Genome Medicine

RESEARCH

Open Access

DeepGAMI: deep biologically guided auxiliary learning for multimodal integration and imputation to improve genotype–phenotype prediction

Pramod Bharadwaj Chandrashekar[1,2], Sayali Alatkar[1,3], Jiebiao Wang[4], Gabriel E. Hoffman[5], Chenfeng He[1,2], Ting Jin[1,2], Saniya Khullar[1,2], Jaroslav Bendl[5], John F. Fullard[5], Panos Roussos[5,6,7] and Daifeng Wang[1,2,3*]

Abstract

Background Genotypes are strongly associated with disease phenotypes, particularly in brain disorders. However, the molecular and cellular mechanisms behind this association remain elusive. With emerging multimodal data for these mechanisms, machine learning methods can be applied for phenotype prediction at different scales, but due to the black-box nature of machine learning, integrating these modalities and interpreting biological mechanisms can be challenging. Additionally, the partial availability of these multimodal data presents a challenge in developing these predictive models.

Method To address these challenges, we developed DeepGAMI, an interpretable neural network model to improve genotype–phenotype prediction from multimodal data. DeepGAMI leverages functional genomic information, such as eQTLs and gene regulation, to guide neural network connections. Additionally, it includes an auxiliary learning layer for cross-modal imputation allowing the imputation of latent features of missing modalities and thus predicting phenotypes from a single modality. Finally, DeepGAMI uses integrated gradient to prioritize multimodal features for various phenotypes.

Results We applied DeepGAMI to several multimodal datasets including genotype and bulk and cell-type gene expression data in brain diseases, and gene expression and electrophysiology data of mouse neuronal cells. Using cross-validation and independent validation, DeepGAMI outperformed existing methods for classifying disease types, and cellular and clinical phenotypes, even using single modalities (e.g., AUC score of 0.79 for Schizophrenia and 0.73 for cognitive impairment in Alzheimer's disease).

Conclusion We demonstrated that DeepGAMI improves phenotype prediction and prioritizes phenotypic features and networks in multiple multimodal datasets in complex brains and brain diseases. Also, it prioritized disease-associated variants, genes, and regulatory networks linked to different phenotypes, providing novel insights into the interpretation of gene regulatory mechanisms. DeepGAMI is open-source and available for general use.

Keywords Deep learning, Cross-modality imputation, Auxiliary learning, Genotype–phenotype prediction, Cell-type gene regulatory networks, Schizophrenia, Alzheimer's disease, Prioritizing disease risk variants

*Correspondence:
Daifeng Wang
daifeng.wang@wisc.edu
Full list of author information is available at the end of the article

Background

The genotype–phenotype association has been found in many biological systems, such as brain-related diseases and behavioral traits. This association is very important as it will help us understand underlying cellular and molecular mechanisms like genes and pathways that causally affect the phenotypes [1, 2]. Many genome-wide association studies (GWAS) determine the association of genetic variants with many heritable diseases [3, 4], including neurodegenerative and psychiatric diseases like Alzheimer's disease (AD) [5–8] and schizophrenia (SCZ) [9, 10]. Despite the ground-breaking findings from these GWAS studies, they have some limitations. Firstly, association studies do not imply causation and require further downstream analysis and validations. Secondly, GWAS studies are independent studies that try to find the relationship between variants and disease individually and ignore the combined effect. Finally, the SNPs having small effect sizes go undetected as they do not meet the threshold criteria of the existing studies [11]. There have been several computational attempts outside the GWAS studies to discover genotype–phenotype association. Most of these attempts involve regression [12, 13]. Polygenic Risk Scores (PRS) [14] is the widely used method that looks at the linear combined effect of several variants on the phenotype. Modern machine learning techniques have been applied to predict the functionality of these phenotypes. For example, Zhang et al. [15] grouped several SNPs based on LD structure and used it as an input to the CNN model to predict AD progression. However, predicting phenotypes from genotypes remains challenging, primarily due to complex underlying molecular and cellular mechanisms. These methods perform genotype-to-phenotype prediction without considering various underlying intermediate phenotypes and mechanisms like gene expression and epigenome regulation.

For a mechanistic understanding from genotype to phenotype, several studies have shown that these variants influence disease risks by altering cell-type regulatory elements that affect the underlying gene expressions, which in-turn affect the disease phenotype [16, 17]. This resulted in studying the effects of genotypes on gene expression. Expression quantitative trait loci (eQTL) studies focus on associating genetic variants to gene expression instead of disease phenotypes [18–23]. They have proved to be a critical step in investigating gene regulation and have identified numerous eQTLs modulating the expression of disease genes. Transcriptome-wide association studies (TWAS) aim at identifying gene-trait interaction by combining GWAS and gene expression. The effect of genetic variation on gene expression is first studied, and then these expression profiles are statistically associated with the traits [24–31]. PrediXcan [32] is another approach imputing gene expression from eQTLs and mapping trait-associated loci based on the imputed gene expression data. A possible drawback in such association studies is that co-expressed gene patterns often lead to prioritizing non-causal genes [33].

It is also necessary to analyze genes and other regulatory elements that impact disease phenotypes. Several methods have attempted to associate genes with disease risks using gene expression profiles directly. For example, one study collected gene expression profiles from three publicly available AD datasets to predict the onset of AD disease using a variational autoencoder [34]. DeepWAS [35] predicted epigenomic functions of the genetic variants using DeepSEA [36] and then applied regression to predict the phenotype. A different approach is using gene regulatory networks (GRNs). GRNs represent a group of genes and various regulatory elements working together to control the expression of other genes and are cell-type specific. They facilitate understanding of various cell operations which allows better understanding of disease initiation and progression [37]. They have proven helpful in mapping molecular interactions [38, 39] and biomarkers for brain diseases [40–42]. However, a major drawback of these methods is that they consider each omics individually and thus miss the relationship between multiple omics.

As biological processes involve a complex interaction with multi-omics, emerging multimodal data enables studying such mechanisms at different scales. Several studies have attempted to integrate multi-omics data for brain disease predictions like SCZ [43–45] and AD [46–48]. Some studies have used known biological findings (e.g., GRNs, eQTLs) to guide feature selection or integrate several omics for disease prediction. For example, Wang et al. [49] used a deep Boltzmann machine (DBM) with GRNs guiding the internal connections to improve predicting neuropsychiatric diseases. Varmole [50] integrated gene expression and genotype (SNPs) using a deep neural network architecture where GRN and eQTLs guide the relationship between the input and the first hidden layer. While most studies consider disease outcomes (case versus control) as the phenotypes, more complex phenotypes, e.g., neuropathological, and cognitive phenotypes for AD, remain understudied. Studying the genetic effects of those disease phenotypes has great potential to deeper understand underlying cellular, molecular, and pathological mechanisms. Also, GRNs and eQTLs provide functional genomic relationship information. Most existing machine learning methods use raw -omics data as inputs but cannot handle these relationship data. Recent advancements in graph learning have led to the use of graph neural networks (GNNs) for integrating multi-omics data and use relationships

like MOGONET [51] and moGCN [52]. The relationship data is converted into graphs and given to GNNs as inputs. GNNs apply different neural network methods on the graphs to extract useful latent features which are then used for various supervised and unsupervised tasks. However, these methods mainly focus on similarity networks (like patient similarity networks) and are not suitable for functional genomic networks as these networks are directed, sometimes bipartite (e.g., eQTL networks) and relationship topology is complicated.

Another challenge for genotype–phenotype studies is multimodal data integration. Due to several factors (e.g., experimental costs, sample availability), multi-omics data will often be partially available for individuals [53], e.g., available genotypes versus limited gene expression, chromatin, or imaging data [54, 55]. Including partially available multimodal datasets in a predictive model will be challenging due to the lack of matched samples and missing modalities. This calls for cross-modal imputation. Cross-modal imputation involves estimating data modalities from available multiple modalities. This will help us infer missing modalities and aid in phenotype predictions. Several computational methods have been developed for cross-modal inference. MOFA [56] uses Bayesian approach for cross-modality estimation by projecting the modalities onto a low-dimensional space. BABEL [57] trains a joint autoencoder on paired multimodal data to impute one modality from another by minimizing a reconstruction and cross-modality loss. scVAEIT [58] proposed a method for mosaic integration that uses a masking procedure to learn a joint representation of cells sequenced across technologies and the distribution of missing modalities. scJoint [59] integrates scRNA-seq and scATAC-seq data using cell-type label information from scRNA-seq through transfer learning and embedding the annotated cells in a lower-dimensional space. There are no existing methods that perform cross-modal imputation alongside predicting disease phenotypes to the best of our knowledge.

Auxiliary learning is a type of learning technique aimed at improving the generalization of the primary task by learning secondary tasks along with the primary task [60–63]. The secondary task also called the auxiliary task is a subtask to be trained along with the primary task where the features are shared between the tasks resulting in additional relevant feature extraction useful for the primary task, and thus is usually defined in terms of estimating entities relevant to solving the primary task [64]. Implementing auxiliary learning involves adding supplementary cost functions to the primary cost of the neural network model [65]. Auxiliary learning has been very successful in reinforcement learning [60, 66, 67], computer vision [62, 68, 69], and autonomous driving assistance [70, 71]. Recently, it has been used in the biomedical domain with applications in screening skin cancer from microscopy images [72], and covid-19 detection from CT images [73]. Although auxiliary learning has not been applied for imputing multimodal data for genotype–phenotype prediction, the closest approach is SCENA [74] which estimates the gene–gene correlation matrix using ensemble learning and auxiliary information for single-cell RNA-seq (scRNA-seq) data where the auxiliary information is used in the form of gene networks and other relevant RNA-seq data. Similarly, DeepDiff [75] predicts cell-type-specific differential gene expression from epigenetics by using cell-type gene expression predictions as auxiliary tasks.

In summary, genotype–phenotype predictions are very important in understanding molecular and cellular mechanisms, but existing genotype–phenotype methods have the following limitations: (I) Statistical methods such as Polygenic Risk Score (PRS) predict phenotypes directly from genotype. They are mostly linear models that cannot tackle genomic variants' nonlinear effects and involve association studies that predict the correlation between genotype and phenotype but cannot explain how the inherited mutations are associated with the phenotype [76, 77]. Moreover, these methods do not consider intermediate phenotypes like molecular activities that significantly contribute to phenotypes; (II) Emerging multi-omics data at the population level enables machine learning which studies such mechanisms at different scales from genotype to phenotype. However, due to the black-box nature of many machine learning techniques, it is challenging to integrate these multiple modalities and interpret the biological mechanisms after prediction, especially when some modality is missing; (III) Functional genomic relationships like GRNs and eQTLs guide us in understanding these molecular mechanisms. However, most existing machine learning methods, including GNNs, cannot handle this kind of relationship data as they do not have a spatial relationship like graphs, and significant effort is required to convert them into a graph-like structure. (IV) Several methods focus on cross-modality estimation for single-cell multi-omics data (e.g., MOFA [56], MultiVI [78], Polarbear [79]) but not in the realm of disease types and clinical phenotypes.

To address these limitations, we propose DeepGAMI: Deep biologically guided auxiliary learning for multimodal integration and imputation to improve phenotype prediction. DeepGAMI is a novel deep learning model that enables cross-modal imputation, predicts clinical phenotypes, and prioritizes phenotypic tissue- or cell-type functional genomics. Our contributions are threefold. Firstly, DeepGAMI provides a biologically guided neural network framework for genotype–phenotype

prediction using biology-guided dropconnect [80]. It integrates genotype and gene expression data guided by prior biological knowledge of QTLs and GRNs. Secondly, it introduces an auxiliary learning layer that performs cross-modal estimation by learning relationships between modalities. This enables the model to take a single modality and use the estimated values of the other modality for disease prediction. Thirdly, DeepGAMI deciphers the black-box nature of the neural networks to prioritize genes and SNPs contributing to disease phenotypes. We applied DeepGAMI to multiple emerging multimodal datasets, including population-level genotype and bulk and cell-type gene expression data for SCZ cohort [49], genotype and gene expression data for AD cohort [55], and single-cell multimodal data comprising transcriptomics and electrophysiology for neurons in the mouse visual cortex [81]. We found that DeepGAMI outperforms existing methods in predicting phenotypes among these datasets and prioritizes genes, SNPs, and electrophysiological features showing biological interpretability. DeepGAMI is open-source and available at https://github.com/daifengwanglab/DeepGAMI.

