BMC Bioinformatics

RESEARCH

Open Access

# DeepMethyGene: a deep-learning model to predict gene expression using DNA methylations

Yuyao Yan[1†], Xinyi Chai[1†], Jiajun Liu[1,2], Sijia Wang[1], Wenran Li[1*] and Tao Huang[1,3*]

[†]Yuyao Yan and Xinyi Chai have contributed equally to this work.

*Correspondence: liwenran@picb.ac.cn; huangtao@sinh.ac.cn

[1] CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China
[2] School of Life Sciences, Shanghai University, Shanghai, China
[3] Department of Artificial Intelligence and Digital Health, CAS Engineering Laboratory for Nutrition, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

## Abstract

Gene expression is the basis for cells to achieve various functions, while DNA methylation constitutes a critical epigenetic mechanism governing gene expression regulation. Here we propose DeepMethyGene, an adaptive recursive convolutional neural network model based on ResNet that predicts gene expression using DNA methylation information. Our model transforms methylation Beta values to M values for Gaussian distributed data optimization, dynamically adjusts the output channels according to input dimension, and implements residual blocks to mitigate the problem of gradient vanishing when training very deep networks. Benchmarking against the state-of-the-art geneEXPLORE model ($R^2 = 0.449$), DeepMethyGene ($R^2 = 0.640$) demonstrated superior predictive performance. Further analysis revealed that the number of methylation sites and the average distance between these sites and gene transcription start sites (TSS) significantly affected the prediction accuracy. By exploring the complex relationship between methylation and gene expression, this study provides theoretical support for disease progression prediction and clinical intervention. Relevant data and code are available at https://github.com/yaoyao-11/DeepMethyGene.

**Keywords:** DNA methylation, Gene expression, Diseases, Deep learning

## Background

Gene expression is the basic process of transforming genetic information into biological functions, which is of central significance for the growth, development, and functional regulation of organisms. Aberrant gene expression plays a pivotal role in the initiation and progression of many diseases, including cancer. For instance, mutations or dysregulated expression of tumor suppressor genes such as P53 and PTEN disrupt critical cellular pathways, including apoptosis and cell cycle control, thereby promoting oncogenesis [1, 2]. By accurately predicting gene expression patterns, researchers can uncover disease-specific molecular mechanisms, offering valuable insights into targeted strategies for disease prevention, early diagnosis, and therapeutic intervention.

DNA methylation is an important epigenetic modification that regulates gene expression by adding methylated groups in DNA cytosine. Changes in methylation patterns, particularly abnormal methylation in gene promoter regions, are often associated with the development of various diseases and can serve as biomarkers for tumors [3]. Thus, in-depth study of the effects of methylation on gene expression not only helps to understand the complex mechanisms of gene regulation but may also provide new perspectives and approaches for treating various diseases, particularly cancer and genetic disorders.

In recent years, machine learning has been extensively applied in the field of biology, including genomics [4–6], gene regulation [7–14], protein structure prediction and functional analysis [15–18], drug discovery [19–22], and biomarker discovery [23–26]. Feature extraction is crucial for understanding biological data, and deep learning provides powerful approaches. Techniques such as Convolutional Neural Networks (CNNs), Graph Convolutional Networks (GCNs), and node2vec address distinct challenges.

In the field of gene regulation and epigenetic analysis, several studies have focused on predicting miRNA-disease associations using diverse architectures: DBNMDA [27], which combines unsupervised pre-training with supervised fine-tuning, implemented deep belief networks to model complex association patterns, which innovatively uses the information of all miRNA-disease pairs during the pre-training process. These approaches were further extended by network embedding techniques, as demonstrated in NCMD [28], which integrated node2vec with neural collaborative filtering to capture topological features in biological networks, demonstrating the potential of integrating network embedding and deep learning techniques.

