D-GPM: a deep learning method for gene promoter methylation inference

Xingxin Pan[a,b], Biao Liu[a,b], Xingzhao Wen[c], Yulu Liu[a,b], Xiuqing Zhang[a,b], Shuaicheng Li[c,*]

[a]BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China.

[b]BGI-Shenzhen, Shenzhen, China.

[c]School of Biological Science and Medical Engineering, Southeast University, Nanjing, China.

[d]Department of Computer Science, City University of Hong Kong, Hong Kong

*To whom correspondence should be addressed.

Abstract: Gene promoter methylation plays critical roles in a wide range of biological processes such as transcriptional expression, gene imprinting, X chromosome inactivation, *etc*. Whole genome bisulfite sequencing can generate comprehensive profiling of the gene methylation levels but suffer from high cost. Recent studies partitioned the genes into landmark genes and target genes, and suggested that the landmark gene expressions capture adequate information to reconstruct the target gene expressions. Besides, the methylation level of the promoter is usually negatively correlated with its corresponding gene expression. These inspire us that the methylation level of the promoters could be adequate to reconstruct the promoter methylation level of target genes, and which eventually reduces the cost of promoter methylation profiling. Here, we developed a deep learning model (D-GPM) to predict whole genome promoter methylation level based on the methylation profile of the landmark genes. We benchmarked D-GPM against three machine learning methods, namely linear regression (LR), regression tree (RT) and support vector machine (SVM) based on two criteria: mean absolute deviation (MAE) and Pearson correlation coefficient (PCC). On profiling dataset MBV from TCGA, D-GPM outperforms LR by 9.59% and 4.34%, RT by 27.58% and 22.96% and SVM by 6.14% and 3.07% on average, with respect to MAE and PCC, respectively. As for the number of better-predicted genes, D-GPM outperforms LR in 92.65% and 91.00%, RT in 95.66% and 98.25% and SVM in 85.49% and 81.56% of the target genes.

Keywords: Promoter methylation, deep neural network, machine learning, prediction method

1 Introduction

By influencing the transcription factor's accessibility to DNA, methylation of the promoter in a gene can regulate various biological process [1]. Enzymatic digestion, affinity enrichment, and bisulfite conversion are methods to capture the DNA methylation level [2]. Despite technological advances, there are still limitations in existing wet-lab methods. The resolution of enzymatic digestion-based approach is restricted to regions adjacent to methylation-sensitive restriction enzyme recognition sites [3]. The methylated DNA immunoprecipitation has its resolution limited to 100-300 base pair long fragments, and it is also biased towards hypermethylated regions [4]. Illumina's 450K bead-chip is the most widely used for profiling DNA methylation in human, but the chip only probes around 450K CpG sites in the human genome and covers partial CpG islands and may bias towards CpG dense contexts [5]. Whole genome bisulfite sequencing is a golden standard protocol; however, it is too costly since genome-wide deep sequencing of bisulfite-treated

fragments needs to generate a compendium of gene methylation level over a large number of conditions, such as retrovirus, activity changes of DNMT and drug treatments [6]. The community awaits more feasible and more economical solutions.

Previous research suggests that there is a low-rank structure in genome-wide gene expression profile [8], i.e, by leveraging the inner correlation between genes, the expression level of a few well-chosen landmark genes captures enough details to reconstruct the expressions of the rest of genes---target genes across the genome. The above was achieved by studying gene regulation networks and conducting principal component analysis on whole genome expression profile [7]. Consequently, scientists created a new technology called L1000, which only acquires expression profiling of landmark genes (~1000) to infer expression profiling of the target genes (~21000) [8].

Inspired by L1000, we proposed a method according to the following rationale to acquiring whole genome promoter methylation level acording to the promoters mythelations of the landmark genes. First, latent associations exist between the expressions of these landmark genes and target genes at the genome-wide level [9]; second, methylation in promoters located upstream of the transcription starting site is usually negatively correlated with its corresponding gene expression level [10]. Hence, it is likely that strong associations present among the methylation levels in the landmark genes and target genes.

To predict the methylation panorama on the whole genome is a large-scale multi-task machine learning problem, with a high-dimensional aim (~21,000) and a low-dimensional attribute (~1,000). Meanwhile, the deep learning method has shown its power in integrating large data scale and capturing the non-linear complexity of input features. In biology, extensive applications include predictions for the splicing activity of individual exons, inferring chromatin marks from DNA sequence, and quantification for the effect of SNVs on chromatin accessibility [11-13].

Here we present a multi-layer deep neural network named Deep-Gene Promoter Methylation Inference (D-GPM). To evaluate our D-GPM model, we benchmarked its performances against linear regression (LR), regression tree (RT) and support vector machine (SVM) with regards to methylation profile data based on Illumina Human Methylation 450k from The Cancer Genome Atlas (TCGA) [14]. The LR is to infer methylation levels of the target genes based on promoter methylation of the landmark genes using linear regression models. However, the linear model may fail to capture the non-linear relations of original data. The SVM can reliably represent complex nonlinear patterns [15], but suffers from poor scalability to big data size. The RT can address interpretability of the biological data and prediction model despite its less accuracy and instability in some predictors.

According to Illumina Human Methylation 450k data, we access methylation information on 902 landmark genes and 21645 target genes. Experiment results show D-GPM outperforms consistently other methods on testing data that under the measurement criteria mean absolute deviation (MAE) and Pearson correlation coefficient (PCC).

2 Methods

In this section, we first specify the gene methylation datasets in this study and formulate gene promoter methylation inference problem. We then propose the D-GPM for this problem and relevant details. Finally, we introduce several machine learning methods served as benchmarks.

2.1 Datasets

The methylation beta value (MBV) datasets are acquired from TCGA [16]. Considering that

Illumina Human Methylation 450k possess more probes and higher coverage rate, we excluded datasets of Illumina Human Methylation 27k, and finally, 9756 records remained for later analysis [17]. After filtering out the records, we calculated the average beta value of all probes located in promoter regions of a certain gene as its promoter methylation level. For more information on data preprocessing, please refer to Data pre-processing section in Supplementary Material.

We randomly partitioned the methylation data into 80% for training, 10% for validation, and 10% for testing, which corresponded to 7,549 samples, 943 samples, and 943 samples, respectively. We denoted them as MBV-tr, MBV-va, and MBV-te in order.

2.2 Multi-task regression model for gene expression inference

In the model, there are *J* landmark genes, *K* target genes and *N* training samples; we denoted the training data as $\{x_i, y_i\}_{i=1}^{N}$, where $x_i \in \Re^J$ is the promoter methylation profiles of landmark genes and $y_i \in \Re^K$ represents the methylation profiles of target genes in the i th sample. Our task is to find a mapping $F: \Re^J \Rightarrow \Re^K$ that can fit $\{x_i, y_i\}_{i=1}^{N}$ well, which can be viewed as a multi-task regression problem.

Let us assume a sample of 9756 individuals, each represented by a 902-dimensional input vector and a 21645-dimensional output vector. Let $\mathbf{X}$ denote the $N \times J$ input matrix, whose column corresponds to observations for the *j*-th input $x_j = \left\{x_j^1, ..., x_j^N\right\}^T$. In genetic association mapping, each element x_j^i of the input matrix takes values from $\{0,1,2,3\}$ according to the number of minor alleles at the *j*-th locus of the *i*-th individual. Let $\mathbf{Y}$ denote the $N \times K$ output matrix, whose column is a vector of observations for the *k*-th output $y_k = \left\{y_k^1, ..., y_k^N\right\}^T$. For each of the K output variables, we assume a linear regression model:

$$y_k = X\beta_k + \varepsilon_k, \forall k = 1, ...K, \tag{1}$$

Where β_k is a vector of *J* regression coefficients $\left\{\beta_k^1, ...\beta_k^J\right\}^T$ for the *k*-th output, and ε_k is a vector of *N* independent error terms having mean 0 and a constant variance. We center the y_k's and x_j's such that $\sum_i y_k^i = 0$ and $\sum_i x_j^i = 0$, and consider the model without an intercept.

Regression coefficients matrix β has been used to take advantage of relatedness across all input variables.

2.3 Assessment Criteria

We adopted MAE and PCC as criteria to evaluate models' performance at each target gene *t* of different samples. We formulated the overall error as the average MAE over all target genes. PCC is used to describe the relationship between real promoter methylation and predicted promoter methylation. Here, definitions of MAE and PCC for evaluating the predictive performance at each target gene *t* are as follows.

$$MAE_{(t)} = \frac{1}{N'}\sum_{i=1}^{N'} | y_{i(t)} - \hat{y}_{i(t)} |, \tag{2}$$

$$Correlation_{(t)} = \frac{\sum_{i=1}^{N'}(y_{i(t)} - \bar{y}_{(t)})(\hat{y}_{i(t)} - \bar{\hat{y}}_{(t)})}{\sqrt{\sum_{i=1}^{N'}(y_{i(t)} - \bar{y}_{(t)})^2 \cdot \sum_{i=1}^{N'}(\hat{y}_{i(t)} - \bar{\hat{y}}_{(t)})^2}}, \tag{3}$$

where N' is the number of testing samples and $\hat{y}_{i(t)}$ is the predicted expression value for target gene t in sample i and $\bar{\hat{y}}_{(t)}$ is the mean predicted expression value for target gene t in N' testing samples.

2.4 D-GPM

D-GPM is a fully connected multi-layer perceptron with one output layer. All the hidden layers consist of H hidden units. In this work, we have a set of Hs, ranging from 1000 to 9000 with step size 1000. A hidden unit j in layer l takes the sum of weighted outputs plus the bias from the previous layer l-1 as the input and produces a single output o_j^l.

$$o_j^l = f\left(\sum_{i=1}^{H} w_{i,j}^{l-1} o_i^{l-1} + b_j^{l-1}\right), \tag{4}$$

where f is a nonlinear activation function, H is the number of hidden units, $\left\{w_{i,j}^{l-1}, b_j^{l-1}\right\}_{i=1}^{H}$ are the weights and the bias of unit j to be found.

