Phenotype prediction using biologically interpretable neural networks on multi-cohort multi-omics data

van Hilten, Arno; van Rooij, Jeroen; Ikram, M. Arfan; Niessen, Wiro J.; van Meurs, Joyce B.J.; Roshchupkin, Gennady V.; More Authors

DOI
10.1038/s41540-024-00405-w

Publication date
2024

Document Version
Final published version

Published in
npj systems biology and applications

Citation (APA)
van Hilten, A., van Rooij, J., Ikram, M. A., Niessen, W. J., van Meurs, J. B. J., Roshchupkin, G. V., & More Authors (2024). Phenotype prediction using biologically interpretable neural networks on multi-cohort multi-omics data. *npj systems biology and applications*, *10*(1), Article 81. https://doi.org/10.1038/s41540-024-00405-w

npj | systems biology and applications

Article

Published in partnership with the Systems Biology Institute

https://doi.org/10.1038/s41540-024-00405-w

Phenotype prediction using biologically interpretable neural networks on multi-cohort multi-omics data

Check for updates

Arno van Hilten[1,19] ✉, Jeroen van Rooij[2,19], BIOS consortium*, M. Arfan Ikram[3], Wiro J. Niessen[1,3], Joyce. B. J. van Meurs[2,4] & Gennady V. Roshchupkin[1]

Integrating multi-omics data into predictive models has the potential to enhance accuracy, which is essential for precision medicine. In this study, we developed interpretable predictive models for multi-omics data by employing neural networks informed by prior biological knowledge, referred to as visible networks. These neural networks offer insights into the decision-making process and can unveil novel perspectives on the underlying biological mechanisms associated with traits and complex diseases. We tested the performance, interpretability and generalizability for inferring smoking status, subject age and LDL levels using genome-wide RNA expression and CpG methylation data from the blood of the BIOS consortium (four population cohorts, $N_{\text{total}} = 2940$). In a cohort-wise cross-validation setting, the consistency of the diagnostic performance and interpretation was assessed. Performance was consistently high for predicting smoking status with an overall mean AUC of 0.95 (95% CI: 0.90–1.00) and interpretation revealed the involvement of well-replicated genes such as *AHRR*, *GPR15* and *LRRN3*. LDL-level predictions were only generalized in a single cohort with an R^2 of 0.07 (95% CI: 0.05–0.08). Age was inferred with a mean error of 5.16 (95% CI: 3.97–6.35) years with the genes *COL11A2, AFAP1*, *OTUD7A*, *PTPRN2*, *ADARB2* and *CD34* consistently predictive. For both regression tasks, we found that using multi-omics networks improved performance, stability and generalizability compared to interpretable single omic networks. We believe that visible neural networks have great potential for multi-omics analysis; they combine multi-omic data elegantly, are interpretable, and generalize well to data from different cohorts.

Over the last decades, association studies have uncovered numerous genes and CpGs to be associated with hundreds of traits and diseases[1]. This has led to tools for identifying high-risk individuals and biomarkers for early disease detection. For example, blood-based methylation biomarkers are currently used for early diagnosis of various forms of cancer[2,3]. However, for most complex diseases and traits, the combined effects, within and between different omics types, are still largely unexplored. For a more comprehensive understanding of human health and diseases and for more accurate prediction models, it is therefore necessary to study omic types in relation to one another. Thanks to recent technological improvements for high throughput sequencing and arrays technologies, the acquisition of multi-omics datasets has become more feasible, providing opportunities for new multi-omics analysis tools[4,5].

Recently, novel statistical frameworks and machine learning techniques have been published that integrate multi-omics data in a single analysis[6,7]. These studies show the potential for multi-omics analysis to improve prediction for various disorders while providing insight into disease biology[4,8]. Integrating different types of omics data in a single analysis is a challenging task, as each type has different, procedures, preprocessing steps and analytical requirements[9]. Combining omics data presents additional challenges, as each omic has unique dimensions, and it is essential to consider correlation structures both within and between the different omics types. Thus, for the combined analysis of multiple omics types, methods

[1]Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands. [2]Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands. [3]Department of Imaging Physics, Delft University of Technology, Delft, The Netherlands. [4]Department of Orthopaedics and Sports Medicine, Erasmus MC, Rotterdam, The Netherlands. [19]These authors contributed equally: Arno van Hilten, Jeroen van Rooij. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: a.vanhilten@erasmusmc.nl

need to be flexible and be able to deal with the high dimensionality of these datasets.

Neural networks have demonstrated such flexibility and have been widely successful in fields such as image classification[10], speech recognition[11], and protein modelling[12]. In contrast to most tasks in image analysis and speech recognition, the focus of multi-omics frameworks is not only on predictive performance but also on understanding the underlying aetiology. To facilitate this, a new field in machine learning, coined visible machine learning[13] emerged, in which prior biological knowledge is embedded in a neural network's architecture to create interpretable neural networks[14–19]. Recent examples of these kinds of neural networks applied in genomics are GenNet[20] and P-net[21]. In the GenNet framework, gene and pathway annotations were used to create interpretable neural networks for genetic risk prediction from genotypes. In P-net, methylation, gene expression and copy number variants were fed to an interpretable neural network to differentiate between primary or metastatic prostate cancers. Importantly, recent work evaluating the reliability of the interpretations of P-net found that the interpretations can be strongly affected by different weight initializations[19]. Other examples of visible neural networks include, PasNet[22], which integrated pathways information to predict survival for glioblastoma multiforme, a primary brain cancer. DrugCell[23] integrated Gene Ontology knowledge in a network to predict drug response for various cancers and ParsVNN[24] continued on this work and pruned the network for increased performance and better interpretability.

In this study, we employ visible neural networks to analyse multi-omics data. We extend the GenNet framework to create interpretable neural networks for multiple omics inputs and apply it to a dataset with transcriptomics and methylomics data. We validate the method using four cohorts in the BIOS consortium for the application of predicting age, low-density lipoproteins (LDL) levels and smoking status. Age prediction from methylation or gene expression data has been an active research area popularized by the work of Hannum et al. and Horvath[25,26]. Additionally, it has been shown that these clocks show an asymptote for older participants and strong biological sex differences, making age prediction particularly interesting to study with neural networks[27]. Smoking status and LDL level predictions are well-suited to evaluate the performance, stability and interpretation of the method. Methylation and gene expression are highly predictive for smoking status and predictive genes are well-documented[28,29]. On the other hand, low lipid lipoprotein cholesterol levels are a complex outcome with both environmental and genetic factors[30].

We present the following contributions to this work. First, we have extended GenNet to create visible neural networks that can predict outcomes from multi-omics data while quantifying the importance of each omic type, gene, and pathway. In the proposed network, we merge each omic at the gene level, enabling gene-wise extraction of the importance of each omic for prediction. Secondly, we investigate the robustness of the performance and interpretations of these visible neural networks across three different phenotypes. By leveraging the multi-cohort setting of the BIOS consortium in cohort-wise cross-validation, we can examine the stability of these networks across different random seeds and cohorts. Finally, we perform additional analyses to identify omic-specific information, gene-covariate interactions, and group-specific patterns learned by the visible neural networks.

Results

In this study, multi-omics data gathered by the Biobank-based Integrative Omics Study (BIOS) consortium was used to predict smoking status, age and low-density lipoprotein levels. Specifically, we used transcriptome and methylome data from BIOS's four largest cohorts: LifeLines (LL), Leiden Longevity Study (LLS), Netherlands Twin Register (NTR), and Rotterdam Study (RS) to evaluate the performance and the interpretations of the created neural networks in a cohort-wise cross-validation. Importantly, all cohorts within the BIOS consortium followed the same procedure in gathering and processing data (see "Methods"). An overview of the characteristics of each cohort can be found in Table 1.