Studies such as Gunasekaran et al. [29] employed CNN architectures combined with bidirectional LSTM networks that increases the accuracy of DNA sequence classification, demonstrating the effectiveness of hierarchical feature learning in genomic applications. DeepRHD [30] combines CNN-based feature extraction with traditional machine learning classifiers, achieving notable improvements in remote homology detection across multiple benchmark datasets through this multi-stage approach. The integration of complementary feature extraction strategies has proven particularly valuable for modeling complex biological interactions. A recent study[10] exemplifies this trend, where the combination of kernel methods with graph convolutional networks (GCNs) yielded high-accuracy predictions of lncRNA-protein interactions. These examples collectively highlight how hybrid architectures that merge different feature extraction paradigms can effectively capture intricate biomolecular relationships.

To better understand the relationship between methylation and gene expression, it is necessary to study transcriptional regulatory regions. In particular, enhancers are critical regulatory elements that can be affected by methylation and play significant roles in aberrant gene expression in cancers [31], and may be located either proximally or distally to their target genes [32]. This highlights the importance of considering the methylation status of distal genomic regions in gene regulation and related disease research. DB Seal et al. [33] utilized deep learning to integrate DNA methylation and copy number variation data to predict gene expression, but the analysis was limited to promoter regions within 1500 BP of the TSS, overlooking methylation in other regions.

The geneEXPLORE [34] emphasized the importance of long-distance DNA methylation in predicting gene expression, proposing the innovative view that methylation at distant locations within a gene may be more significant than proximal methylation, although the prediction accuracy ($R^2$) was only 0.491, indicating room for improvement. In the meantime, the static architecture limits biological data processing. Fixed output channels can't adapt to different input sizes or feature complexities, often resulting in either wasted computing power or lost information.

Therefore, we proposed DeepMethyGene to construct a model using variable convolution kernel and ResNet block in an attempt to further improve the accuracy of predicting gene expression based on methylation. DeepMethyGene predicted the expression levels of 13,982 genes in TCGA breast cancer data, achieving a five-fold cross-validation result $R^2$ of 0.64. The DeepMethyGene model showed significant superiority over the geneEXPLORE model in predicting gene expression, especially when the number of methylation sites within a 1 Mb radius around the gene location was limited and the average distance to the gene transcription start site (TSS) was minimal, leading to notably higher prediction accuracy.

## Methods

### Datasets

We used DNA methylation (450 K array) and RNA sequencing data from several cancer types in TCGA, including breast cancer (BRCA), colon cancer (COAD), glioblastoma (GBM), and lung adenocarcinoma (LUAD). The BRCA dataset, consisting of 873 samples (788 tumor and 85 normal samples), served as the primary training set and the core dataset for downstream analysis. To evaluate the model's generalization performance, we also assessed its performance on the COAD, GBM, and LUAD datasets. All data were obtained from the Xena Public Data Hubs.

### Data preprocessing

Data preprocessing followed the methodology outlined in [34]. Methylation data were filtered and imputed for missing values and converted from beta to M values. Gene expression data were filtered according to expression level and promoter region probe information to obtain highly representative probes and genes.

The BRCA dataset was used as the core for model construction, and the BRCA normal samples ($N = 85$) were used to retrain the model to reduce the effect of potential heterogeneity on model performance. COAD, GBM and LUAD datasets were selected as validation to evaluate the predictive performance in different cancer types. Performance indicators included $R^2$ value and other related evaluation indicators.

### DeepMethyGene model

We present DeepMethyGene, a predictive model built on an adaptive, recursive convolutional neural network architecture inspired by ResNet [35]. This model is specifically designed to process one-dimensional input data, making it ideally suited for time series analysis and signal processing. The network architecture is pivotal in its ability to dynamically adjust its complexity based on the scale of input data, thereby

optimizing computational efficiency and effectiveness. The integration of residual blocks within the architecture significantly enhances the model's ability to train deeply without succumbing to common issues like gradient vanishing or explosion.

