Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Research Article

AutoGGN: A gene graph network AutoML tool for multi-omics research

Lei Zhang[a,#], Wen Shen[a,#], Ping Li[b,#], Chi Xu[a], Denghui Liu[a], Wenjun He[a], Zhimeng Xu[a], Deyong Wang[a], Chenyi Zhang[b], Hualiang Jiang[c], Mingyue Zheng[c], Nan Qiao[a,*]

[a] *Lab of Health Intelligence, Huawei Technologies Co. Ltd.*
[b] *Graph Engine Service, Huawei Technologies Co. Ltd.*
[c] *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences*

ARTICLE INFO

Keywords:
Multi-omics Data
Molecular Interaction Network
Graph Convolution Network
Deep Learning
Single-Cell Stage Classification
Cancer Type Classification
Cancer Subtyping

ABSTRACT

Omics data can be used to identify biological characteristics from genetic to phenotypic levels during the life span of a living being, while molecular interaction networks have a fundamental impact on life activities. Integrating omics data and molecular interaction networks will help researchers delve into comprehensive information hidden in the data. Here, we propose a new multimodal method — AutoGGN — to integrate multi-omics data with molecular interaction networks based on graph convolutional neural networks (GCNs). We evaluated AutoGGN using three classification tasks: single-cell embryonic developmental stage classification, pan-cancer type classification, and breast cancer subtyping. On all three tasks, AutoGGN showed better performance than other methods. This means AutoGGN has the potential to extract insights more effectively by means of integrating molecular interaction networks with multi-omics data. Additionally, in order to provide a better understanding of how our model makes predictions, we utilized the SHAP module and identified the key genes contributing to the classification, providing insight for the design of downstream biological experiments.

Introduction

In recent years, high-throughput biomedical research methods, such as whole genome sequencing (WGS), RNA sequencing (RNA-seq), high-throughput chromosome conformation capture (Hi-C), and liquid chromatography coupled with mass spectrometry (LC-MS), have been widely used in biological research, drug development, and precision medicine [1,2]. By integrating multi-omics data generated from omics assays (especially comprehensive genomic and transcriptomic data), research institutions, hospitals, and companies [3] have been able to boost research and innovation in personalized drug design and precision medication. For instance, multi-omics data is particularly useful in mining potential drug targets and identifying cancer-related genes [4], making it indispensable to the research and development (R&D) process [5] of pharmaceutical companies. There have been examples of integrating gene mutation and expression profiles to identify molecular subtypes of breast cancer, which has the potential to deliver personalized treatment and improve patient care [6]. In general, the integration of multi-omics features helps researchers obtain a comprehensive picture about life development, and gain a deeper understanding of the pathogenesis, development process, and molecular mechanisms of diseases.

Deep neural networks (DNNs) have demonstrated power in mining complex, heterogeneous biological data [7]. For example, the feedforward fully-connected neural network (FFNN) and randomly-wired residual fully-connected neural network (RRFCN) have proven to be efficient in omics data analysis [8–10]. These algorithms can detect complex patterns from omics data through deep architectures and are therefore well suited for diverse biological domains [9,10].

All types of interactions within and across different omics levels form a huge network and function together in a variety of processes. For example, gene networks regulate the synthesis of protease, which further catalyzes metabolic reactions like lipid degradation [11]. Recently, graph-based deep learning (GDL) has provided new insight into analyzing biological network data. The main idea of GDL is message passing, which updates node representations via the aggregation of information from local node neighborhoods. GDL has shown strong performance in analyzing biological networks. However, limited research has been conducted on the integration of biological networks with multi-omics data.

In this paper, we propose AutoGGN, a multimodal method of integrating molecular interaction networks and omics data. GCNs are at the core of AutoGGN. They are designed to explore the hidden patterns in omics data through message passing and aggregating across molecular interaction networks. Specifically, we demonstrate the effectiveness of AutoGGN through three tasks: single-cell embryonic developmental stage classification, cancer type classification, and breast cancer sub-

[*] Corresponding author.
E-mail address: qiaonan3@huawei.com (N. Qiao).
[#] These authors contributed equally

https://doi.org/10.1016/j.ailsci.2021.100019
Received 2 October 2021; Received in revised form 19 November 2021; Accepted 19 November 2021
Available online 22 November 2021

