mEthAE: an Explainable AutoEncoder for methylation data

Sonja Katz[1,2*], Vitor A.P. Martins dos Santos[1,3], Edoardo Saccenti[2], Gennady V. Roshchupkin[4,5]

[1]LifeGlimmer GmbH, Berlin, Germany.
[2]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen WE, The Netherlands.
[3]Laboratory of Bioprocess Engineering, Wageningen University & Research, Wageningen WE, The Netherlands.
[4]Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands.
[5]Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands.

sonja.katz@wur.nl

*Corresponding author

1 Abstract

Despite the wealth of knowledge generated through epigenome-wide association studies our understanding of the relationships of CpG sites is still limited, as analysis of DNA methylation data remains difficult due its high dimensionality. To combat this problem, deep learning algorithms, such as autoencoders, are increasingly applied to capture the complex patterns and reduce dimensionality into latent space. We believe that the way an autoencoder groups together CpGs in its latent dimensions has biological meaning and might reveal novel insights regarding the relationship of CpGs. Therefore, in this work, we propose a chromosome-wise autoencoder for interpretable dimensionality reduction of methylation data (mEthAE). Our framework shows an impressive reduction in dimensions of up to 400-fold compared to the provided input, without compromising on reconstruction accuracy or predictive power in the latent space. Through our perturbation-based interpretability approach we revealed groups of CpGs which are highly connected across all latent dimensions (global CpGs) and were significantly more often reported in EWAS studies, indicating our interpretability method can successfully identify CpGs with biological relevance. In an attempt to gain a deeper understanding of the relationship between individual CpG sites, we focused on interpreting individual latent features and found that CpGs connected to a common feature do not share biological associations, correlation patterns, or are located in close proximity on the chromosome. We conclude that while there is evidence that the autoencoder does not group CpGs randomly, the logic behind the observed CpG relationships can not be delineated easily. With regards to the analyses done in this work, we believe that the autoencoder groups CpGs according to long range non-linear interaction patterns that lack characterisation in the current epigenetic research landscape.

2 Introduction

Epigenetics comprises all changes in gene expression that are independent from DNA sequence alterations. DNA methylation, the most extensively studied epigenetic mechanism [1], involves the addition of a methylation group to a nucleotide, typically the cytosine residue of a cytosine-guanine (CpG) dinucleotide. Epigenome-wide association studies (EWAS) of high-throughput sequencing data have uncovered strong methylation signatures in response to extrinsic factors, such as environmental changes, as well as intrinsic factors, such as stress, proving the reversible conversion between methylated (hypermethylated) and de-methylated (hypomethylated) CpG regions to be highly dynamic [2–6].

EWAS aim at examining the effect of epigenetic alterations on genome function by identifying DNA methylation variants statistically significant associated with phenotypes of interests. Using general linear models EWAS identify differentially methylated positions by conducting pairwise comparison of mean CpG methylation values between different phenotypes [7]. Despite the wealth of knowledge

generated through EWAS [8], such univariate linear analyses do not take into account the whole spatial relationship pattern of high-dimensional methylomics data. As the genomic distribution of DNA methylation is tightly linked to its gene regulatory functions, gaining a deeper understanding of the relationship between CpG sites is crucial for comprehending how epigenetic modifications influence gene expression and thus ultimately have an impact on phenotypes, including disease [9].

One promising approach to overcome the limitations of EWAS in studying the relationship between CpGs is using autoencoders to model genome-wide DNA methylation data. Autoencoders, and variations of it, are versatile deep learning frameworks capable of non-linear dimensionality reduction [10–12], clustering [12], data generation and imputation [13–15], and performing classification and regression tasks [16, 17]. However, the application of autoencoders for whole genome methylation data has proven to be challenging due to their sensitivity to hyperparameter selection [18] and high computational cost [19]. Additionally, autoencoders, as other deep learning frameworks, are inherently not interpretable. Various approaches have been developed to approximate what an autoencoder learns. Most commonly, this involves visualisation of the latent dimension, revealing possible clusters or regions of interest [20–22]. While autoencoders are frequently being applied to DNA methylation data [10, 11, 16] little work has been conducted on interpreting individual latent features and exploring, for example, which CpGs share a relation through common latent features.

We hypothesise that the way an autoencoder groups together CpGs in its latent dimensions has biological meaning and might reveal novel insights regarding the relationship of CpGs. To explore this, we propose a chromosome-wise autoencoder framework for interpretable dimensionality reduction of methylation data (mEthAE). In an attempt to validate the performance of our autoencoder, we conduct supervised prediction of age and sex from the latent embedding. Our interpretability pipeline, which is based on latent feature perturbations, yields groups of related CpGs at the latent-feature specific (local), as well as embedding-wide (global) level, which allows us to study CpG relationships at different granularities. In an attempt to put the obtained local and global CpG groups into a biological context, we compare them to EWAS findings, assess their power to predict phentoypes (age), analyse their genomic location, associated biological pathways, correlation patterns, and genomic distances.

Bettering our understanding of the relationship between CpGs could, in the long term, contribute to our understanding of epigenetic mechanisms and have important implications for the development of novel therapeutic strategies for various diseases, as well as for the identification of biomarkers that can help diagnose and predict disease risk.

3 Methods

3.1 Datasets

The data used derived from a continuous ageing study, publicly obtainable from the Gene Expression Omnibus (GEO) under accession number GSE87571. This study by Johansson et al. assessed the methylome of the white blood cells of 732 participants, ranging in age from 14 to 94 years old, using the Illumina Infinium 450K Human Methylation Beadchip [23]. For comparison to the EWAS studies, all metadata and results of the EWAS catalogue [24] were downloaded (version 2022-8-18). For analyses, only studies with results from measuring whole blood samples were included.

3.2 Preprocessing

We preprocessed the raw data using the *PyMethylProcess* pipeline, following the steps described in the original *PyMethylProcess* publication [25]. This yielded a total of 300.000 DNA methylation values, called beta values, which are continuous variables between 0 and 1, representing the ratio of methylated versus unmethylated probes for the same locus.

3.3 Computational framework

3.3.1 Autoencoder

To reduce the dimensions of the DNA methylation data, our study utilised an autoencoder network, which is an unsupervised neural network consisting of an encoder and a decoder part (see Figure 1). While the encoder maps the high dimensional input data into a lower dimensional latent embedding,

the decoder attempts to reconstruct the original input from said embedding. We trained the network through a mean squared error (MSE) loss function which quantifies the error between the original input and reconstruction. Prior to training models, the data set was divided into 70% training, 20% testing, and 10% validation.

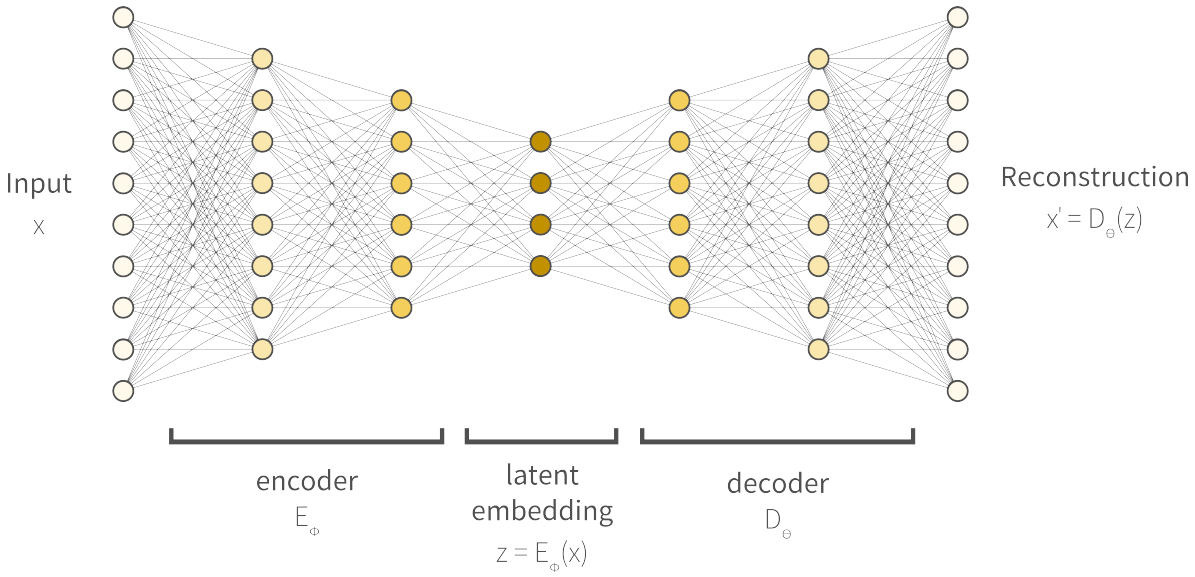


Figure 1: Schematic representation of an autoencoder. The input (x) denotes a set of input CpGs e.g. 29482 CpGs for chromosome 1. The encoder (E_ϕ, parameterised by ϕ) consists of contiguous hidden layers, each with fewer nodes. The latent embedding (z) represents the bottleneck of the autoencoder with the minimum number of nodes. The decoder (D_θ, parameterised by θ) reversely mirrors the layer structure of the encoder, with the final layer featuring the same number of nodes as the input layer as it attempts to reconstruct the original input from the latent embedding.