The loss function is the sum of mean squared error at each output unit; that is:

$$\varsigma = \sum_{t=1}^{T}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_{i(t)} - \hat{y}_{i(t)}\right)^2\right]. \tag{5}$$

D-GPM contains 902 units in the input layer corresponding to the 902 landmark genes, and we also configure D-GPM with 21,645 units in the output layer analogous to the 21,645 target genes. Fig. 1 shows the various architecture of D-GPM.

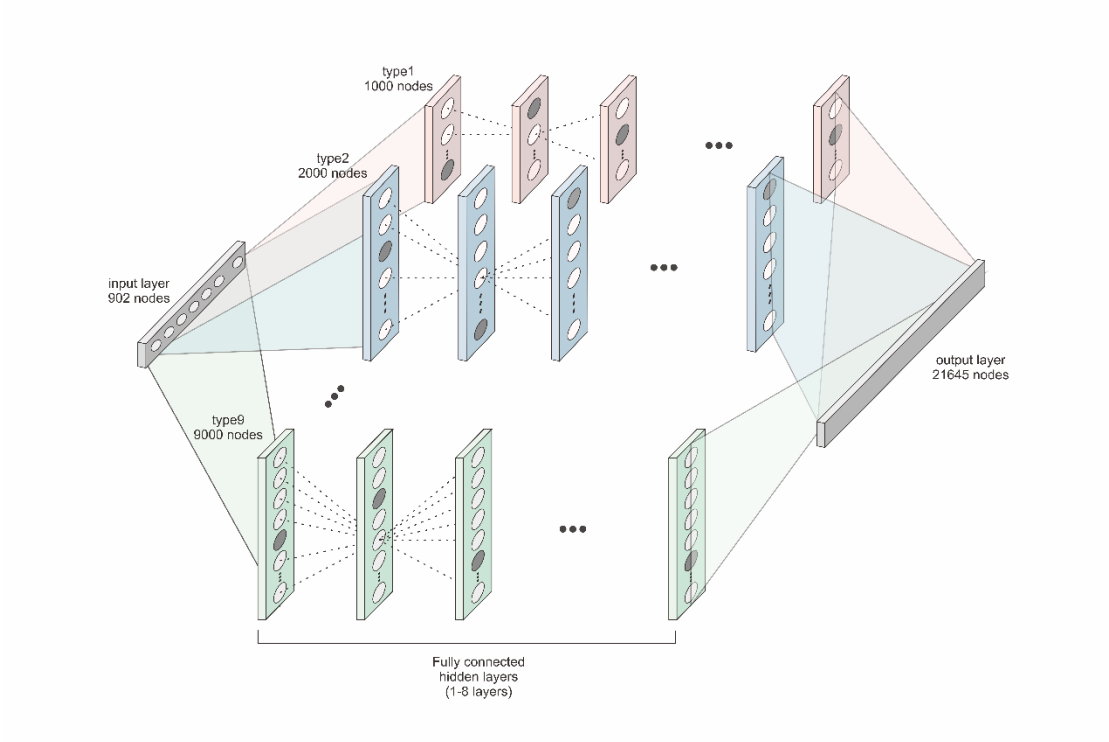


Fig. 1. The architecture of D-GPM. It comprises one input layer, one or multiple hidden layers and one output layer. All the hidden layers have the same number of hidden units.

Here, we briefly describe the training techniques and their significance in training steps:

1. Dropout is a scheme used to perform model averaging and regularization for deep neural networks [18]. Here, we utilize dropout to all hidden layers of D-GPM and dropout rate p that can

steer regularization intensity is set at [from 0% to 50%, with step size 5%] separately to find out the optimum architecture of D-GPM.

2. Normalized initialization can stabilize the variances of activation during epochs. To initialize parameters of deep neural networks, here we set initialized weights to within the range of $\left[-1\times10^{-4},\ \ 1\times10^{-4}\right]$ according to activation function [19].

3. Momentum method is also adopted in our work to speed gradient optimization and improve the convergence rate for deep neural networks [20].

4. Learning rate is initialized to 5×10^{-4}, 2×10^{-4}, 1×10^{-4} or 8×10^{-5} depending on different architecture of D-GPM, and was tuned according to the training error on a subset of MBV-tr.

5. Model selection is implemented based on MBV-va. Models are assessed on MBV-va after each epoch, and the model with the minimum loss function is saved. The maximum epoch for training is set 500-epoch.

Here, we implement D-GPM with Theano and Pylearn2 libraries [21, 22].

2.5 Benchmark methods

To evaluate the performance of the deep learning methods, we adopted LR, RT, and SVM as benchmarks.

Here, we utilize the RT model with rpart [23]. Noticing Gaussian RBF kernel function has high superiority in a large sample and high dimension data, and can reduce computational complexity in our methylation profiling data efficiently, [24]thus we adopt kernlab package to implement SVM for predicting promoter methylation [25].

When training the above three machine models, we harnessed the 5-fold cross-validation method. Our models are modeled using 80% of the randomly sliced data, and the remaining 20% of the data are used for evaluation. After the process of training and evaluating the model is repeated five times independently, we calculated the average performance during these five processes as a final performance index

3 Result

Here, we have introduced MBV datasets from TCGA and defined the methylation profiles inferences as a multi-task regression problem, with MBV-tr being for training, MBV-va being for validation and MBV-te being for testing. We have also illustrated our deep learning method D-PGM and the other three methods, including LR, RT and SVM, to work out the regression problem. Next, we show the predictive performances of the above methods on MBV-te data based on criteria MAE, PCC respectively.

3.1 D-GPM performs the best for predicting promoter methylation.

Back-propagation algorithm, mini-batch gradient descent, and other beneficial deep learning techniques are adopted in training D-GPM [26]. Detailed parameter configurations are shown in Table 1.

Table 1. Detailed parameter configurations are given in D-GPM.

Parameters	
# of hidden layers	[1,2,3,4,5,6,7,8]
# of hidden units in each hidden layer	[1000,2000,3000,4000,5000,6000,7000,8000,9000]
Droupout rate	[0%,5%,10%,15%,20%,25%,30%,35%,40%,45%,50%]
Momentum coefficient	0.5
Initial learning rate	5E-4, 2E-4, 1E-4 or 8E-5
Minimum learning rate	1.00E-05
Learning rate decay factor	0.9
Learning scale	3.0
Mini-batch size	200
Training epoch	500
Weights initial range	$\left[-\frac{\sqrt{6}}{\sqrt{n_i+n_o}}, \frac{\sqrt{6}}{\sqrt{n_i+n_o}}\right]$

According to the parameter configurations, all the combinations of parameters are made during training D-GPM for predicting promoter methylation of the target genes.

Table 2. The MAE-based overall errors of LR, RT, SVM, and D-GPM with partially different architecture and partially different dropout rates on MBV-te. Numerics after "±" are the standard deviations of prediction errors over all target genes. The best performance of D-GPM is underlined.

	# of hidden units	6000	7000	8000
# of hidden layers and dropout rate	**4,10%**	0.0332±0.0253	0.0340±0.0260	0.0343±0.0263
	4,15%	0.0333±0.0253	0.0340±0.0260	0.0344±0.0264
	4,20%	0.0336±0.0255	0.0343±0.0261	0.0344±0.0262
	5,10%	0.0344±0.0264	0.0337±0.0257	0.0346±0.0264
	5,15%	0.0343±0.0260	0.0329±0.0251	0.0343±0.0261
	5,20%	0.0350±0.0267	0.0343±0.0259	0.0347±0.0265
	6,10%	0.0341±0.0259	0.0339±0.0258	0.0339±0.0258
	6,15%	0.0339±0.0259	0.0334±0.0255	0.0331±0.0253
	6,20%	0.0356±0.0269	0.0346±0.0261	0.0351±0.0265
	Linear regression		0.0363±0.0277	
	Support vector machine		0.0341±0.0258	
	Regression tree		0.0454±0.0363	

Table 3. The PCC of LR, RT, SVM and D-GPM with partially different architecture and partially different dropout rates on MBV-te. Numerics after "±" are the standard deviations of PCC over all target genes. The best performance of D-GPM is underlined.

# of hidden layers and dropout rate	# of hidden units	6000	7000	8000
	4,10%	0.8081±0.0964	0.7972±0.0976	0.8058±0.0957
	4,15%	0.8055±0.0968	0.7936±0.0989	0.8041±0.0961
	4,20%	0.8077±0.0964	0.8032±0.0964	0.7968±0.0951
	5,10%	0.7776±0.1022	0.8032±0.0984	0.7990±0.0944
	5,15%	0.7842±0.1012	0.8186±0.0940	0.7828±0.1035
	5,20%	0.7835±0.1001	0.8135±0.0943	0.7933±0.0997
	6,10%	0.7919±0.0987	0.8007±0.0961	0.8010±0.0947
	6,15%	0.7865±0.1002	0.8086±0.0923	0.8106±0.0914
	6,20%	0.7879±0.1006	0.8082±0.0925	0.7952±0.0975
	Linear regression		0.7846±0.1069	
	Support vector machine		0.7942±0.1056	
	Regression tree		0.6658±0.1192	

As Table 2 indicates, D-GPM acquires the best MAE performance on MBV-te with five hidden layers of 7000 units and 15% dropout rate (D-GPM-15%-7000×5) among the 792 (8*9*11) various D-GPMs. Meanwhile, D-GPM has an extraordinary edge over MAE compared with LR, SVM, and RT.

Similarly, D-GPM also obtains the best PCC performance on MBV-te among the 792 prediction models as shown in Table 3. The complete MAE, PCC evaluation of D-GPM armed with other architecture (hidden layer: from 1 to 8, with step size 1; hidden unit: from 1000 to 9000, with step size 1000; dropout rate: from 0% to 50%, with step size 5%) on MBV-te are given in Supplementary Material.

Based on the above MAE and PCC, we can conclude that D-GPM is the best model for predicting promoter methylation among the prediction models.

3.2 Evaluation according to MAE criteria

As D-GPM acquires the best MAE performance on MBV-te with a 15% dropout rate (described as D-GPM-15%) among eleven dropout rates ranging from 0% to 50% with step size 5%. Fig. 2 shows the overall MAE performances of D-GPM-15% and SVM on MBV-te. The bigger architecture of D-GPM-15% (five hidden layers with 7000 hidden units in each hidden layer, described as D-GPM-15%-7000×5) acquires the least MAE on MBV-te. The improvements of D-GPM-15%-7000×5 is 9.59%, 27.58%, and 6.14% over LR, RT, and SVM respectively. A possible explanation is that deep learning can capture complex features [27].