Network design

Neural network architectures were created using principles from the GenNet framework[20]. This framework uses prior knowledge (e.g. gene and pathway annotations) to connect input data to the neurons in the next layer of the neural network. CpG methylation sites were annotated using Genomic Regions Enrichment of Annotations Tool (GREAT)[31] and connected to the closest gene based on genomic distance (in base pairs) resulting in 17,283 gene annotations for 481,388 methylation sites. These gene annotations were intersected with the 14,248 remaining gene expression measurements left after preprocessing, resulting in an overlap of 10,404 genes between both omic types. This set of overlapping genes was used in all analyses. The methylation gene layer was built using these genes and their corresponding 324,295 CpGs. For the creation of pathway layers, the set of overlapping genes was grouped into KEGG's functional pathways[32] from ConsensusPathDB[33]. Out of the 10,404 genes, 4813 genes were annotated for at least one pathway.

The gene expression network (GE network, Fig. 1a) is the simplest network and consists of the gene expression input connected straight to the output node similar as in LASSO regression. The methylation network (ME network, Fig. 1b) consists of the input methylation data, a gene layer with neurons representing gene methylation made and an output node. The methylation and gene expression network (ME + GE network, Fig. 1c) combines both networks. In a similar way as in the ME network, CpGs are fed to the first layer of the network and reduced to one node per gene using gene annotations. In contrast to most other methods, gene expression is not concatenated to the input but is used as a separate input in the gene level of the network. In this layer, gene expression is combined with the neurons representing genes by methylation. Finally, a single node was used to predict the target phenotype.

The activation function transforms the output signal for each neuron. For classification tasks, such as predicting if an individual smokes or not, a sigmoid activation function was used to scale the output to the range [0, 1] in the last neuron. Arctanh activation functions were used for all other layers to introduce non-linearities, increasing the modelling capabilities of the network. For regression tasks, such as predicting continuous traits such as age

Table 1 | Main characteristics for all cohorts used in this study

	Rotterdam study	LifeLines	Leiden longevity study	Netherlands Twin Register	Total of all cohorts
Abbreviation	RS	LL	LLS	NTR	
Individuals	693	727	646	874	2940
Sex, male|female	397|296	421|306	340|306	577|297	1735|1205
Smokers*, current|never	75|231	107|337	75|184	155|500	412|1252
Age [years], mean + 95% CI	67.6 (67.1–68.0)	45.4 (44.4–46.3)	58.8 (58.3–59.3)	38.3 (37.3–39.3)	51.4 (50.9–52.0)
LDL [mmol/L], mean + 95% CI	3.32 (3.26–3.39)	3.19 (3.12–3.25)	3.36 (3.29–3.43)	2.90 (2.84–2.96)	3.17 (3.14–3.2)

Note, the age differences between the cohorts; participants of the NTR were on average 29 years younger than the participants of the RS.
CI confidence interval.
*Former smokers were excluded in this study.

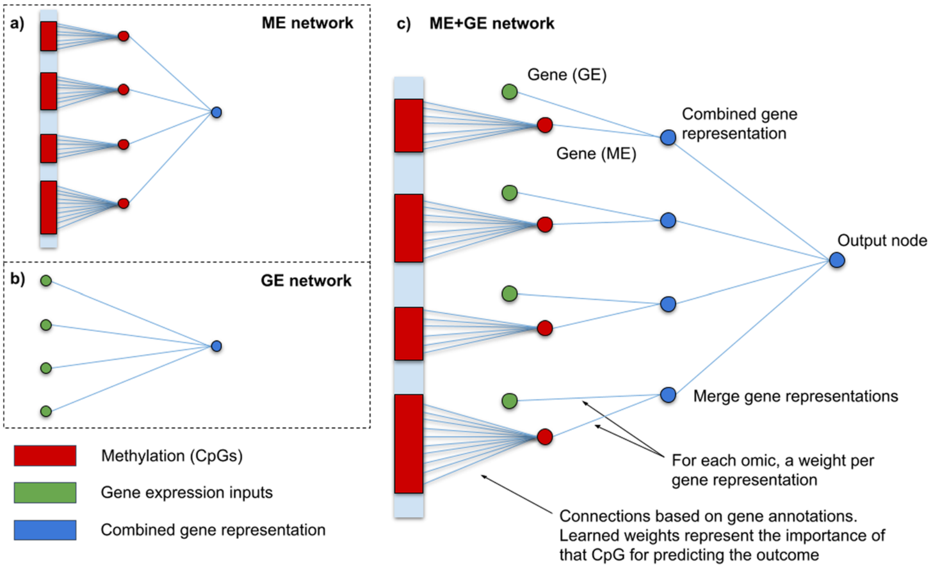
a)
ME network
b)
GE network
🟥 Methylation (CpGs)
🟩 Gene expression inputs
🟦 Combined gene representation
c) ME+GE network
Gene (GE)
Gene (ME)
Combined gene representation
Output node
Merge gene representations
For each omic, a weight per gene representation
Connections based on gene annotations. Learned weights represent the importance of that CpG for predicting the outcome

Fig. 1 | Schematic overview of the main neural network architectures used in this study. First, the methylation data is annotated using GREAT. The intersection of this gene set and genes in the gene expression data (10,404 genes) is used to construct all neural networks. In the ME network (**a**), DNA methylation data (CpGs) are grouped and connected using these gene annotations. The resulting gene nodes are directly connected to the output node. Combining the ME network and the GE network (**b**), results in the ME + GE network (**c**). In the ME + GE network, each gene has a node per omic and a combined gene representation where information from both omics is merged.

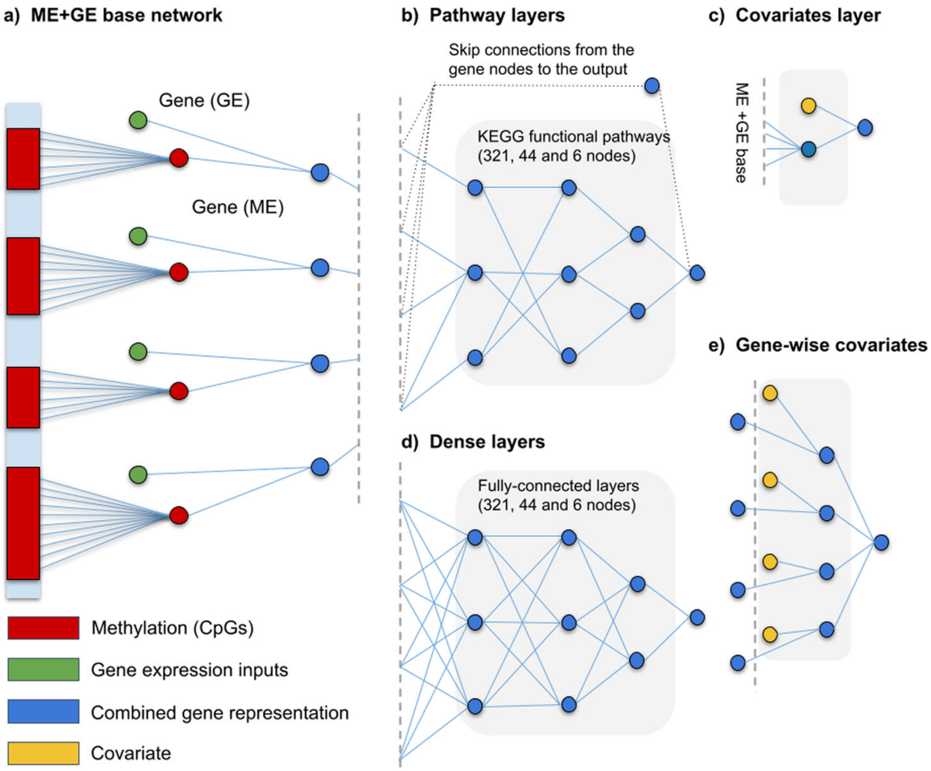

Fig. 2 | Overview of other visible neural network architectures used in the additional analyses. The ME + GE network base (**a**) serves as a basis for all networks displayed. **b** Pathway layers are added to this base by grouping genes and connecting them to their corresponding KEGG pathway. The KEGG functional pathways consist of a three-layer hierarchical structure with 321, 44 and 6 nodes each. To compare this network to an equivalent regular dense network (**d**) fully connected layers with the same dimensions were attached to the ME + GE base. Covariates were integrated in the network by adding a single layer with the covariates (**c**) to the base, or by adding the covariates to each combined gene representation in the ME + GE network (**e**).

