ACCSCIENCE PUBLISHING

Innovative Medicines & Omics

ORIGINAL RESEARCH ARTICLE

Predicting DNA methylation from RNA-sequencing data in renal clear cell carcinoma: A deep variational autoencoder approach

Muhammad Salman*

Kabir Medical College, Gandhara University, Peshawar, Pakistan

Abstract

While RNA-sequencing (RNA-seq) is cost-effective and widely available, obtaining comprehensive DNA methylation profiles remains resource-intensive. To bridge this gap, we developed a deep variational autoencoder (VAE) framework for predicting DNA methylation patterns directly from RNA-seq data in renal clear cell carcinoma. Using The Cancer Genome Atlas–Kidney Renal Clear Cell Carcinoma dataset—sourced from cBioPortal and selected for its clinical relevance in renal clear cell carcinoma and high-quality paired omics data—we implemented a rigorous pre-processing pipeline including quality control, missing data imputation, and RNA feature selection based on variance and predictive power. A patient-wise 80/20 train–test split was employed to prevent data leakage. A deep VAE (NetVAE) with residual encoder blocks was trained to learn a compressed latent representation of the methylome and reconstruct genome-wide beta values directly from RNA-seq input. On the independent test set ($n = 55$ patients), the NetVAE achieved R2 = 0.627, mean absolute error = 0.135, and Pearson correlation coefficient $r = 0.792$ ($p < 0.001$). Spearman correlation was 0.641 ($p < 0.001$). Scatter and density plots confirmed strong predictive power across the full range of beta values. This framework offers a practical, cost-effective method for inferring methylation landscapes from RNA-seq data, with significant potential for biomarker discovery and retrospective epigenetic studies in renal clear cell carcinoma, where direct methylation profiling is unavailable.

Keywords: DNA methylation; RNA sequencing; Machine learning; Renal clear cell carcinoma; Variational autoencoder; Epigenetics

***Corresponding author:**
Muhammad Salman
(Salman1311K@gmail.com)

Citation: Salman M. Predicting DNA methylation from RNA-sequencing data in renal clear cell carcinoma: A deep variational autoencoder approach. *Innov Med Omics*.
doi: 10.36922/IMO025420055

Received: October 17, 2025

Revised: November 30, 2025

Accepted: April 30, 2026

Published online: June 12, 2026

1. Introduction

1.1. Background and clinical significance

The addition of methyl groups to cytosine bases at cytosine–phosphate–guanine (CpG) dinucleotides in DNA constitutes one of the most extensively studied epigenetic modifications in human biological systems.[1] This reversible modification functions as a key regulator of gene expression, genome stability, and cellular differentiation.[2] Aberrant DNA methylation patterns—including hypermethylation of tumour suppressor gene promoters and genome-wide hypomethylation leading to chromosomal instability—are hallmarks of cancer development.[3,4]

Renal cell carcinoma (RCC) accounts for approximately 90% of all kidney cancers; clear cell renal cell carcinoma (ccRCC) is its most prevalent subtype, comprising approximately 75% of cases.[5] Comprehensive genomic and epigenomic characterisation has demonstrated that DNA methylation alterations are fundamental drivers of ccRCC development, disease progression, and patient survival outcomes.[6,7] Identifying methylation-based biomarkers holds considerable promise for improving diagnostic precision, risk stratification, and treatment selection in ccRCC.[8]

1.2. Current challenges and opportunities

Platforms such as the Illumina Infinium HumanMethylation450/EPIC BeadChip arrays and whole-genome bisulfite sequencing provide high-resolution methylation data but remain substantially more expensive than RNA-sequencing (RNA-seq).[9] This cost disparity hinders large-scale epidemiological studies, multi-site clinical trials, and retrospective analyses of biobanked RNA-seq data lacking paired methylation information.[10] The majority of cancer biobanks and public repositories harbour extensive RNA-seq datasets from the past decade that are unaccompanied by methylation profiles.[11]

Integration of machine learning with multi-omics data has emerged as a powerful strategy to address these limitations.[12,13] Gene expression patterns captured by RNA-seq encode information that correlates with DNA methylation states, since both molecular layers reflect underlying regulatory networks and cellular phenotypes.[14,15] Several computational methods have attempted to exploit this relationship, yet most focus on pan-cancer models or single-algorithm strategies that may not fully capture the complex non-linear relationships between transcriptomes and methylomes.[16,17]

