MethylNet: A Modular Deep Learning Approach to Methylation Prediction

Joshua J. Levy[1,2,*], Alexander J. Titus[3], Curtis L. Petersen[1,2,4], Youdinghuan Chen[1,2], Lucas A. Salas[2], Brock C. Christensen[2,5]

Author affiliations:

1. Program in Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756
2. Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756
3. Office of the Under Secretary of Defense for Research & Engineering, Department of Defense, Washington, DC, USA
4. The Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, NH, 03756
5. Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756

* To whom correspondence should be addressed. Contact: joshua.j.levy.gr@dartmouth.edu

Keywords: Deep Learning, Methylation, High Performance Computing, Workflow Automation

Abstract

Background: The use of deep learning in analyses of DNA methylation data is beginning to emerge and distill non-linear relationships among high-dimensional data features. However, a generalized and user-friendly approach for execution, training, and interpreting deep learning models for methylation data is lacking.

Results: We introduce and demonstrate the robust performance of *MethylNet* on downstream tasks of DNA methylation analysis, including cell-type deconvolution, pan-cancer classification, and subject age prediction. We interrogate the learned features from a pan-cancer classification to show high fidelity clustering of cancer subtypes, and compare the importance assigned to CpGs for the age and cell-type analyses to demonstrate concordance with expected biology.

Conclusions: Our findings demonstrate high accuracy of end-to-end deep learning methods on methylation prediction tasks. Together, our results highlight the promise of future steps to use transfer learning, hyperparameter optimization and feature interpretations on DNA methylation data.

Introduction

Deep learning has emerged as a widely applicable modeling technique for a broad range of applications through the use of artificial neural networks (ANN) [1]. Recently, the accessibility of large datasets, graphics processing units (GPUs) and unsupervised generative techniques have made these approaches more accurate, tractable, and relevant for the analysis of molecular data [2–7].

DNA methylation (DNAm) is the addition of a methyl group to cytosine that does not alter the DNA sequence and occurs in the context of cytosine-guanine dinucleotides (CpG). Methylated regions of DNA (hypermethylated), are associated with condensed chromatin, and when present near gene promoters, repression of transcription. Unmethylated regions of DNA (hypomethylated), are associated with open chromatin states and permissive to gene transcription. DNAm patterns are associated with cell-type-specific gene expression programs, and alterations to DNAm have been associated with aging and environmental exposures [8]. Further, it is well-established that DNAm alterations contribute to development and progression of cancer. The hypermethylation of tumor suppressing genes and the hypomethylation of oncogenes can lead to pathogenesis and poor prognosis. Affordable array-based genome-scale approaches to measure DNAm have potentiated Epigenome Wide Association Studies (EWAS) for testing associations of DNAm with phenotypes, exposures, and states of human health and disease. Because DNAm patterns are cell-type specific, EWAS often account for potential confounding from variation in biospecimen cell composition using reference-based, or reference-free approaches to infer cell type proportions [9–12].

Measuring genome-wide DNAm in large numbers of specimens typically uses microarray-based technologies such as the Illumina HumanMethylation450 (450K) and HumanMethylationEPIC (850K) [13] arrays, which yield an approximation to the proportion of DNA copies that are methylated at each specific cytosine locus, and are reported as beta

values. Preprocessing pipelines such as *PyMethylProcess* have simplified derivation and storage of methylation beta values in accessible data formats [14]. The scope of features from DNAm arrays is 20-50-fold higher than that of RNA-sequencing data sets that return normalized read counts for each gene. Though DNAm data can have a similar scope of features as genotyping array data sets, DNAm data are continuous, not categorical. Together, these facets of DNAm data sets pose challenges to analyses such as handling multi-collinearity and correcting for multiple hypothesis testing. To address these challenges, many downstream EWAS analyses have focused on reducing the dimensions into a rich feature set to associate with outcomes. By limiting the number of features through dimensionality reduction and feature selection, analyses become more computationally tractable and the burden of correcting for multiple comparisons is reduced.

An important advancement to methylation-based deep learning analyses was the application of Variational Auto-encoders (VAE). Initial deep learning approaches for DNAm data focused on estimating methylation status and imputation, performing classification and regression tasks, and performing embeddings of CpG methylation states to extract biologically meaningful lower-dimensional features [15–20,20–22]. VAEs embed the methylation profiles in a way that represents the original data with high fidelity while revealing nuances [4,5,23]. Thereafter, researchers attempted to develop similar frameworks for extracting features for a downstream prediction tasks and identify meaningful relationships revealed by VAE latent representations [24]. However, VAE models are sensitive to the selection of hyperparameters [25] and have not been optimized for synthetic data generation, latent space exploration, and prediction tasks. Many auto-encoder approaches represent the data using an encoder, and then utilize a non-neural network model (e.g. support vector machine) to finalize the predictions. Presently, to the best of our knowledge there is no end-to-end training approach that both extracts biologically meaningful features through latent encoding and performs predictions using the derived features. Further, existing frameworks do not output predictions for multi-target regression tasks, such as cell-type deconvolution and subject age prediction.

Here, we leverage deep learning latent space regression and classification tasks through the development of a modular framework that is highly accessible to epigenetic researchers. *MethylNet* is a modular user-friendly deep learning framework for EWAS tasks with automation that leverages preprocessing pipelines. To discover important CpGs for each prediction we use the SHAP (SHapley Additive ExPlanation) approach [26]. We highlight *MethylNet* as an easy-to-use command line interface that utilizes automation to scale, optimize, and simplify deep learning methylation tasks. *MethylNet*'s capabilities are showcased here with cell-type deconvolution, pan-cancer subtype classification, and age regression. These analyses will pave the path for more robust deep learning prediction models for methylation data. Coupled with *PyMethylProcess* [14], we expect the *MethylNet* framework to enable rapid production-scale research and development in the deep learning epigenetic space.

Results

Our approach can be summarized as follows, all of which can executed for any prediction task using a few simple commands:

1. Pre-train deep learning prediction models using variational auto-encoders. The layers of the encoder are used to extract biologically meaningful features. These neural network layers are used to embed the data and extract features.