and LDL levels, ReLu activation functions with output range $[0, \infty)$ were used for all layers. For a better initialization of the network, the bias of the last neuron was set to the mean value of the predicted outcome in the training set.

Deeper neural networks

For more complex modelling of the interactions between expression, methylation and phenotypes, we also evaluated deeper neural networks. First, using KEGG's functional pathways[32,33] as prior knowledge, three hierarchical pathway layers were created (Fig. 2b). The first layer groups genes into 321 functional pathways such as: insulin secretion, thyroid hormone synthesis and PPAR signalling pathway. The aforementioned pathways are all part of the endocrine system group which, in turn, is a subgroup of organismal systems. The mid and global-level pathway layers were created adopting this hierarchical structure, consisting of 44 and 6 groups, respectively. Each pathway is represented by its own neuron resulting in three layers with 321, 44 and 6 nodes each. Not all genes were annotated by the KEGG functional pathway annotations, 5591 genes did not receive a functional pathway annotation. To ensure connectivity to the output for all genes, connections that skip the pathway layers (skip connections) were added from each gene to the output node (see Supplementary Fig. 2 for the distribution).

Additionally, a deeper network was constructed without any additional prior biological knowledge to compare with the KEGG pathway network (Fig. 2d). Similarly, the ME + GE network served as a basis for this network and three densely connected layers, 321, 44 and 6 nodes each, were added

Table 2 | Performance for cohort-wise cross-validation, mean with 95% CI over ten runs

Phenotype	Network type	AUC validation cohorts				AUC test cohort				Mean test AUC over all cohorts
		RS*	LL*	LLS*	NTR*	RS	LL	LLS	NTR	
Smoking	ME + GE	0.92 (0.90–0.94)	0.93 (0.92–0.93)	0.91 (0.90–0.92)	0.93 (0.91–0.94)	0.98 (0.98–0.98)	0.92 (0.92–0.93)	0.95 (0.95–0.96)	0.91 (0.89–0.92)	0.95 (0.90–1.00)
	ME	0.93 (0.92–0.94)	0.94 (0.94–0.95)	0.95 (0.94–0.95)	0.93 (0.93–0.94)	0.97 (0.97–0.98)	0.94 (0.93–0.94)	0.96 (0.95–0.96)	0.95 (0.95–0.96)	0.95 (0.93–0.98)
	GE	0.83 (0.83–0.84)	0.82 (0.82–0.82)	0.83 (0.83–0.83)	0.87 (0.86–0.87)	0.87 (0.87–0.88)	0.85 (0.85–0.85)	0.87 (0.87–0.88)	0.80 (0.80–0.80)	0.85 (0.80–0.90)
Phenotype	**Network type**	**RMSE validation cohorts**				**RMSE test cohorts**				**Mean test RSME over all cohorts**
		RS*	**LL***	**LLS***	**NTR***	**RS**	**LL**	**LLS**	**NTR**	
Age	ME + GE	3.85 (3.72–3.98)	4.13 (4.03–4.23)	3.88 (3.78–3.98)	4.12 (3.99–4.25)	5.44 (5.23–5.65)	6.04 (5.76–6.31)	4.20 (4.11–4.29)	6.33 (5.87–6.8)	5.16 (3.97–6.35)
	ME	14.48 (13.04–15.92)	8.23 (7.26–9.19)	18.06 (15.01–21.11)	11.28 (9.90–12.66)	15.69 (12.37–19.0)	8.14 (6.50–9.79)	13.17 (9.13–17.2)	22.66 (19.67–25.65)	15.00 (4.31–25.69)
	GE	6.79 (6.77–6.81)	6.52 (6.51–6.53)	6.40 (6.38–6.41)	6.16 (6.14–6.17)	9.03 (8.80–9.26)	12.23 (12.04–12.41)	6.87 (6.80–6.95)	17.77 (17.51–18.03)	11.45 (4.21–18.68)
LDL	ME + GE	0.88 (0.87–0.88)	0.88 (0.88–0.88)	0.85 (0.84–0.86)	0.92 (0.91–0.92)	0.92 (0.92–0.93)	0.89 (0.89–0.90)	0.93 (0.93–0.94)	0.93 (0.91–0.95)	0.93 (0.88–0.99)
	ME	1.18 (1.12–1.25)	1.14 (1.06–1.22)	1.10 (1.02–1.18)	1.12 (1.08–1.16)	1.29 (1.15–1.44)	1.17 (1.07–1.26)	1.10 (1.01–1.19)	1.23 (1.15–1.30)	1.27 (0.8–1.75)
	GE	0.88 (0.88–0.88)	0.88 (0.88–0.89)	0.85 (0.85–0.86)	0.90 (0.90–0.91)	0.94 (0.93–0.94)	0.90 (0.89–0.90)	0.93 (0.93–0.93)	1.02 (0.98–1.07)	0.98 (0.79–1.17)

The area under the curve is reported for the classification task (smoking status prediction) and the RMSE for the regression tasks, predicting age and LDL levels.
See Supplementary Table 1 for the performance of out-of-the-box sci-kit-learn implementations for each omic and Supplementary Table 2 for the performance of a baseline neural network.
ME methylation, *GE* gene expression, *ME + GE* both methylation and gene expression as an input for the neural network, *RS* Rotterdam study, *LL* LifeLines, *NTR* Netherlands Twin Register, *LLS* Leiden Longevity Study.
*The name of the test cohort is used to denote the fold. Thus, for the first fold, RS, was used for testing and LL + LLS + NTR were used for training and validation (75% training and 25% validation).

between the gene layer and the output node. The resulting network has thus the same number of neurons as the KEGG pathway network but has fully connected layers instead of layers based on KEGG pathway information.

Cohort-wise cross-validation

An overview of the performance for each cohort and for the three different architectures can be found in Table 2. It shows the mean predictive performance and standard deviation for each fold for ten networks trained with the same hyperparameters but with different random seeds. The corresponding hyperparameters, chosen on the best performance in the validation set, can be found in Supplementary Table 3.