Methods and materials

DeepGAMI overview

DeepGAMI is a multi-view deep learning model that integrates multimodal data for predicting phenotypes (Fig. 1, Methods and materials). Importantly, it uses auxiliary learning to learn the latent space of one modality from another, thereby enabling us to predict phenotypes from the available modalities by cross-modality imputation. We denote the primary task as $f_\theta^{pri}(X_1, X_2)$ that takes available multimodal inputs X_1 and X_2 with parameters θ to predict phenotypes, e.g., X_1 and X_2 can be SNP genotypes and gene expression levels of individuals. The auxiliary task is denoted by $f_\theta^{aux}(X_1, X_2)$ to learn C_{X_2}(latent space of input X_2) from C_{X_1}(latent space of input X_1). Using the trained model with auxiliary task, DeepGAMI can predict the phenotypes of samples with a missing modality. Specifically, it first imputes other modal latent spaces using the auxiliary learning function and then feeds both imputed and input latent spaces into the primary task for phenotype prediction. DeepGAMI uses feed-forward neural networks as it incorporates biological knowledge in terms of directed functional relationship networks like GRNs and eQTLs which aid in deciphering the black box and help us prioritize genes, SNPs, and other biological features for phenotypes. A list of all hyperparameters with its search space is available in Additional file 2: Table S1 and the total number of trainable parameters of DeepGAMI for each dataset is shown in Additional file 2: Table S2.

We compared the classification performance of DeepGAMI with three traditional classifiers: (1) Random Forest classifier (RF), (2) Support Vector Machine SVM), (3) Fully connected neural network classifier (MLP), and four state-of-the-art methods: (4) Logistic Regression (LR) [12], (5) Lasso Regression (Lasso) [13], (6) Varmole [50], (7) Mogonet [51], on several multi-omics datasets. RF, NB, and MLP were trained on the concatenation of multi-omics data.

Model design

The DeepGAMI model consists of four main layers.

Input layer

The input layer consists of two data modalities, e.g., gene expression and SNP genotypes. Each row of the input matrix represents feature vector of a sample. For example, the gene expression matrix contains gene expression profiles of K samples and n TFs and is represented by $X^{GEX} \in R^{K*n}$. Similarly, the genotype matrix consists of dosage information of K samples and l SNPs, $X^{SNP} \in R^{K*l}$.

Biological DropConnect layer

DropConnect is a regularization mechanism that sets random activation units to zero in each layer. It differs from dropout, which sets the random output units to zero while the former sets the connection weights to zero [80]. For our purpose, instead of randomly setting activations to zero, we guide the activations using prior biological knowledge, as shown in Eqs. 1 and 2.

$$C_k^{SNP} = \sigma\left(X_k^{SNP}\left(w_1 \odot m^{eQTL}\right) + b_1\right) \tag{1}$$

and

$$C_k^{GEX} = \sigma\left(X_k^{GEX}\left(w_2 \odot m^{GRN}\right) + b_2\right), \tag{2}$$

where C_k^{SNP} *and* C_k^{GEX} represent intermediate layer with nodes for k^{th} sample, X_k^{GEX} *and* X_k^{SNP} are the k^{th} column of X^{GEX} and X^{SNP} and represent gene expression and dosage data for k^{th} sample, w_1 and w_2 are weight matrices with dimension R^{l*p} and R^{n*p} respectively, $b1$ and $b2$ are the bias vectors of length p, and $\odot$ is the Hadamard product (element-wise multiplication). We do not encode self-loops and inter-connections among the input features (TFs, SNPs, etc.). If there are multiple levels of connections (e.g., $S_1 \rightarrow S_2 \rightarrow T_1$), we treat them as two separate connections ($S_1 \rightarrow S_2$ and $S_2 \rightarrow T_1$). In such cases, the genes (S_2 from the above example) appear under both the input layer and the intermediate layer. The biological DropConnect is applied separately for genotype and gene

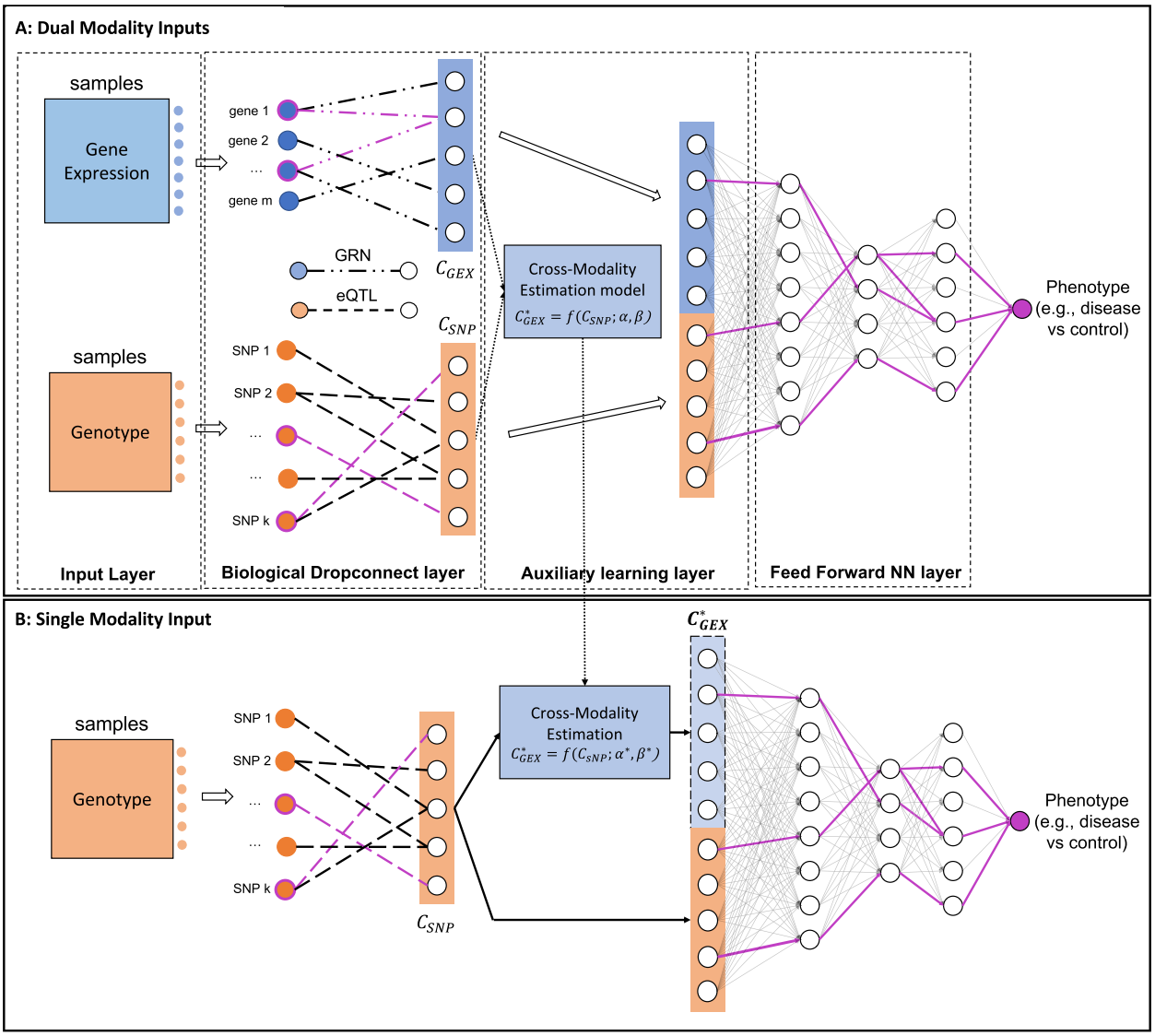

Fig. 1 Architecture of DeepGAMI. **A** DeepGAMI first uses available multimodal features for training the predictive model, e.g., SNP genotypes (orange) and gene expression (blue) of individuals from the major applications in this study. In particular, it learns the latent space of each modality (e.g., consisting of latent features at the first transparent hidden layer). This learning step is also regularized by prior knowledge enabling biological interpretability after prediction, i.e., the input and latent features are connected by biological networks (biological DropConnect). For instance, the input transcription factor genes can be connected to target genes as their latent features (e.g., C_{GEX}) by a gene regulatory network (GRN). The input SNPs can be connected to associated genes as their latent features (e.g., C_{SNP}) by Expression quantitative trait loci (eQTLs). Notably, an auxiliary learning layer is used to infer the latent space of one modality to another, i.e., cross-modality imputation. For instance, DeepGAMI learns a transfer function $f(.)$ to estimate C_{GEX} from C_{SNP}. Finally, the latent features are concatenated and fed to the feed-forward neural networks for phenotype predictions, e.g., classifying disease vs. control individuals. **B** Using the learned predictive model from multimodal input along, DeepGAMI can predict phenotypes from a single modality, e.g., SNP genotypes of new individuals. Specifically, it first imputes other modal latent spaces using the optimal transfer function $f^*(.)$ and then feeds both imputed and input latent features into downstream neural network predictions, e.g., C^*_{GEX} from C_{SNP}

expression. Masking filter (m) encodes biological drop connections (Eqs. 3 and 4).

$$m^{eQTL}_{\{i,j\}} = \begin{cases} 1 \; if \; SNP \; i \; is \; associated \; with \; gene \; j \\ 0 \; othewise \end{cases} \quad (3)$$

$$m^{GRN}_{\{i,j\}} = \begin{cases} 1 \; if \; TF \; i \; regulates \; gene \; j \\ 0 \; othewise \end{cases}, \quad (4)$$

where $m^{eQTL} \in R^{l^*p}$ and $m^{GRN} \in R^{n^*p}$. The underlying idea of this layer is to model the regulatory relationships

among genes such as TFs to genes and SNPs to genes. The input X^{GEX} represents all TFs as features and the intermediate layer nodes (C^{GEX}) which represent genes for all samples, are the output of the biological DropConnect layer. The connections between these TFs and genes are established using GRNs (m^{GRN}). Similarly, the connections between SNPs and genes are established using eQTLs (m^{eQTL}). The model is trained to learn the weights for these connections. This will help us prioritize important features (SNPs, genes, enhancers, etc.) and important interactions (SNP-gene and gene–gene) contributing to the phenotype. The output of this layer is referred to as the latent space of the input matrix.

Auxiliary learning layer

Each data modality from the input layer goes through the biological DropConnect layer producing a set of output nodes of equal dimension (C^{GEX}, C^{SNP}). This layer aims to learn the latent space of one modality from the other. We consider a linear relationship between the two latent spaces, computed using Eq. 5.

$$C^{GEX^{\star}} = f_{\theta}^{aux}\left(C^{SNP}\right) = \alpha C^{SNP} + \beta, \tag{5}$$

where α *and* β are scalar units representing weight and bias. We then concatenate the two latent space vectors and send them to a feed-forward neural network. One can get an average signal of the latent space vectors, but we decide not to pursue it as each latent node can be activated from either both the inputs or only one input.

Feed-forward classification layer

The concatenated gene layer is given to a fully connected feed-forward neural network with multiple hidden layers where each neuron in the hidden layers receives inputs from all previous layer outputs. ReLU activation is applied that forces negative weights to zero and handles non-linearity. ReLU activation is defined as shown in Eq. 6.

$$ReLU(X) = \max(0, X) \tag{6}$$

The final hidden layer is given to a softmax function to predict the output classes from the input provided by Eq. 7.

$$softmax(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{7}$$

where z represents the neuron values from the previous layer.

Training of DeepGAMI model

We split the input data into training (80%) and testing (20%) sets and performed fivefold cross-validation (CV) on the training set for feature selection and identifying the optimal parameter combination. We then pick the best performing model based on the five-fold CV and evaluate the final performance on the test set. Training DeepGAMI model involves minimizing the overall loss function which is a combination of primary task (phenotype prediction) loss and secondary task (cross-modality estimation) loss. The loss function used for the primary task is the cross-entropy loss (Eq. 8) and mean squared error (MSE) loss is used for the secondary task (Eq. 9).

$$\mathcal{L}^{pri}(y, \hat{y}) = \mathcal{L}\left(f_{\theta}^{pri}\left(X^{SNP}, X^{GEX}\right), y\right) = -\frac{1}{K}\sum_{k=1}^{K} y_k \log(\hat{y}_k) \tag{8}$$

$$\mathcal{L}^{aux}\left(C^{GEX}, C^{SNP}\right) = \frac{1}{K}\sum_{k=1}^{K}\left(C_k^{GEX} - f_{\theta}^{aux}\left(C_k^{SNP}\right)\right)^2 \tag{9}$$

The overall objective loss function for the model is computed using Eq. 10.

$$arg\min_{\theta}(\mathcal{L}^{pri} + \lambda\mathcal{L}^{aux}) \tag{10}$$

where i represents the i^{th} training sample, θ is a set of parameters. y *and* $\hat{y}$ represents the ground truth and the predicted labels respectively.