2. Include prediction layers downstream of the encoder. Fine-tune the model's prediction and feature extraction layers end-to-end for the tasks of multi-output regression and classification tasks. Training these layers to optimize the neural network for prediction tasks.
3. Perform autonomous hyperparameter scans to optimize the model parameters for the above two tasks while generating rich visualizations of the data.
4. Determine the contribution of the CpGs to each prediction on varying degrees of granularity through Shapley Feature Attribution methods.

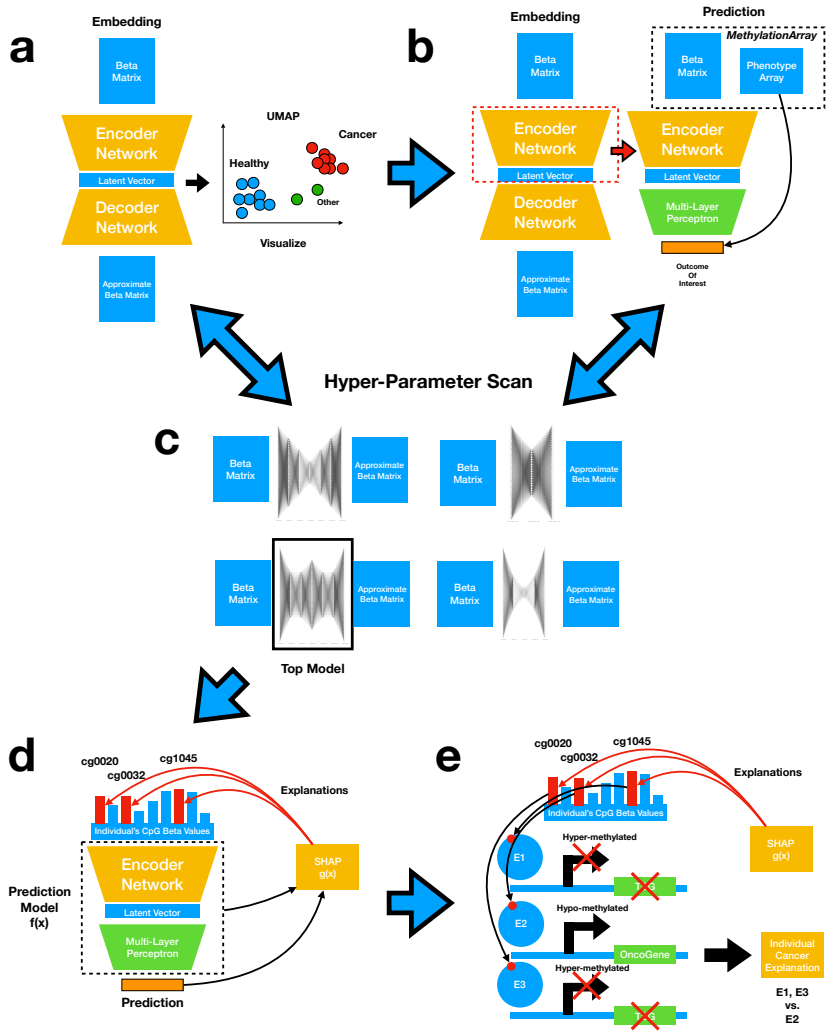
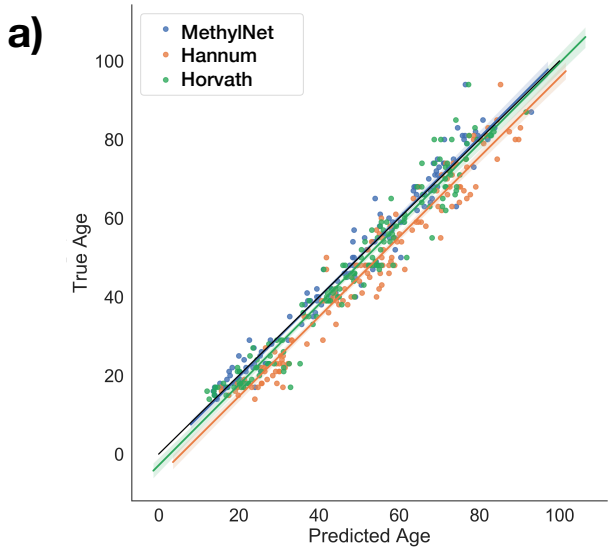


Figure 1: Step-by-Step Description of the Modular Framework: a) Train Feature Extraction Network Using Variational Auto-Encoders; b) Fine-Tune Encoder for Prediction Tasks; c) Perform Hyperparameter Scans for (a) and (b); d) Identify Contributing CpGs; e) Interpret the CpGs.

MethylNet is implemented as a command-line tool that allows for deep learning predictions on methylation data for embedding, classification and regression tasks. With the specification of one command line option, *MethylNet* can be toggled between regression and classification tasks. This makes the pipeline versatile in handling a wide breadth of problems. Its modular

accessible framework makes it easy to train and produce high-quality results across multiple domains.

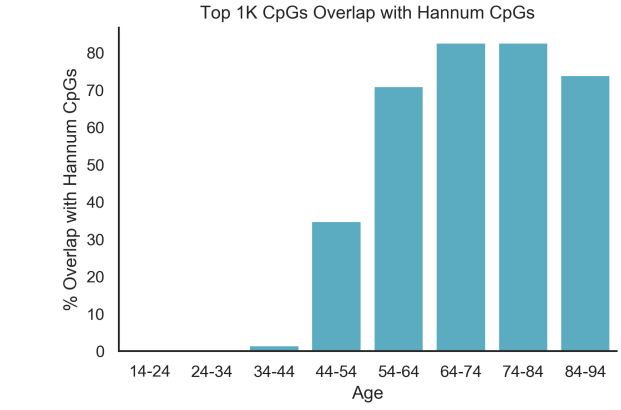
Age Results



b)

Comparison of MethylNet Age Estimates on Test Set to Horvath and Hannum Age Estimators (n=144).