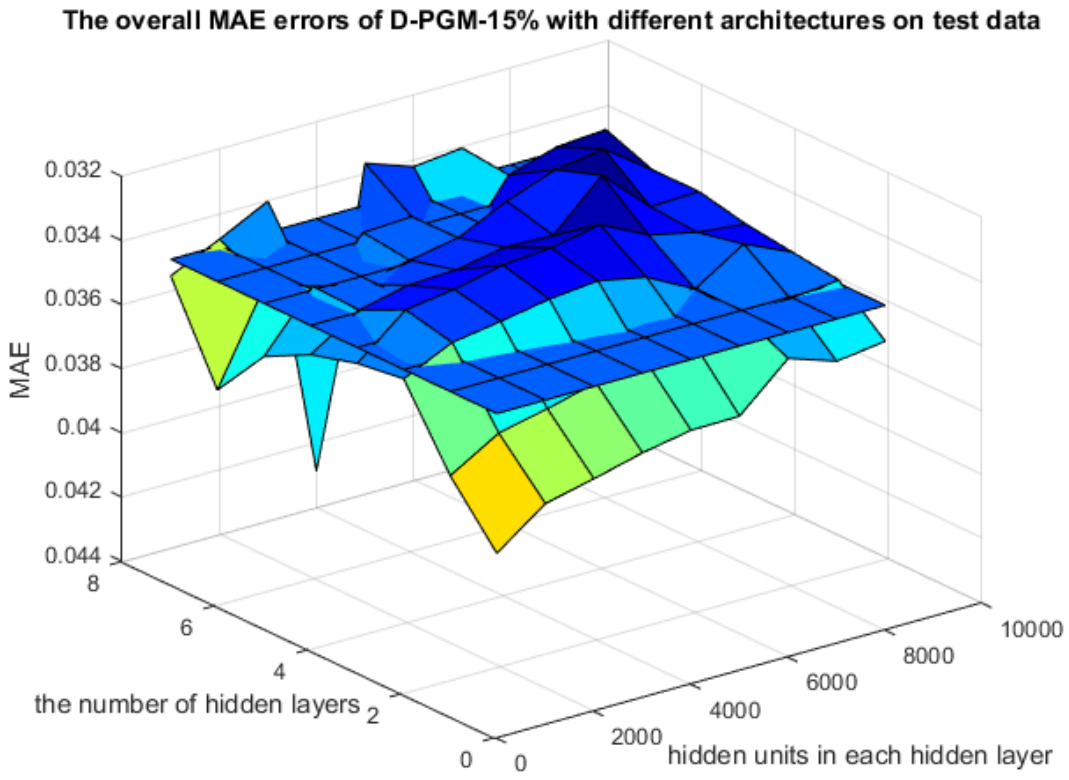


Fig. 2. The overall MAE errors of D-GPM-15% with various architecture on MBV-te.

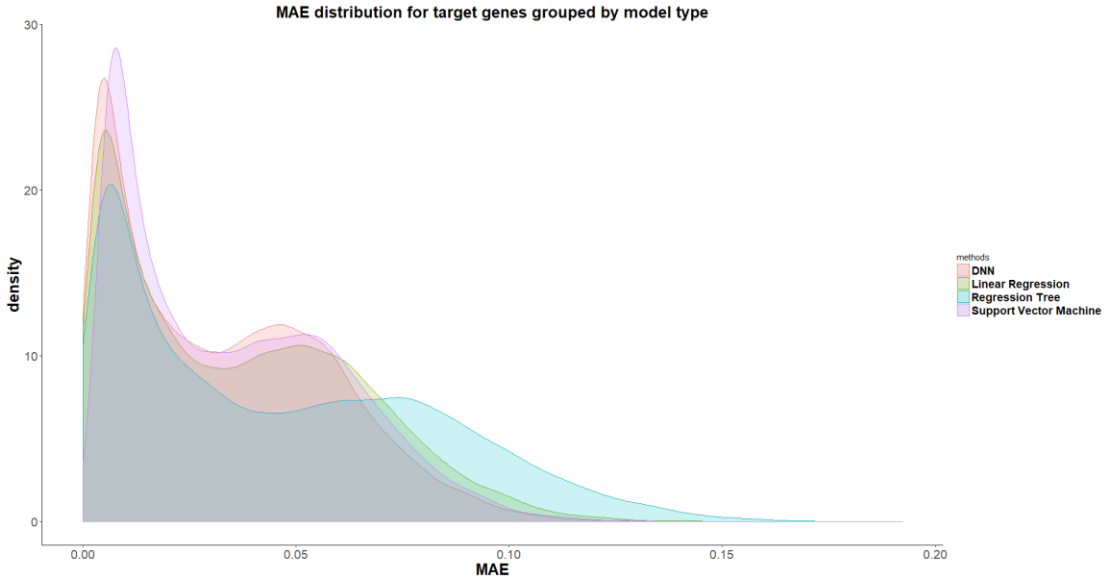


Fig. 3. The density plots of the MAE errors of all the target genes by LR, RT, SVM, and D-GPM on MBV-te.

D-GPM also outperforms LR and RT almost in all target genes on MAE. Fig. 3 shows the density plots of the MAE errors of all the target genes by LR, RT, SVM, and D-GPM. On the whole, we can see D-GPM occupies a larger proportion at the low MAE level and a lower proportion at the high MAE level compared to three machine learning methods, especially RT and LR, attesting to the prominent performance of D-GPM.

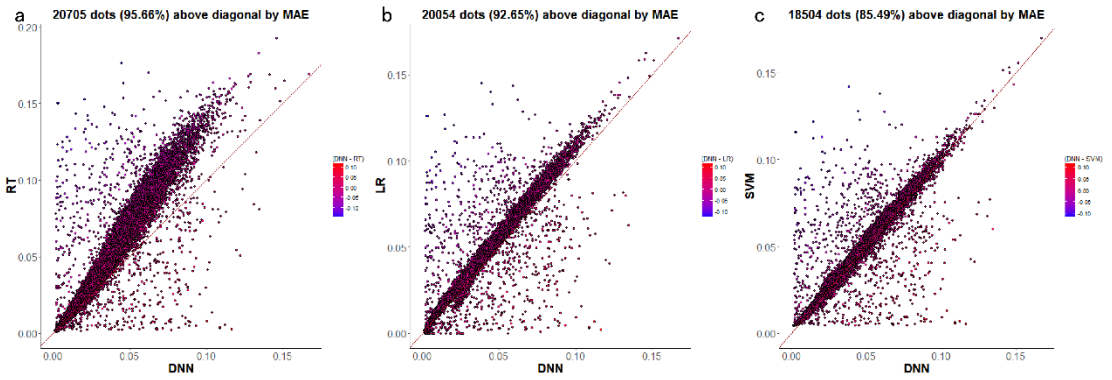


Fig. 4. (a) The MAE errors of each target gene by D-GPM compared with RT on MBV-te. (b) The MAE errors of each target gene by D-GPM compared with LR on MBV-te. (c) The MAE errors of each target gene by D-GPM compared with SVM on MBV-te. Among (a), (b) and (c), each dot represents one out of the 21645 target genes. The x-axis is the MAE of each target gene by D-GPM, and the y-axis is the MAE of each target gene by the other machine learning method. Dots above diagonal means D-GPM achieves lower error compared with the other method.

Fig. 4(a-c) display a gene-wise comparative analysis between D-GPM and the other three methods. In term of MAE, D-GPM outperforms RT in 95.66% (20705 genes) of the target genes, outperforms LR in 92.65% (20054 genes) of the target genes and outperforms SVM in 85.49% (18504 genes) of the target genes. These results can also be viewed by a larger proportion of dots lie above the diagonal, this better performance may suggest that D-GPM can capture some intrinsic nonlinear features of the MBV data which LR, RT, and SVM did not accomplish.

RT performs significantly worse than the other methods in the MAE aspect. One possible reason is that the model is oversimplified to capture essential features between landmark and target genes based on MBV-te [28].

According to the model distribution of the least MAE for each target gene, we find out the best model distribution shown in Fig. 5(a). RT accomplishes the best MAE performance for 305 target genes (1.41%), including *BRD2, GPI, MAF, MICB* genes, implying there is a relatively simple methylation regulation mechanism and promoter methylation of these genes are dominantly regulated by a very few landmark genes. LR can predict 1,242 target genes (5.74%) at best among other three methods, including *ABCD1, HPD, AMH* and *ARAF* genes, laying a solid foundation for pathogenesis on diseases such as Adrenoleukodystrophy, Hawkinsinuria, Persistent Mullerian Duct Syndrome and Pallister-Killian Syndrome using our LR[29-32]. Noticeably, SVM performs at best for a total of 2813 genes (13.00%), including *ACE2, A2M,* and *CA1*. One possible explanation is that there seem to be intricate and complicate interactions among promoter methylation of the landmark genes and these 2813 target genes. Undoubtedly, D-GPM does better than other three methods as far as 17,285 genes (79.86%) are concerned, demonstrating deep neural networks' powerful ability to capture the nonlinear relationship of methylation profiling.

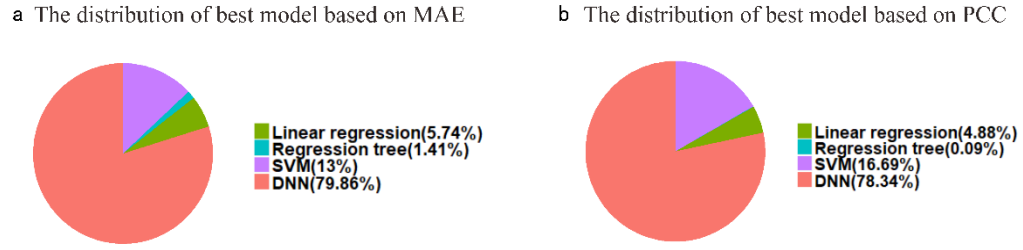


Fig. 5. (a) Distribution of the best model according to MAE for target genes. (b) Distribution of best model according to PCC for target genes.