DeepMethyGene employs an AdaptiveRegressionCNN architecture, consisting of 2 Conv1 d layers followed by fully connected layers for feature mapping and regression prediction. After the first convolution, the network incorporates a Residual Block to enhance feature representation, followed by a second convolution layer and another residual block. The final regression output is computed through a 512-unit fully connected layer and a single-neuron output layer. The model is optimized using Adam (lr = 0.001) with MSELoss as the loss function and is evaluated via fivefold cross-validation, computing the Pearson correlation coefficient and $R^2$. The training process runs for a maximum of 700 epochs, with early stopping (patience = 90) to prevent overfitting. A more detailed discussion of the Residual Block design and its role in deep learning will be presented in the following section.

Additionally, we also utilized Support Vector Machines (SVM) [36] and Deep Forest [37] as core models for modeling and prediction. However, both were found to be less effective compared to DeepMethyGene in handling the complex, high-dimensional bioinformatics data typically involved in this study.

#### Design of residual blocks

Residual blocks are constructed with two layers of one-dimensional convolution (Conv1 d), where each layer maintains the same number of input and output channels, uses a kernel size of 3, and padding of 1 to preserve the dimensionality of the data through the convolution process. This design allows the addition of more convolutional layers without altering the data dimensions. Following each convolution layer, a LeakyReLU activation function with a negative slope of 0.01 is employed to mitigate the issue of information loss due to the activation function's dead zones. Residual blocks enhance the stability of deep network training by establishing direct connections between the output of the convolution layers and the original input, thus alleviating issues related to gradient vanishing or explosion in deep networks.

#### Architecture of the adaptive regression convolutional neural network

The adaptive regression CNN architecture consists of convolutional layers and residual blocks. To handle varying input dimensions effectively, we designed a dynamic channel allocation strategy: the first convolutional layer adapts its output channels based on input size, with a minimum of 64 channels. The subsequent layers reduce this to 32 channels minimum. We incorporated residual blocks between convolutional layers and then the extracted features are processed through fully connected layers, whose dimensions are determined by our _calculate_to_linear function. This design enables robust regression performance across different input scales, as shown in Fig. 1.

#### Support vector machine (SVM)

The Support Vector Machine (SVM) [36] is a widely-used supervised learning algorithm that effectively performs classification and regression analysis by constructing one or

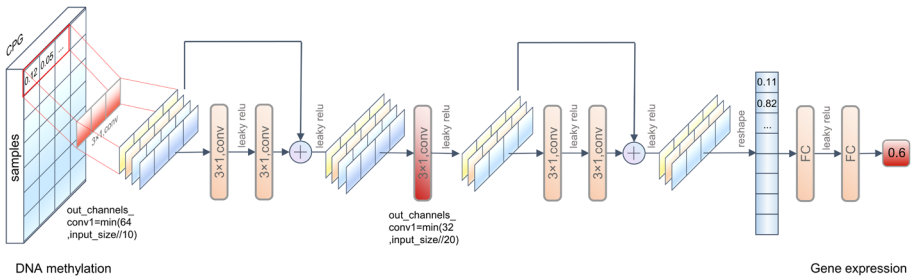


**Fig. 1** Framework of DeepMethyGene. This diagram illustrates the framework of DeepMethyGene. Input Data: The input is gene expression data from 13,892 genes. The number of methylations between the left and right 1 Mb of different gene expression sites is different

more hyperplanes to maximally separate data categories. SVM is particularly effective in small datasets and low-dimensional spaces, as SVM's kernel-based dimensionality transformation enables effective classification even with limited samples.

In our analysis, while SVM performed well for basic classification tasks, it showed notable limitations with our high-dimensional bioinformatics datasets. The computational burden became prohibitive when processing feature vectors in the order of thousands, particularly during kernel optimization. These constraints significantly impacted the model's practical applicability for large-scale genomic analyses. Additionally, SVM faces challenges in processing nonlinear and complex patterns, particularly in time-series data analysis, where data typically includes time-dependent and dynamically changing features.