Architecture optimization To ensure that we do not compromise model design due to memory issues, we split our dataset chromosome-wise, ultimately training and optimising 22 separate models, one for each chromosome. To derive an optimal architecture for our autoencoders, we carried out an exhaustive hyperparameter search assessing the combination of multiple activation functions, layers (batch normalisation, dropout), learning rates, and number of latent features. All tested architectures were based on densely connected autoencoders with 2 hidden layers before the bottleneck (see Figure 1). Between each hidden layer, reduction factors of 3 and 2 were selected, resulting in a total reduction to 6% of the original input size respectively (Supplementary Table 1). Our hyperparameter optimisation strategy comprised two consecutive parts: firstly, a coarse scan to evaluate a broad range of hyperparameters, followed by a fine scan further optimising the best performing settings.

Coarse hyperparameter scan - During the coarse scan we evaluated latent space sizes with a reduction factor of 2, 4, 8, 16, 32, or 64 respective to the last hidden layer. As a regularisation method to avoid over-fitting, dropout layers with probabilities of 0.1, 0.3, and 0.5 were evaluated. Batch normalisation layers [26] were added to the encoder, which was observed to also enhance training speed and stability. Learning rates evaluated ranged from 0.0001 until 0.005. For the activation function, we used the Parametric Rectified Linear Activation Function (PReLU) [27] in all layers except the output layer, which employed a sigmoid activation function, as beta values are finite continuous values between 0 and 1. Models were trained for a fixed period of 500 epochs using an Adam optimiser [28] and a batch size of 64. As loss function mean squared error (MSE) was used. To evaluate and rank the training performance of models the loss on the validation set was monitored as well as the reconstruction accuracies, measured as the Pearson correlation coefficient (ρ) between the original and reconstructed CpGs. To identify an optimal latent space size from the coarse scan, we evaluated and ranked all combinations of hyperparameters for each latent space size. Subsequently, the latent size showing a satisfactory trade-off between validation loss, reconstruction accuracy, and number of dimensions was selected for further refinement in the fine hyperparameter scan.

Fine hyperparameter scan - To design the latent size grid for the fine hyperparameter scan, six latent sizes in close vicinity of the optimal latent space size from the coarse scan were selected. An overview of the final models including hyperparameters can be found in Supplementary Table 1. Subsequently the fine search followed the same strategy as the coarse search, assessing the identical set of hyperparameter combinations. Ultimately, our optimisation procedure yielded one hyperparameter-tuned autoencoder model per chromosome.

Evaluation of reconstruction accuracies To objectively evaluate whether the reconstruction accuracies (measured by the Pearson correlation (ρ) between the input and reconstructed CpGs) of our optimised models are satisfactory, we wanted to compare them to a version of the model without any compression in the latent space (number of latent dimensions = number of input CpGs), as well a version with maximum compression (number of latent dimensions = 1). In theory, the model without any compression should achieve the maximum possible reconstruction accuracy, while the model with only a single latent dimension denotes the minimum achievable accuracy.
All models were built in *pytorch* (version 1.10.2) [29] and trained on using the GPU units RTX 2080 Ti (11GB).

3.3.2 Supervised prediction of age and sex

The amount of biological information encoded by the autoencoders was estimated by trying to predict participants' age and sex from the latent embeddings using a supervised model. Therefore, we used the *scikit-learn* [30] implementation of the Random Forest Regressor and Random Forest Classifier. Hyperparameters of the model were left unchanged, except for the number of trees in the forest, which was raised to 1000. Participants with missing annotation were excluded from the supervised prediction. To avoid data leakage, we used the same train-test splits for training the supervised model as were used for training the autoencoder. Scoring for age regression was done by calculating the coefficient of determination (R^2) from actual and predicted age and for sex using the area under the ROC curve (AUC-ROC).

3.3.3 Interpretability of latent features

To understand how the autoencoder encodes input information (CpGs) in the latent embedding, we implemented a post-hoc interpretation approach based on sequential perturbation of latent features which is outlined in the following sections. A graphical summary of the steps is provided in Figure 2.

Step (1) - Latent feature perturbations: Latent feature perturbations were carried out sequentially one at a time while holding the remaining latent dimensions fixed. The perturbation value was determined to be one standard deviation of the latent feature activation. Subsequently, other forward passes were conducted, this time with the latent feature activation being altered by the addition or subtraction of the perturbation value. To estimate the effect of the introduced perturbation on each CpG, the median absolute difference between the original, unperturbed reconstruction and the new, perturbed reconstruction was calculated. Finally, the differences over all samples and both perturbation values were averaged, resulting in a single perturbation effect value for each CpG. Repeating this procedure for each latent feature yielded a perturbation matrix with a single perturbation effect value for each CpG and latent feature.

Step (2) - Deriving perturbation effect clusters for each latent feature (local CpG groups): For each latent feature we sought to categorise CpGs into different groups based on the perturbation effects ranging from highly, medium, lowly, to not perturbed groups - we refer to those as "local" CpG groups. To categorise CpGs into the aforementioned effect groups, we determined each latent feature's mean perturbation effect and standard deviation and assigned latent-feature specific thresholds (Table 1). Subsequently, for each latent feature the perturbation effect of every CpG was examined and CpGs assigned to an effect group using these thresholds. It is important to highlight that CpG assignments were not unique, as these group assignments were performed for each latent feature sequentially.

Step (3) - Deriving embedding-wide perturbation effect groups (global CpG groups): In addition to the latent feature-wise CpG grouping described above, we also sought to infer CpG groups based on all latent feature perturbations combined. We therefore used the effect groups derived from Step (2) to assess the "global" CpG relationships, which are the perturbation effects of individual

Table 1: Latent-feature specific thresholds for assignment of CpGs to perturbation effect groups. Thresholds were derived through determining each latent feature's mean perturbation effect and its standard deviation.

perturbation effect group	**threshold**	
	lower	upper
local-high	$> 3 * std()$	-
local-medium	$> 2 * std()$	$< 3 * std()$
local-low	$> 1 * std()$	$< 2 * std()$
local-none	-	$< 1 * std()$

CpGs across all latent features. Similar to the thresholding procedure described in Step (2), the mean perturbation effect of all CpG was derived and each CpGs classified into three global relationship groups, termed global CpG connectivitiy groups: globally highly, lowly, and not connected (Table 2).

Table 2: Thresholds for assigning CpGs to global connectivity groups

global connectivity group	**threshold**	
	lower	upper
global-high	$> 1 * std()$	-
global-low	> 0	$< 1 * std()$
global-none	$= 0$	

As a summary, our interpretability approach yields insights on two levels. Firstly, a detailed insight of which CpGs share a relationship by being linked to common individual latent features. Secondly, an estimate of how strongly CpGs are perturbed across the whole embedding, indicating their overall (global) connectivity in the network.

3.4 Characterisation of CpG clusters

3.4.1 Comparison to EWAS findings

A biological validation of the global CpG groups derived by our interpretability pipeline was done through comparison to the EWAS catalogue, calculating how often CpGs from each connectivity group are reported in the EWAS catalogue. To estimate whether the CpGs found by our interpretability method were significantly over- or under-represented with regards to CpGs reported through EWAS, we followed the bootstrapping approach outlined here: (i) all CpGs of the perturbation group under study were checked for significant associations in EWAS studies and the number noted; (ii) a random set of CpGs of the same size as the CpG group under study was drawn and checked for significant associations in EWAS studies; (iii) repeat (ii) for 100 iterations to derive a distribution indicating the chance of a random group of CpGs being mentioned in EWAS; (iv) derive a significance measure by comparing the number of random CpG group configurations that showed the same or higher overlap with EWAS findings than our group of interest.

3.4.2 Investigation of common genomic context, enrichment in biological pathways, correlation, and spatial location

We tested for potential common genetic location or functions of CpGs that are part of the same perturbation group by extracting the genetic information provided in the Illumina Infinium Methylation450K manifest (version 1.2). We applied hypergeometric testing to assess whether CpGs have a common location relative to the nearest gene, association to the same CpG island, location relative to the CpG island, regulatory feature group, DNase I Hypersensitivity Site (DHS), or are located in the same enhancer site. Additionally, using the missMethyl package [31] in R, we tested whether CpGs of the same perturbation groups belong to the same Gene Ontology (GO) terms or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (significance threshold FDR < 0.05). To test for intra group correlation we examined the average Spearman cross-correlation of CpGs in the same perturbation group. To investigate whether CpGs connected to the same perturbation group form LD-like clusters on the chromosome, we assessed their spatial distances by calculating a CpG sparsity ratio. We