Analysis	R^2 (mean±SE)	Mean Absolute Residual (mean±SE)	Explained Variance Proportion (mean±SE)
MethylNet	**0.96±0.0062**	**3.0±0.23**	**0.96±0.0062**
Hannum Clock	0.90±0.013	5.6±0.30	0.95±0.0065
Horvath Clock	0.94±0.0088	3.9±0.27	0.95±0.0078



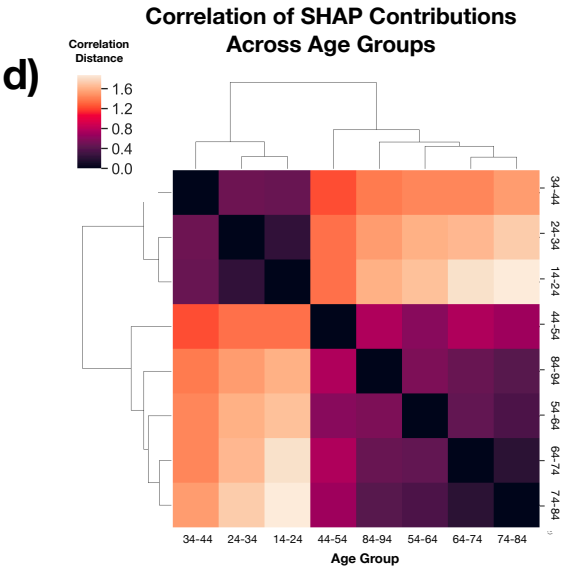


Figure 2: Age Results on Test Set (n=144): a) Age predictions derived using the Horvath, Hannum, and MethylNet estimators are compared to the true age of the individual, the predicted ages are plotted on the x-axis, the actual ages on the y-axis, and a line was fit to the data for each estimator; b) Comparison of MethylNet Age Estimates on Test Set (n=144) to Horvath and Hannum Age Estimators; c) Bar chart depicting the overlap of CpGs important to MethylNet and Hannum age estimators where one thousand CpGs with the highest SHAP scores per 10-year age group are divided by the total number of Hannum CpGs that passed QC; d) Hierarchical clustering using the correlation distance between SHAP CpG scores for age groups across all CpGs. The linkage is found between similar age groups.

MethylNet-predicted age showed excellent concordance with the actual subject age (R^2=0.96, Figure 2a) in the hold-out test set (n=144), and only had 3.0 years mean absolute error (Figure

2b). These results are more accurate than those estimated by the present state of art, Hannum and Horvath clock. The contribution of each CpG to age groups binned by 10-year increments from ages 14 to 94 were measured by Shapley values. The CpGs with the one thousand largest Shapley values for each age group were overlapped with the CpGs of the Hannum clock (Figure 2c). These CpG contributions were compared between age groups using correlation distance, as illustrated in Figure 2d. The connectivity between different age groups' CpG attributions in Figure 2d using hierarchical clustering demonstrates the sharing of important CpGs by similarly aged groups.

We aimed to compare the CpGs highly contributing to age predictions using *MethylNet* and to those calibrated in the Hannum epigenetic clock [27]. The Horvath [28] and Hannum methods used multivariate linear models with elastic net penalization to find a limited set of CpGs strongly associated with age whose degree of methylation that are presumably non-tissue-specific. The CpGs used by the Hannum model were most likely associated with those aged 60-80, the most prevalent ages in the cohort. Since the number of Hannum CpGs rediscovered by *MethylNet* appears to peak around this range, this supports evidence that *MethylNet* is able to recover the defining CpGs of the Hannum cohort.

Cell Type Deconvolution Results

Table 1. Comparison of MethylNet Cell Type Deconvolution Results to EpiDISH Methods.

	Cell Type	R^2 (mean±SE)	Mean Absolute Residual (mean±SE)	Explained Variance (mean±SE)
MethylNet	CD8T	**0.78±0.04**	**0.016±0.0012**	**0.78±0.038**
	CD4T	**0.86±0.018**	**0.014±9.0e-04**	0.88±0.016
	NK	**0.87±0.017**	**0.012±8.5e-04**	**0.87±0.017**
	B Cell	**0.79±0.026**	**0.009±6.3e-04**	0.79±0.025
	Monocytes	0.37±0.067	0.012±7.9e-04	0.38±0.062
	Neutrophil	**0.97±0.0043**	**0.011±7.1e-04**	**0.97±0.0042**
EpiDISH+RPC	CD8T	0.72±0.061	0.019±0.0013	0.76±0.052
	CD4T	0.34±0.091	0.036±0.0014	**0.89±0.018**
	NK	0.024±0.11	0.033±0.0024	0.48±0.049
	B Cell	0.77±0.035	0.01±5.3e-04	**0.93±0.012**
	Monocytes	0.17±0.13	0.015±8.2e-04	**0.64±0.053**
	Neutrophil	0.84±0.025	0.029±0.0013	0.96±0.0073
EpiDISH+Cibersort	CD8T	0.63±0.077	0.023±0.0014	0.75±0.055
	CD4T	0.6±0.058	0.026±0.0014	0.86±0.02
	NK	-0.058±0.12	0.035±0.0025	0.46±0.055
	B Cell	0.76±0.046	0.01±6.0e-04	0.88±0.024
	Monocytes	**0.45±0.089**	**0.012±7.2e-04**	0.54±0.072
	Neutrophil	0.91±0.015	0.02±0.0011	0.96±0.008