### Deep forest

Deep Forest [37], also known as cascade forest, is an emerging machine learning framework based on ensemble learning principles, which constructs a multi-layered model structure of decision trees. Each layer's output serves as the input for the next layer, creating a multi-tiered decision system. This model excels in certain machine learning tasks, especially in handling data with complex structures and inter-feature relationships. In our study, Deep Forest demonstrated certain advantages in processing structured data such as tabular data, particularly when non-linear relationships between features existed. However, when applied to time-series data, despite its ability to handle complex relationships between features to some extent, Deep Forest did not perform as well as deep learning methods like ResNet in capturing the continuity and dynamic changes inherent in time-series data. Time-series data typically involves dependencies over time, which necessitates a model capable of capturing and utilizing these dependencies for effective prediction, an area where the static structure of Deep Forest shows limitations.

## Results

### DeepMethyGene accurately predicts gene expression using DNA methylation

To ensure the validity of data utilization and the equity of result comparison, we employed a five-fold cross-validation method to evaluate multiple computational models, including DeepMethyGene, geneEXPLORE, Support Vector Machine

**Table 1** Comparison of gene expression prediction performance across models

| Method | Mean | Median | Std | Max | Min | *P*-value |
|---|---|---|---|---|---|---|
| SVM | 0.327 | 0.317 | 0.132 | 0.767 | − 0.357 | < 0.0001 |
| RandomForest | 0.374 | 0.365 | 0.139 | 0.936 | − 0.108 | < 0.0001 |
| geneEXPLORE | 0.449 | 0.452 | 0.141 | 0.938 | 0.005 | < 0.0001 |
| NoConvLayer | 0.360 | 0.358 | 0.205 | 0.932 | 0.001 | < 0.0001 |
| DeepMethyGene | 0.640 | 0.641 | 0.133 | 0.957 | 0.165 | |

The table presents the predictive performance of DeepMethyGene compared to DeepMethyGene (NoConvLayer), geneEXPLORE, SVM, and Random Forest. The P-value indicates the significance level of the performance difference between each model and DeepMethyGene, calculated using Wilcoxon signed-rank test

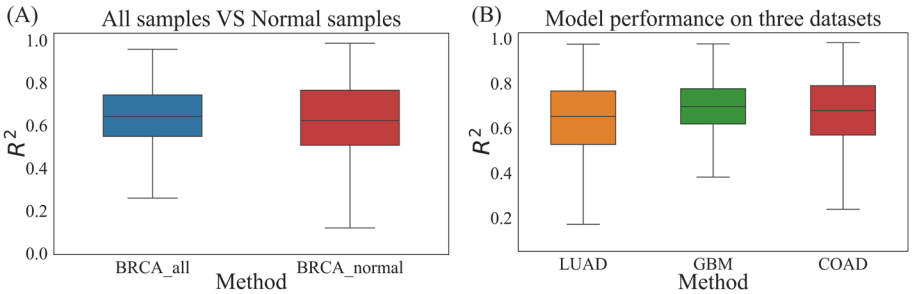


**Fig. 2** Predictive performance of DeepMethyGene on different datasets. (**A**) The predictive performance of the model trained exclusively on the BRCA dataset of normal samples ($R^2 = 0.628$) is comparable to that of the model trained on all samples ($R^2 = 0.640$). (**B**) Predictive performance of the model evaluated across three independent datasets: LUAD ($R^2 = 0.640$), GBM ($R^2 = 0.696$), and COAD ($R^2 = 0.668$)

(SVM), and Random Forest, across the same set of 13,982 genes. During this process, DeepMethyGene demonstrated the highest predictive performance, achieving a mean $R^2$ value of 0.640, which was significantly higher than that of geneEXPLORE (0.449), SVM (0.327), and Random Forest (0.374) (Table 1). These results indicate that DeepMethyGene, leveraging its convolutional neural network architecture, was able to more effectively capture and utilize data features, leading to superior predictive accuracy.