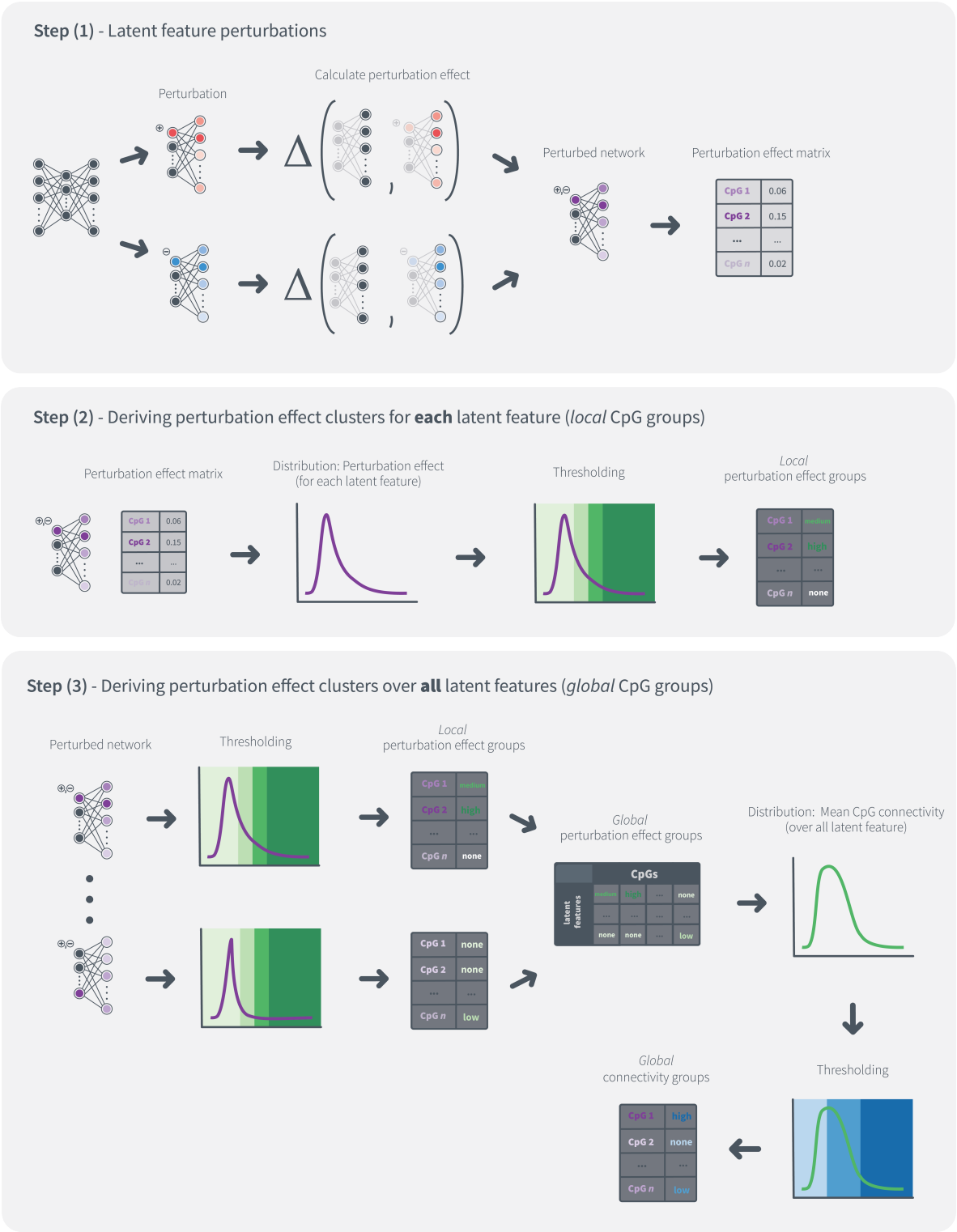


Figure 2: Graphical summary of the interpretability approach utilised in this work which bases on latent features perturbation. An in-depth description of each step can be found in section 3.3.3. An example of the observed perturbation effects can be found in Supplementary Figure 1.

therefore calculated the pairwise distances of CpGs within the same perturbation group and defined a LD-cutoff of 0.25 MB. The sparsity ratio was subsequently calculated by assessing the number of pairwise distances not within the LD-cutoff divided by the total number of distances.

4 Results

4.1 Architecture optimisation

To reduce the dimensionality of genome-wide DNA methylation data we propose mEthAE, a framework based on chromosome-wise autoencoders. We decided to split our dataset per chromosomes and train 22 individual models, as the high number of input CpGs (300.000) led to an explosion of trainable parameters if we attempted to train all chromosomes simultaneously. This parameter explosion disabled us from exploring the use of e.g. multiple hidden layers or larger latent dimensions. With 22 individual models, we were able to thoroughly compare different hyperparameter settings (see section 3.3.1). During optimisation, we focused on achieving a good trade-off between the size of the embedding and performances, measured as validation loss and reconstruction accuracy. An overview of the best models, their latent sizes, and reconstruction accuracies for each chromosome can be found in Table 3. Overall, we found it possible to compress the input information to a factor of up to 400, while simultaneously achieving reconstruction accuracies close to the maximal possible value (Figure 3). We expected the number of optimal latent dimensions to correlate with the size of the input. Interestingly however, although the number of CpGs for each chromosome greatly varies, we observed good performances with latent sizes of 70 across all chromosomes. During optimisation, we often observed a spike in performance around this number of latent dimensions, whereas performances only increased marginally when significantly more latent features were added (Supplementary Figure 2). To validate that our models encode biologically meaningful information, we combined the latent dimensions of all chromosome-wise autoencoders, totalling 1389 latent features, and conducted supervised prediction of age and sex. Age regression yielded a coefficient of determination of 0.82 and a RMSE of 8.72 years, while the classification AUC ROC was 0.95, clearly displaying the combined autoencoders captured biological signal in their embedding (Supplementary Figure 3).

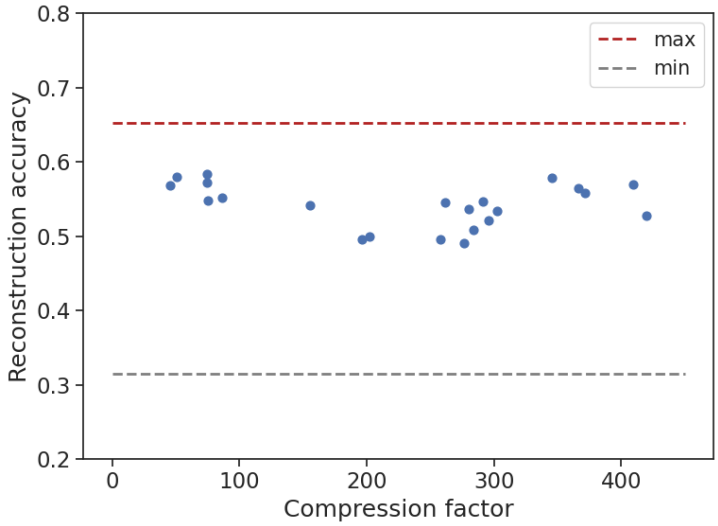


Figure 3: Comparison of reconstruction accuracies in relation to the compression factors for each chromosome; red dashed line: reconstruction of models with a latent size equivalent to the number of input CpGs (no compression), averaged over all chromosomes; grey dashed line: reconstruction of models with a latent size of 1 (maximal compression), averaged over all chromosomes; details on the models with no and maximum compression can be found in Supplementary Table 2.

4.2 Embedding-wide (global) CpG connectivity

By assessing the perturbation effects of CpGs across all latent features, we were able to estimate the embedding-wide (global) CpG relationships. We believe that the more impacted CpGs are by perturbations of the latent embedding, the stronger they are connected to (multiple) latent features. This degree of connectivity between latent features and CpGs - in short the CpG connectivity - reflects

Table 3: Overview of models after hyperparameter optimisation. Number of CpGs: number of input CpGs; Latent Size: number of nodes in latent embedding; Reduction Factor: fold reduction from input size compared to latent size; Reconstruction acuracy: Pearson correlation between input and reconstructed CpGs - indicated for each model by the mean and standard deviation reconstruction accuracy over all CpGs

	Number of CpGs	**Latent Size**	**Reduction Factor**	**Reconstruction accuracy**
chr1	29482	72	409	0.56 ± 0.21
chr2	21984	60	366	0.58 ± 0.21
chr3	15547	45	345	0.58 ± 0.21
chr4	13100	50	262	0.55 ± 0.21
chr5	15965	38	420	0.53 ± 0.22
chr6	23070	78	296	0.52 ± 0.21
chr7	19599	70	280	0.54 ± 0.21
chr8	13704	88	156	0.54 ± 0.22
chr9	6619	88	75	0.55 ± 0.22
chr10	15736	52	303	0.53 ± 0.21
chr11	18570	50	371	0.56 ± 0.21
chr12	15165	52	292	0.55 ± 0.21
chr13	7802	90	87	0.55 ± 0.22
chr14	9417	48	196	0.50 ± 0.23
chr15	9715	48	202	0.50 ± 0.22
chr16	13823	50	276	0.49 ± 0.22
chr17	17044	60	284	0.51 ± 0.23
chr18	3654	72	51	0.58 ± 0.22
chr19	15475	60	258	0.50 ± 0.22
chr20	6567	88	75	0.58 ± 0.21
chr21	2719	60	45	0.57 ± 0.19
chr22	5243	70	75	0.57 ± 0.21
	300000	1389	233 ± 121	0.54 ± 0.21

their importance in the network. CpGs exhibiting similar connectivities form what we refer to as global CpG connectivity groups. We sought to further characterise these global connectivity groups in terms of inter-cluster differences and potential biological relevance.

Table 4: Characteristics of global CpG groups; normalised with regard to the total number of CpGs on each chromosome and average across all chromosomes

Global connectivity group	**Ratio** of associated CpGs	**Total number of CpGs** across all chromosomes
global-high	10.3 ± 1.3%	30949
global-low	60.9 ± 14.6%	181798
global-none	28.8 ± 15.4%	87523

Degree of global connectivity reflects sample and probe heterogeneity

Our interpretability approach (outlined in Section 3.3.3) categorised all input CpGs in three different global connectivity groups, shown in Table 4; details on the observed perturbation effect can be found in Supplementary Figure 1B. Whereas most CpGs show no or a low connectivity (*global-none*, 28.8%; *global-low*, 60.9%), a small percentage of CpGs were highly connected across the embedding (*global-high*, 10.3%). This ratio of highly connected CpGs was similar across all chromosomes, indicated by the low standard deviation, whereas the ratio of lowly and not connected CpGs differed substantially between chromosomes. In-depth comparison of these different global connectivity groups revealed clear differences between them. When inspecting the range of beta values, CpGs in the highly connected

groups showcased a wide range of values, while CpGs in lower global connectivity clusters increasingly converge towards values of 0 or 1 (Figure 4A). The inter-sample heterogeneity, indicated through the standard deviation of beta values across samples, correlated directly with the degree of connectivity, with highly connected CpGs showing the highest standard deviations between samples (Figure 4B).