n=144, RPC: Robust Partial Correlations

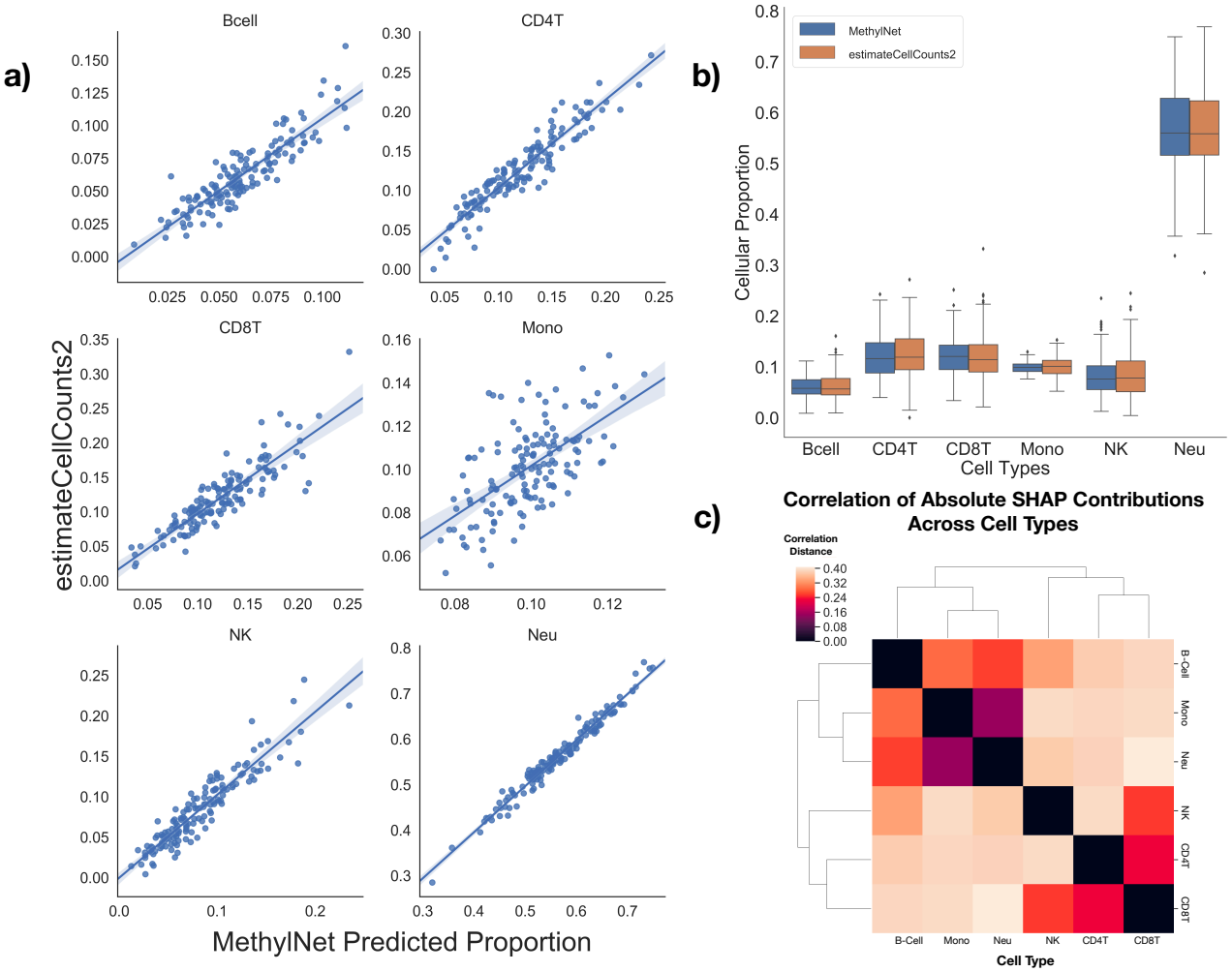


Figure 3: Results on test set (n=144) for cell-type deconvolution: a) For each cell type, the predicted cellular proportion using MethylNet (x-axis) was plotted against the predicted cellular proportion using estimateCellCounts2, which has been found to be a highly accurate measure of cellular proportions and thus serving as the ground truth for comparison, a regression line was fit to the data for each cell type: B-cell, CD4T, CD8T, Monocytes (Mono), NK cells, and Neutrophils (Neu); b) Grouped box plot demonstrating the concordance between the distributions of the MethylNet-estimated proportions of each cell-type and the distributions derived using estimateCellCounts2; c) Hierarchical clustering using the correlation distance between two cell types' SHAP CpG scores across all CpGs. The linkage is found between cell types of similar lineage.

Next, *MethylNet* was tasked with estimating the cell-type proportions for six immune cell-types using the same dataset as supplied for the age analysis. The framework demonstrates exemplary performance on this task, as demonstrated in Table 1. *MethylNet* outperforms all EpiDISH methods in R^2 and mean absolute error across all cell-types save for monocytes (Figure 3a-b, Table 1). Using Shapley attribution, contributions for each of the CpGs for driving the predictions of the cell-types was derived. Figure 3c shows the connectivity of their hierarchical clustering of these CpG attributions.

The hierarchical clustering between the SHAP scores of each of the cell-types is consistent with the known cell lineage, reinforcing that cell lines that have co-evolved similarly share similar driving CpGs that are indicative of their cell-type. Some of the cell-types obtained improved concordance metrics (i.e., R^2) compared to other cell types but had similar absolute errors. This is likely due to the fact that the total range of proportions of monocytes, for instance, from the collected data was small such that these errors could make it difficult to correlate the predicted and true cell type proportions. Alternatively, issues with the purity of the reference monocytes could cause difficulties in calibrating the reference library. A similar overlap test was conducted between the *MethylNet* SHAP CpGs and IDOL-derived DMR CpGs (Supplementals Figure 2). Little overlap was found between the two sets, as only the B-cells were able to capture more than 10% of the IDOL CpGs. This does not indicate that *MethylNet* could not pick up CpGs that are cell-type specific. Rather, it is further indication that models with different objectives differently attribute CpG contributions.

To this point, we still do not know at what point do CpGs, across individuals or larger groupings become statistically significant and thus warrant additional inspection. Some preliminary analysis can be found in the Supplementals Figures 4 and 5. For the Hannum and IDOL analysis, we set this at an arbitrary cutoff value of the top 1000 CpGs per age/cell-type group, but the distribution of these Shapley scores and their fidelity to model predictions is an active area of research [29].

Pan-cancer Prediction Results

a)

Comparison of MethylNet Derived Pan-Cancer Classification of Test Set to UMAP+SVM Method (n=1676).