1.3. Rationale for dataset and model selection

We selected The Cancer Genome Atlas–Kidney Renal Clear Cell Carcinoma dataset (TCGA–KIRC) cohort for several strategic reasons. First, TCGA–KIRC provides high-quality, well-annotated paired RNA-seq and DNA methylation data from the same patient samples, enabling supervised learning with matched ground truth.[18] Second, the cohort contains a sufficient sample size (275 patients after quality control) to train deep learning models while maintaining a fully independent test set.[19] Third, restricting analysis to a single cancer subtype enables the model to learn tissue-specific and disease-specific regulatory patterns rather than averaging across heterogeneous cancer types, potentially improving prediction accuracy.[20,21]

Our choice of a deep variational autoencoder (VAE) architecture—implemented as NetVAE with residual encoder blocks—was motivated by the model's capacity to learn a compact probabilistic latent representation of the high-dimensional methylome while preserving biologically meaningful global structure.[22,23] VAEs excel at generative modelling and can handle the continuous, bounded nature of CpG beta (β) values through their principled probabilistic framework.[24] The residual connections mitigate vanishing-gradient problems in deep networks and enable richer feature hierarchies, which are critical when mapping complex transcriptome–methylome associations.[25] The end-to-end single-model design also eliminates the inter-stage error propagation inherent to two-stage ensemble pipelines.

1.4. Study objectives and novelty

To the best of our knowledge, this study represents the implementation of a deep VAE with residual encoder blocks to predict locus-specific DNA methylation β values from RNA-seq data in renal clear cell carcinoma. This constitutes a methodological advance over existing approaches, which have primarily employed single-algorithm strategies, focused on pan-cancer datasets that may dilute tissue-specific signals, or lacked rigorous evaluation on fully independent test cohorts.[26,27] Our patient-wise data split—which prevents any information from a given patient appearing in both training and test sets—also provides more realistic performance estimates than CpG-level or random splitting strategies used in prior work.

The specific objectives of this study are: (i) To develop and validate a deep VAE framework for predicting DNA methylation β values from RNA-seq expression data in the TCGA–KIRC cohort; (ii) to systematically evaluate prediction performance using multiple complementary metrics on a completely independent test set; (iii) to contextualise the achieved accuracy relative to published cross-omics prediction benchmarks and to assess its practical utility for cost-effective methylation screening and biomarker discovery; and (iv) to provide a transparent, reproducible computational pipeline adaptable to other cancer types or multi-omics integration tasks.

2. Materials and methods

2.1. Data source and patient selection

Paired RNA-seq gene expression and DNA methylation array data were obtained from TCGA–KIRC cohort via the cBioPortal for cancer genomics (https://www.cbioportal.org).[28] The fundamental rationale for using paired data is as follows: The model was first trained on matched RNA-seq and DNA methylation profiles so that it could

subsequently predict methylation from RNA-seq input alone, without requiring methylation assays in prospective or retrospective cohorts. The initial dataset comprised 291 patient samples with available transcriptomic and methylation profiles. To ensure data quality and reliability, we implemented stringent quality control (QC) criteria, resulting in the exclusion of 16 samples (5.5% of the initial cohort) for the following reasons:

- RNA-seq samples with total read counts below 10 million mapped reads (n = 5), indicating insufficient sequencing depth and unreliable expression quantification;
- Methylation array samples with detection p-value failure rate exceeding 5% across probes (n = 4), indicative of technical artefacts or poor sample quality;
- Samples with more than 15% missing values in either RNA-seq or methylation data (n = 4), where imputation would introduce excessive uncertainty;
- Samples lacking critical clinical annotations required for downstream validation analyses, including tumour stage and histological subtype (n = 3).

Sensitivity analyses comparing QC-filtered and unfiltered datasets confirmed that exclusion of these 16 samples reduced noise in both feature spaces without introducing systematic bias in tumour stage or histological subtype distribution. After applying these QC filters, 275 high-quality paired samples were retained for all subsequent analyses.

