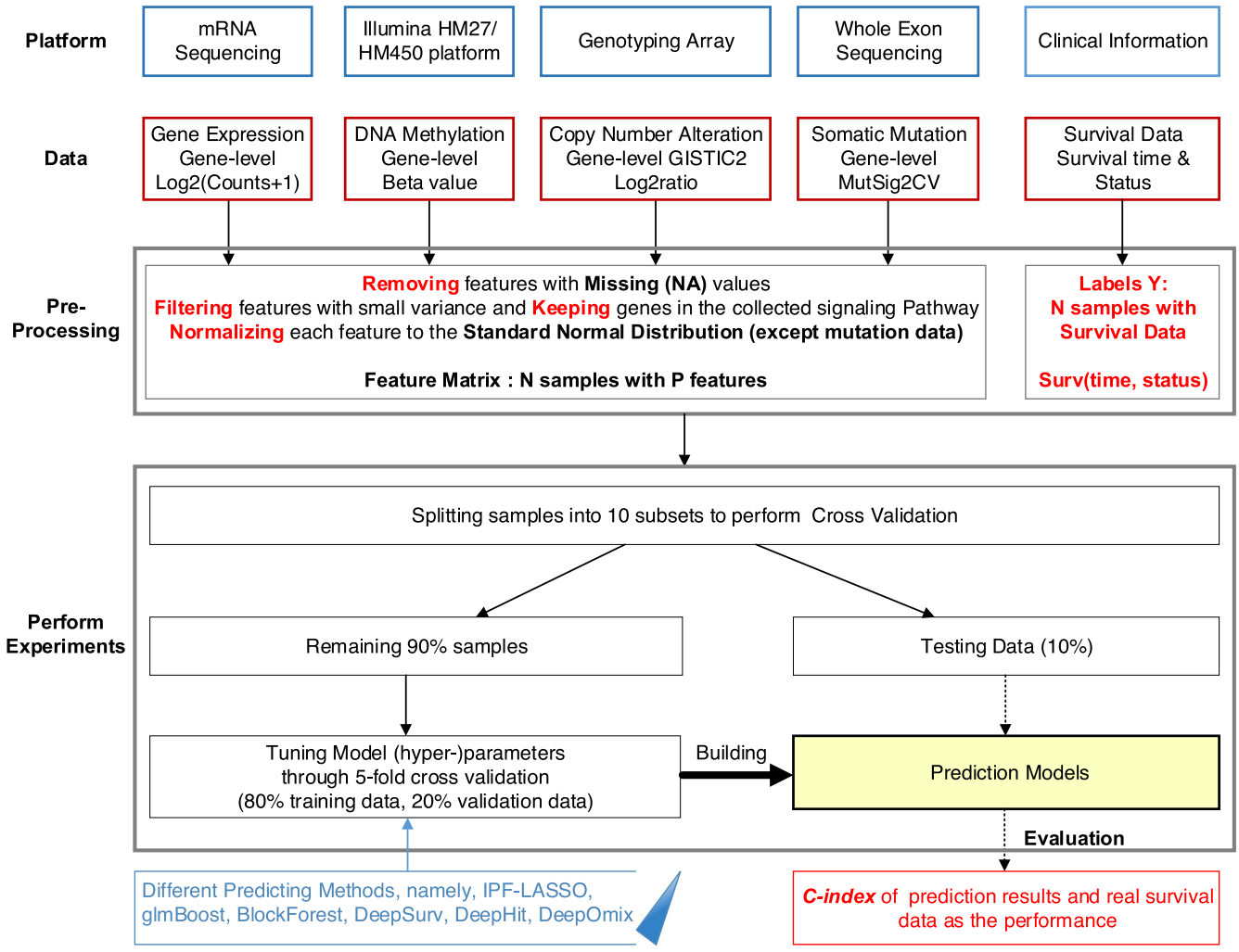

Fig. 4. The pre-process of multi-omics data and the pipeline of performance evaluation experiments.

between the g-th gene and p-th pathway. The features of the pathway layer were constructed through an encoder $q_p(x)$ of the gene layer with a non-fully connected network: $S_{ip} = q_p(x) = \Phi\left(\sum_{g \in pathway_p} (w_{gkp} * x_{igk})\right)$, where $\Phi(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$ as the activation function in our analysis. Then, the pathway features were transformed to the next two hidden layers, and finally to the output layer of the survival data composed of survival time and status with the fully connected networks.