Furthermore, to assess the contribution of the convolutional layers, we conducted an ablation study by constructing a variant of DeepMethyGene without convolutional layers (NoConvLayer). The results showed that removing the convolutional layers led to a considerable drop in performance, with the mean $R^2$ decreasing to 0.360, which is significantly lower than the full model's performance (0.640). This underscores the importance of the convolutional layers in enhancing model performance. Table 1 provides a detailed summary of the predictive performance across different models, further validating the advantages of DeepMethyGene in processing complex bioinformatics data with deep learning technology.

Considering the presence of both cancer and normal samples in the dataset, to minimize the potential impact of sample heterogeneity on model performance and generalization ability, we retrained the model using a dataset that exclusively comprised normal samples, as shown in Fig. 2A. Compared to the predictive performance obtained from training with all samples ($R^2 = 0.640$), DeepMethyGene maintained a high level of predictive accuracy for normal samples ($R^2 = 0.628$). Therefore, we chose all samples

containing cancer and normal samples as the data set, with more sample data and thus better results.

We evaluated its performance on three independent datasets, which encompassed colorectal adenocarcinoma (COAD), glioblastoma multiforme (GBM), and lung adenocarcinoma (LUAD), respectively, as illustrated in Fig. 2B. The results showed that the model's predictive performance on these three datasets was comparable to its performance on the breast invasive carcinoma (BRCA) dataset, with $R^2$ values of 0.668, 0.696, and 0.640 for the COAD, GBM, and LUAD datasets, respectively. We demonstrated that DeepMethyGene exhibits good predictive performance across different TCGA subsets, reflecting its strong generalization and robustness.

#### The prediction performance is influenced by the number of CpGs near a gene

In our analysis of the predictive performance of DeepMethyGene across 13,982 genes, we observed a significant correlation between the accuracy of the predictions and the density of methylation sites within a 1 Mb radius of the gene locus. Specifically, there appears to be a relationship between the concentration of methylation sites and predictive accuracy. Additionally, a comparative analysis was conducted between the predictions of DeepMethyGene and another predictive tool, geneEXPLORE. The results indicated that DeepMethyGene's predictions, as measured by the $R^2$ value, surpassed those of geneEXPLORE for 13,734 genes, while for 248 genes, the performance was inferior. This outcome is directly associated with the number of methylation sites (Fig. 3A).

To further investigate this phenomenon, we analyzed the relationship between model performance and methylation site density. Our results showed that the comparative advantage of DeepMethyGene over geneEXPLORE diminished as methylation site numbers increased (Fig. 3B). This inverse correlation suggests that dense methylation patterns may introduce signal complexity that affects DeepMethyGene's prediction accuracy. These findings highlight how methylation site distribution patterns influence model effectiveness, an important consideration for future algorithm development.