3.3 Evaluation according to PCC

D-GPM accomplishes the best PCC performance on MBV-te with 15% dropout rate. Fig. 6 shows the overall PCC performances of D-GPM-15% and the other methods on MBV-te. Same as MAE, D-GPM-15%-7000 × 5 acquires the most significant PCC on MBV-te. The relative improvement of D-GPM-15%-7000×5 is 4.34% over LR, 22.96% over RT and 3.07% over SVM. Just like MAE, almost all combined architecture of D-GPM-15% outperforms LR and RT in PCC performance.

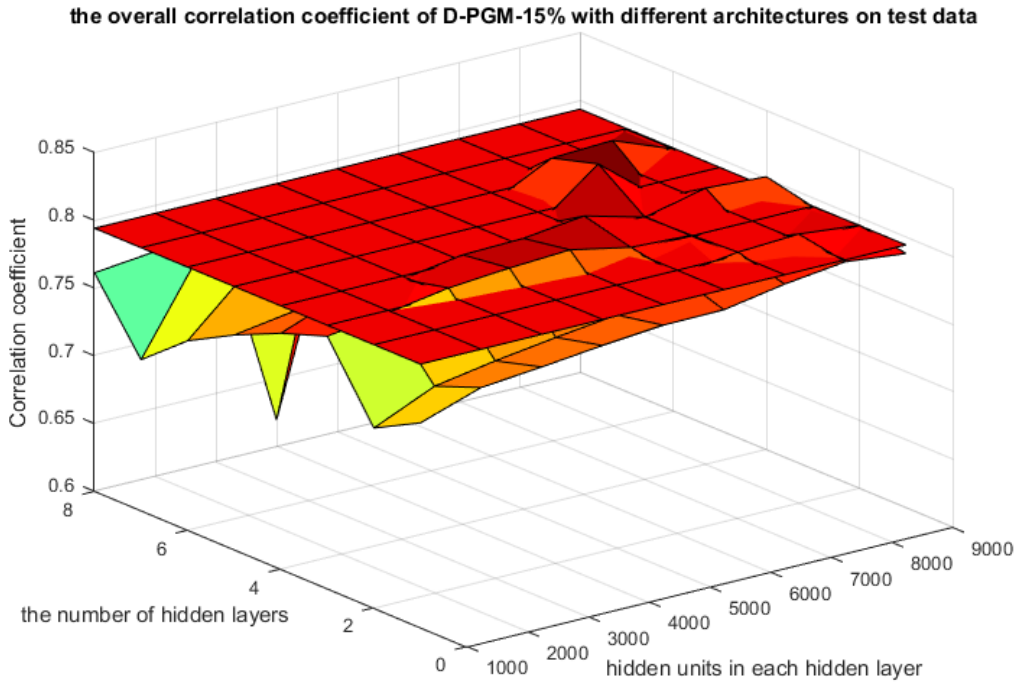


Fig. 6. The overall PCC performance of D-GPM-15% with various architecture on MBV-te.

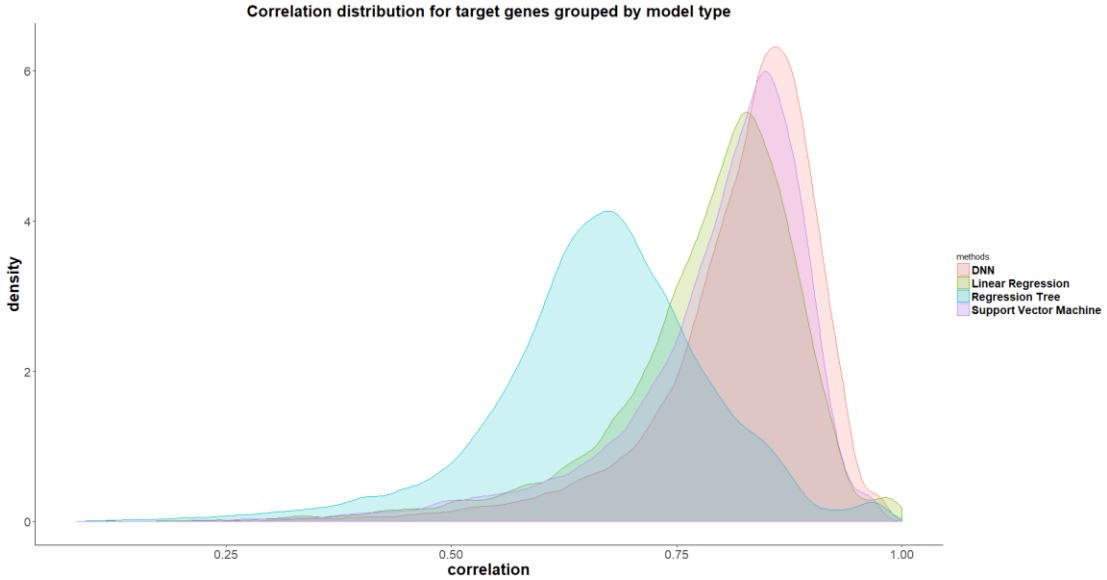


Fig. 7. The density plots of the PCC performance of all the target genes by LR, RT, SVM and D-GPM on MBV-te.

In term of PCC, D-GPM also outperforms RT, LR for almost the target genes. Fig. 7 displays the density plots of the PCC of all the target genes by LR, RT, SVM, and D-GPM. By and large, we can see D-GPM possesses a larger proportion of the high PCC and a lower proportion at the low PCC compared to RT, LR, and SVM.

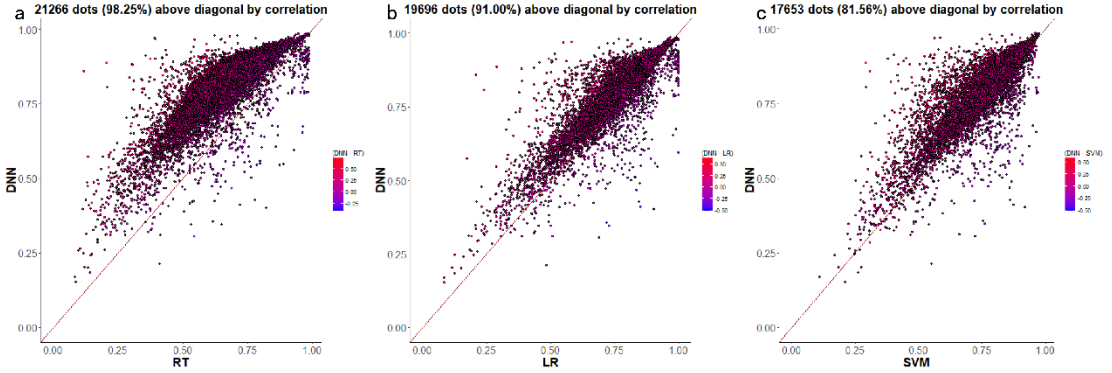


Fig. 8. (a) The PCC performance of each target gene by D-GPM compared with RT on MBV-te. (b) The PCC performance of each target gene by D-GPM compared with LR on MBV-te. (c) The PCC performance of each target gene by D-GPM compared with SVM on MBV-te. Among (a), (b) and (c), each dot represents one out of the 21645 target genes. The x-axis is the PCC of each target gene by the above-mentioned three machine learning technique, and the y-axis is the PCC of each target gene by D-GPM. Dots above diagonal means D-GPM achieves high PCC compared with the other method.

Fig. 8(a-c) show a gene-wise comparative analysis between D-GPM and the other three methods. For PCC, D-GPM outperforms RT in 98.25% (21266 genes) of the target genes, outperforms LR 91.00% (19696 genes) in of the target genes and outperforms SVM in 81.56% (17653 genes) of the target genes. Therefore, D-GPM's powerful prediction for PCC overall target genes preserve just like its effective prediction for MAE. It is obvious that although the prediction property of SVM is still modest in PCC aspect, its PCC on some of the target genes is significantly higher than D-GPM and behaves better than the prediction for the same target genes in MAE aspect. This is probably due to facts that SVM is based on the principle of structural risk minimization, avoiding over-learning problems and having strong generalization ability.

According to the model distribution of maximal PCC, we find out the best model distribution shown in Fig. 5(b). Surprisingly, RT only gains the best PCC performance for 19 target genes (0.09%), including *ALG1*, *NBR2* genes, falling far below the best 305 target genes in MAE aspect. Considering RT's awful prediction power for PCC compared with its better prediction power for MAE, it may be explained that RT makes decision based on over simple assumption. LR predicts the best 1057 target genes (4.88%) among other three methods, including *AASS* and *ACE* genes, almost being the same as 1242 in MAE level. For PCC, SVM is on it best behavior in a total of 3613 genes (16.69%), having an increasing number compared to that of MAE, in contrast to RT. Undoubtedly, D-GPM outperforms other three methods with regards to 16956 genes (78.34%), but behaves badly relative to 79.86% in MAE level, suggesting an inability to predict PCC for target genes.

Noticeably, dropout regularization manages to improve the performance of D-GPM-7000×5 on MBV-te as shown in Fig. 9. With 15% dropout rate, D-GPM-15%-7000×5 consistently achieves the best MAE performance on MBV-te among the models with 0%, 15%, 20%, 25%, and 35% dropout rates, proving over-fitting and under-fitting both result in bad influence on the prediction model.