Evaluation metrics

As traditional accuracy measure can be misleading on skewed datasets, we used balanced accuracy (BACC, Eq. 11) and area under the receiver operating characteristic curve (AUC) to evaluate the performance of our model and other baseline comparison models.

$$BACC = \frac{sensitivity + specificity}{2} \tag{11}$$

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

where TP represents true positive, FP represents false positive, TN represents true negative, and FN is false negative. While the loss function includes both primary and auxiliary task losses, the performance of the model is based on BACC and AUC computed on the primary task.

Hyperparameter tuning for training DeepGAMI

DeepGAMI is trained using Adam Optimizer [82] with default parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The model runs for a maximum of 100 epochs with early stopping enabled to avoid overfitting. Tuning of several hyperparameters (like number of latent dimensions, number

of hidden layers in the feed forward network, dropout rate) is required as it has a huge impact on the prediction performance. All hyperparameters with the search space implemented in DeepGAMI model are present in Additional file 2: Table S1. The optimal hyperparameter combination was selected using a grid search based on the fivefold CV results on the training set. DeepGAMI is coded in python using Pytorch [83] library.

Feature prioritization

Integrated gradient (IG) [84] is a widely used technique for feature prioritization, IG attributes the model's prediction for its input features by computing gradients for each input and measures the change in the output response based on the small changes in the input using Eq. 12. IG is implemented using Captum [85] package in python.

$$IG_i(x) = \left(x_i - x_i'\right) \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha\left(x - x'\right)\right)}{\partial x_i} \partial\alpha \quad (12)$$

where i is the i^{th} feature, x and x' are input and baseline, and $\frac{\partial F(x)}{\partial x_i}$ is the gradient of $F(x)$ along the i^{th} feature.

IG provides feature importance scores where a higher score indicates higher importance of the feature. We first applied IG on the trained model (model with the best performance on the training set with the optimal hyperparameter setting) for generating feature importance scores for the input nodes (SNPs, TFs) and latent space nodes (genes) for the test samples. This will help us identify important SNPs and TFs attributed to the phenotype's outcomes. We then applied IG on the same model to extract link importance scores for the test samples. The link importance score gives us the importance score for the links between the input and the intermediate layer (Biological DropConnect layer). This will provide potential clues in understanding the underlying relationships (SNPs to genes and TFs to genes) for the given phenotype. Using this link importance score, we can fine-tune prioritized regulatory networks for phenotypes.

Enrichment analysis

From the prioritized functional links between SNPs and TFs to genes, we extract the genes having the most important links (top 10% of the link importance scores). We then perform enrichment analysis on these genes using Metascape [86]. The enrichments with binomial FDR p-value < 0.05 were reported in the prioritized genes.

Multimodal datasets and processing

Data preprocessing

The datasets vary between different cohort as they are extracted in different platforms using different protocols. We use the same preprocessing pipeline to process these inputs to the deep learning model and to reduce the effect of curse of dimensionality. We first use Student's t test (for binary phenotypes) and ANOVA (for multi-class phenotypes) for feature selection on the training set. As our intermediate layer consists of genes, we filter and keep the common genes between the two biological networks (e.g., GRNs, eQTLs). We only keep selected features (SNPs, TFs) present in these two networks. Next, we applied StandardScaler() function from scikit-learn [87] for the two modalities separately which scales the data such that they have zero mean and unit variance. StandardScaler() is computed using the following equation:

$$z = \frac{(x - \mu)}{\sigma} \quad (13)$$

where x is the input, μ is the mean, and σ is the standard deviation. We also provide the users an option to choose the standardization of their choice (currently DeepGAMI supports minmax, log, and standard normalization).

Schizophrenia

We used the population-level bulk RNA-seq and genotype data for the human dorsolateral prefrontal cortex (DLPFC) from PsychENCODE [49] for predicting SCZ versus healthy individuals. RNA-seq data consists of normalized gene expression of 14,906 genes for 1818 individuals. We extracted 146,763 eQTLs from GTEx consortium [88] for the human brain frontal cortex (BA9), and used GRNs from the PsychENCODE consortium. We first use Student's t for feature selection (keep significant SNPs and genes). We do not consider LD structure of the SNPs. We then include SNPs and genes which are present in eQTLs and GRN. Based on this pipeline (see Input Data Preprocessing), we ended up with 2080 SNPs, 126 TFs, and 84 intermediate layer genes as features.

We also tested DeepGAMI on the genotype and cell-type specific gene expression data from the CommonMind Consortium imputed using bMIND [89] and a reference panel of 4 cell populations: GABAergic (i.e., inhibitory) neurons, glutamatergic (i.e., excitatory) neurons, oligodendrocytes, and a remaining group composed mainly of microglia and astrocytes. The reference panel for each cell population was constructed by taking the mean log2 counts per million for each gene across 32 brain donors [90]. With the prior information from this reference panel, bMIND adopts a Bayesian approach to impute the cell-type-specific expression of each gene in each bulk sample from gene expression assayed from brain homogenate. We used cell-type-specific eQTLs [23] and applied scGRNom [91] to predict cell-type GRNs.

Alzheimer's disease

We used the bulk RNA-seq data in DLPFC and genotype data from the ROSMAP cohort [55] for our analysis in Alzheimer's disease. We used preprocessed bulk RNA-seq data (quantile normalized and batch effect removed) which contains the FPKM gene expression values.

For genotype data, we extracted SNP array (generated using Affymetrix GeneChip 6.0) dosage information for 1709 individuals. We extracted 146,763 eQTLs from GTEx consortium [88] for the human brain frontal cortex (BA9), and used GRNs from the PsychENCODE consortium [49]. Clinical phenotypes include cognitive diagnosis (COGDX) score ranging between 0 and 6, CERAD score (semi-quantitative measurement of the neuritic plaques useful for determining AD) ranged 0–4, and BRAAK score (semi-quantitative measurement for neurofibrillary tangle pathology) containing six stages. We coded the BRAAK phenotype into two classes (early-stage AD which contains BRAAK stages of 0–3 and late-stage AD containing BRAAK stages of 4–6), CERAD scores into three classes (No AD with scores 3–4, AD probable with score 2, and AD definite with score 0–1), and COGDX into three categories (No cognitive impairment (CI) with scores 0–1, mild CI with scores 2–3, and CI(AD/Dementia) with scores 4–6) using the coding available in ROSMAP. We used analysis of variance test (ANOVA) to filter out SNPs and genes with high variance except for BRAAK, where we used a *t*-test instead. We then intersected the SNPs and genes with eQTLs and GRN. We ended up with 229 early-stage AD individuals and 275 late-stage AD individuals for the BRAAK score phenotype. For the COGDX phenotype, we had no CI ($n=166$), mild CI ($n=130$), and CI (AD/Dementia, $n=208$) individuals. Finally, for the CERAD phenotype, we ended up with no AD ($n=184$), AD probable ($n=171$), and AD definite ($n=149$) individuals.

Mouse visual cortex

Patch-seq dataset includes transcriptomics and electrophysiological (ephys) data for 4435 neuronal cells in mouse visual cortex [81]. We used the cell cortical layers (cell location in the visual cortex) as the cellular phenotype: L1, L2/3, L4, L5, L6. We followed the data extraction and preprocessing in DeepManReg [92] and ended up with 41 ephys features and 1000 genes for 3654 cells. We also used 112 layer4 (L4) neuronal cells Patch-seq data for independent testing [93]. For this application, the inputs contain the gene expression data $X^{GEX} \in R^{K*n}$ of n genes and K samples, and the electrophysiological features $X^{ephys} \in R^{K*l}$ of K samples and l electrophysiological features. The model is trained to optimize the parameters based on the modified loss functions from Eqs. 8 and 9, and the updated loss function is shown in Eqs. 14 and 15.

$$\mathcal{L}^{pri}(y,\hat{y}) = \mathcal{L}\left(f_\theta^{pri}\left(X^{GEX}, X^{ephys}\right), y\right) = -\frac{1}{K}\sum_{k=1}^{K} y_k \log(\hat{y}_k) \tag{14}$$

$$\mathcal{L}^{aux}\left(C^{GEX}, C^{ephys}\right) = \frac{1}{K}\sum_{k=1}^{K}\left(C_k^{ephys} - f_\theta^{aux}\left(C_k^{GEX}\right)\right)^2 \tag{15}$$

The overall objective loss function for the model is computed using Eq. 16.

$$arg\min_{\theta}(\mathcal{L}^{pri} + \lambda * \mathcal{L}^{aux}) \tag{16}$$

Results

Classification of schizophrenia individuals from genotype and bulk gene expression data

We first evaluated the performance of DeepGAMI using population-level genotype and bulk-tissue gene expression data of schizophrenia individuals in the dorsolateral prefrontal cortex (DLPFC). We utilized PsychENCODE consortium [49] data for predicting schizophrenia (SCZ) versus healthy individuals. PsychENCODE contains 1866 individuals from several cohorts with different neuropsychiatric diseases. After filtering for schizophrenia and control samples with multimodal data, we ended up with 1168 samples from three cohorts: CommonMind (CMC, 565 samples), Lieber Institute for Brain Development (LIBD, 511 samples), and BrainGVEX (92 samples). We used the CMC cohort consisting of 343 control and 275 SCZ individuals for tuning and training the model. The CMC data was first split into train and held-out test sets with a ratio of 90:10. We then performed filtering, preprocessing, and feature selection (Methods and materials) on the training samples and ended up with 7433 SNPs, 208 transcription factors (TFs), and 2870 intermediate layer genes.

We performed fivefold cross-validation for selecting the optimal hyper parameters. Figure 2A shows the fivefold cross-validation balanced accuracies of DeepGAMI using both genotype and gene expression as inputs (Dual) and only genotype as input (Single) in comparison to other state-of-the-art classifiers along with DeepGAMI with no biological priors. DeepGAMI dual ($BACC = 0.867 \pm 0.016$) and DeepGAMI single ($BACC = 0.845 \pm 0.042$) achieved the highest performance in comparison with other methods. DeepGAMI with no biological prior ($BACC = 0.835 \pm 0.031$ for dual modalities and $BACC = 0.796 \pm 0.024$ for single modality) was the closest in performance.

A

Method \ Data Type	Single Modality (SNP Only)	Dual Modality (SNP + genes)
Random Forest	0.6 ± 0.026	0.6 ± 0.022
Naïve Bayes	0.6 ± 0.026	0.769 ± 0.05
Logistic Regression (LR)	0.764 ± 0.062	0.79 ± 0.018
Lasso Regression	0.816 ± 0.045	0.815 ± 0.033
MLP	0.795 ± 0.15	0.824 ± 0.032
Varmole	0.767 ± 0.034	0.788 ± 0.036
Mogonet	0.754 ± 0.089	0.819 ± 0.033
DeepGami	0.845 ± 0.042	0.867 ± 0.016
DeepGami (MLP)	0.796 ± 0.024	0.835 ± 0.031