2.2. RNA-sequencing data processing and quality control

RNA-sequencing data were downloaded as fragments per kilobase million (FPKM) normalised expression values generated by the TCGA RNA-seq pipeline using spliced transcripts alignment to a reference and RNA-seq by expectation-maximization quantification.[29] We applied the following sequential QC and preprocessing steps:

- Gene filtering: Genes with zero or near-zero expression (mean FPKM < 0.1) across all samples were removed to eliminate non-expressed or unreliably detected transcripts, reducing noise in downstream analyses.
- Variance-based feature selection: Median absolute deviation was calculated for each gene across all samples; the top 5,000 genes with the highest variance were retained. This threshold was selected because prior studies have demonstrated that highly variable genes capture most biologically informative variance in cancer transcriptomes while maintaining computational tractability.[30,31] Genes with low expression variance carry minimal predictive signal for epigenetic states.
- Log_2 transformation: FPKM values were incremented by 1 prior to log_2 transformation to stabilise variance and approximate a normal distribution, enhancing model performance.
- Z-score normalisation: Each sample's expression profile was standardised (mean = 0, standard deviation = 1) to remove library-size variations and technical batch effects, ensuring comparability across patients.

2.3. DNA methylation data processing and quality control

DNA methylation data were derived from Illumina Infinium HumanMethylation450 BeadChip arrays processed by the TCGA consortium.[32] Methylation levels are represented as $\beta = M/(M + U)$, where M and U denote methylated and unmethylated signal intensities, respectively; values range continuously from 0 (completely unmethylated) to 1 (fully methylated). Comprehensive QC procedures comprised:

- Probe filtering: Probes mapping to sex chromosomes (X and Y) were excluded (n = 11,648 probes removed) to avoid sex-specific confounding effects.[33]
- Cross-reactive probe removal: Probes known to cross-hybridise to multiple genomic locations were excluded based on published annotation lists (n = 29,233 probes removed), preventing spurious methylation signals.[34]
- Detection p-value filtering: Probes with detection p-values > 0.01 in more than 5% of samples were removed, indicating unreliable measurements.
- Missing value imputation: Residual missing β values (<2% of the total data matrix) were imputed using k-nearest neighbours (k = 5) based on Euclidean distance in methylation space, preserving local correlation structure.[35]

After all filtering steps, 385,553 high-quality CpG probes remained for downstream analysis.

2.4. Feature selection for methylation prediction

To refine the methylation feature space and prioritise CpG sites most amenable to prediction from gene expression, a two-stage selection strategy was applied. First, variance filtering removed CpG sites with variance < 0.01 across samples, as low-variance sites provide minimal predictive information. Second, Pearson correlation coefficients were calculated between each of the 5,000 selected RNA-seq genes and all CpG sites; sites with absolute correlation > 0.3 with at least one gene were retained, reflecting potential regulatory links between methylation and expression.[36] This process yielded 50,000 CpG sites with the highest predictive potential, balancing model complexity with computational feasibility.

2.5. Train–test split strategy

To ensure unbiased performance evaluation, a patient-wise 80/20 stratified train–test split was implemented. All CpG sites and genes from a given patient were assigned exclusively to either the training set (n = 220 patients) or the independent test set (n = 55 patients)—no patient contributed data to both partitions. This patient-wise approach is critical for preventing data leakage, which can spuriously inflate performance metrics when predicting multiple CpG sites from the same individual. Stratification was performed based on tumour stage distribution to maintain representative proportions across both sets. The test set remained completely unseen throughout model training, hyperparameter tuning, and feature selection.

2.6. NetVAE architecture and training

The NetVAE was implemented as a deep VAE in PyTorch to learn a compressed probabilistic latent representation of genome-wide methylation patterns and reconstruct CpG β values directly from RNA-seq input.[37] The encoder pathway comprised: Input layer (RNA features) → dense layer (hidden_size neurons, Gaussian error linear unit (GELU) activation, BatchNorm) → dense layer (hidden_size/2 neurons, GELU, BatchNorm) → two residual blocks (BatchNorm → linear → GELU → dropout) → variational bottleneck (separate μ and $\log\sigma^2$ heads, latent_dim = 128 dimensions). The reparameterisation trick was applied during training to sample $z = \mu + \varepsilon\sigma$ ($\varepsilon \sim N(0, I)$); at inference, $z = \mu$ is deterministic. The decoder mirrored the encoder: Latent z → dense (hidden_size/2, GELU, BatchNorm) → two residual blocks → dense (hidden_size) → output layer (CpG features, linear activation; β values clipped to [0,1] post-hoc).