Predicting smoking status. Both gene expression and methylation were highly predictive of smoking status in all folds. The best performance was achieved by the ME + GE network, thus with both methylation and gene expression input, in the fold with the RS as the test cohort (all other cohorts were used for training and validation). In this fold, the network achieved a near-perfect classification with the area under the receiver operating curve (AUC) of 0.98 (95% CI: 0.98–0.98). Overall folds, the ME networks and ME + GE networks performed best with a mean AUC of 0.95 (95% CI: 0.93–0.98) and 0.95 (95% CI: 0.90–1.00) respectively. The GE network, based solely on gene expression input, performed substantially worse with a mean AUC of 0. 0.85 (95% CI: 0.80–0.90). Surprisingly, the mean test performance overall folds for the ME + GE network was lower for deeper networks with three fully connected layers, achieving a mean AUC of 0.91 (95% CI: 0.85–0.96) (see Supplementary Table 4). In general, each fold obtained good predictive performance for predicting smoking status, the GE network in the fold with NTR as the test cohort achieved the worst overall predictive performance with a mean AUC of 0.80 (95% CI: 0.80–0.80). The mean test performance of the multi-omic visible neural network was better than the multi-omic baseline network which achieved an AUC of 0.92 (95% CI: 0.84–1.00), mostly due to poorer generalization to the NTR test cohort (see Supplementary Table 2).

The ME + GE networks exhibit a stable performance, with small CIs for the area under the curve and standard deviations not exceeding 0.03. However, there may be significant variations in the underlying weights due to stochastic processes used for network initialization and training, resulting in different starting points and optimization paths for all weights across runs. As the weights within a neural network operate relative to each other and cannot be directly compared between networks, we compared the relative contribution of each gene instead. Figure 3 demonstrates that certain genes are consistently utilized by the network to differentiate between current smokers and non-smokers across all folds, although there can be notable differences in the percentage of total weight each gene holds. In each fold, *GPR15* is the most or second most predictive gene for smoking status, its signal is mainly driven by gene expression as visualized in Fig. 3. Specifically, 79.8 ± 33.3% (mean and standard deviation over all folds) of the weights that drive the signal for this gene are from the gene expression input. The next gene, *AHRR*, is important for prediction in three out of four cohorts. This signal is driven by both gene expression (44.3%) as well as methylation (55.6%). Other consistently highly predictive genes (i.e. genes with a weight contribution higher than 1% in three out of four cohorts) are *SEMA6B*, *PID1*, *LRRN3*, *P2RY6*, *CDKN1C*, *CLEC10A* and *KCNQ1*. (See Supplementary Table 5 for more details). All these consistently highly predictive genes were found before in association studies for smoking in gene expression and methylation[34–36]. A graphical overview of important pathways for smoking prediction can be found in Supplementary Fig. 3.

To investigate the interplay between the omic types and to find omic-specific information, two additional analyses were conducted where either gene expression or methylation gene representations were penalized (see Supplementary Figs. 4–6). Without penalization, the weights for gene expression and methylation were nearly equally divided after training. Weights connected to gene expression input occupied 51.6 ± 1.3% of the weights over all the ME + GE networks, with the remainder used for

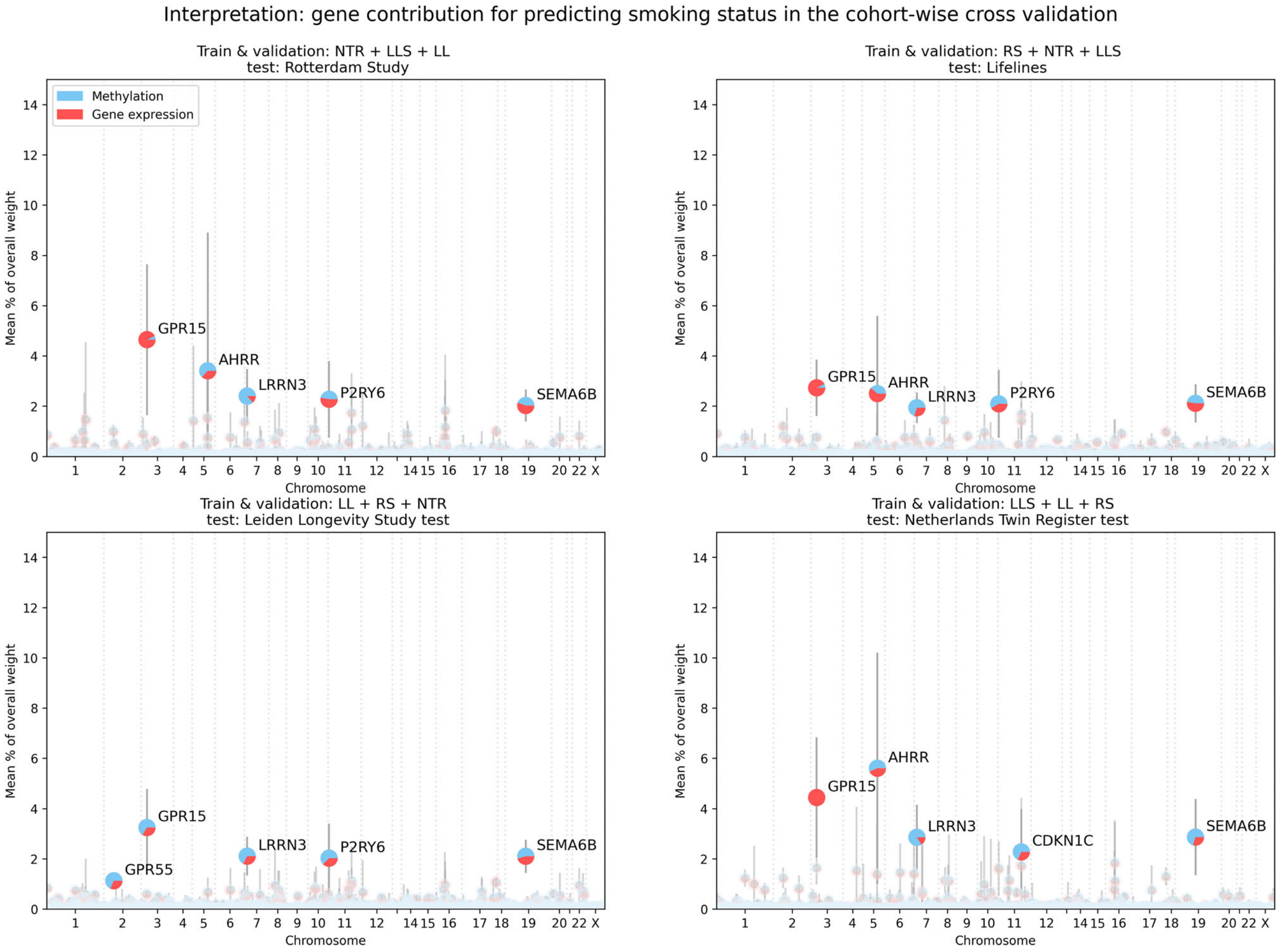

Fig. 3 | Overview of the important genes for predicting smoking status for each fold. For each gene the mean contribution as a percentage of the total weight assigned to each gene is shown together with a bar indicating the standard deviation over ten runs with different weight initializations. For each gene, the pie chart shows the mean contribution of methylation and gene expression. Each quadrant shows the variation of different runs for a single test cohort while each quadrant shows the variation across different cohort splits.

methylation. In these experiments, we found that an omic-specific L1 penalty of 0.01 for gene expression reduced the contribution of the weights associated with gene expression to 0.69 ± 1.16% while a similar threshold reduced the weights associated with methylation to 2.56 ± 1.73%. A more severe omic-specific L1 threshold of 0.001 for methylation reduced the use of methylation in the top genes nearly completely, only for *LRNN3* methylation input is still used in the second and third fold with (respectively ~41% and 16% of the weights for this gene). However, with the same threshold gene expression inputs are responsible for 15% of the weights for AHRR in the first fold, nearly 29% of the *GPR15* weights in the second fold and 39% of *RER1* in the fourth fold (see Supplementary Figs. 3 and 4). Interestingly, the importance of AHRR was severely impacted by the methylation penalty, its gene expression was barely used to predict smoking status when methylation was penalized.