DeepOmix optimized the parameters by minimizing the average negative log partial likelihood with L^2 regularization, and the objective function of the average negative log partial likelihood was defined as:

$$\boldsymbol{l}(\theta) = -\frac{1}{\boldsymbol{n_S}} \sum_{i \in S} \left(\boldsymbol{h_i \delta} - \boldsymbol{\log} \sum_{j \in R(T_i)} \boldsymbol{e^{h_j \delta}} \right) + \boldsymbol{\lambda} \left(\|\theta\|_2^2 \right)$$

where h is the second hidden layer's outputs; $\boldsymbol{S}$ is a set of uncensored samples; and $\boldsymbol{n_S}$ is the total number of uncensored samples. $\boldsymbol{R(T_i)} = \{\boldsymbol{i} | \boldsymbol{T_i} \geq \boldsymbol{t}\}$ is a set of samples at risk of failure at time t. $\|\theta\|_2^2$ is the L^2-norms of $\{\delta, \boldsymbol{W}\}$together. δ is the weight between the last hidden layer and the clinical layer; and W is a union of the weight matrices. λ is a regularization hyper-parameter to control sensitivity ($\lambda > 0$). We optimized the model by partially training small sub-networks with sparse coding. Training a small subnetwork guaranteed feasible optimization, with a small set of parameters in each epoch.

3.4. Performance evaluation

The predictive performance of DeepOmix was evaluated by comparing it with the other five state-of-the-art methods mentioned above, namely glmboost, IPF-LASSO, block forest, DeepHIT, and DeepSurv. To evaluate the survival result predicted by different algorithms, the overall performance was assessed via the concordance index (C-index) [22]. C-index is the most widely used metric that measures the discriminative power of the model by comparing the predicting results with the real survival time. We calculated C-index using function *concordance index* in R package *survcomp* [39].

3.5. Experiments with cross validation

For each method, we repeated 10-fold cross validation ten times to measure the performance on multi-omics data of each type of cancer for the reproducibility of model performance. For each iteration in the repetition of cross validation, first, the samples were randomly grouped into ten subsets, 10% left-out samples for the test data, and the remaining 90% for building the model and tuning the (hyper-) parameters. Then, the remaining 90% samples were

split into five subsets randomly with ensuring the same censoring percentage to perform 5-fold cross validation, and every round, one dataset for validation datasets while four for training data. The model was built on the training data and (hyper-)parameters were tuned by the use of the validation dataset. Finally, once the prediction model was well-trained, the test data were used as input data to perform the prediction and calculate the predictive performance C-index [22]. Finally, the average of the C-index across all the iterations and repetitions was calculated as the final metric.

3.6. Algorithm configurations

IPF-Lasso was performed via R package *ipflasso* [13]. Every type of omics data was input as a block, and the penalty factors for different blocks were selected through 5-fold cross validation.

Block forest was fitted via R package *blockForest* [15]. Block information of different omics data was provided. For blocks with more than 2500 variables, the variables were selected with the smallest *p*-values from the univariate Cox proportional hazards (*cox-PH*) regression model [40] (coxph function in the R package *survival*). The parameters were set as nsets = 300, num.trees.pre = 1500 and nu,.trees = 2000 as the paper [15] suggested.

The method of glmboost was conducted via function *glmboost* in the R package *mboost* [14]. The family argument was set as *CoxPH ()*. The parameter of the number of boosting steps (mstop) was selected on a grid from 1 to 2000 through 5-fold cross validation.

For ~~the~~ two deep learning-based methods, four types of omics data were integrated as one input data X, and each feature from the four datasets was treated as one covariate. Samples were grouped into training, testing, and validation subsets. For DeepSurv [16], we chose ReLU as an optimal activation layer with batch normalization and dropout 40% on the dropout layers. Adam optimizer was used for model training, without setting the initial learning rate value. For the method of DeepHIT, the process of training was performed with mini-batches of the training set over 50,000 iterations. Every 1000 iteration, a prediction was conducted on the validation set and the best model was saved to the specified file path. The evaluation of the models was based on the concordance index. The best result was returned if there was no improvement for the next 6000 iterations (early stopping).

3.7. Functional analysis

We took the LGG as the case study in this part. First, we grouped samples of LGG into two parts according to the node values on the output layer with the threshold of median value, namely high-risk group and low-risk group. The survival difference in these two groups was tested via cox-PH model [40]. Nodes in the functional module layer are the low-dimensional representations for the input gene layer. After grouping, to get the meaningful pathways, *Kolmogorov-Smirnov* test [23] was performed (function *ks.test* in R) to test the distribution difference of values on each node of the functional layer among two groups. Then, Bonferroni adjusted method [41] was applied to perform the multiple correction for the *p*-values.