	Accuracy Score	Recall Score	Precision Score	F1-Score
MethylNet	**0.97±0.0045**	**0.97±0.0045**	**0.97±0.0042**	**0.97±0.0044**
UMAP+SVM	0.84±0.0091	0.84±0.0091	0.814±0.0103	0.82±0.0098

b)

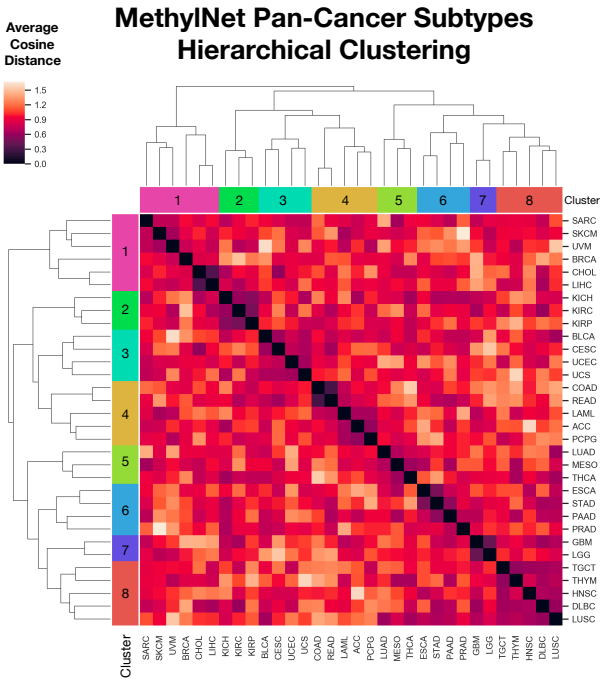


Figure 4: Results on test set for pan-cancer sub-type predictions: a) Comparison of MethylNet derived pan-cancer classification of test set (n=1676) to UMAP+SVM method; b) Hierarchical

clustering of average embedding cosine distance between all pairs of cancer subtypes. Cancer subtypes from both axes are colored by cancer superclasses, derived using the hierarchical clustering method. The clustering of similar MethylNet embeddings is concordant with known biology.

The predictions of 32 cancer subtypes (n=1676) (one removed due to low sample size) across the pan-cancers TCGA cohort yielded 0.97 accuracy, 0.97 precision, 0.97 recall and 0.97 F1-score, averaged across the different subtypes (Figure 4a). These results outperform a support vector machine (SVM)-based classification approach, in which *MethylNet* demonstrated 0.15 increase in F1-score. A breakdown of classification accuracies for each subtype is in the supplemental results (Supplemental Table 2).

The latent profiles derived for pan-cancer subtypes showed clustering with high concordance to biology. Thresholding a hierarchical clustering of the average cosine distance between cancer subtypes from the *MethylNet* derived embeddings (Figure 4b) indicates clustering of the test methylation profiles by eight unsupervised biologically corresponding superclasses. The subtypes that define these larger groupings are concordant with expectations from cancer biology.

Of note are:

- Skin and connective tissue cancers, and bile and liver cancers in Cluster 1.
- All kidney cancers in Cluster 2.
- Bladder, uterine and cervix cancers in Cluster 3.
- Pairing of colon and rectal cancers, both adrenal cancers in Cluster 4.
- A tie between lung adenocarcinoma and mesothelioma in Cluster 5, both of which may develop in similar locations.
- Pairings between stomach and esophagus cancer, and pancreas and prostate cancers in Cluster 6.
- Brain cancers in Cluster 7.
- Thymoma, Diffuse Large B-Cell lymphomas in Cluster 8.
- While the lung cancers were not paired together, they experienced a high degree of embedded similarity. The connectivity between the lung squamous cell cancer and its neighboring types prevented the two cancers from being grouped together.

Taken together, *MethylNet* not only makes highly accurate and robust classification predictions, but also extracts latent features with high fidelity to the actual biology present.

The similarity between some of the subtypes may explain why and how some of the subtypes did not perform as well as others (Supplemental Tables 1 and 3). For instance, we see that 4 KIRC and KIRP cases were conflated with each other. Two cervix cases were predicted to be uterine. There were elevated rates of misclassification between the colon and rectal cancer pairings and Esophageal, head and neck, and stomach cancer pairings. Finally, seven predicted glioblastoma cases were actually low-grade glioma (Supplemental Table 1). Thus, subtypes tended to be misclassified only within each superclass. The exception to this trend was the misclassification of Lung Squamous cell cancers, four of which were predicted to be its adenocarcinoma counterpart, which is consistent with the shared embedding profile. This

more likely reflects similar biology, while misclassifications outside of superclasses may reflect pathological misdiagnoses and technical artifacts. Since *MethylNet* captures and confounds the biology between similar conditions, this presents a unique opportunity to explore similar therapeutic targets and treatments across disease types of similar tissue, within and outside cancer studies. Given the ability of *MethylNet* to capture the differences in the profiles between the cancer subtypes, there is great opportunity for unsupervised clustering and classification of further disease heterogeneity within a particular condition.

For the cancer subtype analysis, we sought to identify concordance between the latent profiles of methylation across cancer types. Because each tumor type has a different baseline DNAm profile for its normal tissues-of-origin and these differences are expected to contribute to the prediction we did not attempt to derive the salient CpGs for each subtype's prediction.

Discussion

Here, we introduce *MethylNet*, a modular deep learning framework that is easy to train, apply, and share. *MethylNet* employs an object-oriented application programming interface (API) and has built-in functionality to easily switch between analyses with respect to embedding, classification, and regression tasks. It has demonstrated the ability to make accurate predictions that conform with expected biology. *MethylNet* extends previous approaches by fine-tuning the feature extractor and adding additional layers for prediction tasks. It also employs a robust hyperparameter search method that optimizes the parameters of the model for generalization to unseen data. The pipeline is flexible to the demands of the user. For instance, if a user only wanted to train a custom machine learning model on the latent features, the data can be extracted before the end-to-end training step. By demonstrating three tasks, age prediction, cell-type deconvolution, and pan-cancer subtype prediction, we present further support of the applicability of VAEs for feature extraction, and more evidence that deep learning presents an opportunity for learning meaningful biology and making accurate predictions from feature-rich molecular data.

Strengths, Limitations, and Future Directions

Interpretation of our high dimensional models still has challenges, partially due to the drawbacks of assigning feature attributions to high dimensional multi-collinear data. While traditional linear models can still be highly predictive, multi-colinearity has the effect of adjusting the coefficients of the predictors such that the results are not as interpretable. Shapley feature attributions are a promising method used to explain predictions estimating complex models with simpler linear ones.