Predicting age. Networks trained with both methylation and gene expression data (ME + GE) achieved a mean error of 5.16 (95% CI: 3.97–6.35) years overall folds for age prediction (see Table 2). Between folds, there were large differences in performance for predicting age. Most notably, networks did not generalize well in folds that have either the RS (ranging between 52 years and 80 years) or the LLS (ranging between 30 years and 79 years) as test cohorts, the two cohorts with the oldest population. For these cohorts, the explained variance in the test set was substantially lower than in the validation set: RS test 0.40 (95% CI: 0.37–0.43), 0.94 (95% CI: 0.93–0.94) validation, LLS test 0.95 (95% CI: 0.95–0.95), 0.61 (95% CI: 0.60–0.63) validation. Aside from being older, these cohorts also have a smaller spread in age distribution compared to the two other cohorts (See Fig. 4a and Supplementary Fig. 7). The NTR cohort ranges between roughly 18-years-old and 80-years-old while individuals from the LL cohort were between 18-years-old and 81-years-old.

Differences between omics and network types were also larger for age prediction than for smoking status prediction. The ME + GE network consistently outperformed the single-omic networks with substantial margins: the mean explained variance over all folds was 0.72 (95% CI: 0.36–1.07) for the ME + GE network, 0.30 (95% CI: −0.26 to 0.86) for gene expression, while the ME networks did not find any predictive pattern that translated to the test cohort. Training and validation performance was generally poor for the ME network, and although the GE network obtained good validation performance in terms of explained variance for each fold, this did not translate in folds with the RS and LLS as test cohorts. The baseline network performed substantially better than the visible neural network for the methylation data with an error of 4.10 years (95% CI: 2.73–5.46) versus that of 15.00 (95% CI: 4.31–25.69) for the ME network. However, the performance of the ME + GE network was slightly better (root mean squared error (RMSE) of 5.16 years, 95% CI: 3.97–6.35) than the multi-omic baseline network (RMSE of 5.90 years, 95% CI: 4.59–7.21).

Interpretation of the ME + GE network revealed that many genes had a small contribution to age prediction (see Supplementary Fig. 7). The neural network found a more multifactorial solution for age prediction than

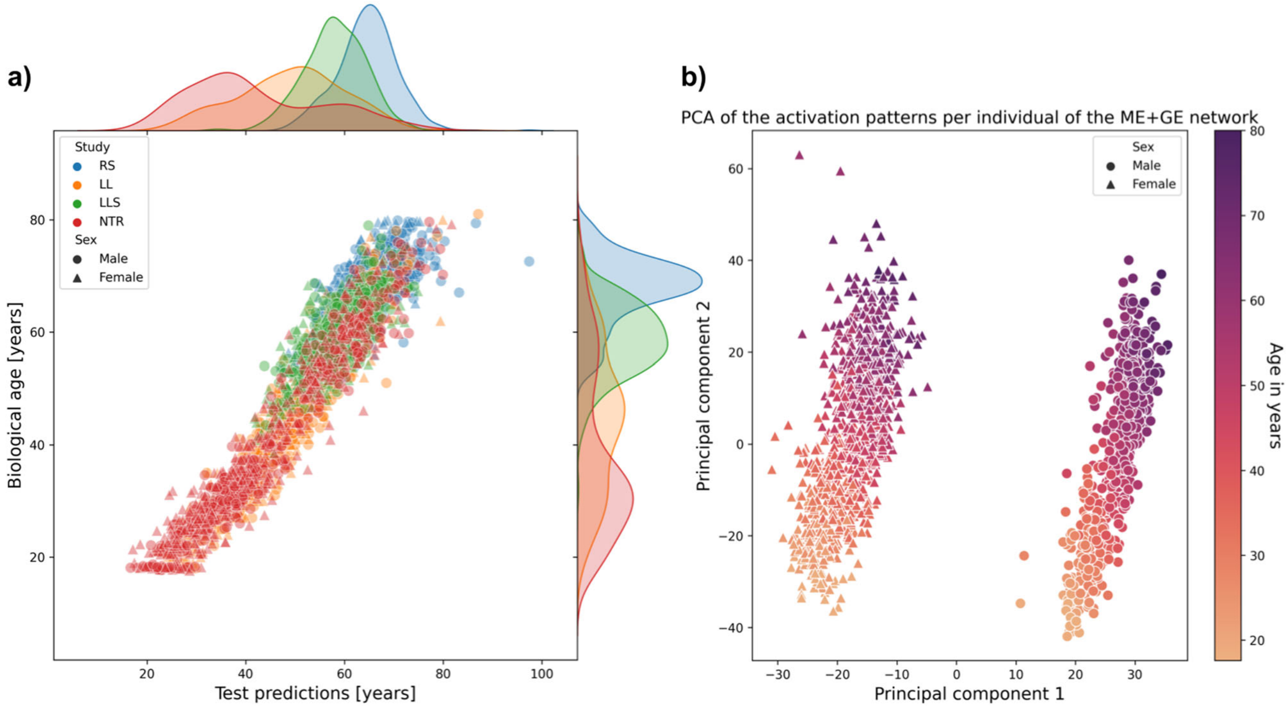

Fig. 4 | Test predictions and activation analysis. **a** Test predictions for the ME + GE network for all folds (each cohort) with corresponding distributions (see Supplementary Fig. 12 for the GE and ME networks). **b** Activation of the ME + GE trained for age prediction. A principal component analysis clearly shows two distinct activation patterns corresponding to the different sexes. Principal component 1 is related to the sex differences, and principal component 2 to the age of the participants.

for smoking, the most important gene over all folds only occupied 0.68% of all weights for predicting age compared to 3.76% for smoking. The most predictive genes with a weight contribution higher than 0.30% of the total weight in three out of the four folds were *COL11A2*, *AFAP1*, *OTUD7A*, *PTPRN2, ADARB2* and *CD34* (Supplementary Table 6). These most predictive genes were not part of Hannum et al. and Horvath's epigenetic clocks[25,26]. The first principal components of the activation patterns of the ME + GE network revealed distinct activation patterns for the different sexes with a gradient in each cluster (see Fig. 4b). However, there is no significant difference in the absolute error between the sexes (Wilcoxon rank-sum, *p*-value of 0.98, Supplementary Figs. 8 and 9), the first principal component clusters perfectly for males and females while the second principal component is strongly related with age. Additional experiments showed that the clustering of the sexes is mainly driven by genes on the X chromosome (see Supplementary Fig. 10). Including sex as a covariate in the last layer of the model did not improve the performance of the model (mean RMSE overall folds of 7.31 [95% CI: 2.89–11.73]). Including sex information to each gene also did not lead to a better performance (mean RMSE overall folds of 10.64 [95% CI: 4.12–17.15]). However, inspecting the weights between the covariate and the genes for the best-performing network revealed strong sex-specific weights for, among others: *KLF13*, *ANO9* and *HECA* (for more details see Supplementary Fig. 11). For these genes the network needed strong weights to model sex-specific effects for age prediction.

After applying an omic-specific L1 penalty for methylation of 0.01, the network only used the methylation input for gene *NEDD1* in the second fold with nearly 33% of the weight contribution for this gene from methylation, while in the third fold *MAD1L1* had a methylation contribution of 23% (see Supplementary Fig. 14). With the same threshold for penalizing gene expression inputs, *DNAJB6* had the largest gene expression use with 31% of the weight for this gene assigned to gene expression input (Supplementary Fig. 15). The deeper neural network architectures quickly overfitted, reaching high performance on the training data which did not generalize to the validation and test set. These networks were consistently outperformed by the ME + GE network (Supplementary Table 7). The best-performing network built with KEGG pathway information had the pathway: "environmental information processing" as the most predictive global pathway because of the high contributions of membrane transport (ABC transporters), signal transduction, and signalling molecules and interaction (see Supplementary Fig. 16).