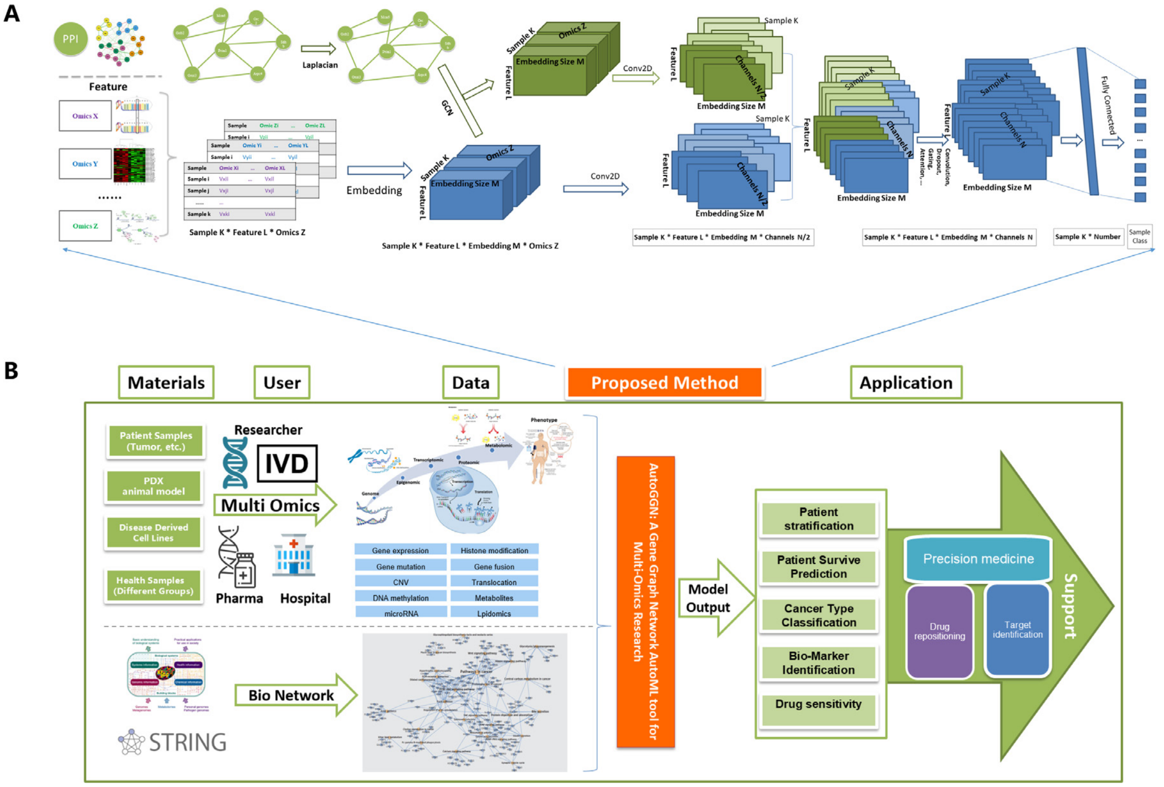

Fig. 1. Model structure and downstream applications of AutoGGN. (A) The detailed architecture of AutoGGN. The input data includes omics data and molecular interactions networks. The GCN module uses molecular interaction network together with omics features (used as node features) as input and outputs a fusion matrix of integrated omics features and network information. The CNN module convolves the embedded omics features directly. The results computed from the two modules are then concatenated. (B) The potential downstream applications of AutoGGN. After researchers obtain multi-omics profiles including genome, transcriptomes, and proteomics data from their experimental samples, they can use AutoGGN to efficiently integrate omics data with a molecular interaction network. The model can be applied to cancer patient stratification, patient survival prediction, bio-marker identification, drug sensitivity, and more tasks.

type classification. AutoGGN provides new insight into the integration of omics data with molecular interaction networks and can be applied to a series of downstream tasks with better performance.

Results

Integrating molecular interaction networks with omics data based on GCNs

DNNs have shown their ability in analyzing omics data, with good performance in some types of tasks [9,10], but they are limited to handling biological network data. AutoGGN takes full advantage of GCNs to efficiently integrate molecular interaction networks with omics data (Fig. 1).

The input of AutoGGN involves multi-omics data plus molecular interactions within and between different omics data types. Two separate modules are set up to effectively extract information from data (Fig. 1A). One is the GCN module, which takes molecular interaction networks together with omics features (used as node features) as input to explore the mutual relationships of omics data. This step results in a fusion matrix of integrated omics features and network information. Subsequently, a 2D-convolutional layer with a 1 × 1 filter is used to further aggregate omics features from the same layer. The other module is the convolutional neural network (CNN), which convolves the embedded omics features with a 1 × 1 filter with the purpose of extracting direct patterns from multi-omics data. The results obtained by the two modules are then concatenated. Depending on the purposes of the downstream tasks, different layers (such as Gating, Attention, Dropout, and Linear) can be stacked to the concatenated layer.

AutoGGN offers an innovative approach to integrating the information of molecular interaction networks with multi-omics data. It can also be applied to various biomedical problems (Fig. 1B), including patient stratification, survival prediction, biomarker identification, and drug sensitivity prediction.

Evaluation of AutoGGN on three classification tasks

We evaluated AutoGGN on three classification tasks (Supplementary Table 1 and Supplementary Table 2) and compared its performance to four baseline methods: XGBoost [21], Multi-Layer Perceptron (MLP), AutoGenome [9], and AutoOmics [10].