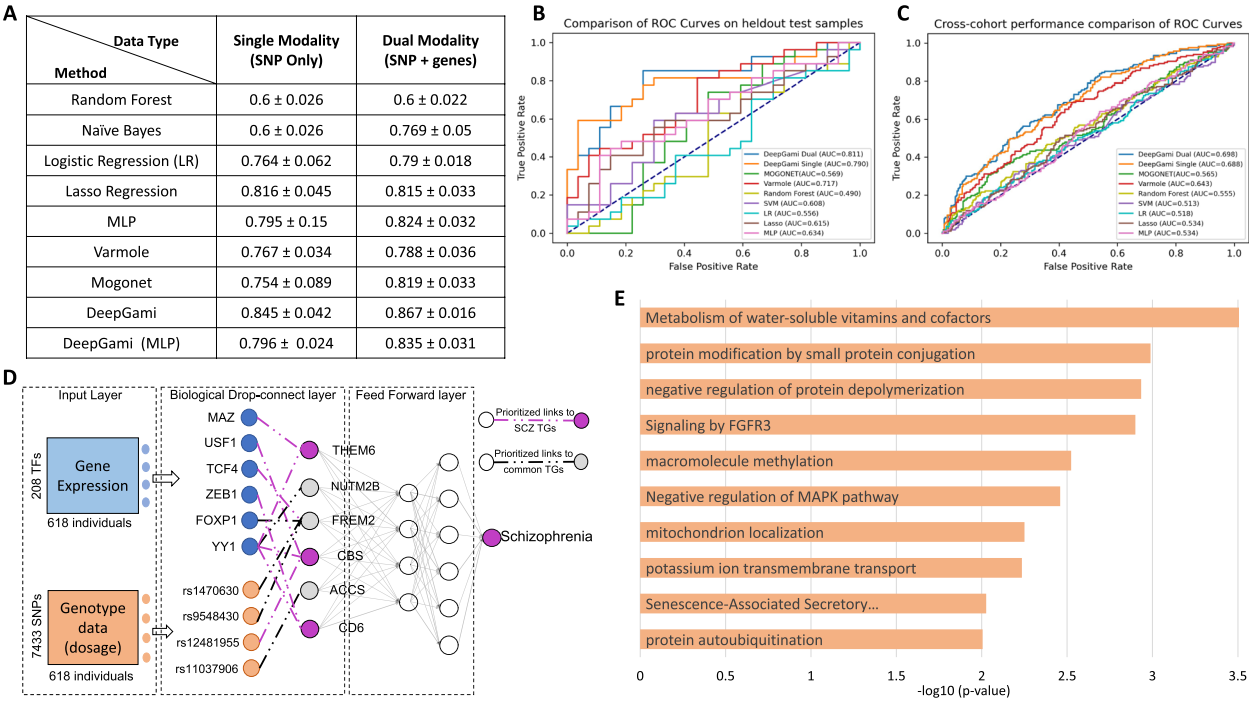

Fig. 2 Schizophrenia classification and functional genomic prioritization using genotype and bulk-tissue gene expression data. The population data was from the PsychENCODE project (Methods and materials). **A** Balanced accuracies from 5-fold cross-validation and **B** receiver operating characteristic (ROC) curves of DeepGAMI dual-modality model (dark blue), DeepGAMI single modality model (orange), Lasso (brown), LR (light blue), Random Forest (yellow), SVM (purple), Multilayer perceptron (MLP, pink), Varmole (red), and MOGONET (green) for classifying schizophrenia vs. control individuals on the held-out test samples. **C** ROC curves of various methods on cross-cohort SCZ prediction. **D** Select examples of prioritized transcription factors, SNPs, target genes (latent features, and functional links (GRNs, eQTLs) for schizophrenia. Purple: known schizophrenia genes. **E** Function and pathway enrichments of prioritized schizophrenia SNPs

After selecting the optimal hyperparameters through fivefold cross-validation, we built a model using these settings on the training samples and evaluated its performance on the held-out test samples within the same CMC cohort. DeepGAMI ($BACC_{DUAL} = 0.796$ and $BACC_{SINGLE} = 0.759$) outperformed other state-of-the-art methods ($BACC_{Varmole} = 0.7296$, $BACC_{MOGONET} = 0.593$, $BACC_{LASSO} = 0.537$, and $BACC_{LR} = 0.5556$, Additional file 1: Fig S1A). To further test the generalizability of DeepGAMI, we used combined SCZ samples from LIBD and BrainGVEX cohorts for independent validation. These cohorts contained 257 SCZ samples and 377 control samples. We trained our model on the CommonMind cohort and evaluated the performance on the LIBD and BrainGVEX samples. DeepGAMI was able to classify with $BACC_{DUAL} = 0.625$ and $BACC_{SINGLE} = 0.623$ outperforming other models: Varmole ($BACC = 0.602$), Mogonet ($BACC = 0.518$), MLP ($BACC = 0.545$), LR ($BACC = 0.519$), and Lasso ($BACC = 0.547$) as shown in Additional file 1: Fig S1A. ROC curves of various models on held-out test samples and independent SCZ samples are shown in Fig. 2B and C respectively, where DeepGAMI has the best performance. Additionally, we evaluated the performance of DeepGAMI on 47 bipolar disorder (BPD) samples from the CMC cohort where we trained DeepGAMI on SCZ samples and tested on BPD samples. As we had only BPD samples, sensitivity measure was used as the performance metric. We found that DeepGAMI ($SENSITIVTY_{DUAL} = 0.6596$ and $SENSITIVTY_{SINGLE} = 0.5957$) performed better than Varmole ($SENSITIVTY_{Varmole} = 0.4681$) and MOGONET ($SENSITIVTY_{MOGONET} = 0.4255$) as shown in Additional file 1: Fig S1B. This demonstrates the application of DeepGAMI for cross-disorder prediction.

We then used integrated gradient approach (Methods and materials) to prioritize SNPs, genes, and functional links on the held-out SCZ samples. Figure 2D shows a few examples of these DeepGAMI's prioritized SNPs, genes, and links. Our model was able to prioritize SCZ-related genes like CBS [94, 95], THEM6 [96], and CD6 [97, 98] among others shown as pink circles in Fig. 2D. CBS (cystathionine beta-synthase) gene plays a significant role in reducing the level of homocysteine which is etiologically linked to SCZ. However, mutations in CBS leads to glia/astrocyte dysfunction which is associated with SCZ pathogenesis [94]. THEM6 gene has shown association with SCZ individual. Similarly, CD6 gene is related

to immune system which in turn is associated with SCZ [97]. A complete list of prioritized SNPs and genes with the importance score is available in Additional files 3 and 4. We then used the prioritized functional links to extract the genes present in these links (213 genes, top-ranked 10% of the link importance scores). We performed enrichment analysis on these 213 genes with 3064 genes (total number of input features) as background using Metascape. We found several known functions and pathways related to SCZ, like signaling by FGFR3 [99, 100], negative regulation of MAPK pathway [101, 102], and senescence-associated secretory phenotype [103] (Fig. 2E).

Clinical phenotype prediction and gene regulatory network prioritization in Alzheimer's disease

To demonstrate the application of DeepGAMI for predicting complex clinical phenotypes, we ran DeepGAMI on Alzheimer's disease (AD) cohort, ROSMAP [55], which contains multi-omics data from the brain DLPFC region. We used its bulk-tissue gene expression and genotype data for this analysis (Methods and materials). As the cohort contains data from the DLPFC brain region, we extracted the eQTL information from the GTEx consortium [88] from the same brain region (human brain frontal cortex, BA9), which contains 146,763 eQTL SNPs, and used the GRN from the PsychENCODE consortium [49]. Clinical phenotypes included in our analysis are cognitive diagnosis (COGDX), CERAD, and BRAAK scores (Methods and materials). Additional file 2: Table S3 summarizes the number of features and class labels for this analysis.

We split the data into training and held-out test sets using an 80:20 ratio. We used a fivefold CV on the training set for tuning the hyperparameters and obtaining the best performance. DeepGAMI outperforms state-of-the-art classifiers ($BACC = 0.806$ for BRAAK, 0.689 for COGDX, and 0681 for CERAD, Fig. 3A, Additional

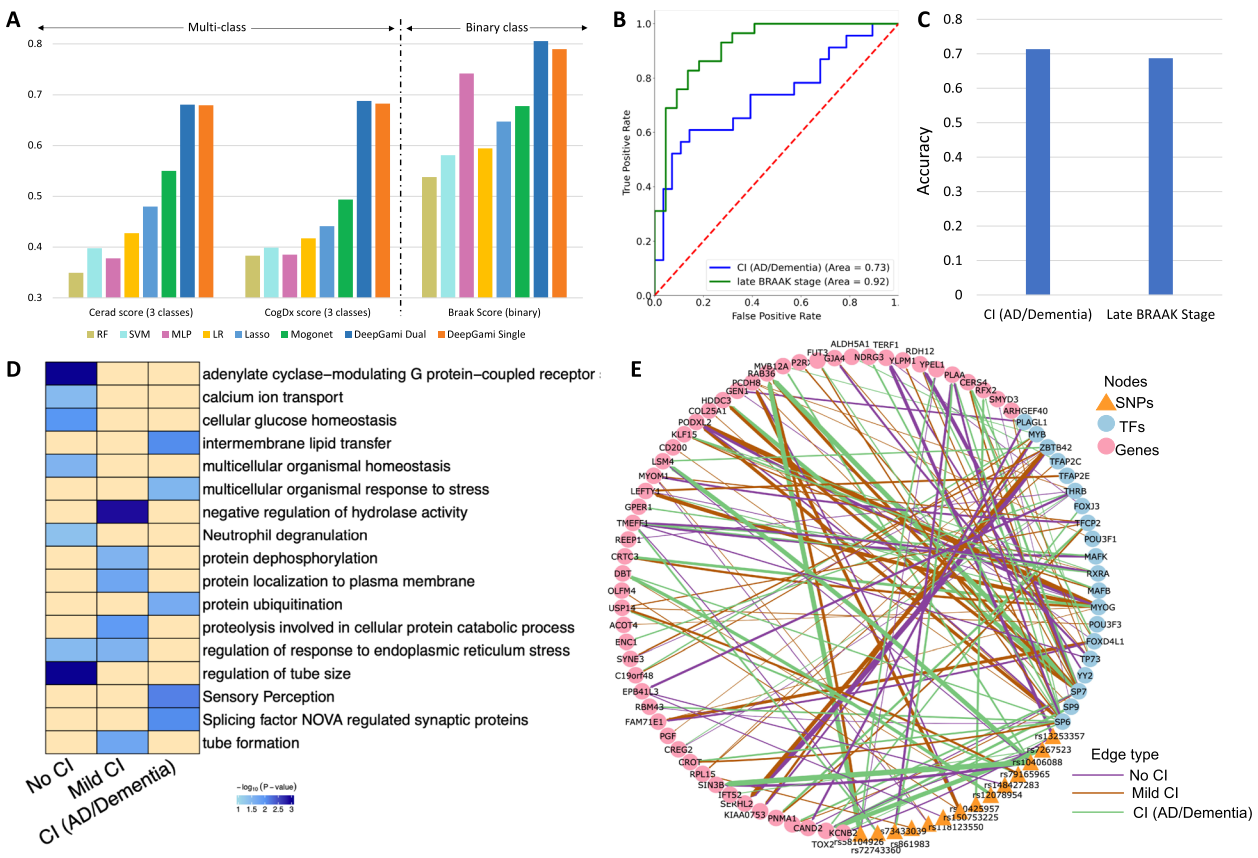

Fig. 3 Multi-class clinical phenotype prediction and regulatory network prioritization in Alzheimer's disease. **A** Fivefold cross-validation performance of DeepGAMI (Dual modality: dark blue, Single modality: orange) on three different phenotypes: neuritic plaque measure (CERAD score, multi-class), cognitive impairment (COGDX score, multi-class) and neurofibrillary tangle pathology (BRAAK stage, binary) in comparison with Lasso (brown), LR (light blue), Random Forest (yellow), SVM (purple), MLP (pink), and MOGONET (green). **B** ROC curves of held-out test samples for cognitive COGDX phenotype (blue) and late BRAAK stage (green). **C** Classification accuracies of the independent dataset for COGDX phenotype and BRAAK stage. **D** Enrichment analysis of prioritized genes for no cognitive impairment, mild cognitive impairment, and cognitive impairment (AD/Dementia) classes of COGDX phenotype. **E** Select an example of a prioritized regulatory network for the cognitive impairment phenotype. The edge thickness between any two nodes corresponds to the prioritized link importance score of the associated nodes. The edge color represents the three classes

file 2: Table S4). For BRAAK phenotype, DeepGAMI ($BACC = 0.806 \pm 0.03$ for dual, $BACC = 0.79 \pm 0.02$ for single) outperforms random guess (BACC = 0.50), Random Forest classifier($BACC = 0.538 \pm 0.01$), SVM ($BACC = 0.5808 \pm 0.13$), MLP ($BACC = 0.742 \pm 0.02$), LR ($BACC = 0.594 \pm 0.018$), Lasso ($BACC = 0.647 \pm 0.057$), and MOGONET ($BACC = 0.678 \pm 0.025$). Looking at a more complex phenotype with multi-class, DeepGAMI with two modalities improved the classification accuracy of COGDX ($BACC = 0.688 \pm 0.07$) compared to the highest accuracy of the best model (MOGONET, $BACC = 0.4938 \pm 0.05$). Also, DeepGAMI with single modality input improved the multi-class classification accuracy ($BACC = 0.6826 \pm 0.07$) compared to MOGONET($BACC = 0.4938 \pm 0.05$). Following this, we tested the performance of the classifiers with optimal hyperparameters on 100 randomly generated training and validation sets. Based on the k.s. test, both DeepGAMI dual and DeepGAMI single has the best performance (Additional file 1: Fig S2A and Fig S2B).