Predicting low-density lipoprotein levels. ME + GE and GE networks explained up to 17% of the phenotypic variance in the validation set but these networks only generalized in the second fold to an explained variance of 0.07 (95% CI: 0.05–0.08) for the ME + GE network and 0.04 (95% CI: 0.04–0.05) for the GE network in the LL test cohort (see Supplementary Table 8). In this fold, the largest gene, *FAM53A* only occupied 0.052% of the total weight (Supplementary Fig. 17). The weights for all genes in the ME + GE network were small and evenly spread, indicating that the network did not find individual genes with a strong effect for predicting LDL levels. Additional layers, be it pathways or densely connected layers, did not improve predictive performance.

Discussion

In this paper, we evaluated the performance, interpretability and stability of visible neural networks for single and multi-omics data. Interpretability was achieved by embedding prior biological knowledge such as gene and pathway annotations in the neural network architecture. We applied these models to predict smoking status, age and low-density lipoprotein levels in a cohort-wise cross-validation using methylation and gene expression data. For smoking, single omic networks and multi-omic networks performed consistently high across all cohorts for predicting smoking status. Predicting smoking status is a relatively simple task, since smoking is a powerful inducer of DNA methylation and gene expression alterations[37]. This is also reflected by the mean AUC of 0.95 overall folds that the ME + GE and ME networks achieved. It is slightly better than the performance of Maas et al. who reported an AUC of 0.90 in an external dataset with a weighted combination of just thirteen CpGs. Inspection of the weights of the ME +

GE network revealed *GPR15*, *AHRR* and *LRRN3* as the most important genes for prediction, which is consistent with existing literature[28,29,37,38]. In the ME + GE network, the contribution of both omics types was nearly equal (in terms of weights), while the gene expression-based network by itself was less predictive than the methylation-based networks. Applying an omic-specific penalty for methylation input showed that the ME + GE network needed some methylation input to achieve similar performance with expression information.

For predicting age, the ME + GE network outperformed the single ME or GE networks. The performance of this network in the test cohorts varied between an R^2 of 0.40 (95% CI: 0.37–0.43) and 0.91 (95% CI: 0.90–0.92). This difference in performance is probably caused by the different distributions in age in the cohorts, depending on the cohorts in the training set the networks are shown fewer examples of older or younger individuals. Similar effects were also seen in traditional methods[9]. Based on the predictive performance shown in Table 2 one could conclude that for age prediction, usage of the two omics types increased stability and performance for these types of neural networks compared to the single omic networks. Additionally, we have evaluated whether the network used sex information in the decision process for age prediction. The first principal component of the activations of the neural network showed a perfect separation between the sexes, mostly caused by genes on the X chromosome, while the second principal component had a clear correlation with age. Owing to the shallowness of the networks, the activation pattern will therefore closely resemble the underlying data, especially if it has some relation with the outcome. For deeper networks, a PCA on the activation may reveal more detailed information (such as different patient subtypes or mediating factors) since the network applies more complex transformations to the data. The inclusion of genes on the X-chromosome allowed the network thus to separate between the sexes but it did not have the capacity to model different effects independently from the input for each sex. To help the model to find a sex-specific effect we modified the network with sex information as an extra input to each gene node. After, training the network found the strongest sex-specific gene effects for *KL13*, *ANO9* and *HECA*. However, this addition to the network architecture did not improve performance.

An earlier EWAS in only the RS did not find significant associations between DNA methylation in blood and LDL cholesterol[39]. Another EWAS using BIOS data found only three significant associations, demonstrating that there is a very weak relation between methylation and LDL measurements from blood which makes the prediction task more complex[40]. The neural networks did find patterns in the training set that were also found in the validation set (up to an R^2 of 0.17 [95% CI: 0.16–0.18]) but this pattern did not generalize to the test cohorts with the exception of the Lifeline cohort. In this cohort the method achieved an R^2 of 0.07 (95% CI: 0.05–0.08) in the test set, substantially lower than the performance of the validation set 0.13 (95% CI: 0.12–0.14). suggesting that the model had trouble generalizing to data from an unseen cohort. Overall, the low prediction performance might also indicate that the studied omic data (gene expression and methylation from blood) might not contain enough information to accurately predict LDL levels.

In general, we found that including multiple omics inputs in the network improved performance. These multi-omic networks had a more stable performance and generalized better to the test cohorts. Surprisingly, deeper networks did not lead to better performance. Generally, one would expect deeper networks to perform better since they can model more complex interactions. Thus, it is possible that the optimal hyperparameter values for deeper networks lie outside the considered hyperparameter range or that more training examples are required to train these deeper networks. Interpreting the ME + GE networks revealed well-known genes such as *GPR15* and *AHRR* for smoking that validate the results. However, we also saw that the interpretation can vary between different random initializations, and it is therefore recommended to train networks with different random seeds for a more complete overview of important predictors. As for all prediction models, it is important to consider that predictive genes and pathways found are not necessarily causal genes and pathways as effects can be mediated. However, these genes and pathways do provide insight into the decision process of the neural network and may be used in follow-up.

For good interpretation, proper regularization is important as it forces the network to use the most predictive input features. For example, an L1 penalty on the weights will force the network to learn sparse weights, resulting in a less complex model. In the absence of an L1 penalty on the weights, the network has more freedom to choose its weights. This does not necessarily harm performance but may harm interpretability. In this work, we use the L1 penalty to regularize the network, but other regularization methods could have been chosen. For example dropout[41], this method drives the network to find a more stable solution by deactivating random sets of neurons during training. Another important factor for interpretation in visible neural networks is the quality of the prior knowledge used in creation. In this study, the annotations for the CpG sites were based on genomic distance. Potential improvements could come from using tissue-specific and functional annotation databases such as ENCODE[42].

We believe that visible neural networks have great potential for genomic applications, especially for multi-omics integration. These interpretable neural networks can combine multi-omics data elegantly in a single prediction model and provide the importance of each gene, pathway and omic input for prediction. Additionally, we found that using multi-omic networks generally improved performance, stability and generalizability compared to interpretable single omic networks.

Methods

BIOS data

Transcriptome and methylome data from the four largest cohorts in the Biobank-based Integrative Omics Study (BIOS) consortium—LL, LLS, NTR and RS—were used to train the models in a cohort-wise cross-validation setting. In the BIOS consortium, all cohorts adhered to the same procedure in gathering and processing data. For each participant, the transcriptome and the methylome were measured in whole blood samples taken from the same visit. DNA methylation was profiled according to the manufacturer's protocol using the Infinium Illumina HumanMethylation 450 k arrays, while blood was first depleted from globin transcripts for RNA sequencing. A detailed description of all data generation and preprocessing steps for the RNA sequencing and DNA methylation data can be found in refs. 43,44. Using the BBMRI-NL's Integrative Omics analysis platform[45], all individuals that had both RNA-seq and methylation data (β-value) available were selected, resulting in a dataset with 2940 individuals. Y-chromosomal data was excluded, and X-chromosomal and autosomal measurements were included. Finally, RNA-seq expression data was filtered using an expression inclusion criterion of one count per million on average across all samples or higher[9].