- XGBoost is a scalable ensemble method based on gradient boosting [21]. It is widely used on many machine learning problems.
- MLP is the most basic DNN type. It consists of several neurons followed by a nonlinear activation function [12].
- AutoGenome proposes residual fully-connected neural network (RFCN) architectures, which introduce fully-connected units into residual networks to deal with genomic profiling data [9].
- AutoOmics is an explainable RFCN-based framework for multi-omics data integration [9,10].

Case 1 – single-cell embryonic developmental stage classification

Gastrulation is an essential developmental process in most animals, after which the embryo will begin differentiation to generate the adult organism [13]. Automated developmental stage classification can improve insight into the complex events and mechanisms at different post-fertilization time points, as well as driving advances *in vitro* fertilization.

The single-cell dataset utilized to evaluate our model is comprised of single-cell RNA-seq (scRNA-seq) profiles of 10,000 samples from mouse embryos covering 10 developmental stages (E6.5, E6.75, E7.0, E7.25, E7.5, E7.75, E8.0, E8.25, E8.5, and mixed gastrulation). A PPI network

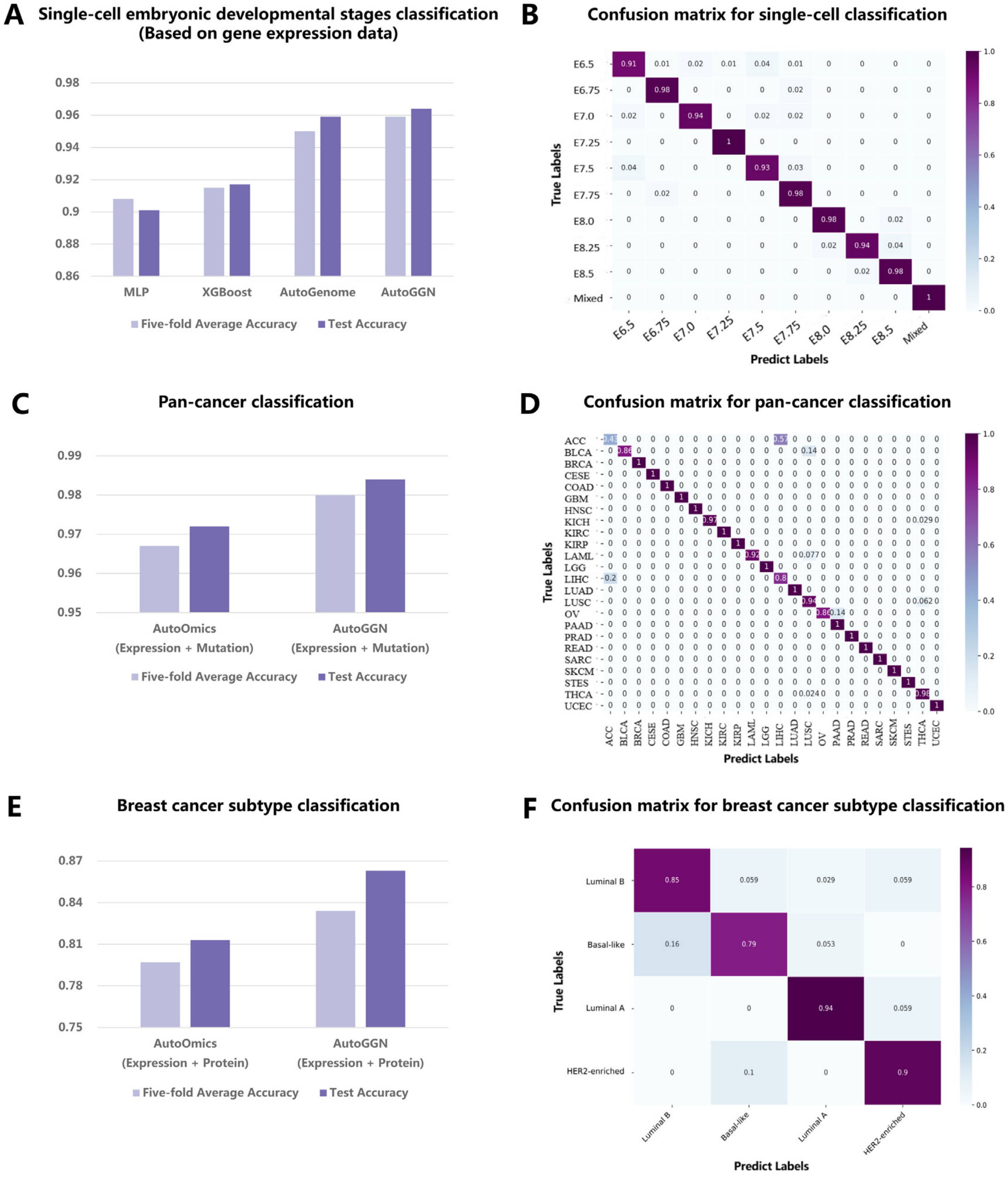

Fig. 2. Experimental results for the three classification task. (A) Accuracy comparison of AutoGGN with MLP, XGBoost, and AutoGenome for single-cell embryonic developmental stage classification. Both the average accuracy of five-fold cross validation and test accuracy are plotted. (B) Confusion matrix for the single-cell embryonic developmental stage classification task. The y-axis indicates the true label, and the x-axis indicates the predicted label. The value in each box is normalized by sample number, indicating the fraction of the samples with an accurately predicted class. Dark purple indicates higher accuracy and light purple indicates lower accuracy. (C) Accuracy comparison of AutoOmics and AutoGGN using gene expression and gene mutation data for pan-cancer classification. Expression: gene expression data; Mutation: gene mutation data. (D) Confusion matrix for pan-cancer classification. (E) Accuracy comparison of AutoOmics and AutoGGN using gene expression and protein expression data for breast cancer subtype classification. Gene: gene expression data; Protein: protein expression data. (F) Confusion matrix for breast cancer subtype classification.

was extracted from STRING [17] based on the 5129 gene features of the scRNA-seq data. The network contains 5129 nodes and 20,691 links in total. For model training and evaluation, we randomly put aside 10% of the data as an independent test set and performed grid search and 5-fold cross-validation on the remaining data to search for the optimal model. The 5-fold cross-validation is repeated five times. Then we evaluated the selected model on the test set. The same training and evaluation processes were applied to other models for comparison.