4. Discussion

The application and integration of different molecular profiling technologies create novel opportunities for personalized medicine; however, they also bring challenges. Heterogeneity of the samples and high dimension are the main bottlenecks to perform the combination of multi-omics data and predictions of the clinical outcomes, such as survival time. This calls for better strategies to build the prediction model and identify the related underlying mechanisms. Although there are many studies for prognosis prediction of cancer patients, embedding multiple omics data sources of the patients by combining the gene functional modules and fitting a non-linear function of the omics data has not been reported, as far as we know.

We addressed this problem by building a scalable and interpretable framework called DeepOmix, an explainable framework for the analysis of highly multiplexed omics data. It leverages functional module information to explore the biological mechanisms that might be associated with the prediction result. DeepOmix outperforms other top ranked cutting-edge methods for integrating multi-omics data and predicting the survival time. It provides an exploratory approach to improve the biological understanding of the prognosis prediction model. In addition, in the case study of LGG, top-ranked identified functional pathways that are associated with prognosis groups are confirmed by previous studies.

The work we presented lays the foundation for further researches. Apart from the used four types of omics data, DeepOmix can be expanded to integrate increasingly complex data, such as proteomic data. This deep learning framework with functional module layer could also be applied to predict other clinical outcomes, including the categorical variables (such cancer subtypes and cancer stages) and continuous variables (such as drug response). In the future, single cell RNA sequencing data or spatial transcriptomics data will be more available for next version. In summary, we believe that DeepOmix is a valuable tool to make an integrative analysis of different resolved-omics data and build prediction models.

5. Code availability

The code of DeepOmix and the scripts to generate the results shown in this paper are available at https://github.com/CancerProfiling/DeepOmix.

CRediT authorship contribution statement

Lianhe Zhao: Methodology, Software, Writing - original draft, Writing - review & editing. **Qiongye Dong:** Data curation, Writing - original draft, Writing - review & editing, Methodology. **Chunlong Luo:** Investigation. **Yang Wu:** Investigation. **Dechao Bu:** Investigation. **Xiaoning Qi:** Investigation. **Yufan Luo:** Investigation. **Yi Zhao:** Conceptualization, Methodology, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank Jintao Li and Rui Zhang for providing feedback on the framework and manuscript.

Funding

This work was supported by National Key R&D Program of China [2019YFC1709801]; Zhejiang Provincial Research Center for Cancer Intelligent Diagnosis and Molecular Technology [JBZX-202003]; National Natural Science Foundation of China [32070670]; Zhejiang Provincial Natural Science Foundation of China [LY21C060003, LY20C060001].

Funding for open access charge: National Key R&D Program of China [2019YFC1709801].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.04.067.