Training and evaluation

The neural networks were evaluated in a cohort-wise cross-validation setup (Supplementary Fig. 1) to assess the generalizability of the models across cohorts. In each fold, one cohort is held out as a test set, while the three other cohorts were used for training and validation (leave-one-out method). From these three cohorts, 75% of the individuals were randomly selected for the training set while the remaining 25% was used in the validation set to tune the hyperparameters. For all methods, the same combinations of hyperparameters were tuned on the validation set. Combinations included learning rates of [0.01, 0.001, 0.005, 0.0001] and L1 penalty on the weights of the combined gene and/or methylation gene layer of [0.01, 0.001, 0.0001]. A higher L1 penalty increases the cost for the network to include more contributors to predict the outcome (see Supplementary Note 1). The L1 penalty thus enforces sparsity over the weights, so that most inputs get assigned a (near) zero weight while important inputs still get assigned a high weight. This L1 regularization on the weights helps prevent overfitting and increases interpretability.

The mean squared error (MSE) was used as a loss function to optimize for regression tasks. For classification tasks, weighted binary cross-entropy, with a weighted inverse to the ratio of the class imbalance, was used as a loss function. The loss function quantifies the difference between the current outcome and the true label and is optimized during training. The performance of the resulting network is evaluated using the AUC for classification tasks, and the RMSE and explained variance (sci-kit learn, explained_variance_score) for regression tasks. For each fold the hyperparameters of the best-performing model in the validation set were selected to evaluate the test cohort. Since neural networks use stochastic processes that can influence the outcome, we trained the network with the best hyperparameters ten times with a different random seed to investigate its stability.

Interpretation

The contribution scores for interpreting the importance of each input and node are calculated using the absolute weights in the network. For each path between the output and the inputs, the product of all absolute weights along the path is computed. The total contribution of an input is retrieved by summing the products of all paths that converge on it. The proportion of each input's contribution relative to the total contribution of all inputs is the final score. To determine the contribution of nodes in hidden (intermediate) layers, this proportion is multiplied by all the weights along the path from the input to the hidden node. Again, the final contribution percentage is calculated as the contribution of the hidden node divided by the total contribution of all hidden nodes in the same layer.

Baselines

To compare the performance of visible neural networks to more traditional neural networks a baseline network with a locally connected layer followed by two dense layers was trained (see Supplementary Note 2). The locally connected 1D layer has been successfully used in various applications in genomics[46–48] and can be seen as a hybrid between convolutional and fully connected layers[49]. For the multi-omics data, the methylation and gene expression data were first concatenated before feeding to the network. Hyperparameters, loss functions and activation functions were optimized and configured in the same manner as for the visible neural networks.

Additional analyses

Neural networks are flexible methods and with the inclusion of prior biological knowledge different architectures can be explored to provide more insight into the interaction between omics types, the contribution of covariates and the gene-specific contribution of covariates. For each of these analyses, we made small changes to the ME + GE networks.

Omic-specific information. Gene expression and methylation data contain redundant information with respect to each other. However, not all information that is present in the one may be present in the other data type. To evaluate the independent contribution of each omics to the prediction we add a L1 penalty for one omics type in the model. This introduces a trade-off for the neural network: the gain in performance for including information on the penalized omic (i.e. RNA expression of a single gene or the methylation representation of a single gene) must outweigh its penalty. If the model uses only the non-penalized omic type without loss of prediction performance, it is likely that there was no omic-specific information. However, if the model still decides to use parts of the penalized omics data, this information is most likely unique to the penalized omic type and was therefore required for prediction.

Covariate-gene interaction. Including covariates in the model, for example, sex and age for smoking can improve performance and interpretation. Commonly, the covariates are included as an extra layer at the end (Fig. 2c). However, by adding a covariate for each gene, more specific information on how a covariate affects a single gene can be obtained (see Fig. 2e). For each phenotype we tested both, a model with covariates in the last layer and a model with covariates for each gene.

Subtyping with activation patterns. In contrast to fully connected neural networks, the visible neural network architectures used in this study are constructed based on prior biological knowledge that can be interpreted by inspecting the weights of the incoming and outgoing connections. The strength of the weights (e.g. between CpGs and genes, expression and genes, genes and pathways), all express the importance of these biological elements for the predicted outcome. The weights of a neural network are a result of an optimization over the population it was trained on and are thus a result of the population characteristics of the training set. However, neural networks may learn different patterns for the same outcome. By inspecting the weights general information is learned about the importance of each element but this does not show differences between groups or individuals. Based on differences between individuals, some neurons can activate for a certain group of individuals, while others do not for others. To gain an overview of the different patterns that are learned by the network we applied principal component analysis[50] (PCA) over all the activations for all (gene-level) nodes for each individual (see Supplementary Note 3). In this PCA, individual-level differences may cluster and provide groups of individuals for which the neural network used a similar activation pattern.

Data availability

BIOS datasets are available from the European Genome-Phenome Archive by accession number EGAS00001001077 (https://ega-archive.org/studies/EGAS00001001077). Alternative options to access the data are available through the BIOS website; https://www.bbmri.nl/acquisition-use-analyze/bios/. All trained networks are available on request.

Code availability

The code is available on GitHub: https://github.com/ArnovanHilten/GenNet-multi-omic.

Received: 14 December 2023; Accepted: 12 July 2024;
Published online: 02 August 2024

References

1. Li, M. et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res.* **47**, D983–D988 (2019).
2. Mikeska, T. & Craig, J. M. DNA methylation biomarkers: cancer and beyond. *Genes* **5**, 821–864 (2014).
3. Taryma-Leśniak, O., Sokolowska, K. E. & Wojdacz, T. K. Current status of development of methylation biomarkers for in vitro diagnostic IVD applications. *Clin Epigenetics* **12**, 100 (2020).
4. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 1–15 (2017).
5. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **16**, 85–97 (2015).
6. Malik, V., Kalakoti, Y. & Sundar, D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genom.* **22**, 1–11 (2021).
7. Singh, A. et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).
8. Bersanelli, M. et al. Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinform.* **17**, 15 (2016).
9. Van Rooij, J. et al. Evaluation of commonly used analysis strategies for epigenome- and transcriptome-wide association studies through replication of large-scale population studies. *Genome Biol.* **20**, 1–15 (2019).
10. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
11. Young, T., Hazarika, D., Poria, S. & Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **13**, 55–75 (2018).
12. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