The classification performances of AutoGGN and other methods (MLP, XGBoost, and AutoGenome) are shown in Fig. 2A (Supplementary Table 3). A confusion matrix demonstrates the per-stage classification performances of AutoGGN on the test set (Fig. 2B). AutoGGN achieved the best performance, with an average cross-validation accuracy of 95.9 ± 0.3% on the evaluation set for all the splits and 96.4% on the test set, outperforming XGBoost [14] and MLP [12] by 4.7% and 6.3%, respectively, on the independent test set. It is particularly remarkable that our method achieved higher accuracy than AutoGenome [9] (95.9%), which is an RFCN-based method designed for genomic profile modeling. This shows that integrating molecular interaction networks with gene expression data enables us to explore more biological information, and can significantly benefit downstream tasks.

Case 2 – cancer type classification

Precise cancer type prediction is essential for cancer diagnosis and therapy [15]. In recent years, machine learning methods have shown great success in cancer prognosis and prediction. Open-source cancer databases such as The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) further provide researchers and clinicians with large-scale multi-omics data for analysis [16]. For example, Mostavi M. etal applied a CNN-based model on pan-cancer RNA-Seq data and achieved accurate cancer type prediction [17]. Ramirez R. et al. built GCNs to predict cancer types or normal tissues using gene expression profiles [18]. Li Y. et al. utilized machine learning algorithms to analyze gene expression data [19]. However, these researches only used single omics data as input; therefore, the extracted information may be limited.

To evaluate the capacity of AutoGGN in dealing with multi-omics data, we collected patient samples with both gene expression and gene mutation data available from TCGA [20] database. The data belongs to 5780 patients covering 24 cancer types (Supplementary Table 1). We obtained gene features that overlapped between the gene expression and gene mutation data, and used them to extract the protein-protein interaction (PPI) network. The network input consists of 5769 nodes and 130,056 edges. Subsequently, we trained and evaluated the model by following the same steps described in Case 1.

The classification performance of AutoGGN and AutoOmics is shown in Fig. 2C (Supplementary Table 4). Overall, AutoGGN achieved an average cross-validation accuracy of 98.0 ± 0.2% on the evaluation set for all the splits and 98.4% on the test set by integrating gene expression data and mutation data with the PPI network. It outperformed AutoOmics (which also used two types of omics data as input) by 1.2% on the test set. This shows that AutoGGN has great potential in integrating multi-omics data with biological interaction networks.

Case 3 – breast cancer subtype classification

Cancer is a very complex disease group where genetic heterogeneity exists not only between but also within cancer types [21]. This may lead to large differences in how patients respond to drugs or immune therapy even when dealing with the same type of cancer [9]. Therefore, accurate cancer subtyping is necessary to provide personalized treatment and better patient care.

We evaluated AutoGGN on breast cancer subtyping using two type of omics data: gene expression profiles and protein expression profiles. We collected patient samples with breast cancer from TCGA and focused on four subtypes: luminal A, luminal B, triple-negative, and HER2-enriched according to PAM50-profiling-test. The dataset consists of 396 patients and the network input contains 14,912 nodes and 3948,022 edges.