Our age and cell-type analyses were conducted to demonstrate the capabilities of the deep learning tool and models were trained on a relatively small study of blood samples, only a subset of those included in the Horvath framework. Our analyses also only presented predictions across one type of tissue without yet accounting for differences in methylation between cell types. Part of the bottleneck for developing robust cell-type proportion estimation methods is the lack of the availability of ground-truth proportions from flow cytometry or other measures. The *MethylNet* model will also benefit from being able to train off of real cell-type proportions rather than that acquired from the highly accurate *estimateCellCounts2.* More

robust and consistent estimators that address these limitations are the focus of future applications of the *MethylNet* method.

Our analyses refrain from uncovering relationship between the discovered CpGs and functional effects because of the difficulties associated with localizing the effect of a small set of CpGs of interest. Once the salient attributions are found, CpG analyses experience common pitfalls when trying to match CpGs to their nearest gene via the found promoter region. Such analyses may ascribe the CpG's effect in the context of what gene they appear to be regulating. However, genes are regulated at a distance in proximity to their enhancers in the 3D topological space [30,31]. This means that interpretation methods such as gometh may not be suitable for interpreting CpGs identified by *MethylNet.* Ideally, downstream approaches to add biological interpretation would take into account chromosome/genome interaction (e.g. through use of Hi-C data) and genome topological structure/organization. For instance, chromatin state and histone modification analyses as used by ChromHMM and LOLA [32,33] might be more warranted. Some model result interpretation issues may be circumvented by building a deep learning mechanism to better predict gene expression from methylation [34].

An important take-away is that as interpretation methods for these high dimensional data are pioneered, VAE-based deep learning models will likely find CpGs that interact in ways we would not traditionally think about. While the other models were trained on a much smaller set of CpGs, *MethylNet* is able to make its predictions on 200-300K CpGs, capturing complex interactions between a much larger set of CpGs. Crucial next steps should address these interpretability and confounding concerns through feature selection, covariate adjustment and more biologically interpretable informatics methods for CpG interpretation.

Finally, to scale up *MethylNet's* deep learning workflows to production grade as well as incorporate information from Whole Genome and Reduced Representation Bisulfite Sequencing, future renditions may utilize common workflow language (CWL) [35]. In addition, new Bayesian search methods may be employed to better automate the selection of model hyperparameters and automate the construction of the ideal neural network architecture [36,37].

Conclusion

We demonstrate a modular, reproducible, and easy-to-use object-oriented deep learning framework for methylation data: *MethylNet*. We illustrate that *MethylNet* achieves high predictive accuracy across age estimation, cell-type deconvolution, cancer subtype prediction tasks. *MethylNet's* accuracy at these tasks was superior, or at least equivalent to, other methods. We hope that *MethylNet* will be used by the greater biomedical community to rapidly generate and evaluate testable biological hypotheses involving methylation data through a scalable, automated, intuitive, and user-friendly deep learning framework.

Methods

Description of Framework

Here, we present a description of a modular and highly accessible framework for deep learning tasks pertaining to unsupervised embedding, supervised classification and multi-output

regression of DNA methylation (DNAm) data. The *MethylNet* pipeline comprises modules and commands specifically pertaining to embedding, prediction, and interpretation.

First, after preprocessing using *PyMethylProcess*. The dataset is split into training, validation, and testing sets using *train_test_val_split* of the preprocessing pipeline utilities.

Training the Feature Extractor to Embed Data

The embedding module is used to pretrain the final prediction model by using Variational Autoencoders to find unsupervised latent representations of the data. Pre-training is an important part of transfer-learning applications. The knowledge extracted from learning unsupervised representation of the data is used towards learning predictive tasks with a lower data requirement. Data fed into these VAEs pass through an encoder network that serves to compress the data and then this compressed representation is fed into a decoder network that attempts to reconstruct the original dataset while attempting to generate synthetic samples. The model attempts to balance the ability to generate synthetic samples with the ability of the data to be accurately reconstructed. The weight given to generation versus reconstruction can be set as a hyperparameter [38]. Generating synthetic training examples are important for adding noise while training a network for prediction tasks, a component which serves as a form of regularization to make the algorithm more generalizable to real-world data. Synthetic data was not generated using *MethylNet*, but during training, the algorithm samples from the latent distribution of the embedded data to regularize. Nevertheless, the ability to reconstruct the original dataset is important because it governs how latent representations of the data are capturing features that properly describe the underlying signal.

In order to run the embedding module on the input *MethylationArray* training and validation objects, *perform_embedding* is executed via the command line interface. Hyperparameters of the autoencoder model can be scanned via the *launch_hyperparameter_scan* command. This randomly searches a grid of hyper-parameters and randomly generates neural network topologies (number of layers, number of nodes per layer). The complexity (network width and depth), of which can be weighted by the user. The framework stores the results of each training run into logs to find the model with the lowest validation loss (Binary Cross Entropy reconstruction loss plus KL-Loss of the validation set). Alternatively, results from the embedding module can be input into any machine learning algorithm of choice. Embedding results are visualized through interactive 3-D plots by running *transform_plot* from *PyMethylProcess*.

Training for Prediction via Transfer Learning

MethylNet can be used to perform classification, regression, and multi-output regression tasks via the prediction module. The prediction module uses *MLPFinetuneVAE* to fine-tune encoding layers of VAE model while simultaneously training a few appended hidden layers for prediction. The *make_prediction* command is run for these prediction tasks, and hyper parameters such as model complexity and learning rate and schedulers are scanned via the *launch_hyperparameter_scan* module. The final model is chosen if it has the lowest validation loss (Mean Squared Error for Regression, Cross-Entropy for Prediction), and the output model is a snapshot at the epoch that demonstrated the lowest validation loss. The test set is also evaluated immediately after the model is trained using the training set. The results from

MethylNet can be immediately benchmarked and compared for performance to other machine learning algorithms, which can be evaluated using the *general_machine_learning* module from *PyMethylProcess*. Furthermore, ROC Curves and classification resorts can be output using *plot_roc_curve* and *classification_report* and regression reports are generated via *regression_report.* A confusion matrix of misclassifications can be generated from *PyMethylProcess*'s *plot_heatmap*. Finally, the training curves for both the embedding and prediction modules can be visualized using the *plot_training_curve* command.