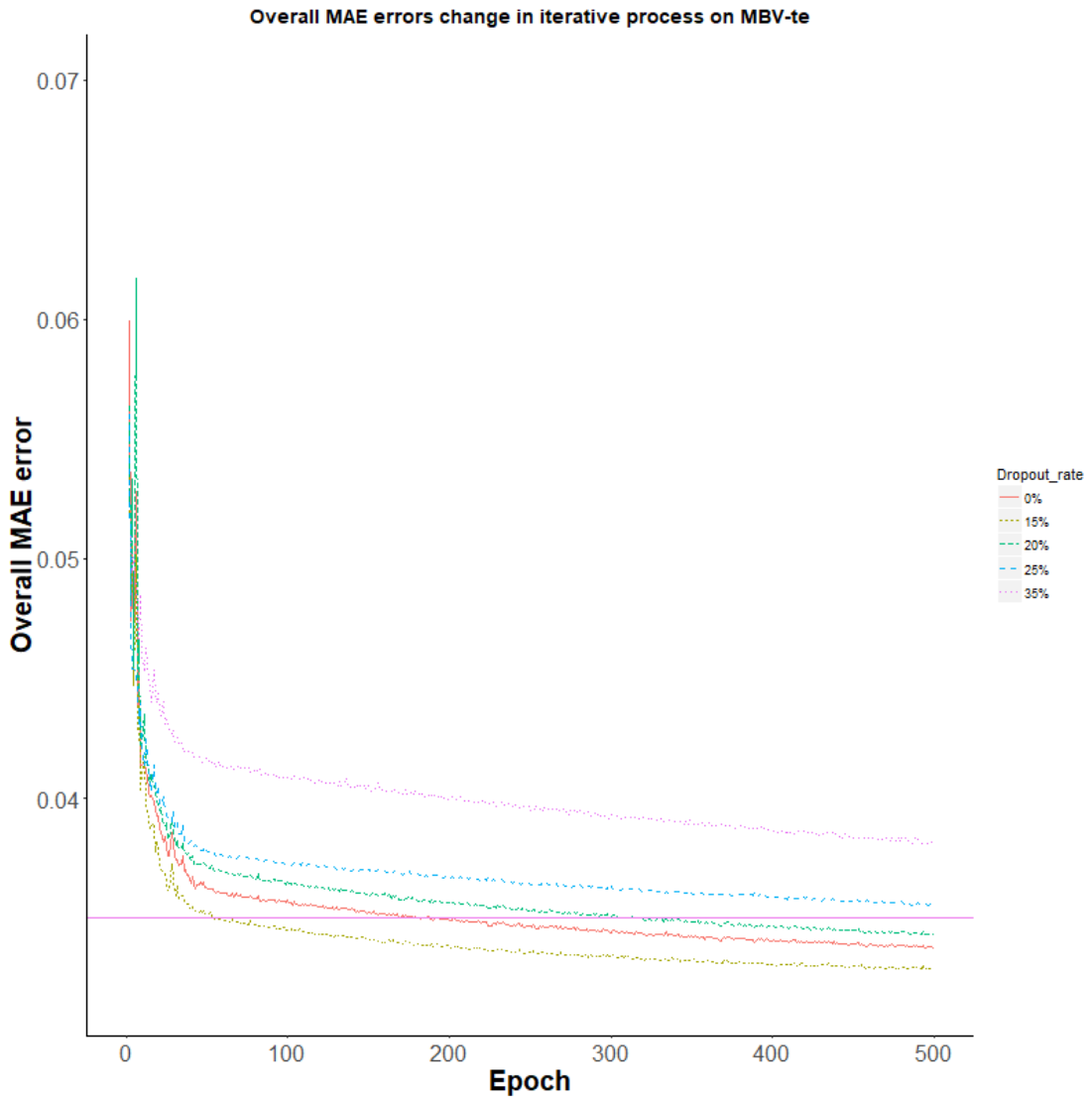


Fig. 9. The overall MAE error is decreasing curves of D-GPM-7000×5 on MBV-te with different dropout rates. The x-axis is the training epoch, and the y-axis is the overall MAE error. The overall MAE error of SVM is also covered for comparison.

4 Discussion

Comprehending intricate regulation modes of promoter methylation profiling under numerous biological states need robust inference frameworks and cost-effective profiling tools. Despite the whole genome bisulfite sequencing being thought of as the best protocol, it is too costly since the genome-wide deep sequencing of bisulfite-treated fragments needs to be implemented.

Promoter methylation of a gene has been confirmed associated with DNA accessibility and binding of transcription factors that can regulate gene expression. Considering there is an underlying relationship between ~1,000 landmark genes and ~21,000 target genes at the genome-wide level of expression and the fact that methylation occurring in promoters located in the promoter region is negatively associated with expression of its corresponding gene. Therefore, we can make good use of methylation levels in the promoter regions of landmark genes to characterize the cellular status of samples under diverse conditions. Here, we also develop three machine learning methods namely LR, SVM and RT and a deep method called D-GPM for inferring promoter methylation of target genes based on only landmark genes. In summary, D-GPM shows as the best model among LR, RT,

and SVM in predicting promoter methylation of target genes from our MBV datasets in multiple aspects containing MAE and PCC.

Different methods have different edges and disadvantages: RT provides us with good interpretability and meanwhile relative lousy accuracy; LR offers better performance compared to RT, even though it ignores the nonlinearity within biological data; SVM can represent complex nonlinear patterns. Unfortunately, it suffers from poor scalability to big data size. To some extent, D-GPM manages to overcome the above drawbacks. These prediction models are not perfect and behave better separately for different target genes. It is instructive to interpret these prediction models and explain the inherent relationship between promoter methylation of target genes and landmark genes according to our results.

Although D-GPM is the best model for predicting most of the target genes, three machine learning methods all have specific edges when predicting some specific target genes. Shortly, we will integrate multiple prediction models as an ensemble tool to ensure it is suitable for predicting target genes as many as possible. Furthermore, we need to conduct a verification experiment to judge whether our conclusion drawn from the relationship among promoter methylation of target and landmark genes (such as *TDP1* and *CIAPIN1*) is sound and persuasive. Furthermore, after obtaining the ensemble prediction model, we will make the most of it impute and revise methylation site that is missing or low-reliability in realistic methylation profiling data.

Author contributions:

S.C.L supervised the research and revised the manuscript; X.X.P designed the study, did the modelling, performed the experiments, carried out the data analyses and wrote the manuscript; B.L collected and prepossessed data; X.Z.W revised the manuscripts; Y.L.L and X.Q.Z participated in literature review; all authors have contributed to the manuscript.

Corresponding author at: Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

E-mail address: shuaicli@gmail.com

Acknowledgement

Computing resource was supported by BGI-Shenzhen.

Competing interests

The authors declare that they have no competing interests.

References

1. Moore, L. D., Le, T. & Fan, G. (2013) DNA methylation and its basic function, *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology.* **38**, 23-38.

2. Hammoud, S. S., Cairns, B. R. & Carrell, D. T. (2013) Analysis of gene-specific and genome-wide sperm DNA methylation, *Methods in molecular biology (Clifton, NJ).* **927**, 451-8.

3. Krygier, M., Podolak-Popinigis, J., Limon, J., Sachadyn, P. & Stanislawska-Sachadyn, A. (2016) A simple modification to improve the accuracy of methylation-sensitive restriction enzyme quantitative polymerase chain reaction, *Analytical biochemistry.* **500**, 88-90.

4. Thu, K. L., Vucic, E. A., Kennett, J. Y., Heryet, C., Brown, C. J., Lam, W. L. & Wilson, I. M. (2009)

Methylated DNA immunoprecipitation, *Journal of visualized experiments : JoVE.* **23**, 935.

5. Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R. & Gunderson, K. L. (2009) Genome-wide DNA methylation profiling using Infinium(R) assay, *Epigenomics.* **1**, 177-200.

6. Li, Q., Hermanson, P. J. & Springer, N. M. (2018) Detection of DNA Methylation by Whole-Genome Bisulfite Sequencing, *Methods in molecular biology (Clifton, NJ).* **1676**, 185-196.

7. Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & di Bernardo, D. (2007) How to infer gene networks from expression profiles, *Molecular systems biology.* **3**, 78.

8. Edgar, R., Domrachev, M. & Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic acids research.* **30**, 207-10.

9. Daura-Oller, E., Cabre, M., Montero, M. A., Paternain, J. L. & Romeu, A. (2009) Specific gene hypomethylation and cancer: new insights into coding region feature trends, *Bioinformation.* **3**, 340-3.

10. Portela, A. & Esteller, M. (2010) Epigenetic modifications and human disease, *Nat Biotechnol.* **28**, 1057-68.

11. Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., Hua, Y., Gueroussov, S., Najafabadi, H. S. & Hughes, T. R. (2015) RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease, *Science.* **347**, 1254806.

12. Kelley, D. R., Snoek, J. & Rinn, J. L. (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks, *Genome Research.* **26**, 990.

13. Zhou, J. & Troyanskaya, O. G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model, *Nature Methods.* **12**, 931.

14. Mclendon, R., Friedman, A., Bigner, D., Meir, E. G. V., Van Meir, E., Brat, D., Mastrogianakis, G., Chin, L. & Network, C. (2008) The Cancer Genome Atlas (TCGA), Comprehensive genomic characterization defines human glioblastoma genes and core pathways.

15. Ye, G., Tang, M., Cai, J. F., Nie, Q. & Xie, X. (2013) Low-rank regularization for learning gene expression programs, *PLoS One.* **8**, e82146.

16. Tomczak, K., Czerwinska, P. & Wiznerowicz, M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, *Contemporary oncology (Poznan, Poland).* **19**, A68-77.

17. Touleimat, N. & Tost, J. (2012) Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation, *Epigenomics.* **4**, 325-41.

18. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. (2012) Improving neural networks by preventing co-adaptation of feature detectors, *Computer Science.* **3**, págs. 212-223.

19. Glorot, X. & Bengio, Y. (2010) Understanding the difficulty of training deep feedforward neural networks, *Journal of Machine Learning Research.* **9**, 249-256.

20. Sutskever, I., Martens, J., Dahl, G. & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. Paper presented at the *International Conference on International Conference on Machine Learning*.

21. Goodfellow, I. J., Wardefarley, D., Lamblin, P., Dumoulin, V., Mirza, M., Pascanu, R., Bergstra, J., Bastien, F. & Bengio, Y. (2013) Pylearn2: a machine learning research library, *Eprint Arxiv*.

22. Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D. & Bengio, Y. (2010). Theano: a CPU and GPU math expression compiler.

23. Therneau, T. M. & Atkinson, B. (2012) rpart: Recursive Partitioning. R package version 3.1-51.

24. Steinwart, I., Hush, D. & Scovel, C. (2006) An Explicit Description of the Reproducing Kernel Hilbert Spaces of Gaussian RBF Kernels, *IEEE Transactions on Information Theory.* **52**, 4635-4643.

25. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. (2004) kernlab - An S4 Package for Kernel Methods in R, *Journal of Statistical Software.* **11**, 721-729.

26. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1988) *Learning representations by back-propagating errors*, Nature.