Fig. 3E (Supplementary Table 5) shows the classification performances of AutoGGN and AutoOmics in breast cancer subtyping. By integrating multi-omics data with molecular interaction networks, AutoGGN achieved an accuracy of 86.3% on the test set, outperforming AutoOmics by 5.0%. This highlights once more AutoGGN's ability to effectively integrate different types of omics profiles with molecular interaction network data and thus improve the performance in patient stratification.

In summary, AutoGGN outperformed other models in the preceding tasks. We also assessed our model using only gene expression data and interaction network data in the latter two tasks and found that the accuracy was lower than when we used multi-omics data (Supplementary Figure 1 and Supplementary Table 6). Model performance varies depending on the degree to which models were able to effectively utilize and analyze biological data. This proves that AutoGGN can extract much more information from biological data through the integration of multi-omics profiles and molecular interaction networks.

Explaining predictions of AutoGGN using SHAP

Many researchers argue that neural networks are "black-boxes"[22–24], which leads to poor model interpretability despite their high performance. Explaining and understanding model predictions is just as important as efforts to improve model performance [22]. To help explain deep learning models, we introduced the SHapley Additive exPlanations (SHAP) [23] module into AutoGGN. Given a deep learning model, SHAP will calculate the marginal contribution of each feature to the overall predictions. This is referred to as the SHAP value, which were used to visualize the feature importance of each gene to the predicted classes.

For the single-cell embryonic developmental stage classification task, we extracted a list of important genes for the classification of developmental stages based on SHAP values (Fig. 3A, Fig. 3B). We conducted an extensive literature review and found that most genes in the top gene list (ranked by SHAP value) were key factors during embryonic development. For example, the protein encoded by Tdgf1 (the top ranked gene in the list) is an extracellular, membrane-bound signaling protein that plays an essential role in embryonic development and tumor growth [24]. The second ranked gene — Pou5f1 — encodes a transcription factor containing a POU homeodomain, which is vital in embryonic development and stem cell pluripotency [25]. Another example is the 4th ranked gene — Fgf5, which encodes a protein possessing broad mitogenic and cell survival activities, and is also involved in embryonic development [26].

From the SHAP value distributions, we can also study in what manner genes contribute to each developmental stage. For instance, Pou5f1, Fgf5, Fgf8, and Tdgf1, the top ranked genes in developmental stage E8.5 (Fig. 3B), are marked in blue for positive SHAP values. This means a low expression level of these genes can increase the probability of the E8.5 class. In contrast, a high expression level can increase the probability of an earlier stage E6.75 (Fig. 3B). Existing reports show that these genes can regulate early embryonic development [24–27]. They are likely to express and regulate activities in the early stages of embryonic development like E6.75 and then exhibit a low expression level in later stages.

Discussion

In this paper, we propose an innovative multimodal approach — AutoGGN — to integrate multi-omics profiles with molecular interaction networks. In AutoGGN, two separate modules, GCN and CNN, are designed to extract information from molecular interaction networks and find hidden patterns from omics data. Then, the feature maps generated from the two modules are integrated to be used in downstream applications. This kind of architecture ensures that the information from both molecular interaction networks and omics data can be obtained and combined for further use.

The three classification tasks proved the robustness of AutoGGN. In the moude single cell embryonic developmental stage classification task, we used gene expression profiles and a PPI network as the model input and achieved an accuracy of 96.4%, outperforming other methods that merely used single-omics data as input. We also proved AutoGGN's great potential in integrating multi-omics data with molecular interaction networks for pan-cancer classification and breast cancer subtyping. More importantly, we found that better performance can be achieved with multi-omics data as input than with single omics data in both tasks. The tasks demonstrated that AutoGGN can integrate information from different levels of omics profiles effectively and thoroughly.

To deal with the black-box challenge of neural networks, we used SHAP to explain and visualize the feature importance of each gene to the predicted class. For example, in the single-cell developmental stage classification task, we found that most of the important genes for the classification task were closely related to embryonic development. The explainable results can help researchers and other types of users better understand deep learning models and also provide insight into the molecular mechanisms behind omics data.