References

[1] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17(6):333–51.
[2] Dandy DS, Wu P, Grainger DW. Array feature size influences nucleic acid surface capture in DNA microarrays. Proc Natl Acad Sci U S A 2007;104 (20):8223–8.
[3] Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics 2011;98 (4):288–95.
[4] Akey JM, Biswas S, Leek JT, Storey JD. On the design and analysis of gene expression studies in human populations. Nat Genet 2007;39(7):807–8.
[5] Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006;7(1):55–65.
[6] Subramanian I et al. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights 2020;14. 1177932219899051.
[7] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18(1):83.
[8] Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. BMC Syst Biol 2014;8(Suppl 2):I1. https://doi.org/10.1186/1752-0509-8-S2-I1.
[9] Dugourd A, Kuppe C, Sciacovelli M, Gjerga E, Gabor A, Emdal KB, et al. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol Syst Biol 2021;17(1). https://doi.org/10.15252/msb.20209730.
[10] Rohart F, Gautier B, Singh A, Lê Cao K-A, Schneidman D. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol 2017;13(11):e1005752. https://doi.org/10.1371/journal.pcbi.1005752.
[11] Hastie TTR, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.
[12] Herrmann M, et al. Large-scale benchmark study of survival prediction methods using multi-omics data. Brief Bioinform; 2020.
[13] Boulesteix AL et al. IPF-LASSO: integrative L1-penalized regression with penalty factors for prediction based on multi-omics data. Comput Math Methods Med 2017;2017:7691937.
[14] Hofner Benjamin, Mayr Andreas, Robinzonov Nikolay, Schmid Matthias. Nikolay Robinzonov, Matthias Schmidmodel-based boosting in R: a hands-on tutorial using the R Package mboost. Comput Stat 2014;29(1-2):3–35.
[15] Hornung R, Wright MN. Block Forests: random forests for blocks of clinical and omics covariate data. BMC Bioinf 2019;20(1):358.
[16] Katzman Jared L, Shaham Uri, Cloninger Alexander, Bates Jonathan, Jiang Tingting, Kluger Yuval. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 2018;18(1). https://doi.org/10.1186/s12874-018-0482-1.
[17] Ryu JY. et al. DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. Bioinformatics 2020;36(10):3049–3055.
[18] Kemmeren Patrick, Sameith Katrin, van de Pasch Loes AL, Benschop Joris J, Lenstra Tineke L, Margaritis Thanasis, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell 2014;157(3):740–52.
[19] Zitnik M, Leskovec J. Predicting multicellular function through multi-layer tissue networks. Bioinformatics 2017;33(14):i190–8.
[20] van Dam S et al. Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform 2018;19(4):575–92.
[21] Lee Eunjung, Chuang Han-Yu, Kim Jong-Won, Ideker Trey, Lee Doheon, Tucker-Kellogg Greg. Inferring pathway activity toward precise disease classification. PLoS Comput Biol 2008;4(11):e1000217. https://doi.org/10.1371/journal.pcbi.1000217.
[22] Uno Hajime, Cai Tianxi, Pencina Michael J, D'Agostino Ralph B, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011;30(10):1105–17.
[23] Marsaglia George, Tsang Wai Wan, Wang Jingbo. Evaluating Kolmogorov's distribution. J Stat Softw 2003;8(18). https://doi.org/10.18637/jss.v008.i18.
[24] Shuck Sarah C, Short Emily A, Turchi John J. Eukaryotic nucleotide excision repair: from understanding mechanisms to influencing biology. Cell Res 2008;18(1):64–72.
[25] Meng Xiangqi, Duan Chunbin, Pang Hengyuan, Chen Qun, Han Bo, Zha Caijun, et al. DNA damage repair alterations modulate M2 polarization of microglia to remodel the tumor microenvironment via the p53-mediated MDK expression in glioma. EBioMedicine 2019;41:185–99.
[26] van Thuijl Hinke F, Mazor Tali, Johnson Brett E, Fouse Shaun D, Aihara Koki, Hong Chibo, et al. Evolution of DNA repair defects during malignant progression of low-grade gliomas after temozolomide treatment. Acta Neuropathol 2015;129(4):597–607.
[27] Patel AN, Jhamandas JH. Neuronal receptors as targets for the action of amyloid-beta protein (A [beta]) in the brain. Expert Rev Mol Med 2012;14.
[28] Kucheryavykh Lilia Y, Ortiz-Rivera Jescelica, Kucheryavykh Yuriy V, Zayas-Santiago Astrid, Diaz-Garcia Amanda, Inyushin Mikhail Y. Accumulation of innate amyloid beta peptide in glioblastoma tumors. Int J Mol Sci 2019;20 (10):2482. https://doi.org/10.3390/ijms20102482.
[29] Milne Elizabeth, Greenop Kathryn R, Bower Carol, Miller Margaret, van Bockxmeer Frank M, Scott Rodney J, et al. Maternal use of folic acid and other supplements and risk of childhood brain tumors 2012;21(11):1933–41.
[30] Schettini Gennaro, Florio Tullio, Meucci Olimpia, Landolfi Elisa, Grimaldi Maurizo, Ventra Carmelo, et al. Somatostatin inhibition of adenylate cyclase activity in different brain areas 1989;492(1-2):65–71.
[31] Warrington NM, Sun T, Rubin JB. Targeting brain tumor cAMP: the case for sex-specific therapeutics. Front Pharmacol 2015;6:153.
[32] Vasaikar SV, et al. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res 2018;46(D1):D956–D963.
[33] Kanehisa M et al. The KEGG databases at GenomeNet. Nucleic Acids Res 2002;30(1):42–6.
[34] Joshi-Tope G et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 2005;33(suppl_1):D428–32.
[35] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27 (12):1739–40.
[36] Reimand Jüri, Isserlin Ruth, Voisin Veronique, Kucera Mike, Tannus-Lopes Christian, Rostamianfar Asha, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc 2019;14(2):482–517.
[37] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
[38] Mermel Craig H, Schumacher Steven E, Hill Barbara, Meyerson Matthew L, Beroukhim Rameen, Getz Gad. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 2011;12(4). https://doi.org/10.1186/gb-2011-12-4-r41.
[39] Schroder MS, et al. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics 2011;27 (22):3206–8.
[40] Andersen PaGR. Cox's regression model for counting processes, a large sample study. Ann Stat 1982;10:1100–20.
[41] Armstrong Richard A. When to use the Bonferroni correction. Ophthalmic Physiol Opt 2014;34(5):502–8.