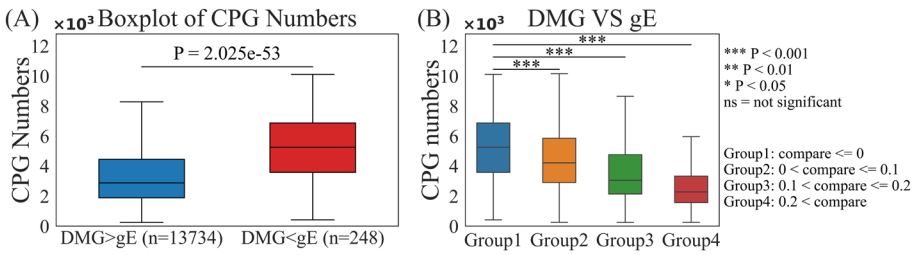


**Fig. 3** Relationship between the number of methylation sites and predictive outcomes of geneEXPLORE and DeepMethyGene. (**A**) Illustrates that the number of methylation sites within a 1 Mb radius of the gene expression loci affects the predictive performance of both DeepMethyGene and geneEXPLORE. As observed, DeepMethyGene demonstrates superior performance compared to geneEXPLORE in regions with relatively lower methylation site density. (**B**) Illustrates the relationship between the differential predictive performance of DeepMethyGene and geneEXPLORE in terms of the $R^2$ value for gene expression predictions and the number of methylation sites. Compare = $R^2$(DeepMethyGene) − $R^2$(geneEXPLORE). The results indicate that a lower number of methylation sites corresponds to a superior predictive performance of DeepMethyGene over geneEXPLORE

### The prediction performance is associated with the distance between CpGs and genes

Our study analyzed gene expression predictions across 13,982 genes using DeepMethyGene, revealing a strong relationship between prediction accuracy and methylation site proximity to transcription start sites (TSS). We compared DeepMethyGene's performance with geneEXPLORE to evaluate the strengths and limitations of different prediction approaches. The comparative analysis provided key insights into how spatial distribution of methylation sites influences prediction accuracy. Our findings indicated that DeepMethyGene surpassed geneEXPLORE in predictive $R^2$ values for 13,734 genes, while it underperformed for 248 genes. This variation in predictive accuracy appears to be associated with the average distance between methylation sites and gene TSS (Fig. 4A).

We further analyzed the relationship between the methylation sites and the average distance of the TSS. Our analysis showed that DeepMethyGene achieved better prediction accuracy ($R^2$) than geneEXPLORE when methylation sites were located closer to the TSS (Fig. 4B). This spatial dependence indicates that the distribution of methylation sites around the TSS significantly affects the performance of the model. The improved prediction accuracy in the proximal region may reflect the enhanced ability of DeepMethyGene to capture local regulatory patterns.

## Conclusions

In this study, we used 450K methylation array data and gene expression data from TCGA dataset to construct DeepMethyGene, a novel deep learning method that can predict gene expression data. One of the key challenges in methylation analysis is the inconsistency in the number and location of methylation sites across different genes. Therefore, we built an adaptive recursive convolutional neural network to solve this problem. Compared to the traditional geneEXPLORE framework, DeepMethyGene can dynamically adjust the number of channels based on the different input dimensions, to efficiently extract features from complex methylation profiles, which not only ensures the prediction accuracy, but also improves the computational efficiency.

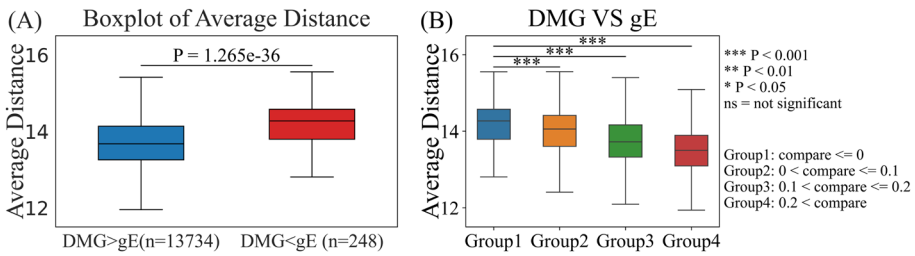


**Fig. 4** The average distance between the transcription start sites and methylation sites influences the difference in $R^2$ values between DeepMethyGene and geneEXPLORE. (**A**) Suggests that the average distance between transcription start sites and methylation sites impacts the differential $R^2$ values between DeepMethyGene and geneEXPLORE. DeepMethyGene tends to outperform geneEXPLORE when the average distance is comparatively smaller. (**B**) Illustrates the relationship between the differential predictive performance of DeepMethyGene and geneEXPLORE in terms of the $R^2$ value for gene expression predictions and the average distance between transcription start sites and methylation sites. Compare = $R^2$(DeepMethyGene) − $R^2$(geneEXPLORE). The results indicate that a lower average distance corresponds to a superior predictive performance of DeepMethyGene over geneEXPLORE