Interpretation of Results

Predictions from *MethylNet* can be interrogated in two ways. The first approach uses SHAPley feature attribution to assign a contribution score to each CpG based on how much it contributed to the prediction. The second approach compares learned clusters of embeddings of methylation samples (and corresponding subtypes), for biological plausibility.

The SHAPley value interpretations, available using *methylnet-interpret* approximate the more complex neural network model using a linear model for each individual prediction, the coefficients of which are Shapley values. Shapley values represent the contributions of each CpG to the individual predictions. They are produced after the prediction model and test *MethylationArray* are input to the *produce_shapley_data* command, which dumps a *ShapleyData* object into memory. The Shapley coefficients can be averaged by condition to yield summary measures of the importance of each CpG to the coarser category, and the coefficients can be clustered to demonstrate the similarity between methylation subtypes and coarser conditions, which can be compared to known biology.

MethylNet was built using Python 3.6 and utilizes the PyTorch framework to run its deep learning models on GPUs using CUDA, although CPUs are also suppored. The workflow is available as an easily installable command line tool and API via PyPI as *methylnet* and on Docker [39] as *joshualevy44/methylnet*. The Docker image contains a test pipeline that requires one line to run through the hyperparameter training and evaluation of all framework components and can run on your local personal computer in addition to high performance computing. Help documentation, example scripts, and the analysis pipeline are available in the *MethylNet* GitHub repository (https://github.com/Christensen-Lab-Dartmouth/MethylNet).

Description of Experiment

We evaluated our *MethylNet* framework (hyperparameter scan, embedding, fine-tuning predictions, interpretation) using 33 datasets from n=9,312 samples for three different prediction tasks: classification (TCGA pan-cancer subtype prediction), regression (age prediction), and multi-output regression (cell-type deconvolution).

PyMethylProcess was used to preprocess the data, and yielded *MethylationArray* objects that contain a matrix of beta values for each individual and the corresponding phenotype information [14]. The *MethylationArray* data for each of these three experiments were split into 70% training, 20% testing, and 10% validation sets. The training set was used to update the parameters of the model. The validation set was used to terminate training early and choose hyperparameters that would be most generalizable to a test set. The test set was used for final model evaluation and interpretation. More information on model training can be found in the

supplementals. For each score, 95% confidence intervals were computed using a 1000 sample non-parametric bootstrap.

The data for the regression tasks were procured from GSE87571 [40], representing a healthy age group of blood samples from individuals aged 15-95, and preprocessed using *PymethylProcess* to yield 300k CpG features.

First, *MethylNet* was configured for regression tasks and applied to derive sample age estimates. These results were compared to those derived from the Hannum and Horvath clocks using *cgageR*[27,28,41]. The Shapley framework was employed to quantify the importance of the CpGs in making predictions for age across 8 different age groups split by 10-year increments. The CpG importance was compared between the groups through hierarchical clustering to find similarities between the age groups. The one thousand most important CpGs from each group were extracted and overlapped with CpGs defined by the Hannum model to depict the concordance of important CpGs between *MethylNet* and the Hannum model.

For a second task, *MethylNet* was configured for multi-target regression to estimate cell-type proportions. First, *estimateCellCounts2*, using the 450K legacy IDOL optimized library [11], was used to deconvolve the cell-type proportions from each sample to develop ground truth outcomes for training the model. The MethylNet model was trained on the *estimateCellCounts2* estimates of cell-type proportions for six different immune cell-types. *MethylNet* was then compared to results derived from the *EpiDISH* framework [42] using 350 IDOL derived CpGs legacy library from FlowSorted.Blood.EPIC[11]. The importance of each CpG to each cell-type was then quantified through SHAP. These Shapley coefficients were compared using hierarchical clustering. A similar clustering profile would indicate these cell-types share similar driving CpGs, and recovery of the cell-lineage dendrogram would demonstrate concordance with known biology. The one thousand most important CpGs from each cell-type were extracted and overlapped with the IDOL CpGs to inspect if the two models picked up similar cell-type-specific CpGs.

The comparison library based cell type proportions were estimated through the use of FlowSorted.Blood.EPIC and EpiDISH [42,43] R packages. The library used was the IDOL optimized CpGs 450k legacy library which contains 350 CpGs. Two methods; Robust Partial Correlations and Cibersort were implemented through EpiDISH all using the same library.

In the final task, *MethylNet* was used to classify samples to cancer types. The data for the classification task are from 8891 TCGA-acquired samples, representing 32 different cancer types, and preprocessed using *PyMethylProcess* to yield a 200k CpG beta matrix. The features with the highest mean absolute deviation across samples were selected to both limit the computational complexity, memory of model training and capture the highest variation in the data. The highly variable sites are assumed to be more biologically meaningful than the lower variable sites. The *MethylNet* analysis pipeline was conducted on the pan-cancer dataset. The results from *MethylNet* were compared to a popular omics classification approach, a uniform manifold approximation and projection (UMAP) embedding of the samples, followed by support vector machine (SVM) classification. UMAP is an effective way to reduce the dimensionality of the data as well as preserve meaningful local and global structure in the data [44,45]. Both were performed using *PyMethylProcess's general_machine_learning* module, which executed a hyperparameter grid search of the SVM model. Finally, the embeddings of the different cancer subtypes were compared by calculating of the average cosine distance

between clusters in the test samples. These distances were clustered using hierarchical clustering to form larger superclasses of cancer that demonstrate a shared embedding profile.