We tested DeepGAMI on additional samples from ROSMAP cohort with missing modalities where these individuals had only genotype information (single modality as input). These samples belonged to late-stage BRAAK individuals and CI (cognitive impairment in AD/Dementia). DeepGAMI was able to achieve AUC scores of 0.92 for late-stage BRAAK and 0.73 for CI (Fig. 3B). We were able to classify these late-stage BRAAK individuals with an accuracy of 0.687 and CI individuals with an accuracy of 0.713 (Fig. 3C). We observed that our model had higher AUC scores in comparison with accuracy scores for late-stage BRAAK samples. This might be due to the imbalance in the training samples. Also, AUC works best in the binary classification setting and has inconsistencies in the multi-class problem [104, 105].

We then performed enrichment analysis on the top-ranked genes that were regulated by the SNPs and TFs for each group (no CI, mild CI, CI) of the COGDX phenotype (Fig. 3D, Additional file 5) and generated a network containing the prioritized SNPs and TFs with prioritized links to the genes (Fig. 3E, Additional file 6). These prioritized genes were enriched with many known cognitive impairment functions and pathways. For example, controls were enriched for adenylate cyclase-modulating G protein-coupled receptors that are known to have a role in the pathological prognosis of AD [106, 107]. Mild CI was associated with protein dephosphorylation [108], response to endoplasmic reticulum stress [109–111] and proteolysis in cellular protein catabolic process [112]. We observed that CI was associated with sensory perception and splicing factor NOVA regulated synaptic proteins. Sensory perception impairment is known to affect cognition [113–115]. NOVA regulates genes critical for neuronal function [116] and known to affect patients with inhibitory motor control dysfunctions [117].

Cortical layer classification for single-cell neuronal cells in mouse visual cortex

We also tested DeepGAMI on additional non-omics modalities using an emerging Patch-seq dataset [81] containing single-cell multimodal data for the visual cortex brain region in neuronal cells of mouse species. This dataset includes transcriptomics and electrophysiological (ephys) data. We used the cell cortical layers (L1, L2/3, L4, L5, L6) that reveal the location of the cells in the visual cortex as the cellular phenotype. We followed the data extraction and preprocessing as done in DeepManReg [92] and ended up with 41 ephys features and 1000 genes for 3654 neuronal cells. Figure 4A depicts the overall architecture of DeepGAMI for this dataset. It demonstrates that DeepGAMI can handle multiple modalities besides genotype and gene expression and can also perform multi-class classification. While Fig. 2C and Fig. 4A looks similar, they serve different purposes, with Fig. 4A showing the model architecture and Fig. 2C highlighting the prioritization results.

To evaluate DeepGAMI's performance on Patch-seq data, we adopted the validation technique used in previous studies [92] and randomly split the cells into training/testing sets with a ratio of 80:20 and obtained 100 different sets. As there was a huge imbalance in the number of cells for each layer (L1: 262 cells; L2/3: 1097 cells; L4: 385 cells; L5: 1176 cells; L6: 734 cells), we applied SMOTE [121] oversampling technique on the training set to have a balanced number of cells for each layer in the training label and ended up with 941 cells in each layer. SMOTE identifies k-nearest neighbors of each sample in the minority class and creates synthetic samples along the line segments joining these neighbors. Biological Dropconnect was not used as there was no prior biological data available. Thus, we instead used full connectivity, where each gene and ephys feature had an association with the intermediate latent space layer. After various parameter tuning, the latent space dimension was set to 500.

We compared the prediction accuracy of DeepGAMI with different methods using pairwise Kolmogorov–Smirnov test (k.s test) on the accuracy distribution over 100 runs. DeepGAMI has higher prediction accuracy for classifying cell cortical layers than other methods (k.s test statistic = 1, $p - value < 2.2e^{-16}$ for dual-modality input, k.s test statistic = 1, $p - value < 1.9e^{-16}$ for single modality gene expression input, Fig. 4B and Additional file 2: Table S5). Furthermore, the average accuracy of DeepGAMI dual-modality mode (0.6571 with a 95% confidence interval (CI) of [0.6249, 0.6892]) and DeepGAMI

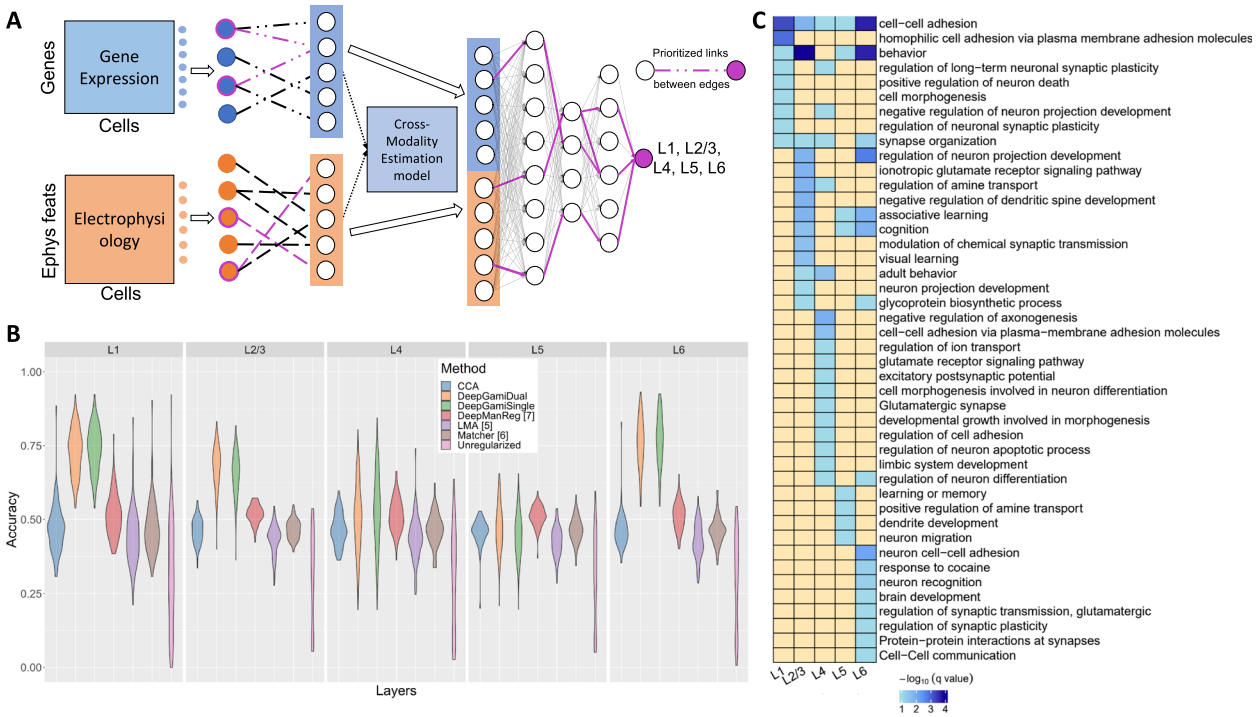

Fig. 4 Classifying cellular phenotype in single neuronal cells of mouse visual cortex. **A** DeepGAMI model for cell layer classification. **B** Balanced accuracies for classifying cell layers in the mouse visual cortex by DeepGAMI dual-modality (orange), DeepgGAMI single-modality (green) versus DeepManReg [92] (dark pink), neural network classification without any regularization (light pink), LMA [118] (violet), CCA [119] (blue), and MATCHEr [120] (brown). **C** Gene enrichment analysis showing the enriched terms for layer-specific prioritized genes

single modality mode (0.6463 [0.6129, 0.6797]) is higher than the random guess baseline of 0.2 (five labels), LMA [118] (0.43 [0.322, 0.496]), CCA [119] (0.462, [0.401, 0.513]), MATCHER [120] (0.465, [0.409, 0.528]), and DeepManReg [92] (0.514, [0.479, 0.548])).

We then compared the performance of DeepGAMI with- and without- oversampling on multi-class classification. We hypothesized that oversampling helps perform better on imbalanced dataset. DeepGAMI with oversampling had higher accuracies for classes L1, L4, and L6 while if performed slightly lesser on layers L2/3 and L5 (Additional file 1: Fig S3A). DeepGAMI with oversampling has two times better accuracy for cell layer L4 which had the least number of samples. We also converted the multi-class labels into binary classification problem. We then developed DeepGAMI for each cell layer without oversampling. As expected, DeepGAMI had higher accuracy on binary classification problem (Additional file 1: Fig S3B). While binary classification has higher accuracies, multiclass classification is useful when classifying new cells into the cell layers.

We extracted 112 (L4) neuronal cells patch-seq data [122] containing gene expression data but only a small set of electrophysiological data for independent testing. We applied DeepGAMI for classifying these 112 cells by giving only gene expression (single modality mode) and allowing the model to estimate the latent space of ephys features. For the negative samples required for performance estimation, we used the predictions on the motor cortex data [81]. The motor cortex dataset contains 1286 genes and 29 electrophysiological features of neuronal cells without the L4 layer: L1, L2/3, L5, and L6. After predicting these cells, we used the predicted values for the L4 layer as the negative samples, combined them with the predictions for the L4 layer for the 112 samples, and computed the AUC score. DeepGAMI classified the cells into L4 layer with an AUC score of 0.73.

Following the prediction, we extracted the top 10% of the prioritized genes and importance scores of the 41 ephys features for each cell layer (Additional file 7). Figure 4C shows the gene set enriched terms [86] for each layer. Enriched terms like cell–cell adhesion and neuron projection development appear in all layers [123]. Layer 4 is enriched with excitatory neurons and their activities [93, 124]. Many groups were enriched for behavior (especially L2 and L6), and synapse organization. L1 and L2/3 groups were enriched for negative regulation of neuron projection development and long-term neuronal synaptic plasticity regulation. The upper layers of the cortex (groups L2/3 and L4) were enriched for amine transport

regulation and adult behavior; the primary input from the thalamus goes to Layer 4, whose input then goes to Layers 2 and 3. Additional file 1: Fig S4 compares the importance scores of the 41 ephys features across six cell layers.

Classification of schizophrenia individuals using genotype and cell-type gene expression data

We also tested if DeepGAMI can prioritize cell-type disease genes and SNPs in major brain cell types for schizophrenia: excitatory neurons, inhibitory neurons, oligodendrocytes, and other glia (microglia and astrocytes). Notably, we used genotype and cell-type-specific gene expression imputed using bMIND [89] from a reference panel of four cell types: GABAergic (i.e., inhibitory) neurons, glutamatergic (i.e., excitatory) neurons, oligodendrocytes, and a remaining group composed mainly of microglia and astrocytes. Additional file 2: Table S6 summarizes the number of features used as input for the model for these four cell types. DeepGAMI classified SCZ individuals with a balanced accuracy of 0.795 ± 0.035 for dual-modality and 0.784 ± 0.024 for single modality for microglia and astrocyte cell type in comparison to 0.563 ± 0.049 for RF, 0.6585 ± 0.079 for SVM, 0.7429 ± 0.058 for MLP, 0.729 ± 0.033 for LR, 0.717 ± 0.031 for Lasso, 0.765 ± 0.026 for Varmole, and 0.6919 ± 0.037 for MOGONET. Similarly, DeepGAMI performs better fivefold classification than other models for the other three cell types (Fig. 5A, Additional file 2: Table S7). Additional file 1: Fig S5A compares the robustness of DeepGAMI with other state-of-the-art methods. DeepGAMI consistently outperformed Varmole, Mogonet, and other machine learning methods in all cell types. DeepGAMI was also able to produce a better classification of schizophrenia samples against control on held-out test data in comparison to existing methods (Additional file 1: Fig S5B).