Another core framework of our model is ResNet, which has many advantages that can be used to better cope with our problem of learning the complex relationship between methylation data and gene expression. The results also show that our model has advantages over the traditional algorithms (elastic network, SVM and random forest) of geneEXPLORE. In addition, in terms of data selection and processing, we converted the methylation β values into m values to achieve a Gaussian-like data distribution, which also improved the statistical properties and predictive stability of the model. Future research may extend the application of this model to a broader range of time series analysis tasks and further optimize the network structure to improve predictive accuracy and generalization ability, which is crucial for understanding the molecular mechanisms of diseases and developing effective treatment strategies.

**Author contributions**
Yuyao Yan: performed experiments and wrote the main manuscript text; Xinyi Chai: performed experiments and wrote the main manuscript text; Jiajun Liu: provided technical support and prepared figures and tables; Sijia Wang: review & editing; Wenran Li: review & editing, Supervision; Tao Huang: review & editing, Supervision.

**Funding**
This work was supported by National Key R&D Program of China [2022YFF1203202], MajorProject of Guangzhou National Laboratory [GZNL2024 A01003], Strategic Priority Research Program of Chinese Academy of Sciences[XDB38050200, XDA26040304], CAS Young Team Program for Stable Support of Basic Research [YSBR- 077], National Natural ScienceFoundation of China [32200472] and China Postdoctoral Science Foundation [2021M693274, BX2021336].

**Availability of data and materials**
The data and code are available at: https://github.com/yaoyao-11/DeepMethyGene.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 25 December 2024 Accepted: 17 March 2025
Published online: 08 April 2025

**References**

1. Di Cristofano A, Pandolfi PP. The multiple roles of PTEN in tumor suppression. Cell. 2000;100(4):387–90.
2. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16(1):6–21.
3. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007;128(4):683–92.
4. Alharbi WS, Rashid M. A review of deep learning applications in human genomics using next-generation sequencing data. Hum Genom. 2022;16(1):26.
5. Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JWR, Fajardo-Flores SB, Gaytan-Lugo LS, Santana-Mancilla PC, Crossa J. A review of deep learning applications for genomic selection. BMC Genom. 2021;22(1):19.
6. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12–8.
7. Gan M, Li W, Jiang R. EnContact: predicting enhancer-enhancer contacts using sequence-based deep learning model. PeerJ. 2019;7:e7657.
8. Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res. 2019;47(10):e60.
9. Shen C, Chen Y, Xiao F, Yang T, Wang X, Chen S, Tang J, Liao Z. BAT-Net: An enhanced RNA Secondary Structure prediction via bidirectional GRU-based network with attention mechanism. Comput Biol Chem. 2022;101:107765.
10. Shen C, Mao D, Tang J, Liao Z, Chen S. Prediction of LncRNA-Protein Interactions Based on Kernel Combinations and Graph Convolutional Networks. IEEE J Biomed Health Inform. 2024;28(4):1937–48.
11. Ha J, Park C. MLMD: metric learning for predicting MiRNA-disease associations. IEEE Access. 2021;9:78847–58.
12. Ha J. SMAP: Similarity-based matrix factorization framework for inferring miRNA-disease association. Knowl-Based Syst. 2023;263:110295.

13. Ha J. MDMF: predicting miRNA–disease association based on matrix factorization with disease similarity constraint. J Pers Med. 2022;12(6):885. https://doi.org/10.3390/jpm12060885.
14. Ha J, Park C, Park C, Park S. IMIPMF: Inferring miRNA-disease interactions using probabilistic matrix factorization. J Biomed Inform. 2020;102:103358.
15. Baek M, Baker D. Deep learning and protein structure modeling. Nat Methods. 2022;19(1):13–4.
16. Bileschi ML, Belanger D, Bryant DH, Sanderson T, Carter B, Sculley D, Bateman A, DePristo MA, Colwell LJ. Using deep learning to annotate the protein universe. Nat Biotechnol. 2022;40(6):932–7.
17. Jisna VA, Jayaraj PB. Protein structure prediction: conventional and deep learning perspectives. Protein J. 2021;40(4):522–44.
18. Pakhrin SC, Shrestha B, Adhikari B, Dukka BKC. Deep learning-based advances in protein structure prediction. Int J Mol Sci. 2021;22(11):5553. https://doi.org/10.3390/ijms22115553.
19. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573–84.
20. Nag S, Baidya ATK, Mandal A, Mathew AT, Das B, Devi B, Kumar R. Deep learning tools for advancing drug discovery and development. 3 Biotech. 2022. https://doi.org/10.1007/s13205-022-03165-8.
21. Pandey M, Fernandez M, Gentile F, Isayev O, Tropsha A, Stern AC, Cherkasov A. The transformational role of GPU computing and deep learning in drug discovery. Nat Mach Intell. 2022;4(3):211–21.
22. Tropsha A, Isayev O, Varnek A, Schneider G, Cherkasov A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. Nat Rev Drug Discov. 2024;23(2):141–55.
23. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. Br J Cancer. 2021;124(4):686–96.
24. Liang J, Zhang W, Yang J, Wu M, Dai Q, Yin H, Xiao Y, Kong L. Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer. Nat Mach Intell. 2023;5(4):408–20.
25. Mandair D, Reis-Filho JS, Ashworth A. Biological insights and novel biomarker discovery through deep learning approaches in breast cancer histopathology. NPJ Breast Cancer. 2023;9(1):21.
26. Steyaert S, Pizurica M, Nagaraj D, Khandelwal P, Hernandez-Boussard T, Gentles AJ, Gevaert O. Multimodal data fusion for cancer biomarker discovery with deep learning. Nat Mach Intell. 2023;5(4):351–62.
27. Chen X, Li T-H, Zhao Y, Wang C-C, Zhu C-C. Deep-belief network for predicting potential miRNA-disease associations. Brief Bioinf. 2021. https://doi.org/10.1093/bib/bbaa186.
28. Ha J, Park S. NCMD: Node2vec-based neural collaborative filtering for predicting MiRNA-disease association. IEEE/ACM Trans Comput Biol Bioinf. 2023;20(2):1257–68.
29. Gunasekaran H, et al. Analysis of DNA Sequence Classification Using CNN and Hybrid Models. Comput Math Methods Med. 2021;2021:1835056.
30. Routray M, Vipsita S, Sundaray A, Kulkarni S. DeepRHD: An efficient hybrid feature extraction technique for protein remote homology detection using deep learning strategies. Comput Biol Chem. 2022;100:107749.
31. Sur I, Taipale J. The role of enhancers in cancer. Nat Rev Cancer. 2016;16(8):483–93.
32. Mora A, Sandve GK, Gabrielsen OS, Eskeland R. In the loop: promoter-enhancer interactions and bioinformatics. Brief Bioinform. 2016;17(6):980–95.
33. Seal DB, Das V, Goswami S, De RK. Estimating gene expression from DNA methylation and copy number variation: a deep learning regression model for multi-omics integration. Genomics. 2020;112(4):2833–41.
34. Kim S, Park HJ, Cui X, Zhi D. Collective effects of long-range DNA methylations predict gene expressions and estimate phenotypes in cancer. Sci Rep. 2020;10(1):3920.
35. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 27–30. 770–778.
36. Suthaharan S. Support vector machine. In: Suthaharan S, editor. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. Boston, MA: Springer US; 2016. p. 207–35. https://doi.org/10.1007/978-1-4899-7641-3_9.
37. Zhou Z-H, Feng J. Deep forest. Natl Sci Rev. 2017;6:74–86.

## Publisher's Note