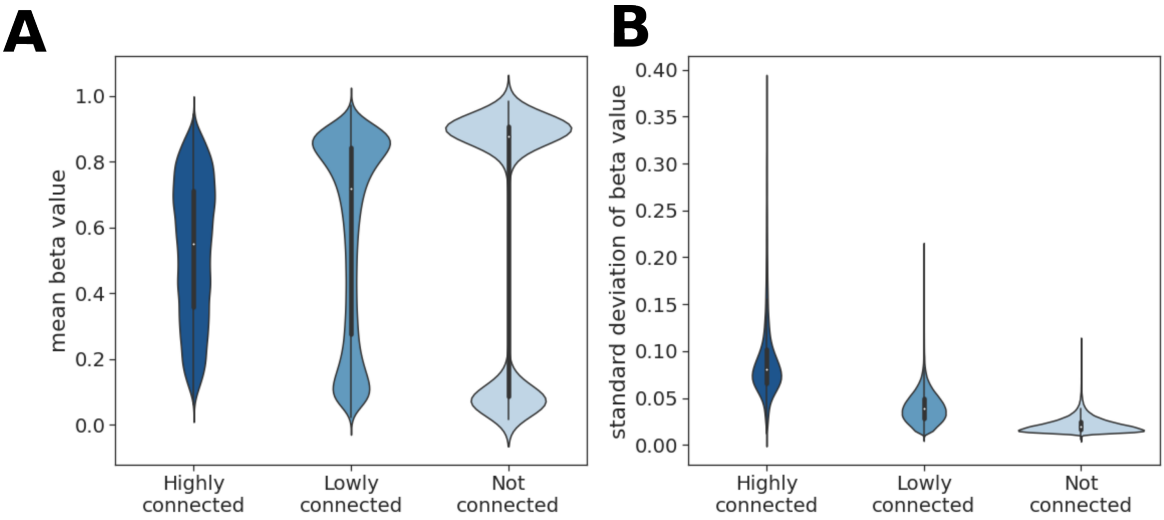


Figure 4: Characterisation of global connectivity groups - (A) average and (B) standard deviation of CpG beta values across all samples for the different clusters

Globally highly connected CpGs show higher biological relevance

To estimate whether a distinction in biological relevance between CpGs from different global connectivity groups can be made we (I) interrogated how frequently CpGs from groups are reported in EWAS studies (Figure 5A) and (II) investigated their predictive power by assessing how accurately participants' age can be predicted using CpGs from the different groups (Figure 5B). Comparison to EWAS studies revealed that CpGs from different groups are reported with differing frequencies (Figure 5A, dashed lines). To put this finding into relation, we employed a bootstrapping strategy (section 3.4.1) (Figure 5A, densities). This combined approach revealed that highly connected CpGs are more often reported in EWAS studies than other groups or in the bootstrapped subset. The observed effect is still present, however less pronounced, with lowly connected CpGs and vanishes completely for not connected CpGs. We additionally conducted supervised age regression (Figure 5B, dashed lines) and evaluated in a similar bootstrapping approach as outlined above, namely through comparison against the prediction accuracies obtained with randomly drawn CpG subsets of the same size (Figure 5B, density). Similar to the picture obtained through the comparison to EWAS findings, it was observed that CpGs with high connectivities are more predictive than other clusters (bootstrap p-value < 0.05), while not connected CpG perform worse than a random subset. As a summary, using two independent approaches, we observed that the degree of connectivity of CpGs to latent features correlates with their biological relevance.

Latent feature-specific (local) CpG groups

In an attempt to gain a deeper understanding of the relationship between CpG sites, we focused on interpreting individual latent features as we hypothesised that the way an autoencoder groups together CpGs in its latent dimensions has biological implications. By subsequently perturbing latent features and categorising CpGs into different groups based on the observed perturbation effects, we are able to delineate how strongly CpGs are linked to respective latent features. This approach yielded groups of high (*local-high*), medium (*local-medium*), low (*local-low*), and not (*local-none*) connected CpGs for each latent feature (see section 3.3.3, Supplementary Figure 1A). An overview of the average number of CpGs assigned to each perturbation group can be found in Table 5. It is evident that a majority

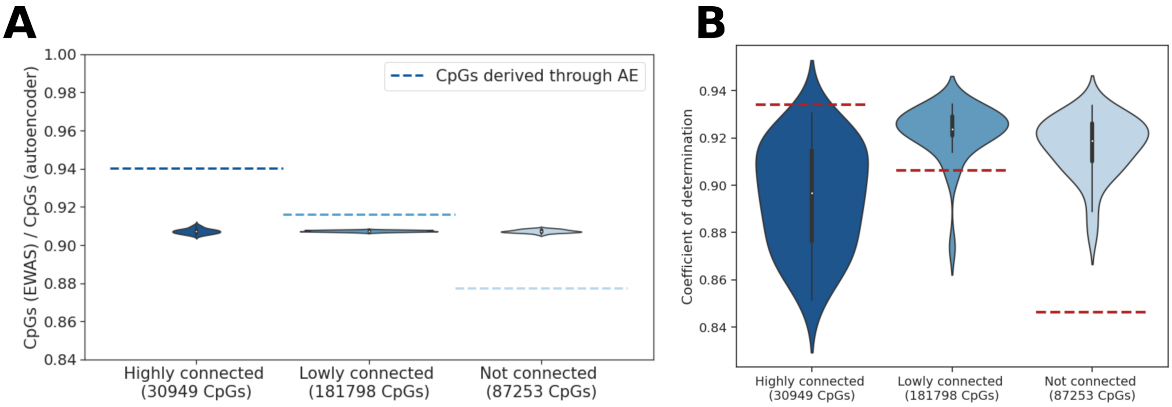


Figure 5: (A) Proportion of CpGs reported as significant in EWAS catalogue. (B) Supervised age prediction using CpGs from different global connectivity groups. Dashed line - CpGs recovered through interpretation approach; distribution - 100 times randomly sampled CpG set of the same size

of CpGs are not affected upon perturbing a latent feature (89.4%); only a small percentage of CpGs show high (1.5%), medium (2.2%), or low (6.8%) effects. The e.g. *local-high* perturbation group only consists of around 208 ± 107 CpG on average per latent feature.

Following the observations from the analyses of global CpG groups, namely that globally highly connected CpGs seem to have biological relevance, we aimed to investigate this further at a latent feature-specific level, hypothesising that CpGs part of the same *local-high* perturbation group are functionally or biologically connected. We thus tested whether these CpGs share genomic context associations, biological pathways, correlation patterns, or are spatially located in close vicinity to each other.

Table 5: Proportion of CpGs in local perturbation groups; averaged across all latent features and chromosomes

Perturbation group	**Total number of CpGs**[1]	**Ratio** of associated CpGs[1]
local-high	208 ± 107	1.5 ± 0.2%
local-medium	296 ± 144	2.2 ± 0.5%
local-low	930 ± 475	6.8 ± 1.8%
local-none	12202 ± 6127	89.4 ± 2.2%

[1]averaged across all latent features

Latent feature-specific CpGs are enriched in specific genomic sites

Inspecting the genomic contexts annotated to the *local-high* CpGs in each latent feature proved that these CpGs are significantly more often located in genomic regions of interest (Table 6). Above 40% of latent features and their associated highly perturbed CpGs can be linked to Enhancer sites or 5'UTR. Between 10-30% were enriched in DNase I Hypersensitivity Sites (DHS), the first exon, in the vicinity of the TSS (TSS1500), or at the 3'UTR. It is to be noted that close to no association to CpG island, shelf or shore regions were found. The most common enriched site, with 62%, were gene bodies.

Biological pathways, correlation patterns, or spatial proximity do not explain CpG grouping

Explanations behind the observed grouping of locally highly perturbed CpGs may include their biological roles, a similar co-correlation pattern, or proximity regarding their genetic locations.

To investigate a potential common biological role of CpGs in the *local-high* perturbation group, we tested for each latent feature if those CpGs are significantly associated with the same Gene Ontology

Table 6: Summary of highly perturbed CpGs for each latent feature significantly associated (p-value < 0.05) with different genetic contexts. Ratio: Number of features normalised over the total number of latent features (n=1389)

	Latent features associated (p-value <0.05)	**Ratio** [%]
Gene body	861	0.62
Enhancer	622	0.45
5'UTR	586	0.42
DHS	383	0.28
1stExon	305	0.22
TSS1500	299	0.22
3'UTR	223	0.16
TSS200	55	0.04
N-Shelf	26	0.02
S-Shelf	19	0.01
N-Shore	17	0.01
Island	13	0.01
S-Shore	9	0.01

(GO) term or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (FDR<0.05) (Supplementary Table 3). Out of 1389 latent features, only 201 showed association to (at least one) GO / KEGG term (14.47%). Notably, we observed big differences between chromosomes - while in most chromosomes not more than 15% of latent features could be linked to biological annotations, the CpG groups associated with every latent feature in chromosome 6 could be matched to a GO term (100%).

Deep learning models organise information following patterns in the data - we therefore examined possible co-correlation patterns between CpGs belonging to the same perturbation group by calculating their Spearman cross-correlation (Supplementary Figure 4). An average cross-correlation of <0.1 was observed, which suggests the absence of prominent correlations of monotonic associations.

Lastly, we hypothesised that CpGs are strongly encoded in a common latent feature due to spatial proximity on the chromosome, forming linkage disequilibrium (LD)-like clusters. To test this, we examined (I) the median pairwise distances of CpGs as well as (II) defined a LD-cutoff of 0.25 megabases (MB) and calculated a sparsity ratio for each latent feature, whereas sparsities of close to 1 denote large spatial distances between CpGs (see section 3.4.2). It is evident that CpGs highly perturbed for the same latent feature are spatially not clustered together on the chromosome, with median pairwise distances of >10 MB even for small chromosomes (Supplementary Figure 5A). This observation is confirmed upon inspecting the sparsity ratios for each chromosome (Figure 6). Even in the case of increasing the LD-cutoff, only a small decrease in the sparsity ratio is observed (Supplementary Figure 5B). An exception to this is chromosome 6 - it shows low median pairwise distances, a low sparsity ratio, as well as high sensitivity to variation in the LD-cutoff.

5 Discussion

The application of deep learning on DNA methylation data is a promising approach having already made an impact in the field, but it remains challenging to delineate the learned biology from models. We therefore set out to build a framework not only capable of efficiently encoding methylation array data, but also enabling interpretation of how CpGs are grouped together in the latent embedding, hoping to shed more light on the decision making of neural networks.

High compression in latent embedding suggests signal sparsity

Although the Illumina 450K array only covers 2% of CpG sites of the human genome [32] and are therefore considered very sparse, our work reveals a high degree of redundancy in the data, demonstrated by the feasibility of compressing the input in a few latent features. We show that just 1389 latent features - as opposed to 300.000 input CpGs, which translates to compression of up to 400-fold - suffice to successfully predict multiple traits (age, sex) and achieve high reconstruction accuracies.

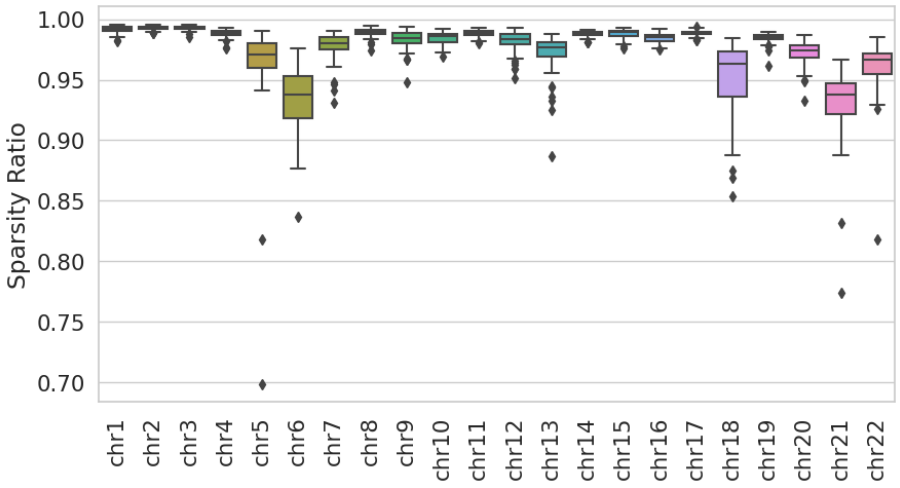


Figure 6: Sparsity ratio of *local-high* CpG perturbation clusters for all latent features with a fixed LD-cutoff of 0.25 MB. A sparsity ratio of 1 denotes distances of CpG sites above the designated cutoff, whereas 0 indicates identical positioning of CpGs.

Our results align with previously published deep learning models, which found comparable compression factors [10, 11, 16]. Our findings encourage a re-evaluation of thresholds applied in EWAS studies; the redundancies observed in the data suggest that the number of true independent tests may be lower than currently assumed for most studies, which would affect the EWAS significance threshold.

Global CpG connectivity groups capture biological diversity

With the aim of interpreting which information is compressed in individual latent features, we developed an interpretability pipeline based on latent feature perturbations. Perturbing or permuting nodes of interest and observing the downstream effects on the network is a widely used attribution method for many data types [33, 34]. Most frequently, input features are altered and effects in the reconstruction observed, e.g. for identifying correlation between input features [35, 36] or for feature selection [37]. However, especially studies from the domain of imaging, also focus on step-wise perturbation of hidden nodes, often referred to as latent traversals [38–40]. A recent publication by Minoura et al. utilised latent traversals to find the correlations of genes of interest with individual latent features, ultimately deducing regulatory programs associated with each latent dimension and used it for interpreting clusters discovered in single-cell sequencing [41]. We took inspiration from all these approaches and applied them to methylation array data. To facilitate a more structured interpretation we divided the obtained perturbations effects in two levels: a local, latent feature-specific level, as well as a global embedding-wide level, obtained by grouping CpGs that show perturbations across multiple latent features.

The analysis of global CpG groups reveals that the stronger they are affected by latent feature perturbation, so the higher their connection to latent features is, the more biological signal they encode. This is demonstrated when comparing different global connectivity groups in terms of methylation beta values - higher connectivity groups tend to include CpGs with beta values of around 0.5. As beta values of around this value typically indicate that sampled cells are highly variable in their methylation pattern, this suggests that the global-high group may contain probes capturing biological variability. The degree of connectivity also correlates with inter-sample variability, revealed by the inspection of beta value standard deviations across samples. Often probes with low standard deviation across samples are not considered biologically meaningful and are discarded during feature selection, indicating that globally highly connected CpGs may capture biological traits differing between samples. We also observe that connectivity is linked to the number of significant reports from EWAS studies, as well as higher predictive values in supervised prediction. It can thus be concluded that our interpretability method grouped together biologically meaningful CpGs from across all chromosomes and does not assemble "random" CpGs into groups.

Local groupings suggest potential long range non-linear CpG interactions

Following the observations that global CpG groups have biological meaning, we attempted to delineate the logic behind local, latent-feature specific groupings by testing if CpGs belonging to the same perturbation group of a latent feature share biological associations, correlation patterns, or spatial locations on the chromosome.

We do not find any of the above postulated assumptions to explain the retrieved CpG groups. CpGs belonging to the same local perturbation group are frequently located in genomic sites such as enhancers, exons or gene bodies. This is of special interest as although initial research mostly focused on the expression regulating role of CpGs located in promoter regions [42], more recent genome-wide studies have proven that CpGs outside of transcription start sites have more impact on gene regulation as previously assumed [43, 44]. For example, while much of the role of methylation in gene bodies is still unclear, it has been shown that increased methylation is positively correlated with gene expression [45, 46]. Also the methylation of the first exon is tightly linked to gene silencing [47]. In this regard, enhancer methylation has received by far the most attention of all genomic regions as it is increasingly recognised to be a key regulatory element [48–50]. While this finding aligns well with results from our embedding-wide analysis, it fails to fully explain the relationships observed in the CpG groupings. Association of latent feature groupings to GO and KEGG terms demonstrates that only around 14% of them can be linked to biological annotations, which does not encourage a general pattern of related biological processes or even common pathways. The speculation that CpGs might group together due to co-expression patterns, with latent features ultimately resembling weighted gene correlation network analysis (WGCNA) modules [51] is disproven due to the low intra-group correlation. With the biological importance behind the genomic distribution of DNA methylation being well known [9], we expected to possibly find short (<2 kb) to middle range (<5 Mb) CpG interactions comparable to an epigenetic equivalent of linkage disequilibrium (LD). Strong associations between nearby CpG sites have been repeatedly reported [52–54, 52] even leading to the development of analytical tools based on the knowledge that spatially close CpGs have the same methylation status [55]. However, our analyses reveal that CpGs of the same perturbation group spread over the length of the chromosome with large median distances between adjacent CpGs (>10-80 Mb), resulting in a very sparse location pattern. An interesting exception to this observation is chromosome 6 - each dimension of the latent embedding can be linked to GO annotations and CpGs of the same perturbation group are clustered more tightly together as indicated by a lower sparsity ratio and median pairwise distance. This is no coincidence, as chromosome 6 contains crucial regulators of immune response such as the human leukocyte antigens (HLA) complex and has repeatedly been shown to be one of the most gene dense and most polymorphic part of the human genome [56, 57].

While there is evidence that the autoencoder does not group CpGs randomly, the logic behind the established CpG relationships can not be delineated easily. Taking into account all analyses done in this study, we suggest that the autoencoder groups CpGs according to long range non-linear interaction patterns that - to the best of our knowledge - lack characterisation in the current research landscape. Few others have attempted to understand dimensionality reduction at such a detailed level, however a comparable work by Titus et al. found it equally challenging to extract novel biological insight for breast cancer from their variational autoencoder [11].

Limitations

During model development, we experienced problems when attempting to train one model on all chromosomes simultaneously, as the high number of optimisable parameters led to GPU memory issues. We therefore decided to split our dataset chromosome-wise and train individual models on each chromosome. This assumes that there are no long range inter-chromosomal interactions of CpGs - although this topic is underrepresented in literature, there is evidence of strong short range correlations between CpGs and methods focusing on co-methylation patterns assume short windows for interaction [55, 58]. While the reconstruction accuracies we achieved are close to the theoretical maximum value - as indicated through models with the same number of latent dimensions as input features - they range around a Pearson correlation of 0.65. This highlights the fact that methylation data is complex and not easy to abstract, even with deep neural networks. We focused on applying a range of conceptually different methods to extract the (biological) meaning behind the CpG groupings uncovered by our interpretability method, however we were unable to test for all possible mechanisms in the frame of

this study. To further research the logic behind latent-feature specific CpG groupings future research should focus on 3D chromatin structures, common meQTL sites, or other epigenetic mechanisms such as histone modifications.

6 Conclusion

In this study, we present chromosome-wise autoencoders using methylation array data that shows impressive compression of size in its embedding of up to 400-fold without compromising on model performances. Furthermore we attempted to delineate how the autoencoders group information using a custom perturbation-based method, investigating embedding-wide as well as latent-feature specific CpG relationships. Our interpretability approach revealed that CpGs highly connected to latent features across the whole embedding displayed a consistently high biological relevance and aligned well with EWAS findings. We additionally aimed to interpret the logic behind CpGs grouped together in individual latent features and found them to be unrelated to common biological associations, correlation patterns, or spatial locations on the chromosome. This suggests the potential presence of long range non-linear CpG interactions that so far lack characterisation in the field of epigenomics.

References

[1] Zachary D. Smith and Alexander Meissner. DNA methylation: Roles in mammalian development. 14(3):204–220.

[2] Elizabeth M. Martin and Rebecca C. Fry. Environmental Influences on the Epigenome: Exposure-Associated DNA Methylation in Human Populations. 39(1):309–333.

[3] Lotte C. Houtepen, Christiaan H. Vinkers, Tania Carrillo-Roa, Marieke Hiemstra, Pol A. van Lier, Wim Meeus, Susan Branje, Christine M. Heim, Charles B. Nemeroff, Jonathan Mill, Leonard C. Schalkwyk, Menno P. Creyghton, René S. Kahn, Marian Joëls, Elisabeth B. Binder, and Marco P. M. Boks. Genome-wide DNA methylation levels and altered cortisol stress reactivity following childhood trauma in humans. 7(1):10967.

[4] Roby Joehanes, Allan C. Just, Riccardo E. Marioni, Luke C. Pilling, Lindsay M. Reynolds, Pooja R. Mandaviya, Weihua Guan, Tao Xu, Cathy E. Elks, Stella Aslibekyan, Hortensia Moreno-Macias, Jennifer A. Smith, Jennifer A. Brody, Radhika Dhingra, Paul Yousefi, James S. Pankow, Sonja Kunze, Sonia H. Shah, Allan F. McRae, Kurt Lohman, Jin Sha, Devin M. Absher, Luigi Ferrucci, Wei Zhao, Ellen W. Demerath, Jan Bressler, Megan L. Grove, Tianxiao Huan, Chunyu Liu, Michael M. Mendelson, Chen Yao, Douglas P. Kiel, Annette Peters, Rui Wang-Sattler, Peter M. Visscher, Naomi R. Wray, John M. Starr, Jingzhong Ding, Carlos J. Rodriguez, Nicholas J. Wareham, Marguerite R. Irvin, Degui Zhi, Myrto Barrdahl, Paolo Vineis, Srikant Ambatipudi, André G. Uitterlinden, Albert Hofman, Joel Schwartz, Elena Colicino, Lifang Hou, Pantel S. Vokonas, Dena G. Hernandez, Andrew B. Singleton, Stefania Bandinelli, Stephen T. Turner, Erin B. Ware, Alicia K. Smith, Torsten Klengel, Elisabeth B. Binder, Bruce M. Psaty, Kent D. Taylor, Sina A. Gharib, Brenton R. Swenson, Liming Liang, Dawn L. DeMeo, George T. O'Connor, Zdenko Herceg, Kerry J. Ressler, Karen N. Conneely, Nona Sotoodehnia, Sharon L. R. Kardia, David Melzer, Andrea A. Baccarelli, Joyce B. J. van Meurs, Isabelle Romieu, Donna K. Arnett, Ken K. Ong, Yongmei Liu, Melanie Waldenberger, Ian J. Deary, Myriam Fornage, Daniel Levy, and Stephanie J. London. Epigenetic Signatures of Cigarette Smoking. 9(5):436–447.

[5] Silvana C. E. Maas, Athina Vidaki, Rory Wilson, Alexander Teumer, Fan Liu, Joyce B. J. van Meurs, André G. Uitterlinden, Dorret I. Boomsma, Eco J. C. de Geus, Gonneke Willemsen, Jenny van Dongen, Carla J. H. van der Kallen, P. Eline Slagboom, Marian Beekman, Diana van Heemst, Leonard H. van den Berg, BIOS Consortium, Liesbeth Duijts, Vincent W. V. Jaddoe, Karl-Heinz Ladwig, Sonja Kunze, Annette Peters, M. Arfan Ikram, Hans J. Grabe, Janine F. Felix, Melanie Waldenberger, Oscar H. Franco, Mohsen Ghanbari, and Manfred Kayser. Validated inference of smoking habits from blood with a finite DNA methylation marker set. 34(11):1055–1074.

[6] Rosa H. Mulder, Alexander Neumann, Charlotte A. M. Cecil, Esther Walton, Lotte C. Houtepen, Andrew J. Simpkin, Jolien Rijlaarsdam, Bastiaan T. Heijmans, Tom R. Gaunt, Janine F. Felix, Vincent W. V. Jaddoe, Marian J. Bakermans-Kranenburg, Henning Tiemeier, Caroline L. Relton, Marinus H. van IJzendoorn, and Matthew Suderman. Epigenome-wide change and variation in DNA methylation in childhood: Trajectories from birth to late adolescence. 30(1):119–134.

[7] Maria Pia Campagna, Alexandre Xavier, Jeannette Lechner-Scott, Vicky Maltby, Rodney J. Scott, Helmut Butzkueven, Vilija G. Jokubaitis, and Rodney A. Lea. Epigenome-wide association studies: Current knowledge, strategies and recommendations. 13(1):214.

[8] Mengwei Li, Dong Zou, Zhaohua Li, Ran Gao, Jian Sang, Yuansheng Zhang, Rujiao Li, Lin Xia, Tao Zhang, Guangyi Niu, Yiming Bao, and Zhang Zhang. EWAS Atlas: A curated knowledgebase of epigenome-wide association studies. 47(D1):D983–D988.

[9] Peter W. Laird. Principles and challenges of genome-wide DNA methylation analysis. 11(3):191–203.

[10] Gregory P. Way and Casey S. Greene. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. 23:80–91.

[11] Alexander J. Titus, Owen M. Wilkins, Carly A. Bobak, and Brock C. Christensen. Unsupervised deep learning with variational autoencoders applied to breast tumor genome-wide DNA methylation data with biologic feature extraction.

[12] Alice R. Owens, Caitríona E. McInerney, Kevin M. Prise, Darragh G. McArt, and Anna Jurek-Loughrey. Novel deep learning-based solution for identification of prognostic subgroups in liver cancer (Hepatocellular carcinoma). 22:563.

[13] Y. L. Qiu, H. Zheng, and O. Gevaert. Genomic data imputation with variational auto-encoders. 9(8).

[14] Joungmin Choi and Heejoon Chae. methCancer-gen: A DNA methylome dataset generator for user-specified cancer type based on conditional variational autoencoder. 21:181.

[15] Lucas Seninge, Ioannis Anastopoulos, Hongxu Ding, and Joshua Stuart. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. 12(1):5684.

[16] Joshua J. Levy, Alexander J. Titus, Curtis L. Petersen, Youdinghuan Chen, Lucas A. Salas, and Brock C. Christensen. MethylNet: An automated and modular deep learning approach for DNA methylation analysis. 21(1):108.

[17] Laura Macías-García, María Martínez-Ballesteros, José María Luna-Romera, José M. García-Heredia, Jorge García-Gutiérrez, and José C. Riquelme-Santos. Autoencoded DNA methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance. 110:101976.

[18] Mostafa Eltager, Tamim Abdelaal, Mohammed Charrout, Ahmed Mahfouz, Marcel J.T. Reinders, and Stavros Makrodimitris. Benchmarking Variational AutoEncoders on cancer transcriptomics data.

[19] Qiwen Hu and Casey S. Greene. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. 24:362–373.

[20] C J Battey, Gabrielle C Coffing, and Andrew D Kern. Visualizing population structure with variational autoencoders. 11(1):jkaa036.

[21] Yongin Choi, Ruoxin Li, and Gerald Quon. Interpretable deep generative models for genomics.

[22] Luca Alessandri, Francesca Cordero, Marco Beccuti, Nicola Licheri, Maddalena Arigoni, Martina Olivero, Maria Flavia Di Renzo, Anna Sapino, and Raffaele Calogero. Sparsely-connected autoencoder (SCA) for single cell RNAseq data mining. 7:1.

[23] Asa Johansson, Stefan Enroth, and Ulf Gyllensten. Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. 8(6):e67378.

[24] Thomas Battram, Paul Yousefi, Gemma Crawford, Claire Prince, Mahsa Sheikhali Babaei, Gemma Sharp, Charlie Hatcher, María Jesús Vega-Salas, Sahar Khodabakhsh, Oliver Whitehurst, Ryan Langdon, Luke Mahoney, Hannah R. Elliott, Giulia Mancano, Matthew A. Lee, Sarah H. Watkins, Abigail C. Lay, Gibran Hemani, Tom R. Gaunt, Caroline L. Relton, James R. Staley, and Matthew Suderman. The EWAS Catalog: A database of epigenome-wide association studies. 7:41.

[25] Joshua J Levy, Alexander J Titus, Lucas A Salas, and Brock C Christensen. PyMethylProcess—convenient high-throughput preprocessing workflow for DNA methylation data. 35(24):5379–5381.

[26] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.

[28] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization.

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[30] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. Scikit-learn: Machine Learning in Python. page 6.

[31] Belinda Phipson, Jovana Maksimovic, and Alicia Oshlack. missMethyl: An R package for analyzing data from Illumina's HumanMethylation450 platform. 32(2):286–288.

[32] Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M. Le, David Delano, Lu Zhang, Gary P. Schroth, Kevin L. Gunderson, Jian-Bing Fan, and Richard Shen. High density DNA methylation array with single CpG site resolution. 98(4):288–295.

[33] Y. Zhang, P. Tino, A. Leonardis, and K. Tang. A Survey on Neural Network Interpretability. 5(5):726–742.

[34] Hai Shu and Hongtu Zhu. Sensitivity Analysis of Deep Neural Networks. 33(01):4943–4950.

[35] Tianwei Yu. AIME: Autoencoder-based integrative multi-omics data embedding that allows for confounder adjustments. 18(1):e1009826.

[36] Rasmus Magnusson, Jesper N. Tegnér, and Mika Gustafsson. Deep neural network prediction of genome-wide transcriptome signatures – beyond the Black-box. 8(1):1–8.

[37] Xiao Li, Jie Ma, Ling Leng, Mingfei Han, Mansheng Li, Fuchu He, and Yunping Zhu. MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. 13:806842.

[38] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE.

[39] Jack Klys, Jake Snell, and Richard Zemel. Learning Latent Subspaces in Variational Autoencoders. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[40] Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. How Important Is a Neuron?

[41] Kodai Minoura, Ko Abe, Hyunha Nam, Hiroyoshi Nishikawa, and Teppei Shimamura. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. 1(5):100071.

[42] Peter A. Jones. Functions of DNA methylation: Islands, start sites, gene bodies and beyond. 13(7):484–492.

[43] Athma A. Pai, Jordana T. Bell, John C. Marioni, Jonathan K. Pritchard, and Yoav Gilad. A Genome-Wide Study of DNA Methylation Patterns and Gene Expression Levels in Multiple Human and Chimpanzee Tissues. 7(2):e1001316.

[44] Peng Yi, Xiongfeng Xu, Jiawei Yao, and Bo Qiu. Analysis of mRNA Expression and DNA Methylation Datasets According to the Genomic Distribution of CpG Sites in Osteoarthritis. 12:618803.

[45] Shasha Zhou, Yihang Shen, Min Zheng, Linlin Wang, Raymond Che, Wanning Hu, and Pin Li. DNA methylation of METTL7A gene body regulates its transcriptional level in thyroid cancer. 8(21):34652–34660.

[46] Aurélie Teissandier and Déborah Bourc'his. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. 36(11):1471–1473.

[47] Fabienne Brenet, Michelle Moh, Patricia Funk, Erika Feierstein, Agnes J. Viale, Nicholas D. Socci, and Joseph M. Scandura. DNA Methylation of the First Exon Is Tightly Linked to Transcriptional Silencing. 6(1):e14524.

[48] Dvir Aran and Asaf Hellman. DNA methylation of transcriptional enhancers and cancer predisposition. 154(1):11–13.

[49] Allegra Angeloni and Ozren Bogdanovic. Enhancer DNA methylation: Implications for gene regulation. 63(6):707–715.

[50] Pier-Luc Clermont, Abhijit Parolia, Hui Hsuan Liu, and Cheryl D. Helgason. DNA methylation at enhancer regions: Novel avenues for epigenetic biomarker development. 21(2):430–446.

[51] Peter Langfelder and Steve Horvath. WGCNA: An R package for weighted correlation network analysis. 9(1):559.

[52] Weiwei Zhang, Tim D Spector, Panos Deloukas, Jordana T Bell, and Barbara E Engelhardt. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. 16(1):14.

[53] Jordana T. Bell, Athma A. Pai, Joseph K. Pickrell, Daniel J. Gaffney, Roger Pique-Regi, Jacob F. Degner, Yoav Gilad, and Jonathan K. Pritchard. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. 12(1):R10.

[54] Florian Eckhardt, Joern Lewin, Rene Cortese, Vardhman K. Rakyan, John Attwood, Matthias Burger, John Burton, Tony V. Cox, Rob Davies, Thomas A. Down, Carolina Haefliger, Roger Horton, Kevin Howe, David K. Jackson, Jan Kunde, Christoph Koenig, Jennifer Liddle, David Niblett, Thomas Otto, Roger Pettett, Stefanie Seemann, Christian Thompson, Tony West, Jane Rogers, Alex Olek, Kurt Berlin, and Stephan Beck. DNA methylation profiling of human chromosomes 6, 20 and 22. 38(12):1378–1385.

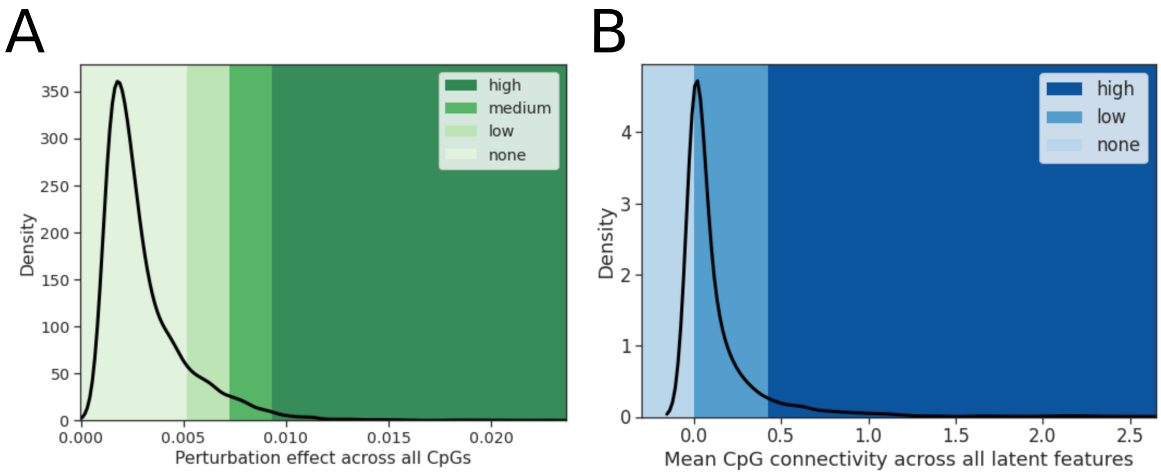
[55] Evan Gatev, Nicole Gladish, Sara Mostafavi, and Michael S Kobor. CoMeBack: DNA methylation array data analysis for co-methylated regions. 36(9):2675–2683.

[56] A. E. Kennedy, U. Ozbek, and M. T. Dorak. What has GWAS done for HLA and disease associations? 44(5):195–211.

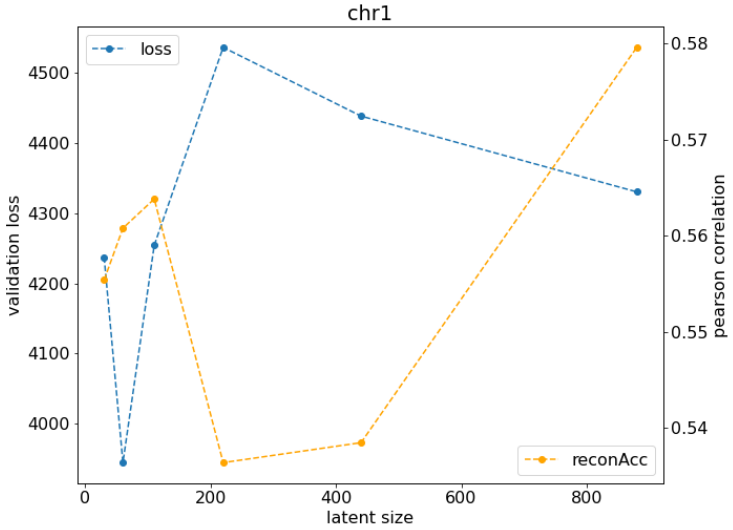
[57] Tao Xie, Lee Rowen, Begoña Aguado, Mary Ellen Ahearn, Anup Madan, Shizhen Qin, R. Duncan Campbell, and Leroy Hood. Analysis of the Gene-Dense Major Histocompatibility Complex Class III Region and Its Comparison to Mouse. 13(12):2621–2636.

[58] Charles E. Mordaunt, Julia S. Mouat, Rebecca J. Schmidt, and Janine M. LaSalle. Comethyl: A network-based methylome approach to investigate the multivariate nature of health and disease. page 2021.07.14.452385.

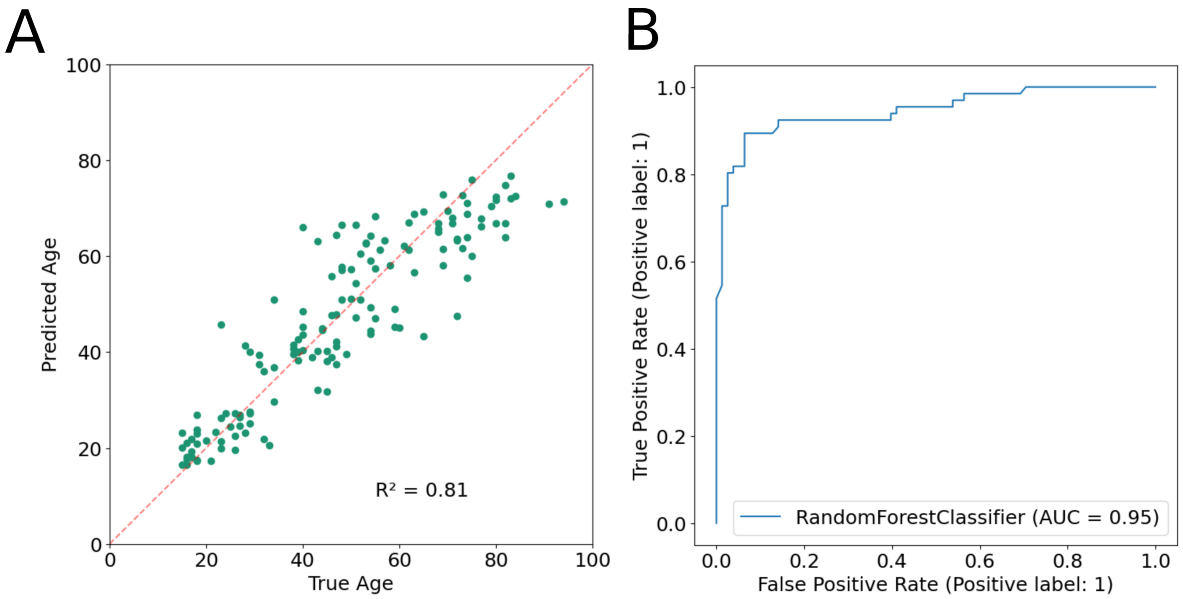
Supplementary Material



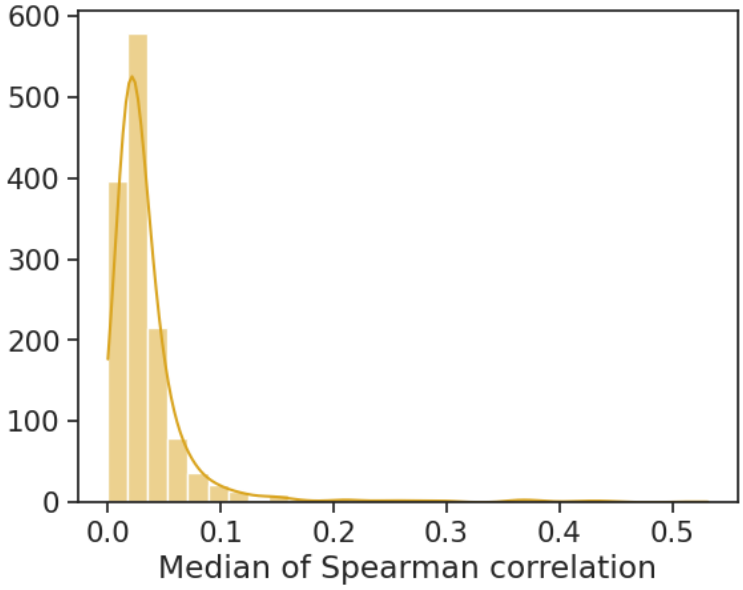
Supplementary Figure 1: Example of observed perturbation effects using chromosome 22;(A) (Local) perturbation effect across all CpGs for a random latent feature. Colours indicate the applied thresholds for cluster assignment, for details see Table 1. (B) (Global) CpG connectivities across all latent features. Colours indicate the applied thresholds for cluster assignment, for details see Table 2



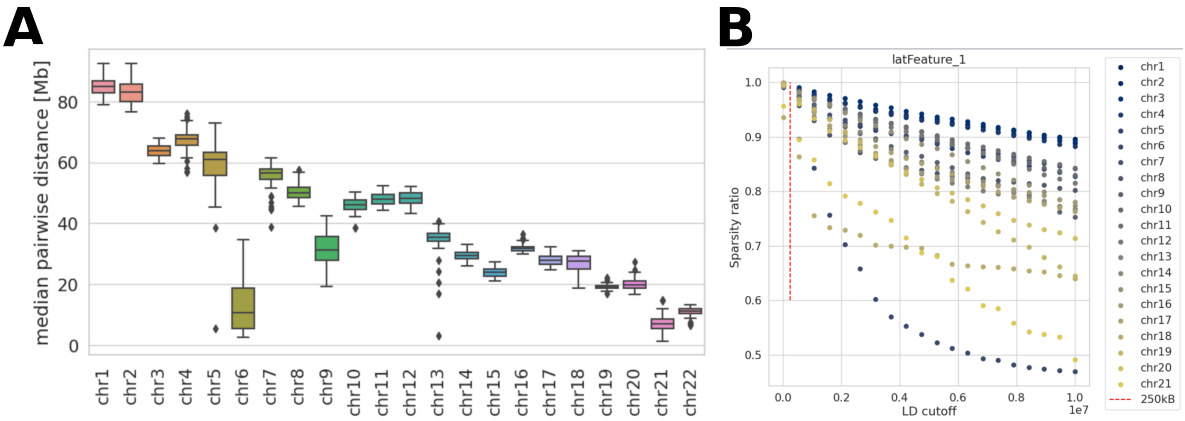
Supplementary Figure 2: Coarse hyperparameter scan results of chromosome 1 (exemplary for other chromosomes); validation loss measured as mean squared error (MSE) loss between input and reconstruction; reconstruction accuracy measured as Pearson correlation between input and reconstruction, averaged over all CpGs



Supplementary Figure 3: Supervised predictions using the combined latent embeddings of all chromosomes, (A) age regression, R^2: coefficient of determination; (B) AUC-ROC curve of classification of participants' sex, AUC: Area under the curve



Supplementary Figure 4: Median spearman cross-correlation over all latent features (n=1389)



Supplementary Figure 5: (A) Median pairwise distances between local-high CpGs of the same latent feature; all latent features per chromosome shown. (B) Effect of the variation of LD-cutoff values on the sparsity ratio for each chromosome. One latent feature is shown as representative for all latent features.

Supplementary Table 1: Details on architecture of all chromosome-wise autoencoders after hyperparameter optimisation

	Number of CpGs	Encoder topology	Latent size	Learning rate	Dropout	Compression factor
chr1	29482	8840, 1770	72	0.0001	0.1	409
chr2	21984	6600, 1320	60	0.0001	0.1	366
chr3	15547	4660, 930	45	0.0001	0.1	345
chr4	13100	3930, 790	50	0.0001	0.1	262
chr5	15965	4790, 960	38	0.0001	0.1	420
chr6	23070	6920, 1380	78	0.0005	0.1	296
chr7	19599	5880, 1180	70	0.0001	0.1	280
chr8	13704	4110, 820	88	0.0005	0.3	156
chr9	6619	1990, 400	88	0.0005	0.3	75
chr10	15736	4720, 940	52	0.0001	0.1	303
chr11	18570	5570, 1110	50	0.0001	0.1	371
chr12	15165	4550, 910	52	0.0001	0.1	292
chr13	7802	2340, 470	90	0.0005	0.3	87
chr14	9417	2830, 570	48	0.0001	0.1	196
chr15	9715	2910, 580	48	0.0001	0.1	202
chr16	13823	4150, 830	50	0.0001	0.1	276
chr17	17044	5110, 1020	60	0.0001	0.1	284
chr18	3654	1100, 220	72	0.0005	0.3	51
chr19	15475	4640, 930	60	0.0001	0.1	258
chr20	6567	1970, 390	88	0.0005	0.3	75
chr21	2719	820, 160	60	0.0005	0.1	45
chr22	5243	1570, 310	70	0.0005	0.3	75

Supplementary Table 2: Reconstruction accuracies (measured by Pearson correlation) of models with a latent size equivalent to the number of input CpGs (no compression) and of models with a latent size of 1 (maximal compression)

	No compression	**Maximal compression**
chr1	0.64 ± 0.18	0.33 ± 0.24
chr2	0.65 ± 0.18	0.32 ± 0.25
chr3	0.67 ± 0.17	0.33 ± 0.25
chr4	0.64 ± 0.18	0.33 ± 0.25
chr5	0.64 ± 0.18	0.32 ± 0.25
chr6	0.65 ± 0.18	0.29 ± 0.25
chr7	0.64 ± 0.18	0.31 ± 0.23
chr8	0.65 ± 0.18	0.32 ± 0.24
chr9	0.65 ± 0.18	0.33 ± 0.24
chr10	0.64 ± 0.18	0.31 ± 0.26
chr11	0.66 ± 0.17	0.32 ± 0.24
chr12	0.65 ± 0.18	0.31 ± 0.23
chr13	0.65 ± 0.18	0.30 ± 0.24
chr14	0.65 ± 0.18	0.33 ± 0.24
chr15	0.65 ± 0.18	0.32 ± 0.25
chr16	0.64 ± 0.17	0.30 ± 0.23
chr17	0.67 ± 0.17	0.31 ± 0.25
chr18	0.68 ± 0.18	0.33 ± 0.27
chr19	0.65 ± 0.18	0.30 ± 0.23
chr20	0.66 ± 0.17	0.32 ± 0.26
chr21	0.69 ± 0.17	0.31 ± 0.24
chr22	0.66 ± 0.17	0.31 ± 0.25
average	0.65 ± 0.18	0.32 ± 0.25

Supplementary Table 3: Details on the number of latent features significantly (FDR <0.05) associated with GO or KEGG terms per latent feature

	Total number of latent features	**Latent features associated with at least one GO / KEGG term** (FDR <0.05)	**Percentage** [%]
chr1	72	10	13.9
chr2	60	3	5
chr3	45	2	4.4
chr4	50	1	2
chr5	38	34	89.5
chr6	78	78	100
chr7	70	0	0
chr8	88	7	8.0
chr9	88	1	1.1
chr10	52	0	0
chr11	50	12	24
chr12	52	7	13.5
chr13	90	4	4.4
chr14	48	6	12.5
chr15	48	0	0
chr16	50	6	12
chr17	60	9	15
chr18	72	0	0
chr19	60	19	31.7
chr20	88	2	2.3
chr21	60	0	0
chr22	70	0	0
	1389	**201**	**14.5**