27. Bengio, Y. (2009) Learning Deep Architecture for AI. **2**, 1-127.

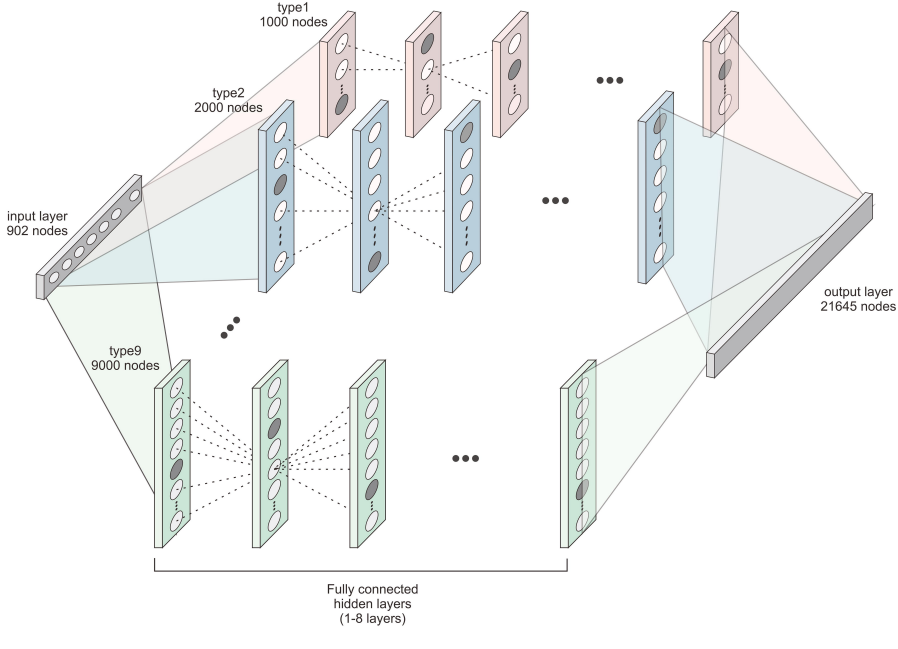
28. Timofeev, R. (2004) Classification and Regression Trees (CART) Theory and Applications, *Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät*.

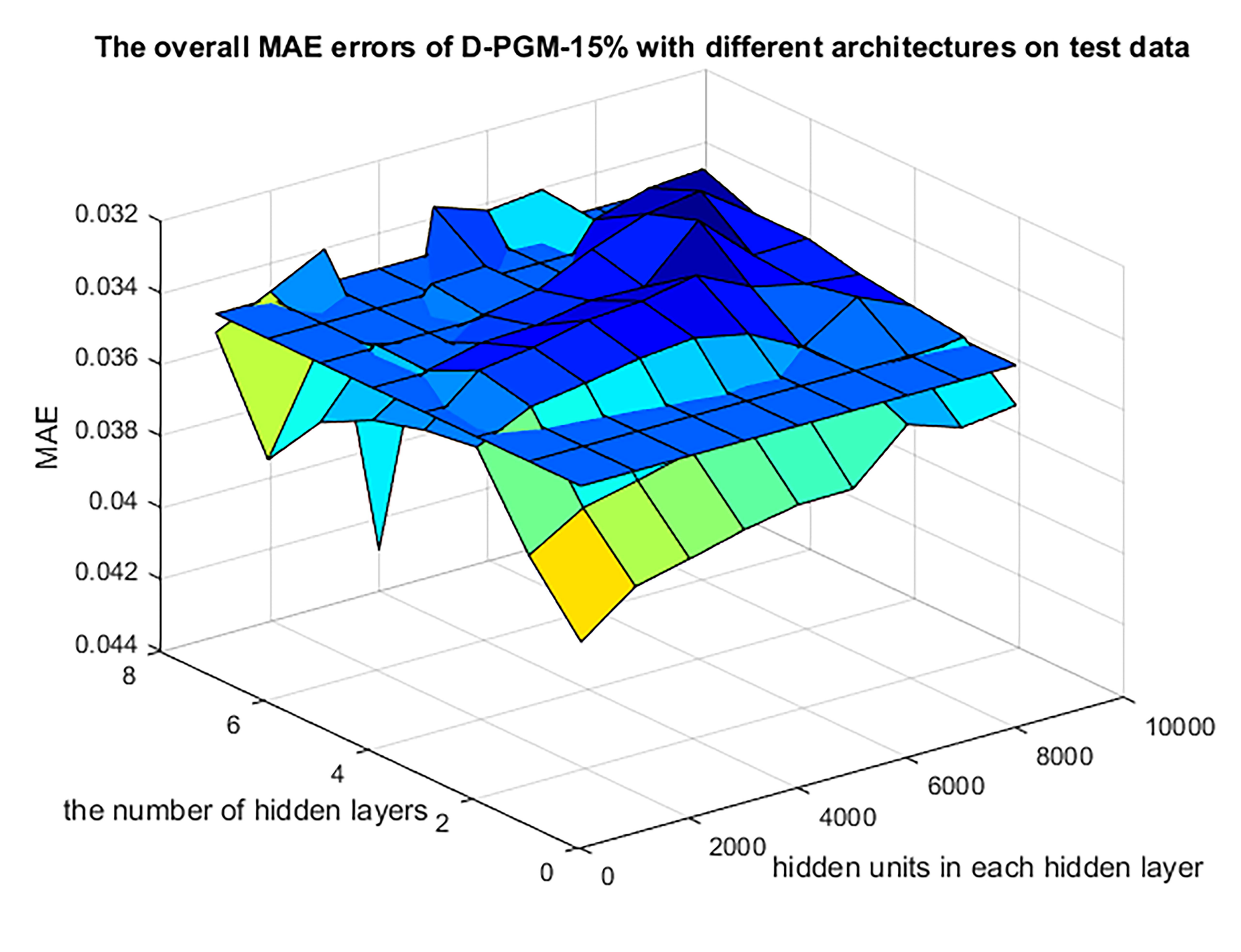
29. Park, J. A., Jun, K. R., Han, S. H., Kim, G. H., Yoo, H. W. & Hur, Y. J. (2012) A novel mutation in the ABCD1 gene of a Korean boy diagnosed with X-linked adrenoleukodystrophy, *Gene.* **498**, 131-3.

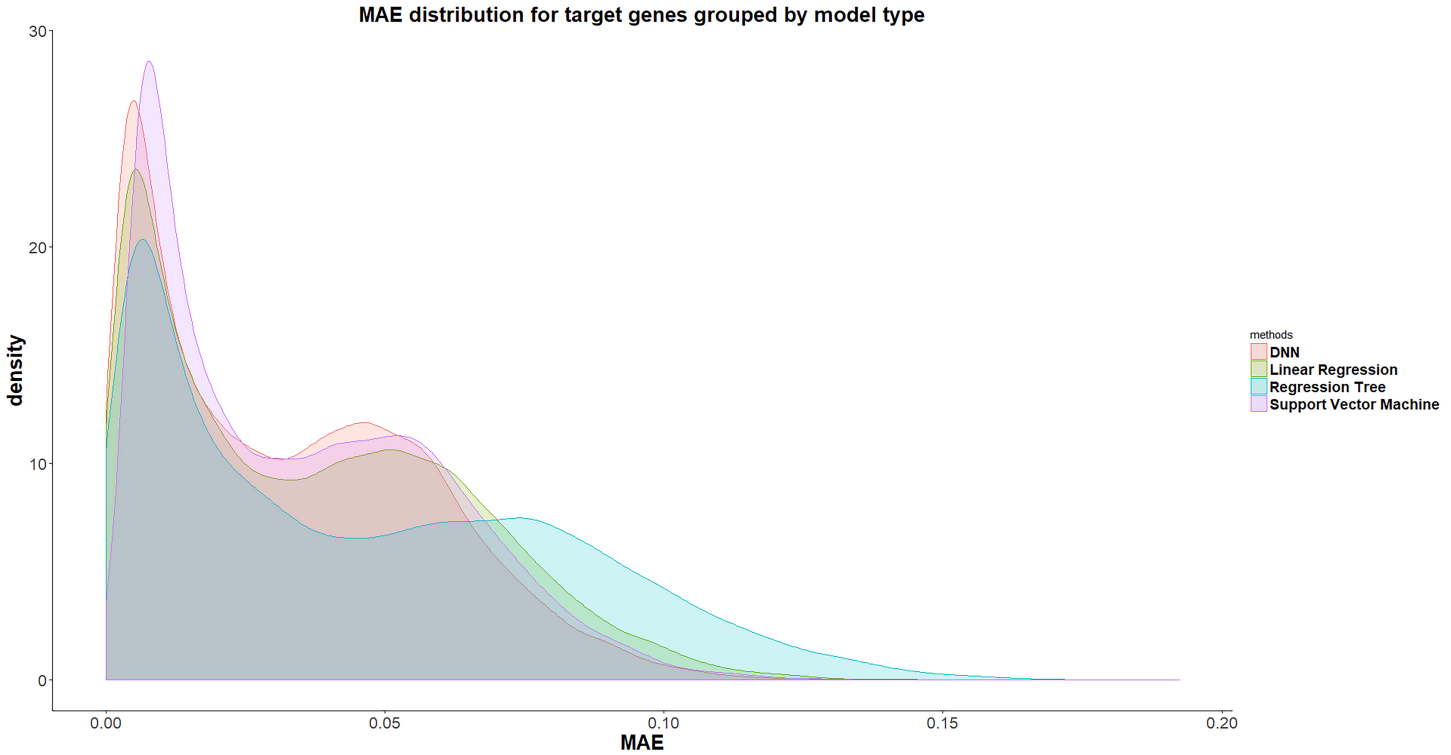
30. Thodi, G., Schulpis, K. H., Dotsikas, Y., Pavlides, C., Molou, E., Chatzidaki, M., Triantafylli, O. & Loukas, Y. L. (2016) Hawkinsinuria in two unrelated Greek newborns: identification of a novel variant, biochemical findings and treatment, *Journal of pediatric endocrinology & metabolism : JPEM.* **29**, 15-20.