We then performed enrichment analysis for the prioritized SNPs for each cell type. We found various known cell-type pathways and functions associated with SCZ (Fig. 5B) like organelle localization in inhibitory neurons [125], structural molecule activity in microglia and

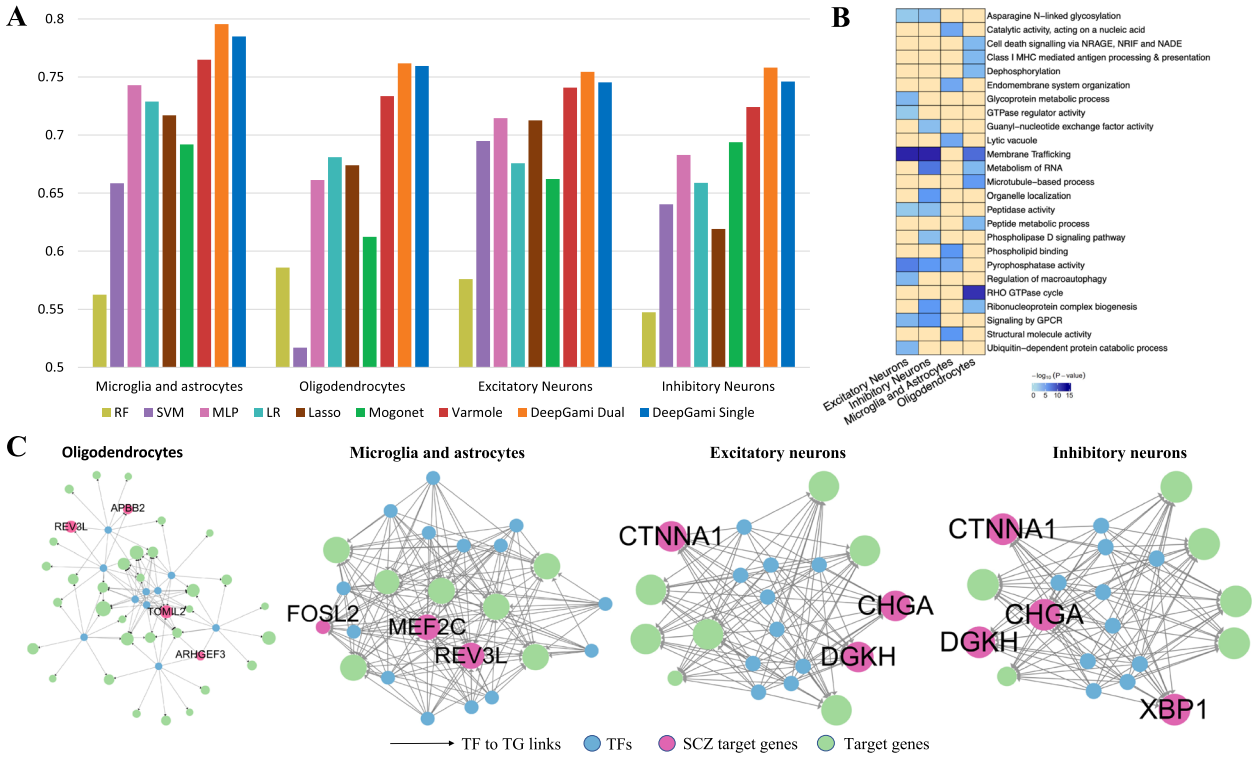

Fig. 5 Classification of schizophrenia individuals and prioritization of genes, SNPs, and regulatory network using genotype and gene expressions of four cell types (microglia and astrocytes, oligodendrocytes, inhibitory neurons, and excitatory neurons). **A** Balanced accuracies from 5-fold cross-validation of DeepGAMI dual-modality model (dark blue), DeepGAMI single modality model (orange) in comparison with Lasso (brown), LR (light blue), Random Forest (yellow), SVM (purple), MLP (pink), Varmole (red), and MOGONET (green). **B** Pathway enrichment of prioritized schizophrenia SNPs for four cell types. The blue shade gives the $-\log(p\text{-values})$ of the term. **C** Prioritized cell-type gene regulatory networks with pink circles representing schizophrenia genes. The size of the target gene is defined by the number of prioritized links between the SNPs and the associated gene

astrocytes [126], RHO GTPase cycle in oligodendrocytes [127], and regulation of macroautophagy in excitatory neurons [128]. We also found common SCZ-associated functional pathways across cell types like asparagine N-linked glycosylation [129, 130], membrane trafficking [131], and signaling by GPCR [132].

Figure 5C visualizes subnetworks of prioritized cell-type regulatory networks showcasing the ability of DeepGAMI to prioritize cell-type genes like APBB2 and TOMIL2 for oligodendrocytes, FOSL2 and MEF22C for microglia and astrocytes, and XBP1 for inhibitory neurons, and some common genes like REV3L (oligodendrocytes and microglia and astrocytes) and CTNNA1 (excitatory and inhibitory neurons). A complete list of importance scores of genes and SNPs for each cell type is available in Additional files 8 and 9.

Assessing the impact of different networks as input on the classification performance

One of the major contributions of this study is the use of prior biological knowledge to guide the neural network model for phenotype prediction. To test the effectiveness of these prior networks (GRNs and eQTLs), we performed an ablation study of DeepGAMI, where we compared our proposed model with two additional variants: (1) DeepGAMI Random—We generated random networks in place of prior biological networks and used these as the basis for DropConnect and (2) DeepGAMI Bernoulli—Based on the number of edges in the GRNs and eQTLs, we used Bernoulli distribution to generate networks with the same distribution as the original networks where existing edges has an 80% chance of being selected where other links have 20% chance of being selected. We compared the performance of these three variations on the bulk tissue SCZ cohort, inhibitory neuron cell type SCZ cohort, and AD cohort with CERAD score phenotype. We performed 5-fold cross-validation to assess the performance of all the variations. As shown in Additional file 1: Fig S6, DeepGAMI, with prior biological knowledge, outperformed its variations in all classification tasks. DeepGAMI Bernoulli variations had a better performance than DeepGAMI random, indicating that a biology-guided neural network helps improve phenotype prediction and aids in prioritizing molecular and cellular features.

Discussion

DeepGAMI is a novel interpretable deep learning model for improving genotype–phenotype prediction from multimodal data. Its auxiliary learning layer enables cross-modal imputation to predict phenotypes still when some modalities are unavailable. The model also takes prior biological information for aiding in prioritizing multimodal features (e.g., SNPs, genes) and feature networks (e.g., gene regulatory networks) related to the phenotypes.

As brain phenotypes involve complex cellular and molecular mechanisms, genotype and gene expression are a few of the many factors associated with mechanisms. We have demonstrated that DeepGAMI can handle various multimodal data as input in two scenarios. In the first scenario, DeepGAMI was able to accurately predict cortical layers using gene expression and electrophysiological features in mouse visual cortex (Fig. 4). In the second scenario, we used DNA methylation and gene expression as inputs to DeepGAMI to predict AD phenotype (COGDX scores: No CI vs Mild CI vs CI-AD/Dementia) from the ROSMAP cohort. After preprocessing and filtering, we used 1198 CpG sites as features from the methylation data and 183 TF gene expressions as inputs. The intermediate gene layers consist of 1013 target genes. We used CpG island sites for each gene and GRN as biological priors. Notably, DeepGAMI achieved the best multi-class accuracy of 0.524 on the held-out test samples and 0.557 on the independent samples that only had methylation data (Additional file 1: Fig S7). Integrating additional modalities into DeepGAMI enables a more profound understanding of such mechanisms. For example, several studies have tried integrating copy number variations with DNA methylation [133], gene expression [134], and clinical data [135]. Trevino et al. [136] integrated RNA-seq and ATAC-seq over a period of time, studying various genetic activities and disease susceptibility in various neuropsychiatric disorders. MVIB [137] integrated gut microbial markers and abundance scores to classify various diseases. Furthermore, DeepGAMI currently integrates two data modalities. We plan to extend DeepGAMI to integrate more than two modalities in the future.

DeepGAMI uses fivefold cross-validation balanced accuracies and AUC scores to compare the performance of the model on various datasets. While cross-validation can be misleading when used for model selection, it can still be a useful technique for estimating the expected performance of the model on the test set and comparing results when the sample size is not large enough to split the data into separate train and test sets [138, 139]. Moreover, despite the relatively lower sample sizes, DeepGAMI has demonstrated accurate performance on held-out test samples for three cohort datasets: SCZ (Fig. 2B, C), AD (Fig. 3B, C), and mouse visual cortex (Fig. 4C) that only contain genotype information.

We evaluated the scalability of DeepGAMI by varying the number of input features to the model and recording the runtime. For this analysis, we used PsychENCODE consortium [49] data for predicting schizophrenia (SCZ)

versus healthy individuals. Additional file 1: Fig S8 gives a detailed comparison of three models: DeepGAMI, Varmole, and MOGONET. We varied the number of input features from 100 to 100,000 and recorded the total runtime. We see that the runtime (training time) scales as the number of features increases. While the runtime performance is similar among all three models, MOGONET has the least runtime. We believe the cross-modal imputation layer of DeepGAMI might be the cause for relatively slower runtime.

One of the major contributions of DeepGAMI is its ability to make accurate phenotype predictions even when some modalities are missing, which is achieved using cross-modal imputation. DeepGAMI differs from traditional cross-modal imputation methods. Firstly, it aims to integrate multimodal data for improving phenotype prediction, rather than focusing solely on cross-modal imputation. Secondly, it can handle both lower sample size population-level datasets and single-cell multimodal datasets while the latter methods are mainly based on single-cell datasets. Lastly, existing cross-modal imputation techniques can be extended to perform various supervised and unsupervised learning tasks. However, it is important to note that this process involves building two separate models: one for modality estimation and another for prediction. As a result, there is a risk of missing phenotype-related shared features between modalities, which can potentially impact the accuracy of the predictions. Furthermore, traditional cross-modal imputation methods often do not allow for the incorporation of prior biological knowledge into the models. This can limit the interpretability of the results. DeepGAMI uses linear regression for cross-modal imputation. Even though our applications have shown the imputations work well, DeepGAMI can use nonlinear functions for cross-modality imputation, aiming to have nonlinear auxiliary learning. It is also possible to integrate existing cross-modal imputation methods into DeepGAMI's auxiliary task of cross-modality imputation. This could potentially lead to further improvements in performance. The results from our analysis on AMP-AD data may be susceptible to batch effects and data-source-specific effects as we were able to only extract the ROSMAP cohort. Additional cohorts from the same study can further enhance the generalizability of our model.

Disease variation is affected by both genetic and non-genetic factors which can include covariates like age, batch, sex, and ethnicity. In many cases, the covariate information is either missing or partially available. The genotype data can be adjusted by these covariates by the users. We test the effect of three covariates (age, gender, and ethnicity) on the CMC cohort. For each SNP, we regress out its genotype data across individuals by covariates and then input the residuals to DeepGAMI and the classification performance decreased ($AUC_{dual} = 0.748$ and $AUC_{single} = 0.723$). DeepGAMI has higher performance with original genotype data, suggesting that the nodes in the hidden layers of DeepGAMI can capture the hidden effects of these covariates.

We showcase the ability of DeepGAMI in predicting phenotypes with genotype data alone. PRS is a popular regression-based method to quantify genotype to phenotype association. Thus, we calculated PRS for SCZ (binary trait) using three different methods: PLINK [140], LDpred2 [141], and PRSice [142] on our data. We used 7433 SNPs along with age, gender, and ethnicity as covariates for this analysis. PRS was able to explain moderate percentages of variations ($R^2_{PLINK} = 0.584$, $R^2_{LDpred2} = 0.567$, and $R^2_{PRSice2} = 0.762$). As AUC is typically used for evaluating classification problems, out result show that DeepGAMI dual ($AUC = 0.895$) and DeepGAMI single ($AUC = 0.867$) perform better in comparison to the recently reported PRS score for SCZ ($AUC = 0.61$) [76] and heritability of SCZ (0.8) [143]. In future, DeepGAMI can be extended to integrate regression and predict continuous phonotypes like PRS.

Conclusions

In this study, we presented DeepGAMI, an interpretable biology-guided deep learning framework for phenotype prediction using multi-modal data. We demonstrated that DeepGAMI improves prediction of disease types and clinical phenotypes and prioritizes phenotypic genomic features and regulatory networks in AD and SCZ, especially at the cell-type level. We envision DeepGAMI can be used to decipher functional genomics and gene regulation for other complex diseases.

Abbreviations

GRN	Gene regulatory network
eQTL	Expression quantitative trait loci
GWAS	Genome-wide association studies
SCZ	Schizophrenia
AD	Alzheimer's disease
CI	Cognitive impairment
MLP	Multilayer perceptron
PRS	Polygenic Risk Score
SNP	Single-nucleotide polymorphism

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-023-01248-6.

Additional file 1: Fig S1. Independent validation performance comparison on schizophrenia cohort with genotype and bulk tissue gene expression. **Fig S2.** Kolmogorov-smirnov (k.s.) test comparison of classification accuracy for Alzheimer's disease cohort. **Fig S3.** Performance comparison of DeepGAMI with oversampling, without- oversampling, and binary classification on Patch-seq mouse visual cortex data. **Fig S4.**

Integrated Gradient results for Patch-seq mouse visual cortex data. **Fig S5.** Independent validation performance comparison on schizophrenia cohort with genotype and celltype gene expression. **Fig S6.** Performance of DeepGAMI with its variations on ablation study across all classification tasks. **Fig S7.** Multiclass classification of AD phenotype (COGDX score: No CI, Mild CI, and CI) using methylation and gene expression data from ROSMAP cohort. **Fig S8.** Runtime comparison of DeepGAMI with MOGONET and Varmole on varying input feature sizes.