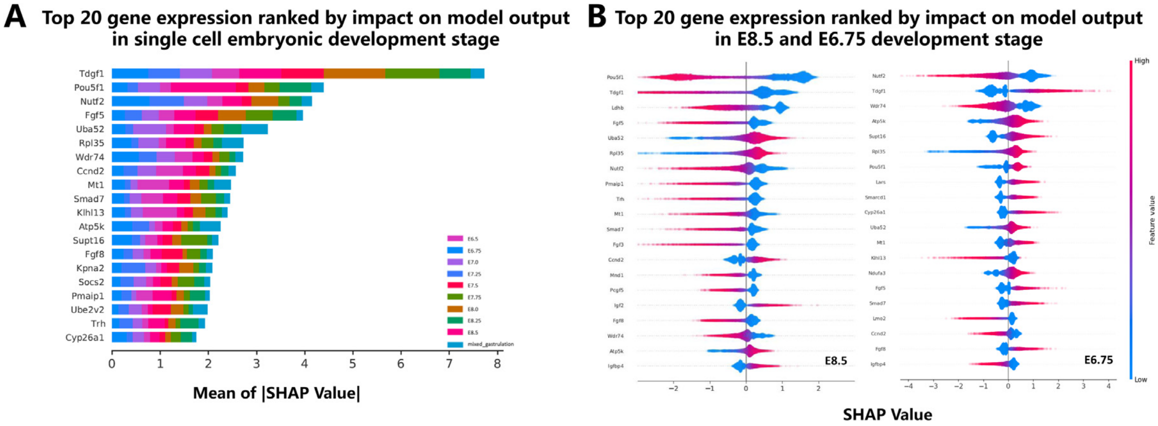

Fig. 3. Explainable results of single cell embryonic developmental stage classification based on the SHAP module. (A) The top 20 genes ranked by feature importance values for the 10 single-cell embryonic development stages. The contribution of genes to each stage are marked in different colors. The x-axis indicates the mean of the absolute SHAP value. (B) The contribution of the top 20 gene expression ranked by the SHAP value in the E8.5 and E6.75 stages. It indicates how gene expression influences single-cell development in different stages. Red shows that high gene expression will increase the probability of the predicted class, while blue shows that low gene expression will increase the probability of the predicted class.

To enable researchers with limited machine learning expertise to train high-quality models more efficiently, we have developed AutoGGN into an automated machine learning (AutoML) tool. We also built an explainer based on SHAP in our tool to ensure that researchers can obtain accurate results along with clear explanations using AutoGGN. Model training, evaluation, prediction, and explanation can all be done with only five lines of code.

Besides the preceding classification tasks, AutoGGN can also be applied to a wide range of biomedical tasks. For example, AutoGGN can be integrated with the Cox proportional hazards model [28] to predict the life expectancy of patients. Accurate estimates will help doctors develop more appropriate treatment plans and provide better patient care [29]. Researchers have found the expression of some biomarkers to vary considerably in tissue- and age-specific manners [30]. AutoGGN can tackle this challenge by utilizing omics data from different tissues of donors of various ages and the molecular interaction networks to find the most important genes that contribute to tissue and age predictions. Other applications may include bio-marker identification and drug sensitivity prediction. Further work will be continuously conducted to extend the application scope of AutoGGN.

Methods

Datasets

In this study, we performed three classification tasks to evaluate AutoGGN. Each task utilizes two types of data, single/multi-omics data and molecular interaction network data. All the datasets used in the experiments are publicly available and subsequently described in this paper. The sample distribution by the class of each dataset is shown in Supplementary Table 1.

Mouse single-cell transcriptomes dataset: The original dataset [https://github.com/MarioniLab/EmbryoTimecourse2018] consists of the single-cell RNAseq (scRNA-seq) profiles of 116,312 single cells from mouse embryos covering 10 developmental stages (E6.5, E6.75, E7.0, E7.25, E7.5, E7.75, E8.0, E8.25, E8.5, and mixed gastrulation), which were collected at nine sequential time points during post-fertilization. We randomly selected 1000 single-cell samples from 10 stages, and the final dataset contains the corresponding scRNA-seq profiles of 10,000 single cells in total.

Pan-cancer dataset: We downloaded all the pan-cancer patient samples with both gene expression and somatic mutation profiles available from TCGA. Patient samples with more than one cancer was removed. The final dataset consists of 5780 samples covering 24 cancer types, and their gene expression and somatic mutation profiles were used in the task.

Breast cancer subtype dataset: We obtained gene expression and protein expression profiles of patients with breast cancer from TCGA. PAM50-based subtypes for patients were downloaded from a published paper [31].

Protein-protein interaction network: The protein-protein interaction (PPI) network data of humans (Homo sapiens) and mice (Mus musculus) was obtained from the STRING [32] database, which collects both known and predicted PPIs. Interactions with a combination score lower than 0.98 were excluded for the first two tasks. In the breast cancer subtyping task, we excluded interactions with a combination score lower than 0.1 to retain as many node features as possible due to the limited size of the protein expression data.

Data preprocessing

We applied data preprocessing steps to omics data and PPI network data. The detailed preprocessing procedures for each task are as follows:

Case 1 - single-cell embryonic developmental stage classification. The original dataset was already pre-processed with quality control, counts normalization, highly variable gene selection, and batch correction. We selected a random subset of the original dataset, which contained the scRNA-seq profiles of 10,000 single cells covering 10 embryonic developmental stages (each stage consists of 1000 samples). We further applied 0–1 scaling on the dataset. Then, we extracted the corresponding mouse PPI network based on the gene features. In total, 18,379 genes and 167,188 protein-protein interactions were gathered for analysis.

Case 2 - Cancer type classification. We obtained 5780 patient samples with mRNA-seq data and somatic mutation profiles covering 24 cancer types after excluding those with more than one cancer type. For RNA-seq data, transcripts per million (TPM) was log2 transformed and used as gene expression values. For somatic mutation data, the features had two types of values: 0 and 1, representing unmutated and mutated genes, respectively. Features that were missing expression or somatic mutation data were removed. Afterwards, we extracted a subset of the human PPI network based on the remaining gene features. In total, 5769 gene features and 260,104 interactions were collected for analysis.

Case 3 – Breast cancer subtyping. The dataset contained 396 patients with overlapping feature data and subtype data. Gene expression values were log2-transformed, and protein expression data was originally corrected by median centering across antibodies. No further preprocessing steps were applied. Due to the limited size of protein features, we retained all the features from two types of omics data. The value of features missing gene expression or protein expression data was set to 0. A PPI network that consists of 14,912 nodes and 3948,022 edges was extracted based on the features from omics data.

Model training and evaluation

For all the tasks, we put aside 10% of the dataset for testing and applied 5-fold cross validation to the rest of the data. For each of these five folds, we constructed AutoGGN as well as the baseline models (MLP, XGBoost, AutoGenome, and AutoOmics) using the training set and predicted the labels of the evaluation set. This method of splitting the dataset helped us evaluate the robustness of the models. Five-fold cross-validation is repeated 5 times to avoid accidental error and their mean and standard deviation of the accuracy of the classification accuracy are reported. To find the best hyperparameters of different models, we performed grid search. The search space for different models are listed as follows:

AutoGGN search space. 1) The number of GCN layers, searching value [1, 2, 3, 4]. 2) The embedding dimension, searching value [4, 8, 16, 32, 64]. 3) The embedding channel, searching value [4, 8, 16, 32, 64];

AutoOmics and AutoGenome. 1) The number of ResNet blocks, searching value [1, 2, 3, 4, 5, 6]. 2) The number of neurons in each layer, searching value [8, 16, 32, 64, 128, 256, 512, 1024, 2048]. 3) The drop-out ratio of the first layer compared with the input layer, searching value [0.6, 0.8, 1.0].

MLP. 1) The number of hidden layers, searching value [1, 2, 3, 4]. 2) The number of neurons in each layer, searching value [128, 256, 512, 1024, 2048]. 3) The activation function, searching value [“relu”, “logistic”, “tanh”].

XGBoost. 1) The maximum depth of the tree, searching value [3, 4, 5, 6]. 2) The minimum sum of instance weights needed in a child, searching value [1, 3, 5].

Hyperparameter settings with the highest average accuracy on the five-fold split training sets are the best hyperparamters. Then we evaluated the best model obtained from the grid search on the independent test set.

Feature importance estimation

We used the SHAP package [23] to conduct the feature importance estimation and explain how the model works. SHAP supports GradientExplainer (an implementation of expected gradients), which can approximate SHAP values for AutoGGN. The best AutoGGN model identified and the training set were used as input for the SHAP module. After obtaining the SHAP value of each feature for each sample, we summed all the SHAP values within each class and obtained the feature importance score for each class. This is how we identified important features for the classification.

AutoGGN AutoML tool

AutoGGN is built on Tensorflow [33] and implemented as a Python package, the usage of AutoGGN includes six main modules. 1) Package loader module. AutoGGN can be imported as a python package. 2) Data loader module: The paths of the training, evaluation and test datasets can be specified in a JSON configuration file. Also the number of threads and capacity can be set when loading data. 3) Model trainer module: In this module, users can specify training parameters, like number of channels, batch size, number of GPUs, learning rate range, activation function and so on. The searching space for main parameters is specified at the same time. 4) Model evaluation module: The detailed evaluation reports, including classification reports and confusion matrix, are output in the specified path. And the best model is saved based on the best metrics. 5) Model prediction module: The best model is used to predict on the independent test dataset. Predicted class and softmax value of samples are output in this module. 6) Explanation module: To interpret the reason underlying the classification, the important features which contributes more to each category are reported based on SHAP [23].

Data availability

All the data sets utilized in our study are public data. The omics data used in the pan-cancer classification and breast cancer subtype classification tasks is from TCGA. The data used for single-cell classification data is from accessions: Atlas: E-MTAB-6967 and the processed data was downloaded following the instructions at https://github.com/MarioniLab/EmbryoTimecourse2018.

Software availability

The software could be accessed from the website http://autoggn.com.cn, which contains the software introduction, installation method and the tutorial. For the experiments, the protocols also will be provided as notebook examples on the website.

Author contributions

N.Q. designed and conceived the project. L.Z., W.S., and P.L. developed the algorithm of AutoGGN and performed the necessary experiments under the guidance of N.Q. C.X., D.L, W.H, Z.X., D.W., M.Z. and H.J. discussed and contributed ideas. L.Z., W.S., and P.L. wrote the paper. N.Q. revised the manuscript. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2021.100019.