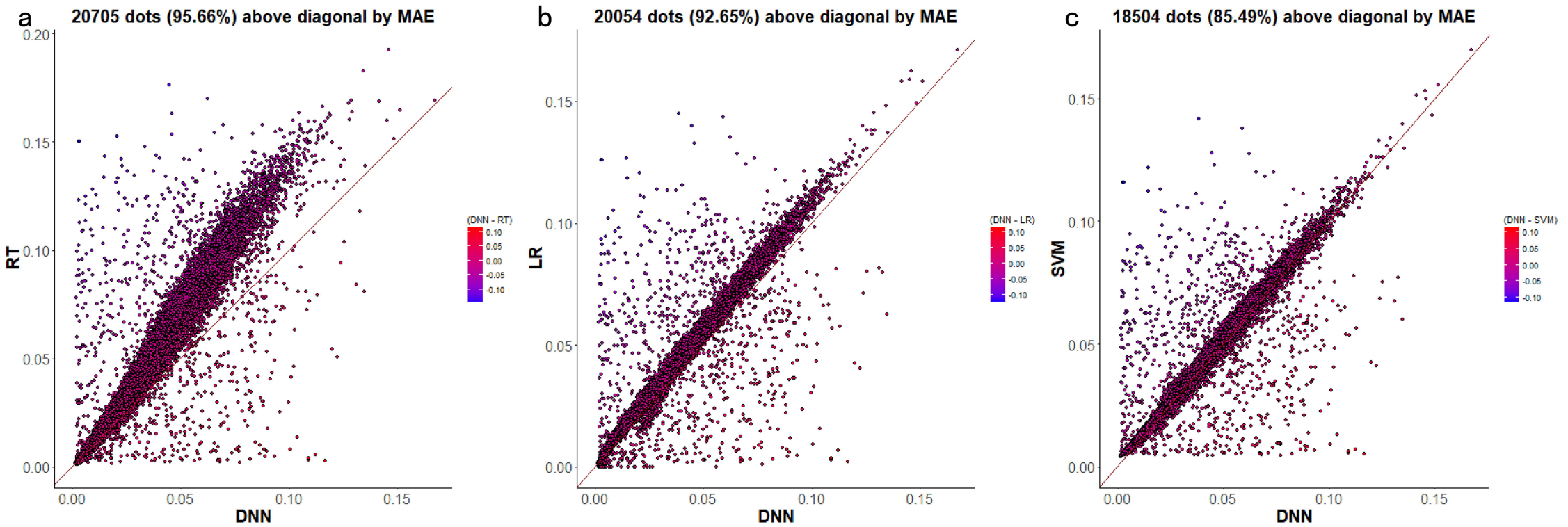
31. Wongprasert, H., Somanunt, S., De Filippo, R., Picard, J. Y. & Pitukcheewanont, P. (2013) A novel mutation of anti-Mullerian hormone gene in Persistent Mullerian Duct Syndrome presented with bilateral cryptorchidism: a case report, *Journal of pediatric urology.* **9**, e147-9.

32. Ticho, B. H. (2010) Iris transillumination defects associated with pallister-killian syndrome, *Journal of pediatric ophthalmology and strabismus.* **47**, 58-9.

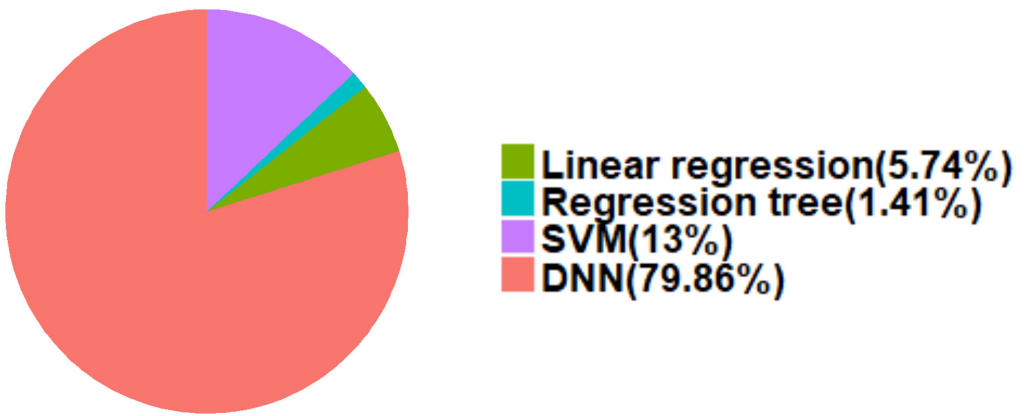

type1
1000 nodes
type2
2000 nodes
input layer
902 nodes
type9
9000 nodes
output layer
21645 nodes
Fully connected
hidden layers
(1-8 layers)


The overall MAE errors of D-PGM-15% with different architectures on test data
MAE
0.032
0.034
0.036
0.038
0.04
0.042
0.044
the number of hidden layers
8
6
4
2
0
hidden units in each hidden layer
0
2000
4000
6000
8000
10000

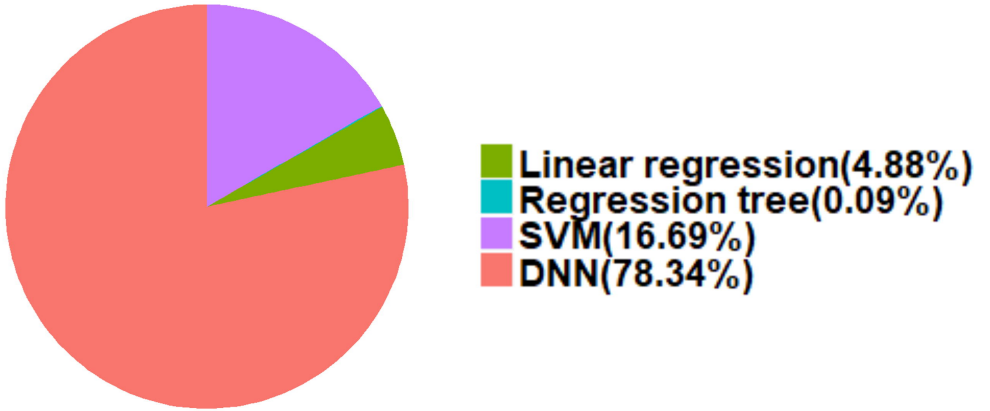

MAE distribution for target genes grouped by model type
density
MAE
30
20
10
0
0.00
0.05
0.10
0.15
0.20
methods
DNN
Linear Regression
Regression Tree
Support Vector Machine

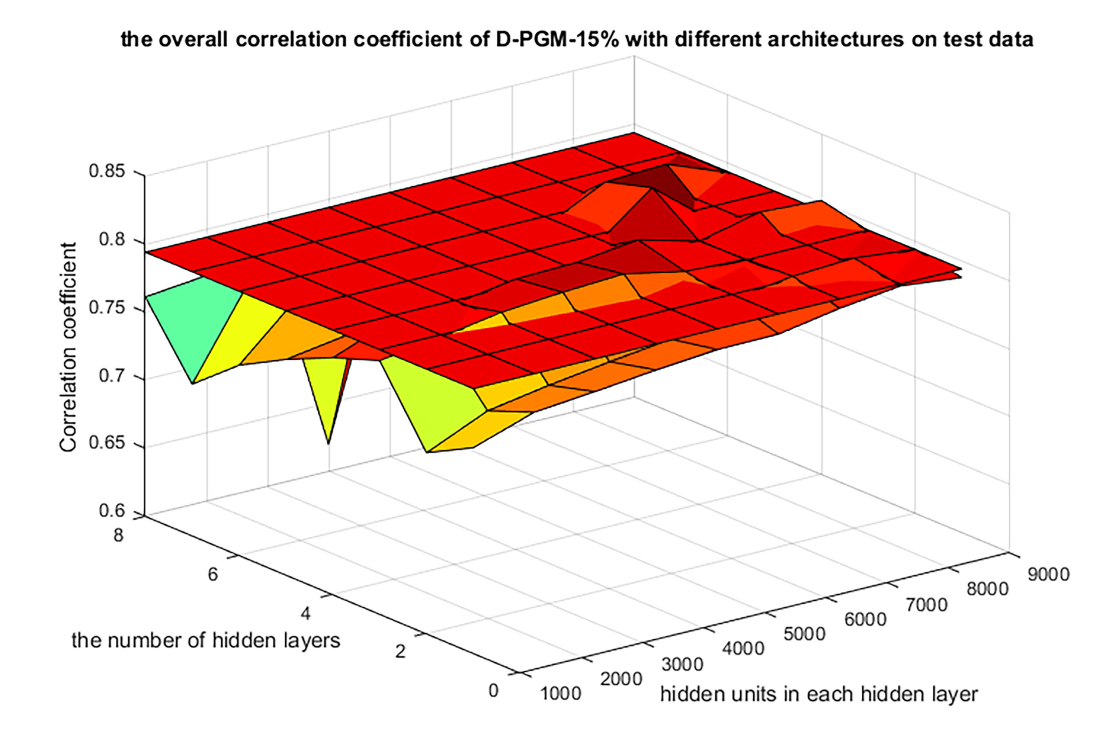

a
20705 dots (95.66%) above diagonal by MAE
RT
DNN
(DNN - RT)
0.10
0.05
0.00
-0.05
-0.10
b
20054 dots (92.65%) above diagonal by MAE
LR
DNN
(DNN - LR)
c
18504 dots (85.49%) above diagonal by MAE
SVM
DNN
(DNN - SVM)
0.00
0.05
0.10
0.15
0.20