Abbreviations

450K – HumanMethylation450
850K - HumanMethylationEPIC
ANN – Artificial Neural Networks
CpG – Cytosine-Guanine Dinucleotides
CWL – Common Workflow Language
DNAm – DNA Methylation
EWAS – Epigenome-Wide Association Studies
SHAP – Shapley Additive Feature Explanations
SVM – Support Vector Machine
UMAP - Uniform Manifold Approximation and Projection
VAE – Variational Auto-encoders

Contributions

The conception and design of the study were contributed by JJL and BCC. Implementation, programming, data acquisition, and analyses were by JJL. JJL and BCC wrote the manuscript and all authors contributed to writing and editing of the manuscript. CLP performed the EpiDISH comparisons. AJT, YC, CLP contributed towards refining the analytic plan and direction. AJT, YC, CLP, and JJL tested the pipeline. LAS provided technical support to streamline and debug important aspects of the pipeline.

Funding:

This work was supported by NIH grants R01CA216265, R01DE022772, and P20GM104416 to BCC, a Dartmouth College Neukom Institute for Computational Science CompX award to BCC, and training fellowship support for AJT from T32LM012204. CLP is supported through the Burroughs Wellcome Fund Big Data in the Life Sciences at Dartmouth.

Declarations/Competing Interests

The views expressed in this article are solely those of the authors and do not necessarily represent the views of the DoD or its components.

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

2. Tian, T., Wan, J., Song, Q. & Wei, Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nat. Mach. Intell.* **1**, 191 (2019).

3. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).

4. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **23**, 80–91 (2018).

5. Titus, A. J., Wilkins, O. M., Bobak, C. A. & Christensen, B. C. Unsupervised deep learning with variational autoencoders applied to breast tumor genome-wide DNA methylation data with biologic feature extraction. *bioRxiv* 433763 (2018). doi:10.1101/433763

6. Ching Travers *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).

7. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).

8. Christensen, B. C. *et al.* Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet.* **5**, e1000602 (2009).

9. Titus, A. J., Gallimore, R. M., Salas, L. A. & Christensen, B. C. Cell-type deconvolution from DNA methylation: a review of recent applications. *Hum. Mol. Genet.* **26**, R216–R224 (2017).

10. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

11. Salas, L. A. *et al.* An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol.* **19**, (2018).

12. Houseman, E. A. *et al.* Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics* **17**, 259–259 (2016).
13. Moran, S., Arribas, C. & Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* **8**, 389–399 (2016).
14. Levy, J. J., Titus, A. J., Salas, L. A. & Christensen, B. C. PyMethylProcess - highly parallelized preprocessing for DNA methylation array data. *bioRxiv* 604496 (2019). doi:10.1101/604496
15. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **18**, 67 (2017).
16. Ni, P. *et al.* DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics* doi:10.1093/bioinformatics/btz276
17. Qiu, Y. L., Zheng, H. & Gevaert, O. A deep learning framework for imputing missing values in genomic data. *bioRxiv* 406066 (2018). doi:10.1101/406066
18. Wang, Y. *et al.* Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. *Sci. Rep.* **6**, 19598 (2016).
19. Zeng, H. & Gifford, D. K. Predicting the impact of non-coding variants on DNA methylation. *Nucleic Acids Res.* **45**, e99 (2017).
20. Korfiatis, P. *et al.* Residual Deep Convolutional Neural Network Predicts MGMT Methylation Status. *J. Digit. Imaging* **30**, 622–628 (2017).
21. Yu, H. & Ma, Z. Deep Neural Network for Analysis of DNA Methylation Data. *ArXiv180801359 Q-Bio Stat* (2018).
22. Islam, Md. M., Tian, Y., Cheng, Y., Wang, Y. & Hu, P. A deep neural network based regression model for triglyceride concentrations prediction using epigenome-wide DNA methylation profiles. *BMC Proc.* **12**, (2018).

23. Titus, A. J., Bobak, C. A. & Christensen, B. C. A New Dimension of Breast Cancer Epigenetics - Applications of Variational Autoencoders with DNA Methylation. in 140–145 (2018).

24. Wang, Z. & Wang, Y. Exploring DNA Methylation Data of Lung Cancer Samples with Variational Autoencoders. in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1286–1289 (2018). doi:10.1109/BIBM.2018.8621365

25. Hu, Q. & Greene, C. S. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. in *Biocomputing 2019* 362–373 (WORLD SCIENTIFIC, 2018). doi:10.1142/9789813279827_0033

26. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).

27. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol. Cell* **49**, 359–367 (2013).

28. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).

29. Joseph, A. Shapley regressions: A framework for statistical inference on machine learning models. *ArXiv190304209 Cs Econ Stat* (2019).

30. Nothjunge, S. *et al.* DNA methylation signatures follow preformed chromatin compartments in cardiac myocytes. *Nat. Commun.* **8**, 1667 (2017).

31. Geeleher, P. *et al.* Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics* **29**, 1851–1857 (2013).

32. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2016).

33. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).

34. Wang, Y., Franks, J. M., Whitfield, M. L. & Cheng, C. BioMethyl: an R package for biological interpretation of DNA methylation data. *Bioinformatics* doi:10.1093/bioinformatics/btz137

35. Amstutz, P. *et al.* Common Workflow Language, v1.0. (2016). doi:10.6084/m9.figshare.3115156.v2

36. Tim Head *et al. scikit-optimize/scikit-optimize: v0.5.2*. (Zenodo, 2018). doi:10.5281/zenodo.1207017

37. Kandasamy, K., Neiswanger, W., Schneider, J., Poczos, B. & Xing, E. P. Neural Architecture Search with Bayesian Optimisation and Optimal Transport. in *Advances in Neural Information Processing Systems 31* (eds. Bengio, S. et al.) 2016–2025 (Curran Associates, Inc., 2018).

38. Higgins, I. *et al.* beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. (2016).

39. Boettiger, C. An Introduction to Docker for Reproducible Research. *SIGOPS Oper Syst Rev* **49**, 71–79 (2015).

40. Johansson, Å., Enroth, S. & Gyllensten, U. Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. *PLOS ONE* **8**, e67378 (2013).

41. metamaden/cgageR: version 0.1.0 from GitHub. Available at: https://rdrr.io/github/metamaden/cgageR/. (Accessed: 10th June 2019)

42. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* **18**, (2017).

43. FlowSorted.Blood.EPIC. *Bioconductor* Available at: http://bioconductor.org/packages/FlowSorted.Blood.EPIC/. (Accessed: 20th June 2019)

44. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).

45. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2018).