References

[1] Fuentes-Pardo AP, Ruzzante DE. Whole-genome sequencing approaches for conservation biology: advantages, limitations and practical recommendations. Mol Ecol 2017;26:5369–406.

[2] Chen R, Snyder M. Promise of personalized omics to precision medicine. WIREs Syst Biol Med 2013;5:73–82.

[3] John A, Qin B, Kalari KR, Wang L, Yu J. Patient-specific multi-omics models and the application in personalized combination therapy. Future Oncol 2020;16:1737–50.

[4] Subramanian KAI, Verma S, Kumar S, Jere A. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights 2020;14.

[5] P T. Big data in pharmaceutical R&D: creating a sustainable R&D engine. Pharm Med 2015;29:87–92.

[6] Prat A, et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. The Breast 2015;24:S26–35.

[7] Xu SA, Chunming Jackson. Machine learning and complex biological data. Genome Biol 2019;20.

[8] Mahmud M, Kaiser MS, Hussain A, Vassanelli S. Applications of Deep Learning and Reinforcement Learning to Biological Data. IEEE Trans Neural Netw Learn Syst 2018;29:2063–79.

[9] Liu D, et al. AutoGenome: An AutoML Tool for Genomic Research. bioRxiv 2019. doi:10.1101/842526.

[10] Xu, C. et al. AutoOmics: An AutoML Tool for Multi-Omics Research. bioRxiv (2020) doi:10.1101/2020.04.02.021345.

[11] Schulze H, Kolter T, Sandhoff K. Principles of lysosomal membrane degradation: Cellular topology and biochemistry of lysosomal lipid degradation. Biochim Biophys Acta BBA - Mol Cell Res 2009;1793:674–83.

[12] Jin, H., Song, Q. & Hu, X. Auto-Keras: An Efficient Neural Architecture Search System. (2019).

[13] Peng G, Suo S, Cui G, Yu F, Jing N. Molecular architecture of lineage allocation and tissue organization in early mouse embryo. Nature 2019;572:1–5.

[14] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 785–94. doi:10.1145/2939672.2939785.

[15] Lee K, Jeong HO, Lee S, Jeong WKCPEM. Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network. Sci Rep 2019;9:16927.

[16] Zhu W, Xie L, Han J, Guo X. The Application of Deep Learning in Cancer Prognosis Prediction. Cancers 2020;12:603.

[17] Mostavi, M., Chiu, Y.C., Huang, Y. & Chen, Y. Convolutional neural network models for cancer type prediction based on gene expression. (2019).

[18] Ramirez R, Chiu YC, Hererra A, Mostavi M, Jin YF. Classification of Cancer Types Using Graph Convolutional Neural Networks. Front Phys 2020;8:203.

[19] Li Y, et al. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. BMC Genomics 2017.

[20] Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol 2015;19:A68–77.

[21] Zhao L, Lee VHF, Ng MK, Yan H, Bijlsma MF. Molecular subtyping of cancer: current status and moving toward clinical applications. Brief Bioinform 2021;2.

[22] Coley CW, et al. A graph-convolutional neural network model for the prediction of chemical reactivity. Chem Sci 2019;10:370–7.

[23] Lundberg SM, Lee S-I, et al.Guyon I, et al., editors A Unified Approach to Interpreting Model Predictions. Advances in neural information processing systems 30 2017:4765–74.

[24] Strizzi L, et al. Postovit LM, Margaryan NV. Emerging roles of nodal and Cripto-1: from embryogenesis to breast cancer progression. Breast Dis 2008;29:91–103.

[25] Tantin D. Oct transcription factors in development and stem cells: insights and mechanisms. Development 2013;140:2857–66.

[26] Allerstorfer S, et al. FGF5 as an oncogenic factor in human glioblastoma multiforme: autocrine and paracrine activities. Oncogene 2008;27:4180–90.

[27] Abu-Issa R, Smyth G, Smoak I, Yamamura K, Meyers EN. Fgf8 is required for pharyngeal arch and cardiovascular development in the mouse. Development 2002;129:4613–25.

[28] Cai J, Zeng D. Cox Proportional Hazard Model. Wiley statsref: statistics reference online. American Cancer Society; 2014. doi:101002/9781118445112stat06880.

[29] Bashiri A, Ghazisaeedi M, Safdari R, Shahmoradi L, Ehtesham H. Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. Iran J Public Health 2017;46.

[30] Hudgins AD, et al. Age- and Tissue-Specific Expression of Senescence Biomarkers in Mice. Front Genet 2018;9:59.

[31] McCullough & AComprehensive molecular portraits of human breast tumours. Yearb Pathol Lab Med 2013:286–8 2013.

[32] Szklarczyk D, von MC, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47.

[33] Abadi, M. et al. TensorFlow: A system for large-scale machine learning. arXiv:160508695 Cs (2016).