Additional file 2: Table S1. List of all hyperparameters used in DeepGAMI. **Table S2.** Summary table of the total number of trainable parameters for each dataset. **Table S3.** Summary table showing features and class labels for different available phenotypes for ROSMAP AD dataset. **Table S4.** Balanced accuracy comparison for ROSMAP AD dataset. **Table S5.** Balanced accuracy comparison for Patch-seq dataset. **Table S6.** Summary table of the features for cell-type-specific Schizophrenia dataset. **Table S7.** Binary Classification results for cell-type-specific Schizophrenia dataset.

Additional file 3. Prioritized bulk genes and SNPs for schizophrenia.

Additional file 4. Prioritized bulk transcription factors to genes and SNPs to gene links for schizophrenia.

Additional file 5. Prioritized genes and SNPs for cognitive impairment phenotype in Alzheimer's disease.

Additional file 6. Prioritized transcription factors to genes and SNPs to gene links for cognitive impairment phenotype in Alzheimer's disease.

Additional file 7. Prioritized genes and electrophysiological features for cell cortical layers in mouse visual cortex.

Additional file 8. Prioritized cell-type genes and SNPs for schizophrenia.

Additional file 9. Prioritized cell-type transcription factors to genes and SNPs to gene links for schizophrenia.

Acknowledgements
The authors wish to thank all members of the Wang lab and Roussos lab for insightful discussions on the topic.

Authors' contributions
D.W. conceived the study. D.W. and P.B.C. designed the methodology and experiments. P.B.C., C.H., T.J. J.B., and J.F. curated and processed data required for the analysis. J.W. and G.H. imputed cell-type gene expression data of individuals. P.B.C., C.H., T.J., and S.K. performed analysis and visualization. P.B.C. and S.A. implemented the software. P.B.C., J.W., G.H., P.R. and D.W. wrote and edited the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by National Institutes of Health grants, R01AG067025 (to P.R. and D.W.), RF1MH128695 (to D.W.), R03NS123969 (to D.W.), R21NS127432 (to D.W.), R21NS128761 (to D.W.), U01MH116492 (to D.W.), U01MH116442 (to P.R.), R01MH110921 (to P.R.), R01MH109677 (to P.R.), P50HD105353 (to Waisman Center), National Science Foundation Career Award 2144475 (to D.W.), Simons Foundation Autism Research Initiative pilot grant 971316 (to D.W.), and the start-up funding for D.W. from the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison. The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation.

Availability of data and materials
The PsychENCODE bulk gene expression file for schizophrenia disorder can be downloaded from http://resource.psychencode.org/Datasets/Derived/DER-01_PEC_Gene_expression_matrix_normalized.txt [49], and the genotype data can be accessed from http://resource.psychencode.org [49]. The cell-type-specific reference panel used for gene expression imputation is available at https://www.synapse.org/#!Synapse:syn22321061 [144], and the imputed gene expression is available at https://www.synapse.org/#!Synapse:syn23234712 [89, 145]. ROSMAP Alzheimer's disease gene expression data is available at https://doi.org/10.7303/syn3388564 [55] and genotype can be downloaded from https://doi.org/10.7303/syn3157329 [55]. The processed gene expression and electrophysiological data from Patch-seq in mouse visual cortex is available at https://github.com/daifengwanglab/scMNC [146]. DeepGAMI was implemented in Python using PyTorch as the deep learning package and the source code is publicly available at https://github.com/daifengwanglab/DeepGAMI [147].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Waisman Center, University of Wisconsin-Madison, Madison, WI 53705, USA. [2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53076, USA. [3]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53076, USA. [4]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA. [5]Center for Disease Neurogenomics, Department of Psychiatry and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [6]Mental Illness Research, Education and Clinical Centers, James J. Peters VA Medical Center, Bronx, NY 10468, USA. [7]Center for Dementia Research, Nathan Kline Institute for Psychiatric Research, Orangeburg, NY 10962, USA.

Received: 5 December 2022 Accepted: 16 October 2023
Published online: 31 October 2023

References

1. Manolio TA, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.
2. Lehner B. Genotype to phenotype: lessons from model organisms for human genetics. Nat Rev Genet. 2013;14:168–78.
3. Visscher PM. Sizing up human height variation. Nat Genet. 2008;40:489–90.
4. Nicolae DL, et al. Trait-associated snps are more likely to Be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6:e1000888.
5. Coon KD, et al. A high-density whole-genome association study reveals that *APOE* is the major susceptibility gene for sporadic late-onset Alzheimer's disease. J Clin Psychiatry. 2007;68:8183.
6. Marioni RE, et al. GWAS on family history of Alzheimer's disease. Transl Psychiatry. 2018;8:1–7.
7. Jansen IE, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet. 2019;51:404–13.
8. Kunkle BW, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat Genet. 2019;51:414–30.
9. Ripke S, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421–7.
10. Ikeda M, et al. Genome-wide association study detected novel susceptibility genes for schizophrenia and shared trans-populations/diseases genetic effect. Schizophr Bull. 2019;45:824–34.
11. Gandhi S, Wood NW. Genome-wide association studies: the key to unlocking neurodegeneration? Nat Neurosci. 2010;13:789–94.
12. Wang H, et al. From phenotype to genotype: an association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs. Bioinformatics. 2012;28:i619–25.
13. Yang T et al. Detecting genetic risk factors for Alzheimer's disease in whole genome sequence data via Lasso screening. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). 2015. p. 985–989. https://doi.org/10.1109/ISBI.2015.7164036.
14. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19:581–90.

15. Zhang Y, Zhan L, Thompson PM, Huang H. Biological knowledge guided deep neural network for brain genotype-phenotype association study. In: International Workshop on Multimodal Brain Image Analysis 2019 Oct 10 (pp. 84-92). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-33226-6_10.
16. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22:1748–59.
17. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.
18. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 2008;24:408–15.
19. Li G, Jima D, Wright FA, Nobel AB. HT-eQTL: integrative expression quantitative trait loci analysis in a large number of human tissues. BMC Bioinformatics. 2018;19:95.
20. Cai L, et al. Implications of newly identified brain eQTL genes and their interactors in schizophrenia. Mol Ther Nucleic Acids. 2018;12:433–42.
21. Sieberts SK, et al. Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. Sci Data. 2020;7:340.
22. Võsa U, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53:1300–10.
23. Zeng B et al. Trans-ethnic eQTL meta-analysis of human brain reveals regulatory architecture and candidate causal variants for brain-related traits. 2021:2021.01.25.21250099 Preprint at https://doi.org/10.1101/2021.01.25.21250099.
24. Gamazon ER, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47:1091–8.
25. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. PLoS Genet. 2017;13:e1006646.
26. Barbeira AN, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9:1825.
27. Gusev A, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat Genet. 2018;50:538–48.
28. Zhang Y, et al. PTWAS: investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis. Genome Biol. 2020;21:232.
29. Li B, et al. Tissue specificity-aware TWAS (TSA-TWAS) framework identifies novel associations with metabolic, immunologic, and virologic traits in HIV-positive adults. PLoS Genet. 2021;17:e1009464.
30. Tang S, et al. Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia. PLoS Genet. 2021;17:e1009482.
31. Brain transcriptome wide association study (TWAS) implicates 8 genes across 6 loci in Alzheimer's disease - Gockley - 2020 - Alzheimer's & Dementia - Wiley Online Library. https://alz-journals.onlinelibrary.wiley.com/doi/abs/https://doi.org/10.1002/alz.044839.
32. Li B et al. Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. In: Biocomputing (World Scientific, 2017). 2018. p. 448-459. https://doi.org/10.1142/9789813235533_0041.
33. Wainberg M, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;51:592–9.
34. Lee T, Lee H. Prediction of Alzheimer's disease using blood gene expression data. Sci Rep. 2020;10:3485.
35. Arloth J, et al. DeepWAS: multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. PLoS Comput Biol. 2020;16:e1007616.
36. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods. 2015;12:931–4.
37. Liu E, Li L, Cheng L. Gene regulatory network review. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, Editorrs. Encyclopedia of bioinformatics and computational biology. Oxford: Academic Press; 2019.
38. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nat Genet. 2001;27:167–71.
39. Basso K, et al. Reverse engineering of regulatory networks in human B cells. Nat Genet. 2005;37:382–90.
40. Potkin SG, et al. Identifying gene regulatory networks in schizophrenia. Neuroimage. 2010;53:839–47.
41. Kawalia SB, et al. Analytical strategy to prioritize alzheimer's disease candidate genes in gene regulatory networks using public expression data. J Alzheimers Dis. 2017;59:1237–54.
42. Yazdani A, Mendez-Giraldez R, Yazdani A, Kosorok MR, Roussos P. Differential gene regulatory pattern in the human brain from schizophrenia using transcriptomic-causal network. BMC Bioinformatics. 2020;21:469.
43. Cabral C, et al. Classifying schizophrenia using multimodal multivariate pattern recognition analysis: evaluating the impact of individual clinical profiles on the neurodiagnostic performance. Schizophr Bull. 2016;42:S110–7.
44. Liang S, et al. Classification of first-episode schizophrenia using multimodal brain features: a combined structural and diffusion imaging study. Schizophr Bull. 2019;45:591–9.
45. Salvador R, et al. Multimodal integration of brain images for MRI-based diagnosis in schizophrenia. Front Neurosci. 2019;13:1203.
46. Lee G, Nho K, Kang B, Sohn K-A, Kim D. Predicting Alzheimer's disease progression using multi-modal deep learning approach. Sci Rep. 2019;9:1952.
47. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. Sci Rep. 2021;11:3254.
48. Zhao L, et al. DeepOmix: a scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis. Comput Struct Biotechnol J. 2021;19:2719–25.
49. Wang D, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362:eaat8464.
50. Nguyen ND, Jin T, Wang D. Varmole: a biologically drop-connect deep neural network model for prioritizing disease risk variants and genes. Bioinformatics. 2021;37:1772–5.
51. Wang T, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12:3445.
52. Li X, et al. MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. Front Genet. 2022;13:806842.
53. Conesa A, Beck S. Making multi-omics data accessible to researchers. Sci Data. 2019;6:251.
54. Martin KR, et al. The genomic landscape of tuberous sclerosis complex. Nat Commun. 2017;8:15816.
55. De Jager PL, et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci Data. 2018;5:180142.
56. Argelaguet R, et al. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14:e8124.
57. Wu KE, Yost KE, Chang HY, Zou J. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. Proc Natl Acad Sci U S A. 2021;118:e2023070118.
58. Du J-H, Cai Z, Roeder K. Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scVAEIT. Proc Natl Acad Sci. 2022;119:e2214414119.
59. Lin Y, et al. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat Biotechnol. 2022;40:703–10.
60. Jaderberg, M. et al. Reinforcement learning with unsupervised auxiliary tasks. 2016. arXiv:1611.05397 [cs].
61. Goyal P, Mahajan D, Gupta A, Misra I. Scaling and benchmarking self-supervised visual representation learning. 2019. p. 6391–6400.
62. Nediyanchath, A., Paramasivam, P. & Yenigalla, P. Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. p. 7179–7183. https://doi.org/10.1109/ICASSP40776.2020.9054073.
63. Lyu S, et al. Auxiliary learning for relation extraction. IEEE Trans Emerg Topics Comput Intell. 2022;6:182–91.
64. Suddarth SC, Kergosien YL. Rule-injection hints as a means of improving network performance and learning time. In: Almeida LB, Wellekens CJ, editors. Neural networks. 1990. p. 120–129. https://doi.org/10.1007/3-540-52255-7_33.
65. Sutton RS et al. Horde: a scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In: The 10th

international conference on autonomous agents and multiagent systems-volume 2. 2011. p. 761–768.
66. Hernandez-Leal P, Kartal B, Taylor ME. Agent modeling as auxiliary task for deep reinforcement learning. Proc AAAI Conf Artific Intell Interact Digit Entertain. 2019;15:31–7.
67. Lin X, Baweja H, Kantor G, Held D. Adaptive auxiliary task weighting for reinforcement learning. In: Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NIPS2019), Vancouver, BC, Canada; 2019.
68. Zhang Y, Tang H, Jia K. Fine-Grained Visual Categorization Using Meta-learning Optimization with Sample Selection of Auxiliary Data. InComputer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII 15. Springer International Publishing; 2018. p. 241–56.
69. Chen S, Wang J, Chen Y, Shi Z, Geng X, Rui Y. Label Distribution Learning on Auxiliary Label Space Graphs for Facial Expression Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA; 2020. p. 13984–993.
70. Chen Y, Praveen P, Priyantha M, Muelling K, Dolan J. Learning On-Road Visual Control for Self-Driving Vehicles With Auxiliary Tasks. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). 2019. p. 331–338. https://doi.org/10.1109/WACV.2019.00041.
71. Mehta A, Subramanian A, Subramanian A. Learning end-to-end autonomous driving using guided auxiliary supervision. In: Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing. 2018; p. 1–8.
72. Situ N, Yuan X, Zouridakis G. Assisting main task learning by heterogeneous auxiliary tasks with applications to skin cancer screening. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings; 2011. p. 688–697.
73. Hu K, et al. Deep supervised learning using self-adaptive auxiliary loss for COVID-19 diagnosis from imbalanced CT images. Neurocomputing. 2021;458:232–45.
74. Gan L, Vinci G, Allen GI. Correlation Imputation for single-cell RNA-seq. J Comput Biol. 2022;29:465–82.
75. Sekhon A, Singh R, Qi Y. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. Bioinformatics. 2018;34:i891–900.
76. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12:44.
77. Plomin R, von Stumm S. Polygenic scores: prediction versus explanation. Mol Psychiatry. 2022;27:49–52.
78. Ashuach T, Gabitto MI, Jordan MI, Yosef N. MultiVI: deep generative model for the integration of multi-modal data. 2021. 2021.08.20.457057 Preprint at https://doi.org/10.1101/2021.08.20.457057.
79. Zhang R, Meng-Papaxanthos L, Vert JP, Noble WS. Semi-supervised single-cell cross-modality translation using Polarbear. Bioinform, preprint. 2021. https://doi.org/10.1101/2021.11.18.467517.
80. Wan L, Zeiler M, Zhang S, Le Cun Y, Fergus R. Regularization of neural networks using DropConnect. Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research. 2013;28:1058–1066.
81. Gouwens NW, et al. Integrated morphoelectric and transcriptomic classification of cortical GABAergic cells. Cell. 2020;183:935-953.e19.
82. Kingma DP, Ba J. Adam: a method for stochastic optimization. 2017. arXiv:1412.6980 [cs].
83. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A. Pytorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Sys. 2019;32.
84. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia. 2017;70:3319–28.
85. Kokhlikyan N et al. Captum: a unified and generic model interpretability library for PyTorch. 2020. arXiv:2009.07896 [cs, stat].
86. Zhou Y, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10:1523.
87. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
88. GTEx Consortium, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.
89. Wang J, Roeder K, Devlin B. Bayesian estimation of cell type–specific gene expression with prior derived from single-cell data. Genome Res. 2021;31:1807–18.
90. Hoffman GE, et al. Sex differences in the human brain transcriptome of cases with schizophrenia. Biol Psychiat. 2022;91:92–101.
91. Jin T, et al. scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. Genome Med. 2021;13:95.
92. Nguyen ND, Huang J, Wang D. A deep manifold-regularized learning model for improving phenotype prediction from multi-modal data. Nat Comput Sci. 2022;2:38–46.
93. Yoshimura Y, Dantzker JLM, Callaway EM. Excitatory cortical neurons form fine-scale functional networks. Nature. 2005;433:868–73.
94. Golimbet V, Korovaitseva G, Abramova L, Kaleda V. The 844ins68 polymorphism of the cystathionine beta-synthase gene is associated with schizophrenia. Psychiatry Res. 2009;170:168–71.
95. Sundararajan T, Manzardo AM, Butler MG. Functional analysis of schizophrenia genes using GeneAnalytics program and integrated databases. Gene. 2018;641:25–34.
96. Garg P, Sharp AJ. Screening for rare epigenetic variations in autism and schizophrenia. Hum Mutat. 2019;40:952–61.
97. Gardiner EJ, et al. Gene expression analysis reveals schizophrenia-associated dysregulation of immune pathways in peripheral blood mononuclear cells. J Psychiatr Res. 2013;47:425–37.
98. Wagh VV, et al. Peripheral blood-based gene expression studies in schizophrenia: a systematic review. Front Genet. 2021;12:736483.
99. van Scheltinga AFT, Bakker SC, Kahn RS. Fibroblast growth factors in schizophrenia. Schizophr Bull. 2010;36:1157–66.
100. Klimaschewski L, Claus P. Fibroblast growth factor signalling in the diseased nervous system. Mol Neurobiol. 2021;58:3884–902.
101. Funk AJ, McCullumsmith RE, Haroutunian V, Meador-Woodruff JH. Abnormal activity of the MAPK- and cAMP-associated signaling pathways in frontal cortical areas in postmortem brain in schizophrenia. Neuropsychopharmacology. 2012;37:896–905.
102. Crisafulli C, Drago A, Calabrò M, Spina E, Serretti A. A molecular pathway analysis informs the genetic background at risk for schizophrenia. Prog Neuropsychopharmacol Biol Psychiatry. 2015;59:21–30.
103. Solana C, Pereira D, Tarazona R. Early senescence and leukocyte telomere shortening in SCHIZOPHRENIA: a role for cytomegalovirus infection? Brain Sci. 2018;8:188.
104. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manage. 2009;45:427–37.
105. Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Brief Bioinform. 2012;13:83–97.
106. Zhao J, Deng Y, Jiang Z, Qing H. G protein-coupled receptors (GPCRs) in Alzheimer's disease: a focus on BACE1 related GPCRs. Front Aging Neurosci. 2016;8:58.
107. Azam S, et al. G-Protein-coupled receptors in CNS: a potential therapeutic target for intervention in neurodegenerative disorders and associated cognitive deficits. Cells. 2020;9:506.
108. Reese LC, Laezza F, Woltjer R, Taglialatela G. Dysregulated phosphorylation of Ca2+/calmodulin-dependent protein kinase II-α in the hippocampus of subjects with mild cognitive impairment and Alzheimer's disease. J Neurochem. 2011;119:791–804.
109. Popugaeva E, Bezprozvanny I. Role of endoplasmic reticulum Ca2+ signaling in the pathogenesis of Alzheimer disease. Front Mol Neurosci. 2013;6:29.
110. Wu J, et al. Endoplasmic reticulum stress and disrupted neurogenesis in the brain are associated with cognitive impairment and depressive-like behavior after spinal cord injury. J Neurotrauma. 2016;33:1919–35.
111. Liu Y, Yu J, Shi Y-C, Zhang Y, Lin S. The role of inflammation and endoplasmic reticulum stress in obesity-related cognitive impairment. Life Sci. 2019;233:116707.
112. Kepchia D, et al. Diverse proteins aggregate in mild cognitive impairment and Alzheimer's disease brain. Alzheimers Res Ther. 2020;12:75.
113. Fischer ME, et al. Age-related sensory impairments and risk of cognitive impairment. J Am Geriatr Soc. 2016;64:1981–7.

114. Rong H, et al. Association of sensory impairments with cognitive decline and depression among older adults in China. JAMA Netw Open. 2020;3:e2014186.
115. Rhodus EK, et al. Sensory processing abnormalities in community-dwelling older adults with cognitive impairment: a mixed methods study. Gerontol Geriatr Med. 2022;8:23337214211068290.
116. Zhang C, et al. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Science. 2010;329:439–43.
117. Licatalosi DD, Darnell RB. Splicing regulation in neurologic disease. Neuron. 2006;52:93–101.
118. Wang C, Mahadevan S. Alignment without correspondence. In: In Proceedings of the 21st International Joint Conferences on Artificial Intelligence. 2009.
119. Hotelling, H. Relations between two sets of variates. In Breakthroughs in statistics: methodology and distribution, pp. 162-190. New York, NY: Springer New York, 1992. https://doi.org/10.1007/978-1-4612-4380-9_14.
120. Welch JD, Hartemink AJ, Prins JF. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol. 2017;18:138.
121. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. Jair. 2002;16:321–57.
122. Scala F, et al. Layer 4 of mouse neocortex differs in cell types and circuit organization between sensory areas. Nat Commun. 2019;10:4174.
123. Leone DP, Srinivasan K, Chen B, Alcamo E, McConnell SK. The determination of projection neuron identity in the developing cerebral cortex. Curr Opin Neurobiol. 2008;18:28–35.
124. Khibnik LA, Cho KKA, Bear MF. Relative contribution of feedforward excitatory connections to expression of ocular dominance plasticity in layer 4 of visual cortex. Neuron. 2010;66:493–500.
125. Morris JA, Kandpal G, Ma L, Austin CP. DISC1 (Disrupted-In-Schizophrenia 1) is a centrosome-associated protein that interacts with MAP1A, MIPT3, ATF4/5 and NUDEL: regulation and loss of interaction with mutation. Hum Mol Genet. 2003;12:1591–608.
126. Mallya AP, Deutch AY. (Micro)Glia as effectors of cortical volume loss in schizophrenia. Schizophr Bull. 2018;44:948–57.
127. Sarowar T, Grabrucker AM. Rho GTPases in the amygdala—a switch for fears? Cells. 2020;9:1972.
128. Vucicevic L, Misirkic-Marjanovic M, Harhaji-Trajkovic L, Maric N, Trajkovic V. Mechanisms and therapeutic significance of autophagy modulation by antipsychotic drugs. Cell Stress. 2018;2:282–91.
129. Mueller TM, Haroutunian V, Meador-Woodruff JH. N-Glycosylation of GABAA receptor subunits is altered in schizophrenia. Neuropsychopharmacol. 2014;39:528–37.
130. Williams SE, Mealer RG, Scolnick EM, Smoller JW, Cummings RD. Aberrant glycosylation in schizophrenia: a review of 25 years of post-mortem brain studies. Mol Psychiatry. 2020;25:3198–207.
131. Schubert KO, Föcking M, Prehn JHM, Cotter DR. Hypothesis review: are clathrin-mediated endocytosis and clathrin-dependent membrane and protein trafficking core pathophysiological processes in schizophrenia and bipolar disorder? Mol Psychiatry. 2012;17:669–81.
132. Boczek T, et al. The role of G Protein-Coupled Receptors (GPCRs) and calcium signaling in schizophrenia. Focus on GPCRs activated by neurotransmitters and chemokines. Cells. 2021;10:1228.
133. Tong L, Wu H, Wang MD. Integrating multi-omics data by learning modality invariant representations for improved prediction of overall survival of cancer. Methods. 2021;189:74–85.
134. Zhang L, et al. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Front Genet. 2018;9:477.
135. Sun D, Wang M, Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. IEEE/ACM Trans Comput Biol Bioinf. 2019;16:841–50.
136. Trevino AE, et al. Chromatin accessibility dynamics in a model of human forebrain development. Science. 2020;367:eaay1645.
137. Grazioli F, et al. Microbiome-based disease prediction with multimodal variational information bottlenecks. PLoS Comput Biol. 2022;18:e1010050.
138. Forman G, Scholz M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. SIGKDD Explor Newsl. 2010;12:49–57.
139. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform. 2014;6:10.
140. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
141. LDpred2: better, faster, stronger | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/36/22-23/5424/6039173.
142. Choi SW, O'Reilly PF. PRSice-2: polygenic risk score software for biobank-scale data. GigaScience. 2019;8:giz082.
143. He D, et al. Prioritization of schizophrenia risk genes from GWAS results by integrating multi-omics data. Transl Psychiatry. 2021;11:1–12.
144. Hoffman GE. RNA-seq from 4 cell populations. https://www.synapse.org/#!Synapse:syn22321061. Accessed 25 Oct 2022.
145. Hoffman GE. Imputed celltype gene expression using bMIND and FANS4 reference panel. https://doi.org/10.7303/syn23234712. https://www.synapse.org/#!Synapse:syn23234712.
146. Huang J, Sheng J, Wang D. Manifold learning analysis suggests strategies to align single-cell multimodal data of neuronal electrophysiology and transcriptomics. Commun Biol. 2021;4(1):1308. https://github.com/daifengwanglab/scMNC.
147. Chandrashekar PB. DeepGAMI: Deep biologically guided auxiliary learning for multimodal integration and imputation to improve genotype-phenotype prediction. GitHub; 2023. https://github.com/daifengwanglab/DeepGAMI.

Publisher's Note