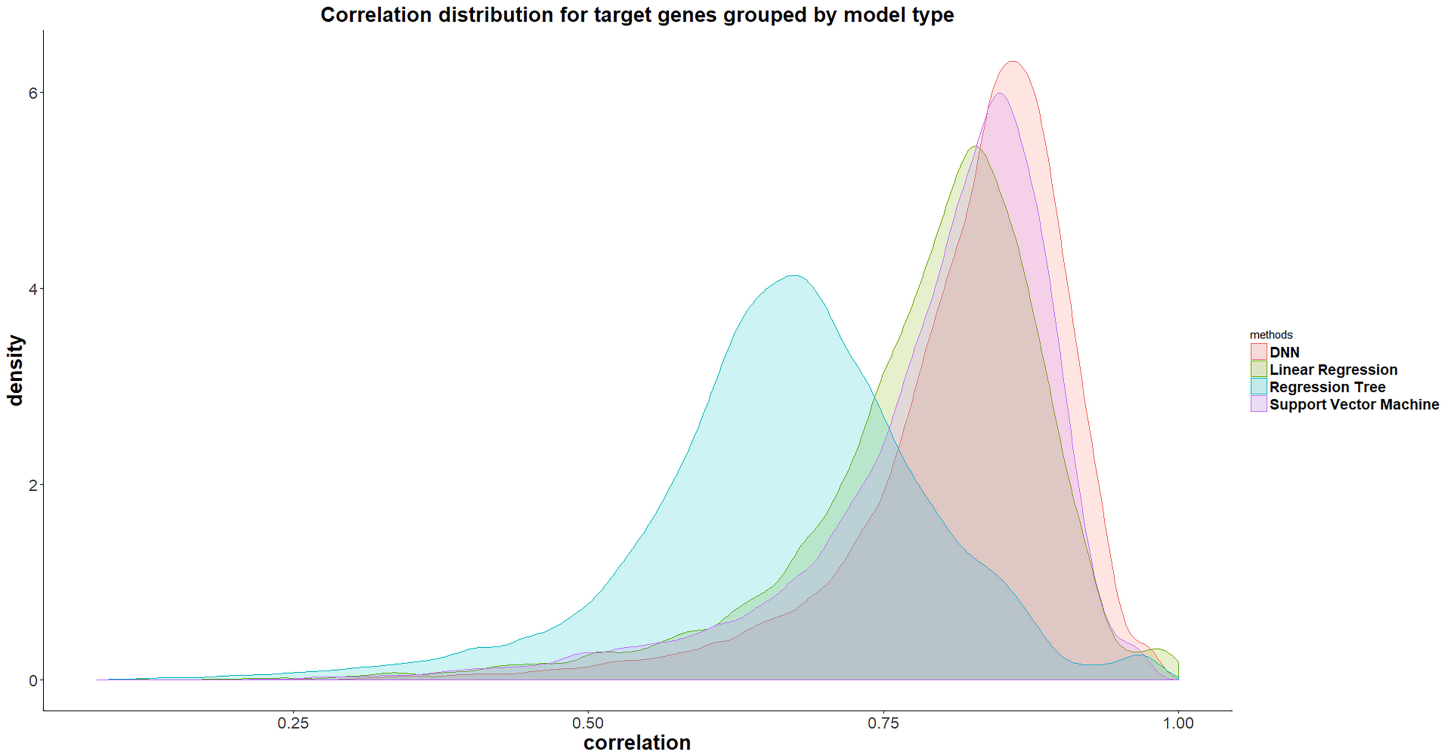
a The distribution of best model based on MAE

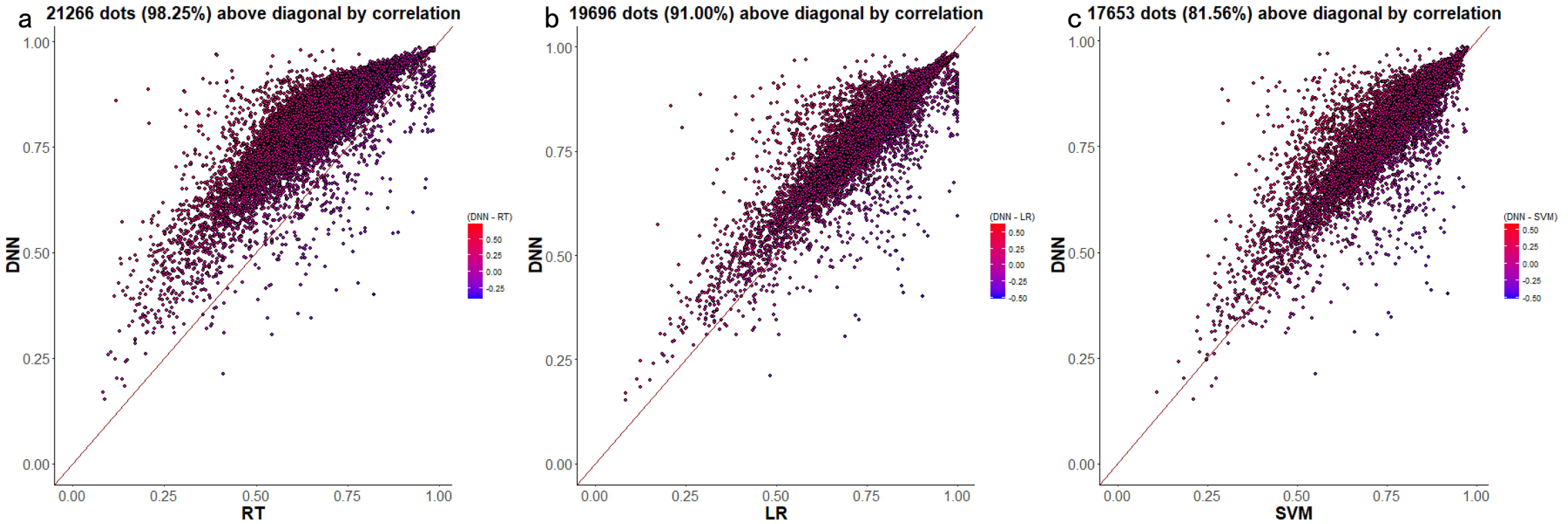


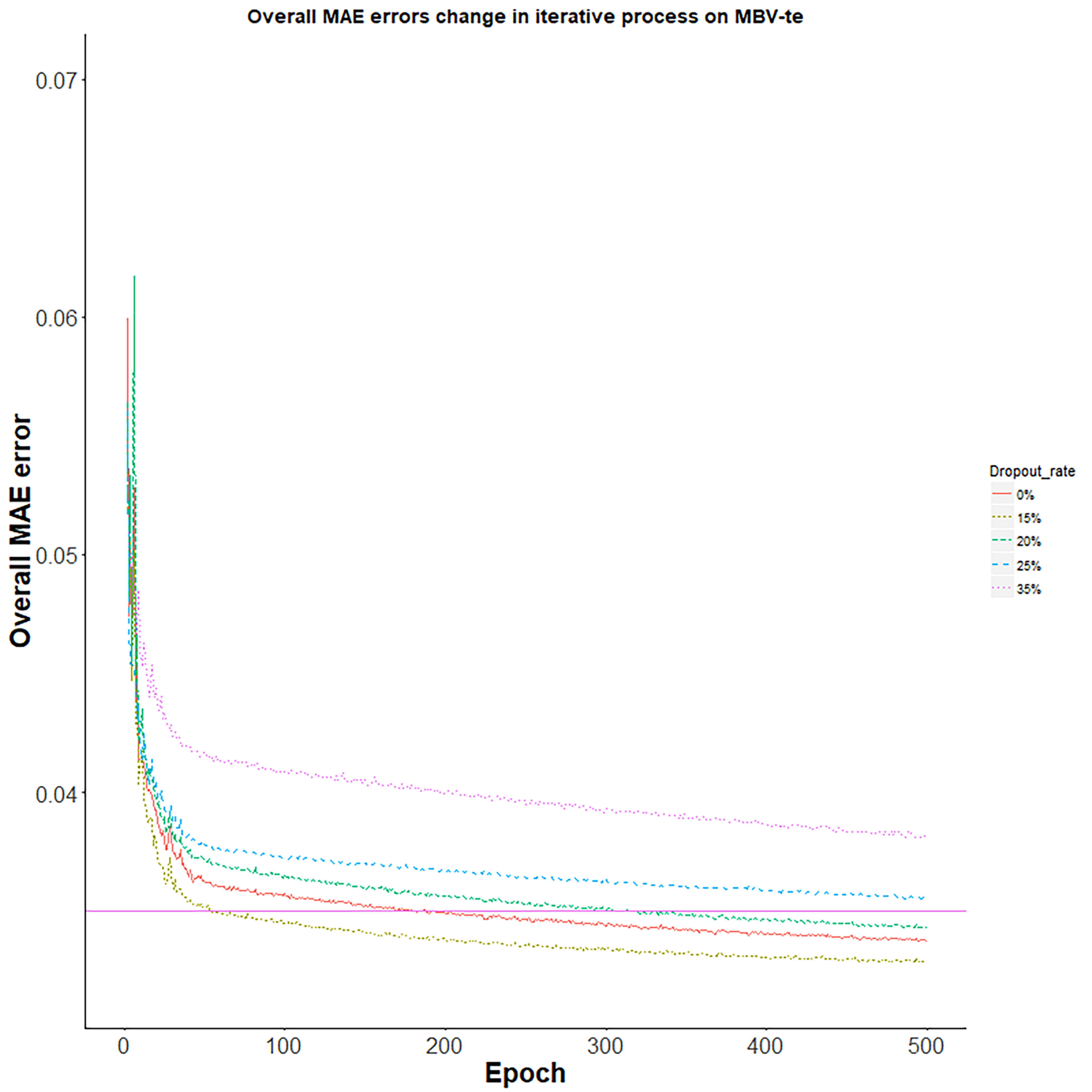
b The distribution of best model based on PCC




the overall correlation coefficient of D-PGM-15% with different architectures on test data
Correlation coefficient
0.85
0.8
0.75
0.7
0.65
0.6
the number of hidden layers
8
6
4
2
0
hidden units in each hidden layer
1000
2000
3000
4000
5000
6000
7000
8000
9000


Correlation distribution for target genes grouped by model type
density
correlation
0
2
4
6
0.25
0.50
0.75
1.00
methods
DNN
Linear Regression
Regression Tree
Support Vector Machine


a
21266 dots (98.25%) above diagonal by correlation
DNN
RT
(DNN - RT)
0.50
0.25
0.00
-0.25
b
19696 dots (91.00%) above diagonal by correlation
DNN
LR
(DNN - LR)
0.50
0.25
0.00
-0.25
-0.50
c
17653 dots (81.56%) above diagonal by correlation
DNN
SVM
(DNN - SVM)
0.50
0.25
0.00
-0.25
-0.50
0.00
0.25
0.50
0.75
1.00


Overall MAE errors change in iterative process on MBV-te
0.07
0.06
0.05
0.04
Overall MAE error
0
100
200
300
400
500
Epoch
Dropout_rate
0%
15%
20%
25%
35%