13. Michael, K. Y. et al. Visible machine learning for biomedicine. *Cell* **173**, 1562–1565 (2018).
14. Bourgeais, V., Zehraoui, F., Ben Hamdoune, M. & Hanczar, B. Deep GONet: self-explainable deep neural network based on gene ontology for phenotype prediction from gene expression data. *BMC Bioinform.* **22**, 1–24 (2021).
15. Ma, J. et al. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* **15**, 290–298 (2018).
16. Wang, D. et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
17. Moon, S. & Lee, H. MOMA: a multi-task attention learning algorithm for multi-omics data interpretation and classification. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btac080 (2022).
18. Fortelny, N. & Bock, C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome Biol.* **21**, 190 (2020).
19. Esser-Skala, W. & Fortelny, N. Reliable interpretability of biology-inspired deep neural networks. *Npj Syst. Biol. Appl.* **9**, 50 (2023).
20. van Hilten, A. et al. GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Commun. Biol.* **4**, 1–9 (2021).
21. Elmarakeby, H. A. et al. Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348–352 (2021).
22. Hao, J., Kim, Y., Kim, T. K. & Kang, M. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC Bioinform.* **19**, 1–13 (2018).
23. Kuenzi, B. M. et al. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* **38**, 672–684.e6 (2020).
24. Huang, X. et al. ParsVNN: parsimony visible neural networks for uncovering cancer-specific and drug-sensitive genes and pathways. *NAR Genom. Bioinform.* **3**, 1–12 (2021).
25. Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
26. Horvath, S. DNA Methylation Age of Human Tissues and Cell Types. *Genome Biol* http://genomebiology.com//14/10/R115 (2013).
27. Bell, C. G. et al. DNA methylation aging clocks: challenges and recommendations. *Genome Biol*. **20**, 249 (2019).
28. Langdon, R. J., Yousefi, P., Relton, C. L. & Suderman, M. J. Epigenetic modelling of former, current and never smokers. *Clin. Epigenetics* **13**, 206 (2021).
29. Maas, S. C. E. et al. Validated inference of smoking habits from blood with a finite DNA methylation marker set. *Eur. J. Epidemiol.* **34**, 1055–1074 (2019).
30. Sayols-baixeras, S., Irvin, M. R., Elosua, R., Arnett, D. K. & Aslibekyan, S. W. Epigenetics of lipid phenotypes. *Curr. Cardiovasc. Risk Rep*. https://doi.org/10.1007/s12170-016-0513-6 (2016).
31. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
32. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
33. Kamburov, A., Wierling, C., Lehrach, H. & Herwig, R. ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic Acids Res.* **37**, D623–D628 (2009).
34. Ligthart, S. et al. Tobacco smoking is associated with DNA methylation of diabetes susceptibility genes. *Diabetologia* https://doi.org/10.1007/s00125-016-3872-0 (2016).
35. Baiju, N., Sandanger, T. M., Sætrom, P. & Nøst, T. H. Gene expression in blood reflects smoking exposure among cancer—free women in the Norwegian Women and Cancer (NOWAC) postgenome cohort. *Sci. Rep*. https://doi.org/10.1038/s41598-020-80158-8 (2021).
36. Vink, J. M. et al. Differential gene expression patterns between smokers and non-smokers: cause or consequence? Addict Biol. https://doi.org/10.1111/adb.12322 (2015).
37. Silva, C. P. & Kamens, H. M. Cigarette smoke-induced alterations in blood: a review of research on DNA methylation and gene expression. *Exp. Clin. Psychopharmacol.* **29**, 116–135 (2021).
38. Kõks, S. & Kõks, G. Activation of GPR15 and its involvement in the biological effects of smoking. *Exp. Biol. Med.* **242**, 1207–1212 (2017).
39. Braun, K. V. E. et al. Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. *Clin. Epigenetics* **9**, 1–11 (2017).
40. Dekkers, K. F. et al. Blood lipids influence DNA methylation in circulating cells. *Genome Biol*. https://doi.org/10.1186/s13059-016-1000-6 (2016).
41. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
42. Davis, C. A. et al. The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
43. Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
44. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
45. van Iterson, M. & Cats, D. BBMRIomics: R utilities for BBMRI omics data analysis. R package version 3.4.2. (2020).
46. Sigurdsson, A. I. et al. Deep integrative models for large-scale human genomics. *Nucleic Acids Res.* **51**, e67 (2023).
47. Lee, H.-J. et al. GpNet: genomic prediction network using locally connected layers in korean native cattle. Preprint at https://doi.org/10.21203/rs.3.rs-622476/v1 (2021).
48. Kassani, P. H., Lu, F., Le Guen, Y., Belloy, M. E. & He, Z. Deep neural networks with controlled variable selection for the identification of putative causal genetic variants. *Nat. Mach. Intell.* **4**, 761–771 (2022).
49. Ngiam, J. et al. Tiled convolutional neural networks. In *Proc of the Advances in Neural Information Processing Systems* (NIPS, 2010).
50. Jolliffe, I. T. *Principal Component Analysis for Special Types of Data* https://doi.org/10.1007/978-1-4757-1904-8_11 (1986).

Acknowledgements

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. Samples used in this study were contributed by the LL, the LLS, the NTR and the RS. We thank all BIOS consortium members and participants of these cohorts for their contributions to this study. This work was funded by the Dutch Technology Foundation (STW) through the 2005 Simon Steven Meester grant 2015 to W.J. Niessen. G.V. Roshchupkin was supported by the ZonMw Veni grant (Veni, 1936320). The RS is supported by the Netherlands Organization for Scientific Research (NWO, 91203014, 175.010.2005.011, 91103012). This research was financially supported by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO, numbers 184.021.007 and 184.033.111).

Author contributions

A.H. and J.R. are shared first authors. A.H., J.R. and G.V.R. conceived and designed the method. A.H. performed experiments and implemented the method. G.R., W.N. and J.M. supervised the work. Data set generation and quality control of the BIOS datasets was done by the BIOS Consortium, more details can be found at: http://www.bbmri.nl/acquisition-use-analyse/bios/, including details on contributions of all consortium members. A.H., J.R., G.V.R., M.A.I., W.N. and J.M. wrote, revised and approved the paper.

Competing interests

W.J.N. is a co-founder and shareholder of Quantib BV. Other authors declare no competing interests.

Ethics approval and consent to participate

This study was approved by the BIOS Data Access Committee. All data was generated by the bios consortium and the ethical approval for this study lies with the individual participating cohorts (RS, LL, LLS and NTR). Institutional review boards of all cohorts approved this study (LL, Ethics Committee of the University Medical Centre Groningen; LLS, Ethical Committee of the Leiden University Medical Center; NTR, Central Ethics Committee on Research

Involving Human Subjects of the VU University Medical Centre; RS, Institutional review board (Medical Ethics Committee) of the Erasmus Medical Center). In addition, informed consent was provided by all participants.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-024-00405-w.

Correspondence and requests for materials should be addressed to Arno van Hilten.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

BIOS consortium

Bastiaan T. Heijmans[5], Peter A. C. 't Hoen[6], Joyce van Meurs[2], Rick Jansen[7], Lude Franke[8], Dorret I. Boomsma[9], René Pool[9], Jenny van Dongen[9], Jouke J. Hottenga[9], Marleen M. J. van Greevenbroek[10], Coen D. A. Stehouwer[10], Carla J. H. van der Kallen[10], Casper G. Schalkwijk[10], Cisca Wijmenga[8], Sasha Zhernakova[8], Ettje F. Tigchelaar[8], P. Eline Slagboom[5], Marian Beekman[5], Joris Deelen[5], Diana van Heemst[11], Jan H. Veldink[12], Leonard H. van den Berg[12], Cornelia M. van Duijn[13], Bert A. Hofman[14], Aaron Isaacs[13], André G. Uitterlinden[2], P. Mila Jhamai[2], Michael Verbiest[2], H. Eka D. Suchiman[5], Marijn Verkerk[2], Ruud van der Breggen[5], Jeroen van Rooij[2], Nico Lakenberg[5], Hailiang Mei[15], Maarten van Iterson[5], Michiel van Galen[6], Jan Bot[16], Peter van 't Hof[15], Patrick Deelen[8], Irene Nooren[16], Matthijs Moed[5], Martijn Vermaat[6], René Luijk[5], Marc Jan Bonder[8], Freerk van Dijk[17], Michiel van Galen[6], Wibowo Arindrarto[15], Szymon M. Kielbasa[18], Morris A. Swertz[18] & Erik. W. van Zwet[18]

[5]Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands. [6]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [7]Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands. [8]Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands. [9]Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands. [10]Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands. [11]Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands. [12]Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands. [13]Department of Genetic Epidemiology, Erasmus MC, Rotterdam, The Netherlands. [14]Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands. [15]Sequence Analysis Support Core, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands. [16]SURFsara, Amsterdam, The Netherlands. [17]Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [18]Medical Statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands.