NCAE: data-driven representations using a deep network-coherent DNA methylation autoencoder identify robust disease and risk factor signatures

N.B. When citing this work, cite the original published paper.

(article starts on next page)

NCAE: data-driven representations using a deep network-coherent DNA methylation autoencoder identify robust disease and risk factor signatures

David Martínez-Enguita 🆔, Sanjiv K. Dwivedi, Rebecka Jörnsten and Mika Gustafsson

Corresponding author: M. Gustafsson, Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University SE-581 83, Linköping, Sweden.
Tel.: +46-132-82138; Fax: +46-132-88909; E-mail: mika.gustafsson@liu.se

Abstract

Precision medicine relies on the identification of robust disease and risk factor signatures from omics data. However, current knowledge-driven approaches may overlook novel or unexpected phenomena due to the inherent biases in biological knowledge. In this study, we present a data-driven signature discovery workflow for DNA methylation analysis utilizing network-coherent autoencoders (NCAEs) with biologically relevant latent embeddings. First, we explored the architecture space of autoencoders trained on a large-scale pan-tissue compendium ($n = 75\ 272$) of human epigenome-wide association studies. We observed the emergence of co-localized patterns in the deep autoencoder latent space representations that corresponded to biological network modules. We determined the NCAE configuration with the strongest co-localization and centrality signals in the human protein interactome. Leveraging the NCAE embeddings, we then trained interpretable deep neural networks for risk factor (aging, smoking) and disease (systemic lupus erythematosus) prediction and classification tasks. Remarkably, our NCAE embedding-based models outperformed existing predictors, revealing novel DNA methylation signatures enriched in gene sets and pathways associated with the studied condition in each case. Our data-driven biomarker discovery workflow provides a generally applicable pipeline to capture relevant risk factor and disease information. By surpassing the limitations of knowledge-driven methods, our approach enhances the understanding of complex epigenetic processes, facilitating the development of more effective diagnostic and therapeutic strategies.

Keywords: deep learning, autoencoders, DNA methylation, transfer learning, biomarkers, systems medicine

INTRODUCTION

Knowledge-driven methods for data analysis in systems medicine involve the use of prior biological understandings to guide and inform the analysis of large datasets. One of the most common of these approaches are network models, which represent biological entities, such as proteins or genes, and their functional relationships as nodes and edges within the interactome network. The interconnected nature of disease genes and their protein products has been exploited by algorithms that can detect so-called disease modules from omics data, validated using disease-associated single nucleotide polymorphisms (SNPs) from genome-wide analyses [1, 2]. However, despite curation efforts, human protein–protein interaction (PPI) networks are often incomplete and may not reflect the full complexity of biological systems. They are prone to research biases, as studies tend to focus on well-known proteins or on interactions that are easier to detect [3–5]. Furthermore, network inference tools often use simplifications in order to construct networks, which can affect the certainty of

their predictions [6, 7]. They are also limited by the coverage and quality of omics data: incomplete or noisy data lead to inaccurate networks [8]. Therefore, there is a need for robust data-driven approaches with the potential to mitigate knowledge biases and identify novel and meaningful signatures.

DNA methylation (DNAm) modifications are well-established biomarkers due to their capacity to capture long-term environmental effects. DNAm is an ideal modality for large-scale analyses due to its molecular stability, continuously variable nature and accessibility. It has proven useful in multiple studies of aging, cancer and other diseases [9]. Changes in the DNAm at specific positions, known as 5′-cytosine-phosphate-guanine-3′ (CpG) sites, located in genes involved in inflammation and DNA damage responses, have been shown to occur alongside age in a predictable manner [10]. This has allowed researchers to develop algorithms named DNAm clocks, which can accurately estimate chronological age from epigenetic profiles [11–13]. Similarly, distinct alterations in DNAm caused by tobacco exposure can be

used as precise biomarkers of the effects of smoking habits on gene expression and lung function [14–16]. DNAm homeostasis dysregulation also plays a key role in certain autoimmune diseases like systemic lupus erythematosus (SLE), neurodegenerative disorders and cancer. For instance, in SLE patients, several genes involved in immune response mechanisms such as inflammation and antibody production present characteristic DNAm patterns [17] that have been associated with different degrees of disease activity, severity and susceptibility [18–20].

The emergence of deep learning techniques, based on the concept of artificial neural networks (ANNs), has revolutionized the scientific panorama due to their superb capacity to model complex big data [21–23]. Autoencoders (AEs) are a type of ANN trained to efficiently compress and reconstruct unlabeled feature sets by learning their internal representations. AEs of different types and configurations have been successfully applied to omic research. For instance, deep undercomplete or variational AE architectures, such as scETM [24], VEGA [25], LDVAE [26], scMVAE [27], scIAE [28] or VASC [29], have been used to analyze single-cell transcriptomic data and identify cellular and gene signatures. Likewise, AEs have also been used to determine disease progression [30, 31], to cluster cancer subtypes [32, 33], to investigate protein variants [34] or to integrate spatial modalities in tissue samples [35]. More recently, the interpretation of the internal embeddings of deep gene expression AE models has revealed that biologically functional modules can be captured in their encoding, either by incorporating an orthogonality constraint on a single-cell transcriptomic AE [36] or autonomously within an unconstrained bulk transcriptomic AE [37]. Generative frameworks such as MethylNet [38] or siVAE [39] have further demonstrated the potential of interpretable feature embeddings in methylation and genomic research. Altogether, this capacity of AEs to capture such biological patterns suggests a partial understanding of the underlying functional and regulatory omic context, including pathway-level interactions and functional gene sets. Among AE types, deep undercomplete models are highly efficient at pattern learning and relatively simple to implement and train. In addition, their hidden representations are deterministic, with inputs being clustered into discrete vectors within a compressed space that is in principle not constrained into a prior distribution. Another benefit of deep AE generic representations is that they are transferable across tasks, meaning that task-specific features can be extracted on demand for any given purpose. These properties make deep AEs particularly attractive for scenarios where large amounts of labeled data may not be available, as the AE can be first pre-trained on a much larger and diverse sample collection, and then fine-tuned for particular purposes.

Here, we present a novel data-driven approach for the functional analysis of DNAm data, in which we introduce the concept of 'network-coherent autoencoders' (NCAEs). NCAEs are deep AEs with biologically relevant embeddings that prioritize genes involved in co-localized regions of the human interactome within their latent representation, which we refer to as 'network coherence'. By capturing and preserving functional relationships among genes, akin to PPI modules, in an interpretable latent space, NCAEs offer a comprehensive understanding of multi-tissue DNAm data. We present a robust pipeline for training and leveraging NCAEs, showcasing the potential of their network-coherent latent representation in mitigating knowledge biases and discovering novel DNAm signatures. Notably, supervised neural network models can be efficiently trained using these NCAE embeddings in disease and risk factor modeling tasks, exhibiting a superior performance that often surpasses other

DNAm-based estimators. Our versatile workflow facilitates the identification of data-driven epigenetic signatures and can be applied to any task involving DNA methylation data and a supervised training objective.

MATERIALS AND METHODS
Data pre-processing
Human DNA methylation profiles and metadata ($n = 75\ 326$) were downloaded from the EWAS (Epigenome-Wide Association Study) Data Hub public repository (https://ngdc.cncb.ac.cn/ewas/datahub, accessed on 25 January 2021). Sources for this database include Gene Expression Omnibus, ArrayExpress, The Cancer Genome Atlas and Encyclopedia of DNA Elements. Methylation profiles from EWAS Data Hub were generated by Illumina Infinium HumanMethylation450 or MethylationEPIC arrays and were normalized and corrected for batch effects using Gaussian Mixture Quantile Normalization [40]. After sample quality control, 75 272 samples were left (50 623 non-cancer and 24 649 cancer profiles). Additional filtering was performed using the ChAMP package (version 2.26.0) for the R programming environment (version 4.2.1). Non-CpG probes, probes related to SNPs, multi-hit probes and probes located on the X or Y chromosomes were filtered out. Missing beta (β) values for probes were imputed using the k-nearest neighbors method ($k = 10$) from the bnstruct R package (version 1.0.12). After filtering and excluding probes that are not shared by both Illumina 450K and MethylationEPIC arrays, a total of 384 629 CpG sites were left. The pre-processed methylation data consisted in a beta-value matrix of 75 272 methylation profiles (Table S1 available online at http://bib.oxfordjournals.org/) by 384 629 CpG sites.

Design and training of AE models
Artificial neural network models were trained using Keras 2.4.3 library with TensorFlow 2.4.0 and TensorFlow-GPU 2.2.0 back-end, implemented for Python 3.8.10. Normalized DNAm samples were used to train and evaluate constant-width and hourglass deep methylation AEs (DMAEs), sparse deep AE (spDMAEs) and methylation variational AE (MVAE) models, with a training/validation/test split ratio of 64:16:20, balanced for tissue and sample group proportions using multivariate stratified sampling. To inspect the impact of the number of hidden nodes on reconstruction performance, we chose to use constant-width AE models, following the rationale of Dwivedi et al. [37]. We benchmarked their performance against classical hourglass-shaped AEs with either 2 or 16 hidden nodes in the narrowest layer of three, and 32 to 1024 hidden nodes in the others. Prior to training, hyperparameter fine-tuning was performed using balanced sample subsets of 10–20% of the original population. We conducted learning rate range tests on a three-layered AE of 256 hidden nodes to identify the learning rate (between 1 and 10e−6) and optimizer (Adam, Adadelta, Adagrad, Adamax, RMSprop, stochastic gradient descent) combination that yielded the best balance between convergence speed and reconstruction. The optimal configuration used the Adam optimizer to minimize the mean squared error, with learning rate $= 9.0e-5$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e-8$ and decay $= 1e-6$. We selected the leaky rectified linear unit (leaky ReLU, $\alpha = 0.3$) function as hidden layer activation function and the sigmoid function for the output layer. Dropout regularization did not improve performance and was therefore not applied in the final model. To avoid overfitting, models were trained using early stopping, with a patience of 10 epochs. The batch size for training was 128. The architecture and hyperparameters of the spDMAE,

MVAE and NCAE-like models trained on cancer sets or on sets with different proportions of whole blood samples were matched to that of the network-coherent AE (3 hidden layers, 128 hidden nodes per layer). The MVAE had 3 dense encoding layers of 128 nodes per layer and a latent Gaussian space of size 2, plus a dense decoding layer. The spDMAE included an L1 activity regularizer constraint = 1e−3 to induce sparsity in each hidden layer. Warm-start training for the pre-trained NCAE on cancer samples was conducted with a reduced learning rate = 9.0e−6.

Design and training of supervised neural network models

Supervised deep neural network (DNN) models trained on the NCAE latent representations were designed to be feed-forward, fully connected neural networks with an input layer of the same dimension as the feature embeddings from the NCAE ($k = 128$). We utilized the NCAE to compress the high-dimensional DNAm sample sets into lower-dimensional latent embeddings of 128 dimensions, extracted from the third hidden layer. These embeddings are then used for training, validating and testing the supervised DNNs in each case. To better take advantage of the number of available samples, the DNN training strategy applied was as follows: (i) split the NCAE embeddings into training, validation and held-out test sets (64:16:20); (ii) use grid searches to explore model depths (one to three layers), widths (16 to 128 hidden nodes) and regularization options (L1, L2, L1/L2) until an optimal configuration that minimizes the respective objective function in each case is found; (iii) train a DNN with identical architecture and hyperparameters on the full training and validation sets (80:20). We used the Adam optimizer with the same hyperparameters as previously to minimize the mean squared error (MSE) (NCAE-Age), categorical cross-entropy (NCAE-Smoke) or binary cross-entropy (NCAE-SLE, SLE principal component analysis (PCA)-based DNN). The multi-tissue and whole blood NCAE-Age models used ReLU as hidden layer activation function and leaky ReLU as output layer function. For NCAE-SLE, SLE PCA-based DNN and NCAE-Smoke, we opted for leaky ReLU in the hidden layers, and sigmoid or softmax function in the output layer, respectively. In all cases, L1 kernel regularization was applied on the third hidden layer to prevent overfitting, with a scale factor $\lambda = 0.01$. NCAE-DNNs were trained with early stopping (patience = 1e3). The batch sizes used were 1024 for NCAE-Age and 256 for NCAE-Smoke, NCAE-SLE and SLE PCA-based DNN. To determine whether sample age and gender could be relevant DNN covariates for smoking status prediction, we trained three additional NCAE-Smoke models using the sample embeddings plus each covariate or both as inputs of the DNN. However, we did not observe an increase in classification recall. Thus, age and gender were not used as DNN covariates.

Performance evaluation

Reconstruction, regression and classification performance of ANN models were assessed using metrics from the Python library scikit-learn 0.24.2. Test set local (CpG-wise) reconstruction performance for AEs was measured using the coefficient of determination (R^2) computed as

$$R^2_{pi} = 1 - \frac{(w_i \mathbf{MSE}_{pi})}{(w_i \sigma^2_{pi})}$$

where p, i, w_i and σ_i^2 correspond to the pth CpG probe, the ith data set, the weight (number of samples divided by the total number

of samples in the test set) of the ith data set and the variance of the pth CpG probe of the ith dataset.

Light-up analysis from hidden nodes

To determine the association of AE hidden layers with input CpGs, we retrieved the output layer activations computed from the recursive light-up activation of each hidden node in every layer [37]. That is, we forward-propagated an activation vector x^a consisting of the maximum activation value for a single hidden node h of a hidden layer k, while keeping the rest of the nodes at the minimum activation. Maximum (1) and minimum (α) activation values used corresponded to the derivative of the AE hidden layer activation function (i.e. leaky ReLU). The following Equation (1) defines the activations x^k of the kth layer from the activations x^{k-1} at the $(k-1)$th layer with the initial activation vector x^a, for a node h in the pth hidden layer:

$$x^k = \begin{cases} f^k\left(W^k x^{(k-1)} + b^k\right) \text{ if } p < k \leq L \\ x^a \text{ if } p = k \end{cases}$$

where f^k, W^k and b^k correspond to the kth layer activation function, weight matrix and bias term, respectively. The activations at the output layer x^L have the same dimensions as the model input and are then used to rank the CpGs in terms of their association with the maximally activated hidden node at layer $k-1$. The process is then repeated for every hidden node and layer.

Light-up analysis from inputs

To prioritize CpGs in supervised NCAE-DNN models by their contribution to the output, we applied a perturbation-based forward propagation analysis for feature importance ranking. Since both hyper- and hypomethylation are viable states for a CpG site, we first recursively altered input CpG values to either a complete hyper- ($\beta = 1$) or hypomethylation ($\beta = 0$) level before forward-propagating them through the trained NCAE and DNN. Then, we compared the output of the concatenated models to that of an average methylation profile (mean beta value across samples) representative of a specific tissue and condition, e.g. CD4$^+$ T cells from SLE patients. After the signal propagation was iterated through every input feature (CpG), their contribution to the regression or classification objective of the DNN can be measured by the observed changes in the model outcome, i.e. estimated age or predicted disease classification probability.

Age estimation using DNAm clocks

We used the getAgeR() function from the R package cgageR to obtain age estimates from Horvath and Hannum DNAm clocks for samples in the multi-tissue ($n = 24\,676$) and the whole blood ($n = 13\,647$) sets. Evaluation metrics were calculated using the functions provided by the R package Metrics. Biologically meaningful age bins were established based on the classification of age categories by the Medical Subject Headings controlled vocabulary thesaurus of the U.S. National Library of Medicine (http://www.ncbi.nlm.nih.gov/mesh).

Gene annotation and enrichment analyses

Genome-wide annotation of CpG probes was performed using Infinium HumanMethylation450 BeadChip probe annotation files and the R package org.Hs.eg.db (version 3.15.0). Pathway and gene ontology enrichment analyses were performed using the respective functions from the R package clusterProfiler (version

4.4.4). Enrichments in disease-associated genes of the top light-up CpG-associated genes per tissue were calculated using a Kolmogorov–Smirnov-like statistic, similar to the enrichment score in gene set enrichment analysis [41]. First, we computed one-sided Fisher's exact tests for overlap with disease-associated genes retrieved from DisGeNET v7.0 (accessed on 2 June 2022) over the cumulative ranked gene lists until the position 1000. To increase the stringency of the outcome, we avoided the overestimation of the maximal enrichment scores (minimum P) of the light-up ranked gene lists by limiting the minimum cumulative set size to 101 genes. Then, we generated a null distribution of P-values by performing enrichment testing in the same interval across gene lists obtained by randomly assigning gene labels from all CpG-associated genes included in the model input. The procedure was repeated for 1e4 permutations. Finally, we calculated the permutation test P-value by dividing the number of cases in which the null distribution P-values are more extreme than the light-up ranked gene list P-value between the number of permutations. Average permutation P-values were obtained using the function hmp.stat() of the R package harmonicmeanp to compute the harmonic mean P-value for dependent tests.

RESULTS
Deep AEs can accurately reconstruct low-dimensional embeddings of methylation data

We searched in an unsupervised manner for a functional data representation that could encompass and learn the complete feature space, while simultaneously being dimensionally reduced to facilitate predictive modeling and decrease noise. For this aim, we downloaded and pre-processed a multi-tissue compendium of 75 272 human DNAm profiles and metadata from the EWAS Data Hub repository [42] from Illumina 450K or EPIC arrays. Case–control samples from 315 diseases and 471 tissues or cell types are represented in the collection (Figure 1). Non-cancer samples ($n = 50\,623$) were included in the model training, while cancer samples ($n = 24\,649$) were kept as an independent test set to avoid introducing bias. DNAm profiles were randomly split into training, validation and test sets balanced for tissue and disease proportions (Methods). We trained and evaluated 24 different undercomplete AE architectures with depths of one to four hidden layers, and constant widths ranging from 32 to 1024 hidden nodes per layer. We benchmarked their performance against classical hourglass-shaped three-layered AE architectures with either 2 or 16 hidden nodes in the narrowest layer, and 32 to 1024 hidden nodes in the others. Reconstruction performance was measured by the coefficient of determination (R^2) calculated over the global and local (CpG-specific) variances on the test set.

We observed global R^2 values between 0.978 and 0.993 for constant-width models, with decreasing reconstruction error primarily associated with an increase in model width, not depth, until our maximum tested 1024 nodes per layer (Figure 2A). Since some CpGs are prone to have stable methylation levels, we analyzed the proportion of explained variance per CpG (Figure 2B), which showed a similar pattern. AE models of 32 nodes per layer achieved median local R^2 values between 0.767 and 0.749, while wider configurations such as 512 and 1024 nodes per layer performed better (median local R^2 values between 0.901 and 0.891). Specifically, we observed that 85.9% of the variance could be explained already for models with 128 nodes, corresponding to

a 3000-fold dimensionality reduction (Table S2 available online at http://bib.oxfordjournals.org/). Models with limited width (2 nodes) in the narrowest layer struggled to compress DNAm data accurately (median local $R^2 = 0.368$ to 0.508), while models with 16 nodes demonstrated noticeably better reconstruction capabilities (median local $R^2 = 0.748$ to 0.871), which were close to but still below their matched constant-width architectures.

We then investigated whether fractions of CpGs existed that were consistently well or poorly predicted. Notably, we found that 4.7% of the probes ($n = 18\,118$) were systematically predicted with a local error below Q1 ($R^2 = 0.922$) for every constant-width AE. These CpGs were more variable than average (mean $\sigma^2 = 4.63e{-}2$, average mean $\sigma^2 = 1.65e{-}2$, Wilcoxon $P < 2.2e{-}16$) and included a higher proportion of hypermethylated probes (57.1% versus 50.5%, chi-square $P < 2.2e{-}16$) than expected. Conversely, 12.9% of the CpGs ($n = 49\,465$) were consistently reconstructed with a local error above Q3 ($R^2 = 0.751$). These probes were less variable than average (mean $\sigma^2 = 4.73e{-}3$, Wilcoxon $P < 2.2e{-}16$) and predominantly hypomethylated (74.0% versus 43.4%, chi-square $P < 2.2e{-}16$).

CpGs from the third layer of a deep AE are associated with highly co-localized genes

We then investigated the large-scale associations of the learned representations to the human PPI network. Our hypothesis was that a functional representation should cluster CpGs associated with functionally related genes together, thus associating genes close in the PPI to the same latent variable. We ranked CpGs by their association with each hidden node using light-up analysis [37], a type of perturbation-based forward propagation technique that allows to interpret the non-linear embeddings of an ANN by relating components of the internal layers to the model output (Methods). The resulting activation signal can be used to rank features, in this case CpG sites. We repeated the procedure for all hidden nodes and layers across AE architectures and inspected the relationships between the top prioritized CpG-associated genes, in terms of their co-localization and centrality.

Thus, we mapped the top CpGs to their associated genes and analyzed their co-localization in the human PPI network defined by STRING v11 [43] high-confidence interactions (combined score > 0.7). We calculated the harmonic average distance (HAD) between the gene lists and compared it with the average HAD within the PPI network (HAD = 3.48). Gene sets with low HAD cluster together in the interactome, i.e. they co-localize, suggesting their participation in the same biological processes. Our analysis revealed that top-ranked genes exhibited a higher degree of co-localization than expected (Wilcoxon rank sum $P < 2.2e{-}16$). This decrease in HAD was more pronounced as the light-up signal propagated deeper into the AE layers, until the third layer of AE models with up to four hidden layers (Figure 3 and Figure S1 available online at http://bib.oxfordjournals.org/). Genes linked to the first and second layers showed a progressive increase in co-localization with respect to the average HAD. Notably, the third hidden layer of three-layered AEs exhibited the lowest HAD between their top associated genes, while the effect diminished at the fourth layer (Figure 3A–E). Differences in average HAD across AE widths were highly significant (Kruskal–Wallis adj. $P < 2.2e{-}16$) between every hidden layer except the third and fourth (Figure 3F). Importantly, this co-localization pattern was predominantly observed in the latent representation of architectures with widths of 128, 256 and 512 hidden nodes per layer, while it was less pronounced or absent in other AE configurations. Likewise, it was more prominent for the top 100
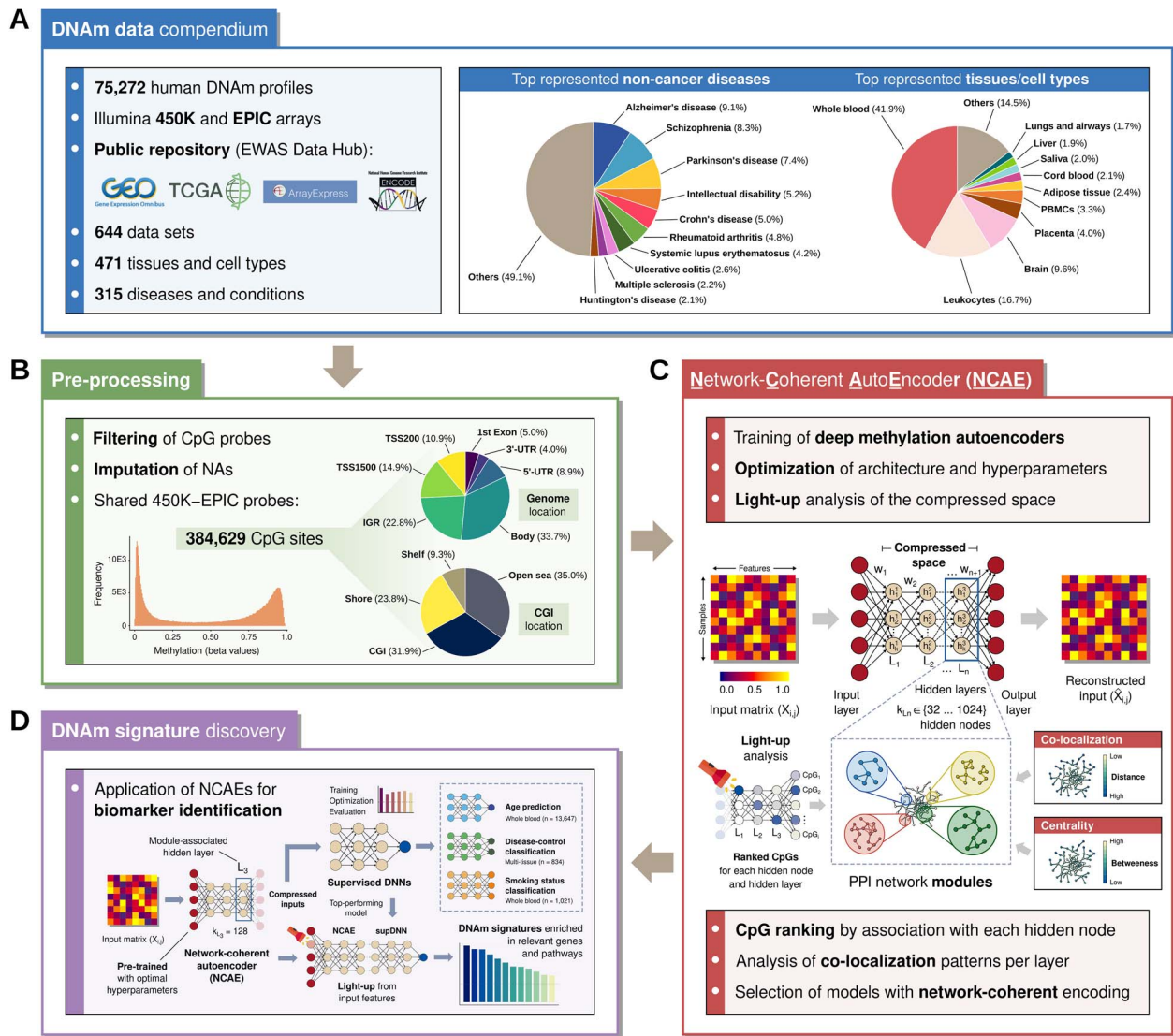
Figure 1. Schematic of the workflow for training and functionalization of network-coherent autoencoders (NCAEs). (**A**) Summary of DNAm datasets included in the compendium, top represented non-cancer diseases, and tissue and cell types. (**B**) Pre-processing steps and description of probe set genomic and CpG island (CGI) locations. (**C**) Training and selection of deep autoencoders based on the network coherence of their latent space. (**D**) Functionalization of deep NCAE compressed representations for the identification of DNAm signatures using concatenated task-specific deep supervised neural networks (DNNs).

to 400 ranked genes, gradually vanishing further in the gene ranking (Figure 3A–E). To provide additional insights, we also analyzed the betweenness centrality of the associated CpGs, which showed a similar trend (Figure S2 available online at http://bib.oxfordjournals.org/). Genes with high betweenness centrality are frequently situated on shortest network paths between other genes. Thus, they regulate the flow of information through their regions of the interactome and usually correspond to proteins involved in signaling pathways. Third-layer latent variables of three-layered AEs with 128, 256 and 512 nodes were associated with central genes, whereas the effect was not as pronounced for other layers and depths.

In summary, we observed that latent variable representations differ substantially in terms of their relation to the PPI network. CpGs associated with the third layer of the three-layered deep AE of 128 hidden nodes showed the strongest co-localization pattern (HAD = 2.9) and above-average centrality (betweenness = 3.1e−4) within the human PPI network. Therefore, we subsequently focused our analysis on the three-layered

deep AE of 128 nodes, which will henceforth be referred to as network-coherent AE (NCAE). We replicated the analyses in sparse AEs and variational AEs with architectures equivalent to the NCAE (Methods), but we did not find a significant increase in co-localization.

The deep NCAE reconstructed most common non-cancer diseases and tissues as well as or better than controls

Next, we tested if the high-level compression of the NCAE was biased toward certain tissues or diseases. To do so, we calculated the CpG-wise R^2 across test set samples from the 10 most frequent tissues or cell types, and non-cancer diseases and healthy controls. Interestingly, well-explained tissues were related to circulating blood cells, such as cord blood ($R^2 = 0.930$), whole blood ($R^2 = 0.918$) and peripheral blood mononuclear cells (PBMCs) ($R^2 = 0.917$), whereas localized tissues, such as liver ($R^2 = 0.818$), brain ($R^2 = 0.884$) or placenta ($R^2 = 0.918$), showed relatively lower but adequate performance (Figure 4A). We found a small but
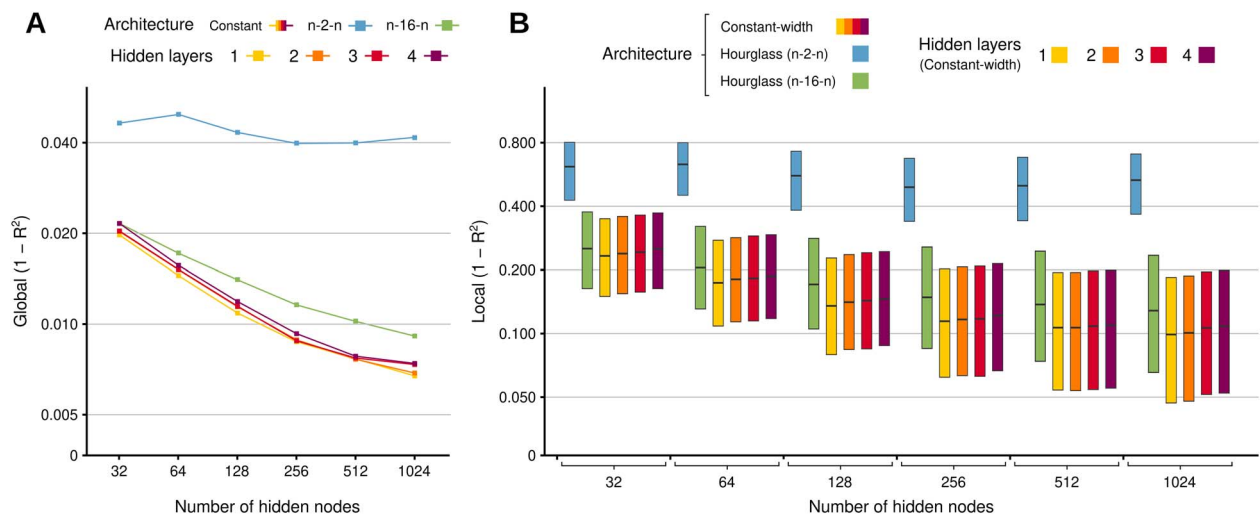
Figure 2. Performance evaluation of DNA methylation autoencoder architectures. Global (**A**) and local (**B**) reconstruction error, expressed as 1—coefficient of determination (R^2) on the test set data, across constant-width DNAm AEs of one to four hidden layers, from 32 to 1024 hidden nodes per layer, and hourglass DNAm AEs of three hidden layers and 2 or 16 hidden nodes in the narrowest layer.
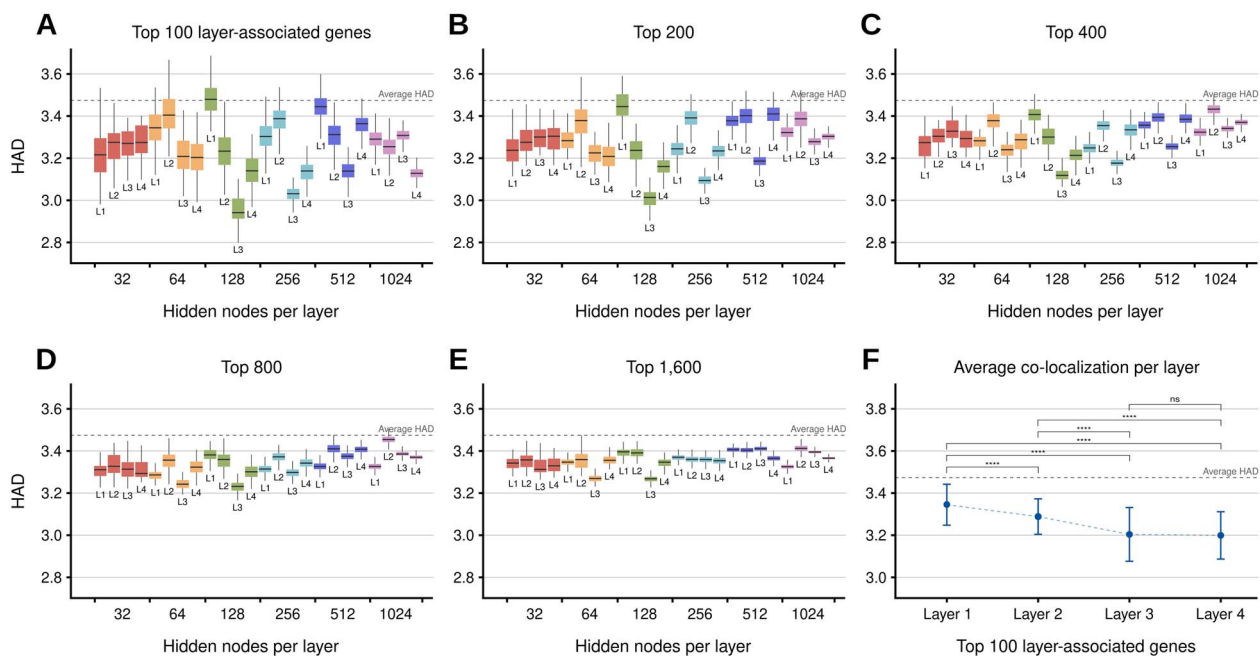

Figure 3. The latent representations from deep methylation AEs show increasing gene co-localization patterns in the human interactome until the third hidden layer. Harmonic average distance (HAD) in the human PPI network of the top 100 to 1600 (**A–E**) ranked layer-associated genes from the first, second, third and fourth hidden layers of deep methylation AEs with up to four hidden layers and 32 to 1024 hidden nodes per layer. (**F**) Average HAD per deep AE hidden layer of the top 100 layer-associated genes.

non-significant correlation (Pearson $r = 0.36$, $P = 0.304$) between the NCAE local performance on top represented tissues and their training set sample sizes. To further address possible performance biases with respect to tissue sample size in the training data, we trained five new NCAE-like models (three hidden layers, 128 nodes per layer) using modified training and validation sets. We held every tissue type constant, except for whole blood samples, which were gradually removed until proportions of 1/4, 1/8, 1/16, 1/32 and 0 were left, with respect to the original sample size (Figure S3A available online at http://bib.oxfordjournals.org/). The reconstruction performance of these models on the whole blood test set gradually decreased as the proportion of training whole blood samples decreased, with $R^2 = 0.918$ (original set), 0.892 (1/4), 0.883 (1/8), 0.873 (1/16), 0.855 (1/32) and 0.847 (0). This outcome

suggests that the learning of tissue-specific patterns is positively influenced by available training sample sizes. Nevertheless, it is worth noting that even when the training set contained no whole blood samples, the performance of the NCAE-like model remained relatively good.

Top represented non-cancer diseases showed median R^2 values between 0.951 and 0.898, whereas controls were predicted with $R^2 = 0.865$ (Figure 4B), likely due to a higher variability across healthy individuals. The NCAE performance and the proportion of training samples per disease were not significantly correlated (Pearson $r = -0.01$, $P = 0.978$). Lastly, we analyzed the NCAE performance on samples from common cancer types and their respective test set adjacent normal tissue samples (Figure 4C). Cancer samples ($n = 24\,649$) were explained poorly,
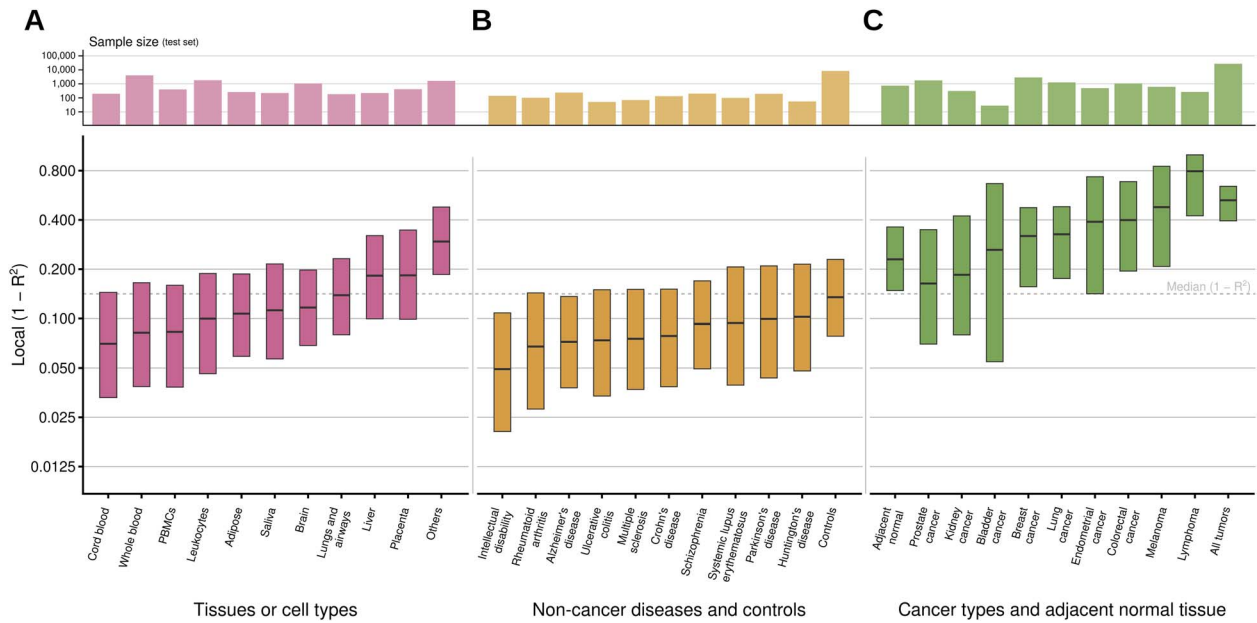
Figure 4. Local reconstruction error of the NCAE of three hidden layers and 128 hidden nodes per layer on the top represented tissues or cell types (**A**), top represented non-cancer diseases and controls (**B**), and common cancer types and adjacent normal tissue samples (**C**). Test set sample sizes of the evaluated categories are shown on the top barplot.

achieving only $R^2 = 0.471$, compared to $R^2 = 0.769$ for adjacent normal tissue. Certain cancer types, such as prostate and kidney, were accurately predicted, with $R^2 = 0.836$ and 0.815, respectively, while others such as lymphoma were especially difficult to reconstruct ($R^2 = 0.203$). To test whether the NCAE performance on cancer samples would improve if included in the training set, we warm-start trained the non-cancer NCAE on cancer samples, and we compared it with NCAE-like models with the same architecture as our selected NCAE, but trained exclusively in cancer (Figure S3B available online at http://bib.oxfordjournals.org/). The results revealed an improvement in the reconstruction accuracy on cancer in both cases, with R^2 values increasing from 0.471 (non-cancer NCAE) to 0.742 (cancer NCAE) and 0.735 (warm-start NCAE). However, neither model surpassed the performance of the non-cancer NCAE on non-cancer samples ($R^2 = 0.857$). Simultaneously, we observed a decrease in the performance of these cancer-trained models on the non-cancer set, with $R^2 = 0.487$ (cancer NCAE) and 0.417 (warm-start NCAE). These outcomes could be attributed to the aberrant methylation features unique to cancer DNAm profiles [44], which may form independent clusters within the NCAE latent space and compromise the model's ability to learn fainter biological signals. Overall, these results indicate that the NCAE latent representation can capture disease DNAm patterns with low error while achieving a proficient reconstruction of key cell types for biomarker identification. This led us to examine how well the latent variables could be re-purposed for phenotypic predictions associated with known epigenetic signatures.

An NCAE-age model accurately estimates chronological age and identifies relevant aging DNAm signatures

A highly suitable task for DNAm data modeling is the estimation of age using DNAm clocks. Two of the most popular are the hallmark clocks of Horvath [11] and Hannum [12], which can predict chronological age with high accuracy (reported test $R^2 = 0.922$, $R^2 = 0.963$, respectively). Remarkably, they use only 353 and 71 CpGs, respectively, although aging-related processes are

likely spread across many more DNA regions. Thus, a broader and more robust DNAm clock could potentially serve better to understand aging. For this purpose, we utilized the NCAE to compress DNAm data into their low-dimensional representations of 128 features, which we then used to train a deep supervised neural network (NCAE-Age) to predict chronological age, comparing its performance with Horvath and Hannum clocks.

First, we fed the NCAE a total of 13 647 whole-blood DNAm samples from healthy individuals with ages between 0 and 112 years (mean = 40.5 ± 23.6 years, Table S1 available online at http://bib.oxfordjournals.org/) to extract the latent embeddings and train the NCAE-Age model. Notably, the NCAE-Age achieved highly accurate results on the test set ($n = 2729$, $R^2 = 0.965$), followed by Horvath ($R^2 = 0.936$) and Hannum clocks ($R^2 = 0.929$) (Figure 5A and B). Restricting the analysis to test samples in the age range of 19–101 years to match Hannum's training set did not improve its performance ($R^2 = 0.859$). The NCAE-Age model was also better than the DNAm clocks in terms of MAE and RMSE (Figure 5B). To better assess performance across age ranges, we binned the test samples into four biologically meaningful age groups (Newborn-Adolescent: 0–18 years, $n = 746$; Adult: 19–44 years, $n = 638$; Middle-age: 45–64 years, $n = 901$; and Aged: ≥ 65 years, $n = 444$). NCAE-Age estimates were more accurate than DNAm clocks in the four groups (Figure 5C). All models performed well on Newborn-Adolescent and Adult groups, with a subsequent decrease in accuracy, especially for Horvath clock. In general, the performance of the NCAE-Age model was comparable or superior to the DNAm clocks.

To assess the leverage of particular CpG sites on the NCAE-Age and determine its capacity to prioritize CpGs associated with aging mechanisms, we applied a light-up analysis from the NCAE-Age input. Then, we tested whether the resulting CpG-associated genes were enriched in genes linked to nine chronological and five biological DNAm clocks ($n = 1328$, Table S3 available online at http://bib.oxfordjournals.org/) [45]. We found that genes associated with the top 1000 CpGs identified by hyper- or hypo-perturbation significantly overlapped the combined age clock gene list ($P = 3.49e{-}13$, OR = 2.23, 95% CI [1.83, 2.72]; and
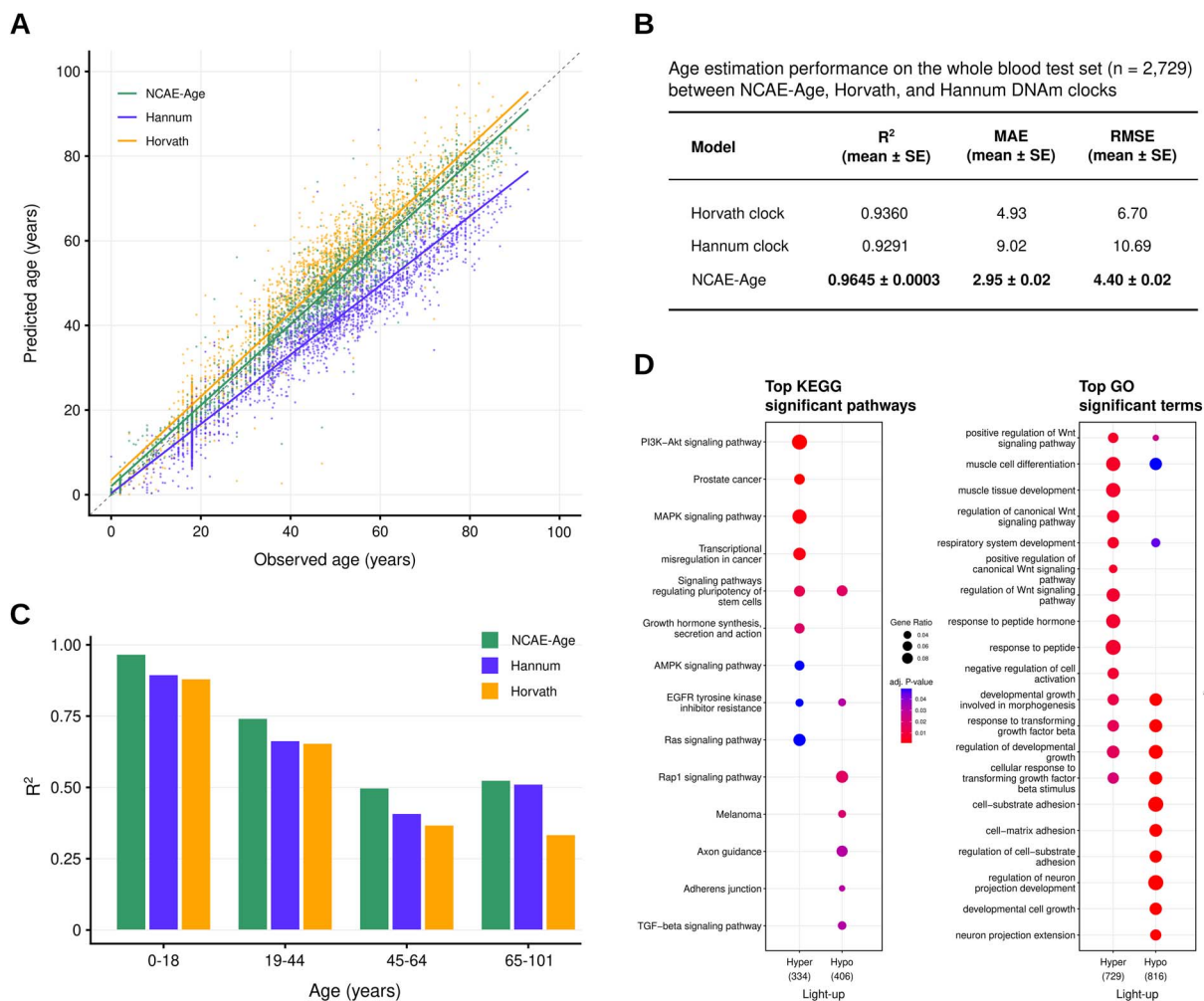
Figure 5. A non-linear DNAm age estimator (NCAE-Age) trained on whole blood NCAE-compressed inputs can predict age more accurately than hallmark DNAm clocks. (**A**) Comparison of true and predicted ages of test set samples (*n* = 2729) as estimated by NCAE-Age, Horvath and Hannum DNAm clocks. Prediction performance of the DNAm estimators on the full test set (**B**) and on the binned samples by age group (0–18: Newborn-Adolescent, 19–44: Adult, 45–64: Middle-age, 65–101: Aged) (**C**). (**D**) Top significantly overrepresented KEGG pathways and GO terms for the top-ranked CpG-associated genes after the light-up analysis.

$P = 2.27e{-}5$, OR = 1.59, 95% CI [1.29, 1.96], respectively). In particular, top genes obtained by hypermethylation light-up were significantly enriched in every DNAm clock gene list containing more than four genes ($P = 1.71e{-}8$ to 0.04, OR = 1.59 to 80.02, 95% CI [1.00, 2.54] to [21.44, 298.60]) (Table S3 available online at http://bib.oxfordjournals.org/). We also investigated if the NCAE-Age signatures (Table S6 available online at http://bib.oxfordjournals.org/) were associated with biological pathways linked to aging in humans. The Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses revealed a significant overrepresentation of phosphoinositide 3-kinase (PI3K)-Akt, mitogen-activated protein kinase (MAPK), adenosine monophosphate-activated protein kinase (AMPK) and Ras signaling (adjusted $P = 1.19e{-}3$ to 0.049, OR = 2.08 to 2.27, 95% CI [1.30, 3.33] to [1.41, 2.85]), and cancer-related pathways (adj. $P = 1.19e{-}3$ to $2.82e{-}3$, OR = 2.19 to 3.78, 95% CI [1.37, 3.47] to [2.17, 6.58]), among others. Interestingly, both lists were highly enriched in Gene Ontology (GO) terms associated with responses to transforming growth factor (TGF) beta (adj. $P = 2.52e{-}5$ to $3.78e{-}2$, OR = 2.48 to 3.02, 95% CI [1.42, 4.33] to [2.08, 4.37]) and positive regulation of the Wnt signaling pathway (adj. $P = 3.73e{-}3$ to $1.93e{-}2$, OR = 2.56 to 3.58, 95% CI [1.52, 4.31] to [2.23, 5.75]) (Figure 5D).

An NCAE-Smoke model determines smoking status and defines associated smoking DNAm signatures

Besides aging, another DNAm research application concerns the modeling of alterations due to tobacco smoking. Smoking DNAm signatures can be identified even long after cessation. We hypothesized that these long-lasting changes in the DNAm landscape may be imprinted on the NCAE representations, thus being able to derive epigenetic signatures for smoking from them. We retrieved 1021 whole blood DNAm samples from individuals enrolled as controls in five studies with self-reported smoking status information (Table S1 available online at http://bib.oxfordjournals.org/). We obtained their NCAE embeddings and trained a supervised DNN (NCAE-Smoke) for multi-class classification of 'current' (*n* = 408), 'former' (*n* = 185) and 'never smokers' (*n* = 428), where 20% of samples in each group were used as test set. We observed the best performance in an NCAE-Smoke with leaky ReLU activation [area under the curve $(AUC)_{current} = 0.87$, $AUC_{former} = 0.91$, $AUC_{never} = 0.69$, average AUC = 0.80, Figure 6A].

To evaluate the classification performance of the NCAE-Smoke, we compared it with another available DNAm-based smoking
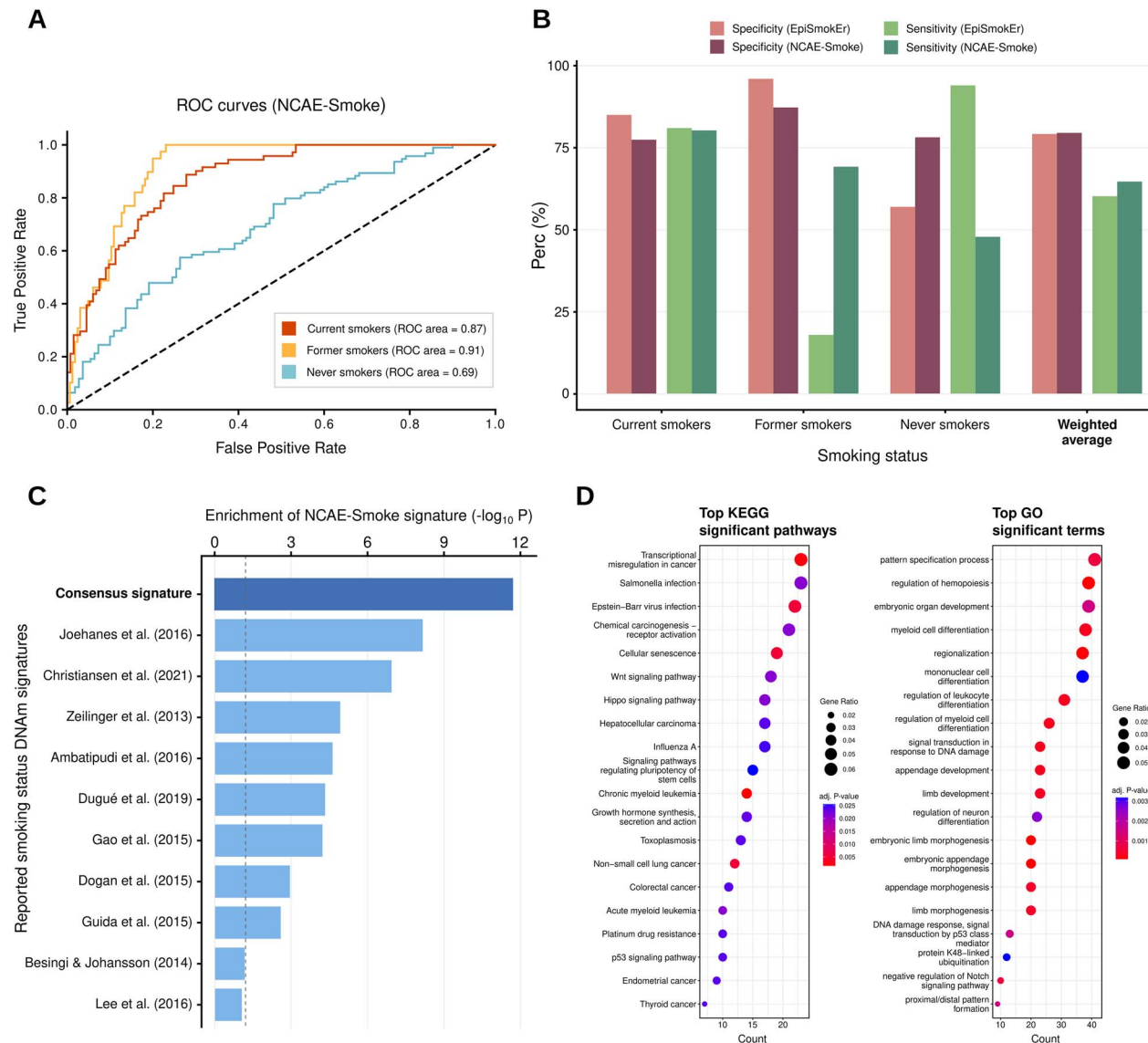
Figure 6. A non-linear DNAm smoking status classifier (NCAE-Smoke) trained on NCAE-compressed inputs accurately separates current, former and never smokers. (**A**) Area under the receiver operating characteristic curves for NCAE-Smoke classification performance of smoking status classes. (**B**) Comparison between NCAE-Smoke and EpiSmokEr reported average per class specificity and sensitivity on the test set, and weighted average by sample size across classes. (**C**) Significance of the enrichment of top-ranked CpG-associated genes from the NCAE-Smoke, for 'current' versus 'never smokers', in reported DNAm signatures for smoking status. The consensus signature includes all genes appearing in at least two of these studies. (**D**) Most significantly overrepresented KEGG pathways and GO terms for the top-ranked NCAE-Smoke CpGs for 'current' versus 'never smokers'.

status predictor trained on whole blood data. EpiSmokEr [46] uses 121 CpGs identified via a Least Absolute Shrinkage and Selection Operator (LASSO)-penalized generalized linear model, plus a sex and an intercept coefficient, to determine smoking status. In comparison with the reported performance of EpiSmokEr on its test set data (Figure 6B), NCAE-Smoke was able to correctly categorize 'current smokers' (EpiSmokEr Spec = 85%, Sens = 81%; NCAE-Smoke Spec = 77%, Sens = 80%), as well as reliably rule out samples that are not 'former smokers' (EpiSmokEr Spec = 96%, NCAE-Smoke Spec = 87%). However, NCAE-Smoke achieved a higher true positive rate for 'former smokers' (Sens = 69%) than EpiSmokEr (Sens = 18%). Regarding 'never smokers', NCAE-Smoke had better specificity (77% versus 57% for EpiSmokEr), whereas EpiSmokEr had higher sensitivity (94% versus 48% for NCAE-Smoke). On average, NCAE-Smoke performed similar to or above EpiSmokEr (NCAE-Smoke Spec = 79%, EpiSmokEr Spec = 79%; NCAE-Smoke Sens = 65%, EpiSmokEr Sens = 60%).

We next retrieved a list of CpGs prioritized by differences between the 'never smoker' and 'current smoker' classes using light-up, selecting the genes associated with the top 1000 CpGs as the NCAE-Smoke signature (Table S6 available online at http://bib.oxfordjournals.org/). We evaluated the significance of their overlap with 10 available smoking status DNAm signatures (n = 95 to 3978 CpGs, associated with 43 to 1632 unique genes, Table S4 available online at http://bib.oxfordjournals.org/). We observed that CpG-associated genes from 8 out of 10 signatures were significantly overrepresented (Fisher's exact P = 6.60e−9 to 2.48e−3, OR = 1.69 to 2.31, 95% CI [1.32, 2.17] to [1.62, 3.30]) in the NCAE-Smoke signature (Figure 6C). Moreover, the consensus list of reported DNAm signatures (genes appearing at least twice across the signatures, n = 1462) showed the most significant overlap (Fisher's exact P = 1.87e−12, OR = 2.09, 95% CI [1.73, 2.53]). Examining the biological context of the multi-tissue NCAE-Smoke signature (Figure 6D), we found that it was enriched

in KEGG pathways associated with cancer, such as myeloid leukemia (adj. $P = 1.57e{-}3$ to $2.11e{-}2$, OR = 3.35 to 4.13, 95% CI [1.71, 6.56] to [2.32, 7.38]), non-small cell lung cancer (adj. $P = 5.22e{-}3$, OR = 3.74, 95% CI [2.01, 6.95]), p53 signaling pathway (adj. $P = 2.38e{-}2$, OR = 3.07, 95% CI [1.57, 6.00]) and chemical carcinogenesis receptor activation (adj. $P = 2.11e{-}2$, OR = 2.22, 95% CI = [1.40, 3.52]). Furthermore, it was also enriched in GO terms such as signal transduction in response to DNA damage (adj. $P = 3.03e{-}4$, OR = 3.30, 95% CI [2.12, 5.13]), signal transduction by p53 class mediator (adj. $P = 1.55e{-}3$ to $2.85e{-}2$, OR = 2.35 to 4.51, 95% CI [1.40, 3.94] to [2.49, 8.17]) and mitotic DNA damage checkpoint signaling (adj. $P = 2.47e{-}2$, OR = 3.39, 95% CI [1.80, 6.39]).

An NCAE-SLE model for tissue-specific disease biomarker discovery using latent space features

We further hypothesized that epigenetic disease biomarkers could also be detected in the NCAE-compressed feature space. SLE, considered the prototypical autoimmune disorder, presents a characteristic pattern of DNAm alterations in patients' immune cells. We compiled 834 DNAm profiles from six case–control studies that included SLE patients ($n = 476$) and healthy controls ($n = 358$) (Table S1 available online at http://bib.oxfordjournals.org/) across 11 tissues or cell types, and we trained a multi-tissue DNN (NCAE-SLE) for SLE patient-control classification. We compressed these DNAm profiles with the NCAE and used their latent embeddings to train the NCAE-SLE DNN model to separate SLE cases from healthy individuals. To benchmark this embedding-based model, we also constructed a DNN with identical architecture and objective function using the principal components of the beta-value matrix of the DNAm profiles. This PCA-based DNN was trained on the first 128 PCs, aligning with the number of dimensions of the NCAE embeddings.

The NCAE-SLE performed above the PCA-based DNN in the complete multi-tissue set ($AUC_{NCAE} = 0.89$, $acc_{NCAE} = 0.78$ versus $AUC_{PCA} = 0.52$, $acc_{PCA} = 0.51$) and in five out of seven single-tissue sets ($AUC_{NCAE} = 0.70$ to 0.95, $acc_{NCAE} = 0.67$ to 0.87 versus $AUC_{PCA} = 0.33$ to 0.65, $acc_{PCA} = 0.43$ to 0.55) (Figure 7A). The NCAE-SLE showed the best results for DNAm embeddings from T cells ($AUC_{NCAE} = 0.95$, $acc_{NCAE} = 0.87$), PBMCs ($AUC_{NCAE} = 0.93$, $acc_{NCAE} = 0.77$) and monocytes ($AUC_{NCAE} = 0.85$, $acc_{NCAE} = 0.86$). The two single-tissue sets where the PCA-based model achieved higher AUC than the NCAE-SLE model (granulocytes and neutrophils, $AUC_{PCA} = 0.78$ and 1 versus $AUC_{NCAE} = 0.67$ for both, $acc_{PCA} = 0.83$ and 0.43 versus $acc_{NCAE} = 0.67$ and 0.43, respectively) had the lowest sample sizes, which may decrease the statistical robustness compared to the rest of single-tissue sets with higher number of individuals (Table S5 available online at http://bib.oxfordjournals.org/).

Using the trained NCAE-SLE model, we performed a light-up analysis to identify whether CpGs prioritized by the model are linked to known SLE-associated genes retrieved from DisGeNET v7.0 [47]. Since 11 tissues and cell types were present in the SLE data set compilation, tissue- and condition-specific DNAm profiles were recursively perturbed to obtain lists of CpGs ranked by their association to SLE. The effect of CpG perturbations in the output of the model was measured as the absolute change in the predicted probability of a DNAm profile to be assigned as a SLE case or control, depending on the profile. We computed the cumulative enrichment in SLE-associated genes for the genes associated with the top 1000 light-up ranked CpGs per tissue (Table S6 available online at http://bib.oxfordjournals.org/), assessing the significance of the observed enrichments

using a permutation test ($n = 1e4$ permutations). We found statistically significant enrichments in SLE-associated genes in the hypermethylation light-up CpG lists for all the tissue-specific sets (Figure 7B). The top tissue-specific signatures were obtained from naive T cells ($P = 1.75e{-}4$, OR = 1.97, 95% CI [1.43, 2.72]) and whole blood ($P = 1.83e{-}4$, OR = 1.52, 95% CI = [1.23, 1.89]). Furthermore, we observed that a multi-tissue list containing the first 1000 genes by frequency across tissue-specific lists achieved the highest possible enrichment ($P = 1e{-}4$, OR = 1.96, 95% CI [1.43, 2.68]). The subset of 184 genes appearing at least once in every single-tissue NCAE-SLE signature was also highly enriched in SLE genes (Fisher's exact test $P = 2.70e{-}4$, OR = 2.00, 95% CI [1.41, 2.85]).

To identify the biological context of the NCAE-SLE signatures in relation to known pathological mechanisms in SLE, we performed pathway enrichment analyses. Considering the tissue-specific signatures (Figure 7C), we found significant (adj. $P < 0.05$) enrichments in KEGG pathways linked to transcriptional misregulation in cancer, Epstein–Barr virus infection, Th17 cell differentiation, antigen processing and presentation, and Hippo and FoxO signaling pathways, among others. In like manner, they were highly enriched in GO terms such as antigen processing and presentation, morphogenesis-related terms, and T-cell-mediated immunity and differentiation. With regard to the multi-tissue signature, we observed strong enrichments in GO terms linked with muscle tissue development (adj. $P = 1.50e{-}3$ to $4.74e{-}2$, OR = 4.56 to 6.89, 95% CI [2.70, 7.69] to [2.48, 19.11]) and morphogenetic processes (adj. $P = 4.41e{-}3$ to $4.33e{-}2$, OR = 11.16 to 15.18, 95% CI [3.38, 36.88] to [5.90, 39.05]), adaptive immune response regulatory processes (adj. $P = 4.41e{-}3$ to $1.68e{-}2$, OR = 3.43 to 8.39, 95% CI = [1.85, 6.37] to [4.03, 17.48]), TGF-beta (adj. $P = 4.41e{-}3$ to $3.76e{-}2$, OR = 4.03 to 5.03, 95% CI [1.87, 8.71] to [2.70, 9.39]) and T-cell cytokine production (adj. $P = 4.41e{-}3$ to $4.97e{-}2$, OR = 6.78 to 15.18, 95% CI [2.45, 18.82] to [5.90, 39.05]), and mesenchymal cell proliferation (adj. $P = 5.06e{-}3$ to $3.94e{-}2$, OR = 11.93 to 14.07, 95% CI [3.60, 39.56] to [5.49, 36.05]) (Figure 7D).

DISCUSSION

Here, we introduced a deep learning workflow for the identification of NCAEs, which encode a biologically meaningful latent space that can be used for DNAm signature discovery. Assessing the performance of an AE is a non-trivial task since simply comparing the reconstructed data to the original input may not be sufficient to determine the usefulness of the learned representation. Our study aimed at demonstrating whether the interpretable embeddings of a deep AE trained on large DNAm data can capture complex, non-linear relationships of biological relevance that could be used for selecting an NCAE, from which data-driven epigenetic signatures can then be identified. We examined multiple architectures and hyperparameter configurations of deep AEs trained on a multi-tissue DNAm compendium to determine configurations that balanced reconstruction performance and co-localization within the human protein interactome in its encoding. Wider AE models explained DNAm data better than deeper ones, with the slight decrease in performance for subsequent layers probably due to the higher risks of information loss in complex network architectures.

The latent space analysis of the trained AEs revealed that CpGs associated with genes co-localizing in the PPI network were increasingly prioritized alongside model depth, until the third layer. The observed co-localization gradient suggests that different hidden layers encode different biological signals, as shown previously for a deep transcriptomic AE [37]. Furthermore, the
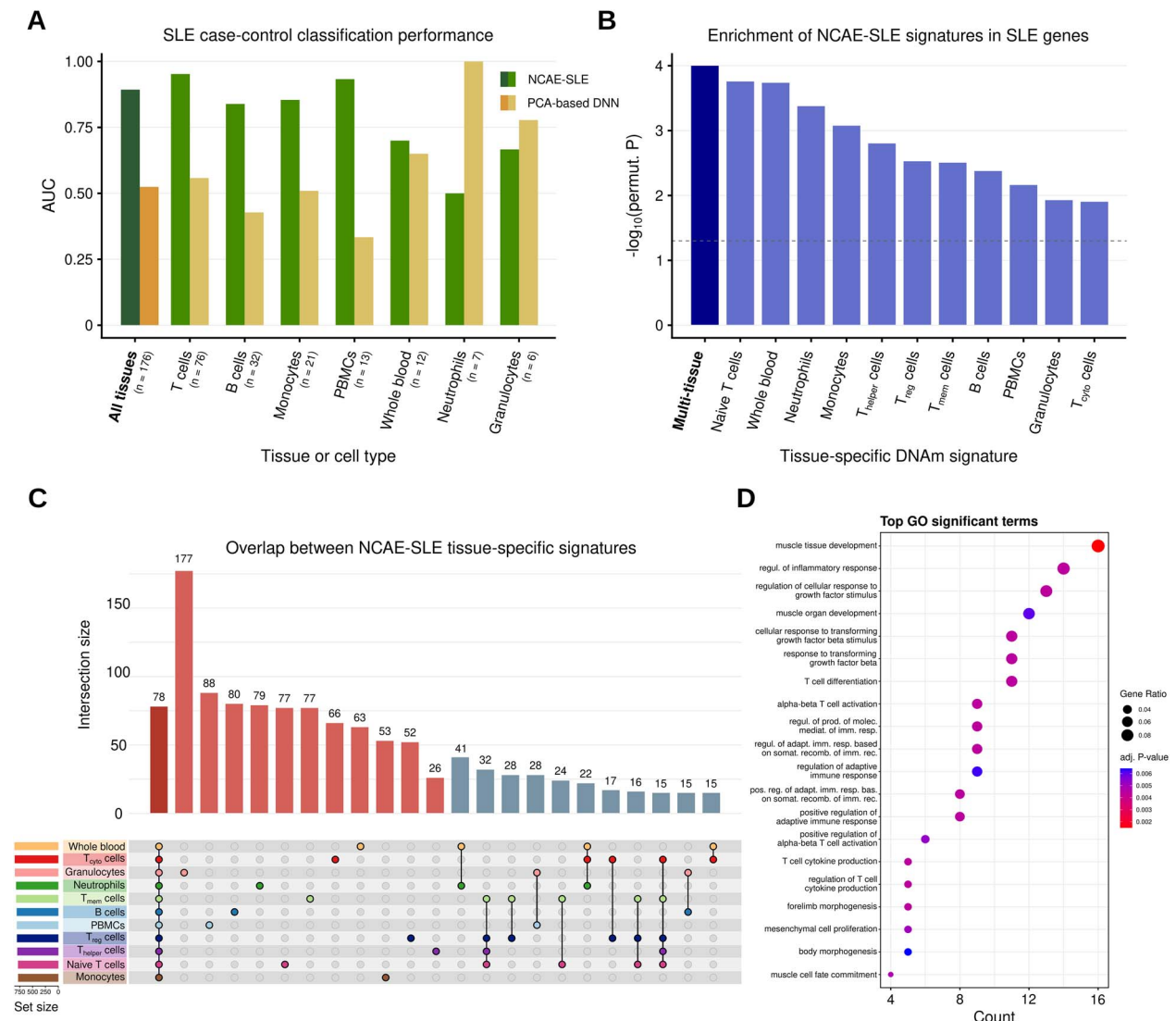
Figure 7. A non-linear DNAm SLE case–control classifier (NCAE-SLE) trained on NCAE embeddings allows the identification of single- and multi-tissue disease-associated methylation signatures. (**A**) Comparison of AUC values for embedding-based NCAE-SLE and PCA-based DNN test set case–control classification performance across tissues or cell types. (**B**) Significance of the enrichment in SLE-associated genes from DisGeNET of top-ranked CpG-associated genes by tissue and cell type from the trained NCAE-SLE. Multi-tissue refers to the top 1000 most frequent genes across the NCAE-SLE rankings. (**C**) Overlap between NCAE-SLE CpG-associated genes from the top light-up ranked 1000 CpGs per tissue or cell type. (**D**) Top significantly enriched GO terms for the multi-tissue SLE DNAm signature.

late emergence of the co-localization signal may indicate that processes of a higher order are modeled before the association to the PPI is decoded, particularly since an increase in central genes was observed in parallel. We selected the deep AE with three hidden layers and 128 hidden nodes per layer as our NCAE. We showed that the latent embeddings of the pre-trained NCAE can be functionalized for transfer-learning-based signature discovery, by first mapping DNAm data to the biologically relevant compressed space, before feeding the new feature set to a concatenated supervised deep ANN. Ranking CpGs by their association to the training objective of this NCAE-DNN allows obtaining task-specific epigenetic signatures.

We validated this approach on three use cases: age estimation, smoking status and SLE patient-control classification. The NCAE-Age performed as well or above Horvath and Hannum DNAm clocks across age groups. Genes linked to CpGs from a list of DNAm age estimators were significantly overrepresented in the NCAE-Age DNAm signature, also associated with pathways

known to regulate key aging mechanisms [48–50]. Regarding the NCAE-Smoke classifier, its performance was on par with the smoking status predictor EpiSmokEr. The NCAE-Smoke signature was validated against other existing DNAm signatures for smoking status, and was strongly enriched in pathways related to smoking effects [51–53]. Thirdly, the NCAE-SLE classifier was used to identify tissue-specific and multi-tissue SLE DNAm signatures, validated using DisGeNET disease–gene associations, and enriched in processes related to the course of autoimmune diseases [54–56].

CONCLUSIONS

Overall, we have demonstrated the utility of a deep learning workflow based on NCAEs that encode a biologically meaningful latent space, which can be used for DNAm signature discovery. Through the evaluation of multiple architectures of AEs trained on a large multi-tissue DNAm dataset compendium, we determined a

configuration that balanced reconstruction performance and coherence with the human protein interactome within its latent space. Our findings indicate that this compressed representation can be functionalized for efficient DNN training and CpG prioritization, leading to the identification of task-specific epigenetic signatures. We showcased the robustness of this approach on three use cases, including age estimation, smoking status and SLE patient-control classification, with NCAE-DNN models outperforming or achieving similar performance as alternative estimators or classifiers. Their DNAm signatures were significantly enriched in biological processes related to the respective condition of interest. In summary, we provide a generally applicable data-driven biomarker discovery workflow for DNAm data that can help pave the way for the development of diagnostic and therapeutic opportunities for a variety of diseases and conditions.

Key Points

- Our study introduces a novel deep learning workflow based on network-coherent autoencoders (NCAEs) that encode a biologically meaningful latent space for efficient DNA methylation signature discovery.
- The latent space analysis of AEs trained in large DNA methylation data reveals that different hidden layers encode different biological signals. We identified the optimal configuration that balances reconstruction performance and coherence with the human protein interactome within the autoencoder embeddings.
- We show that this approach can prioritize CpGs associated with genes co-localizing in the protein–protein interactome, leading to the identification of task-specific epigenetic signatures.
- We validated the workflow on three use cases, including age estimation, smoking status and SLE patient-control classification, with NCAE-based classifiers outperforming or achieving similar performance than available methods.
- This data-driven biomarker discovery workflow presents a promising opportunity for the development of diagnostic and therapeutic applications in a range of diseases and conditions.

ACKNOWLEDGEMENTS

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

FUNDING

DATA AVAILABILITY

The code used in this study is available at the GitLab repository https://gitlab.com/Gustafsson-lab/deep_methylation_ncaes. The normalized DNA methylation data are available for download at the EWAS Data Hub repository (https://ngdc.cncb.ac.cn/ewas/datahub/repository). The trained models, including the deep methylation network-coherent AE and supervised deep neural networks NCAE-Age, NCAE-Smoke and NCAE-SLE, are available at https://figshare.com/projects/071350_network_coherent_autoencoders/155090.

AUTHORS' CONTRIBUTIONS

D.M-E., S.K.D. and M.G. designed the study. D.M-E. performed data processing and computational analyses, which were supervised by M.G. with inputs from S.K.D. and R.J. D.M-E. and M.G. drafted the initial manuscript. All authors contributed to the writing of the manuscript. The authors read and approved the final manuscript.

REFERENCES

1. Barrenäs F, Chavali S, Alves A, *et al.* Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. *Genome Biol* 2012;**13**(6):R46.
2. Choobdar S, Ahsen ME, Crawford J, *et al.* Assessment of network module identification across complex diseases. *Nat Methods* 2019;**16**(9):843–52.
3. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. *Sci Rep* 2018;**8**(1):1362.
4. Hakes L, Pinney JW, Robertson DL, Lovell SC. Protein-protein interaction networks and biology—what's the connection? *Nat Biotechnol* 2008;**26**(1):69–72.
5. Gillis J, Pavlidis P. The impact of multifunctional genes on guilt "by association" analysis. *PloS One* 2011;**6**(2):e17258.
6. Barbosa S, Niebel B, Wolf S, *et al.* A guide to gene regulatory network inference for obtaining predictive solutions: underlying assumptions and fundamental biological and data constraints. *Biosystems* 2018;**174**:37–48.
7. Zhao M, He W, Tang J, *et al.* A comprehensive overview and critical evaluation of gene regulatory network inference technologies. *Brief Bioinform* 2021;**22**(5):bbab009.
8. Krassowski M, Das V, Sahu SK, *et al.* State of the field in multi-omics research: from computational needs to data mining and sharing. *Front Genet* 2020;**11**:610798.
9. Yousefi PD, Suderman M, Langdon R, *et al.* DNA methylation-based predictors of health: applications and statistical considerations. *Nat Rev Genet* 2022;**23**(6):369–83.
10. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet* 2018;**19**(6):371–84.
11. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013;**14**(10):R115.
12. Hannum G, Guinney J, Zhao L, *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* 2013;**49**(2):359–67.
13. Levine ME, Lu AT, Quach A, *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)* 2018;**10**(4):573–91.

14. Zeilinger S, Kühnel B, Klopp N, *et al.* Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PloS One* 2013;**8**(5):e63812.

15. Lee KWK, Pausova Z. Cigarette smoking and DNA methylation. *Front Genet* 2013;**4**:132. https://doi.org/10.3389/fgene.2013.00132.

16. Langdon RJ, Yousefi P, Relton CL, Suderman MJ. Epigenetic modelling of former, current and never smokers. *Clin Epigenetics* 2021;**13**(1):206.

17. Hedrich CM, Mäbert K, Rauen T, Tsokos GC. DNA methylation in systemic lupus erythematosus. *Epigenomics* 2017;**9**(4):505–25.

18. Hedrich CM, Crispin JC, Tsokos GC. Epigenetic regulation of cytokine expression in systemic lupus erythematosus with special focus on T cells. *Autoimmunity* 2014;**47**(4):234–41.

19. Ballestar E. Epigenetic alterations in autoimmune rheumatic diseases. *Nat Rev Rheumatol* 2011;**7**(5):263–71.

20. Teruel M, Sawalha AH. Epigenetic variability in systemic lupus erythematosus: what we learned from genome-wide DNA methylation studies. *Curr Rheumatol Rep* 2017;**19**(6):32.

21. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**(7553):436–44.

22. He K, Zhang X, Ren S, *et al.* Deep residual learning for image recognition. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2016;**2016**:770–8.

23. Brown TB, Mann B, Ryder N, *et al.* Language models are few-shot learners. *Adv Neural Inf Process Syst* 2020;**28**:2020.

24. Zhao Y, Cai H, Zhang Z, *et al.* Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat Commun* 2021;**12**(1):5261.

25. Seninge L, Anastopoulos I, Ding H, Stuart J. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nat Commun* 2021;**12**(1):5684.

26. Svensson V, Gayoso A, Yosef N, Pachter L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* 2020;**36**(11):3418–21.

27. Zuo C, Chen L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Brief Bioinform* 2021;**22**(4):bbaa287.

28. Yin Q, Wang Y, Guan J, Ji G. scIAE: an integrative autoencoder-based ensemble classification framework for single-cell RNA-seq data. *Brief Bioinform* 2022;**23**(1):bbab508.

29. Wang D, Gu J. VASC: dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. *Genom Proteom Bioinform* 2018;**16**(5):320–31.

30. Chen L, Saykin AJ, Yao B, *et al.* Multi-task deep autoencoder to predict Alzheimer's disease progression using temporal DNA methylation data in peripheral blood. *Comput Struct Biotechnol J* 2022;**20**:5761–74.

31. Kmetzsch V, Becker E, Saracino D, *et al.* Disease progression score estimation from multimodal imaging and microRNA data using supervised variational autoencoders. *IEEE J Biomed Health Inform* 2022;**26**(12):6024–35.

32. del Amor R, Colomer A, Monteagudo C, Naranjo V. A deep embedded refined clustering approach for breast cancer distinction based on DNA methylation. *Neural Comput Applic* 2022;**34**(13):10243–55.

33. Wang Z, Wang Y. Extracting a biologically latent space of lung cancer epigenetics with variational autoencoders. *BMC Bioinform* 2019;**20**(S18):568.

34. Ward MD, Zimmerman MI, Meller A, *et al.* Deep learning the structural determinants of protein biochemical properties by comparing structural ensembles with DiffNets. *Nat Commun* 2021;**12**(1):3023.

35. Zhang X, Wang X, Shivashankar GV, Uhler C. Graph-based autoencoder integrates spatial transcriptomics with chromatin images and identifies joint biomarkers for Alzheimer's disease. *Nat Commun* 2022;**13**(1):7480.

36. Kinalis S, Nielsen FC, Winther O, Bagger FO. Deconvolution of autoencoders to learn biological regulatory modules from single cell mRNA sequencing data. *BMC Bioinformatics* 2019;**20**(1):379.

37. Dwivedi SK, Tjärnberg A, Tegnér J, Gustafsson M. Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder. *Nat Commun* 2020;**11**(1):856.

38. Levy JJ, Titus AJ, Petersen CL, *et al.* MethylNet: an automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinform* 2020;**21**(1):108.

39. Choi Y, Li R, Quon G. Interpretable deep generative models for genomics. *bioRxiv* 2022. https://doi.org/10.1101/2021.09.15.460498.

40. Xiong Z, Li M, Ma Y, *et al.* GMQN: a reference-based method for correcting batch effects and probe bias in HumanMethylation BeadChip. *Front Genet* 2022;**12**:810985. https://doi.org/10.3389/fgene.2021.810985.

41. Subramanian A, Tamayo P, Mootha VK, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;**102**(43):15545–50.

42. Xiong Z, Li M, Yang F, *et al.* EWAS data hub: a resource of DNA methylation array data and metadata. *Nucleic Acids Res* 2020;**48**(D1):D890–5.

43. Szklarczyk D, Gable AL, Lyon D, *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**(D1):D607–13.

44. Takeshima H, Yamada H, Ushijima T. Cancer epigenetics: aberrant DNA methylation in cancer diagnosis and treatment. In Dammaco F, Silvestris F (Eds.), *Oncogenomics: From Basic Research to Precision Medicine*. Academic Press, London, 2018, 65–76.

45. Bergsma T, Rogaeva E. DNA methylation clocks and their predictive capacity for aging phenotypes and Healthspan. *Neurosci Insights* 2020;**15**:263310552094222.

46. Bollepalli S, Korhonen T, Kaprio J, *et al.* EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. *Epigenomics* 2019;**11**(13):1469–86.

47. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 2020;**48**(D1):D845–55.

48. Méndez-Pertuz M, Martínez P, Blanco-Aparicio C, *et al.* Modulation of telomere protection by the PI3K/AKT pathway. *Nat Commun* 2017;**8**(1):1278.

49. Long HZ, Cheng Y, Zhou ZW, *et al.* PI3K/AKT signal pathway: a target of natural products in the prevention and treatment of Alzheimer's disease and Parkinson's disease. *Front Pharmacol* 2021;**12**:648636. https://doi.org/10.3389/fphar.2021.648636.

50. Hu HH, Cao G, Wu XQ, *et al.* Wnt signaling pathway in aging-related tissue fibrosis and therapies. *Ageing Res Rev* 2020;**60**:101063.

51. Künzi L, Holt GE. Cigarette smoke activates the parthanatos pathway of cell death in human bronchial epithelial cells. *Cell Death Dis* 2019;**5**(1):127.

52. Pfeifer GP, Denissenko MF, Olivier M, *et al.* Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* 2002;**21**(48):7435–51.

53. Gridelli C, Rossi A, Carbone DP, *et al.* Non-small-cell lung cancer. *Nat Rev Dis Primers* 2015;**1**(1):15009.

54. Bentham J, Morris DL, Cunninghame Graham DS, *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat Genet* 2015;**47**(12):1457–64.

55. Sang A, Danhorn T, Peterson JN, *et al.* Innate and adaptive signals enhance differentiation and expansion of dual-antibody autoreactive B cells in lupus. *Nat Commun* 2018;**9**(1): 3973.

56. Li MO, Sanjabi S, Flavell RAA. Transforming growth factor-β controls development, homeostasis, and tolerance of T cells by regulatory T cell-dependent and -independent mechanisms. *Immunity* 2006;**25**(3):455–71.