Training parameters included: Loss function = combined VAE loss (mean squared error reconstruction loss + β-weighted Kullback–Leibler [KL] divergence, β = 0.01); optimiser = AdamW with learning rate 3×10^{-4} and weight decay 10^{-4}; learning rate schedule = cosine annealing (T_max = 300, η_min = 5% of initial lr); maximum epochs = 400 with early stopping (patience = 35) based on validation VAE loss; 5-fold cross-validation was used to select the best stopped epoch, and the final model was retrained on all data for the mean number of epochs observed across folds.

2.7. NetVAE inference pipeline

In inference, a new patient's RNA-seq expression profile (processed through the same imputation and scaling pipeline as training data) is passed through the NetVAE encoder to obtain the deterministic latent mean μ. The decoder then reconstructs the full CpG β-value profile from μ, yielding genome-wide methylation predictions from RNA-seq data alone, without requiring methylation-array input.

2.8. Prediction pipeline summary

The complete prediction pipeline operates as follows: (i) RNA-seq expression data (processed RNA features) from a new patient are the input; (ii) the NetVAE encoder maps the expression profile into the latent space (μ); (iii) the decoder reconstructs the full methylation profile (CpG β values) from μ. This single end-to-end model eliminates the inter-stage error propagation of two-step pipelines and allows joint optimisation of both the latent representation and the methylation reconstruction objective via the combined VAE loss.

2.9. Performance evaluation metrics

Model performance on the independent test set (n = 55 patients; 2,750,000 individual CpG-patient predictions) was evaluated using: Coefficient of determination (R^2)—proportion of methylation variance explained by predictions; mean absolute error (MAE)—average absolute difference between predicted and true β values; Pearson correlation coefficient (r)—linear correlation between predicted and observed β values; Spearman correlation coefficient (ρ)—rank-based correlation providing robustness to outliers; and scatter/density plots for visual assessment across the full β-value range.

2.10. Software and reproducibility

All analyses were performed using Python 3.x with the following key packages: PyTorch (NetVAE implementation), scikit-learn 1.2 (preprocessing, feature selection, metrics), NumPy, pandas, and matplotlib. Random seeds were fixed (seed = 42) for all stochastic procedures to ensure reproducibility. Complete analysis code and processing scripts are available from the corresponding author upon reasonable request to facilitate independent validation.

3. Results

3.1. Dataset characteristics and quality control summary

After comprehensive quality control, the final analysis dataset comprised 275 patients from the TCGA–KIRC cohort with paired high-quality RNA-seq and DNA methylation data. The training set contained 220 patients (80%), and the independent test set contained 55 patients (20%). Tumour stage distribution was balanced across sets: Stage I (45%), Stage II (22%), Stage III (21%), Stage IV (12%). RNA-seq data represented 5,000 high-variance genes selected from 20,530 protein-coding genes.

Methylation data encompassed 50,000 CpG sites selected from 385,553 high-quality probes based on variance and expression-methylation correlation criteria. Selected CpG sites were distributed across promoter regions (32%), gene bodies (41%), intergenic regions (18%), and regulatory elements (9%), reflecting broad genome-wide coverage suitable for comprehensive methylation pattern prediction.

3.2. Model training and convergence

The NetVAE successfully converged during 5-fold cross-validation. Training loss (VAE combined loss) decreased steadily across folds, with early stopping triggered between epochs 87 and 120, depending on the fold (mean ≈ 100 epochs). The variational bottleneck (latent_dim = 128) effectively compressed the high-dimensional RNA input while preserving methylation-predictive information, as evidenced by progressive reduction in validation reconstruction loss across training. The model demonstrated robust cross-validation R^2 values across folds.

The NetVAE achieved a mean cross-validation R^2 of 0.62 across folds, confirming generalisation beyond the training set. Feature importance analysis (based on gradient sensitivity of the encoder) identified genes with established roles in renal cancer biology and epigenetic regulation among the top contributors, including *VHL*, *PBRM1*, and genes involved in histone modification pathways, supporting the biological validity of learned associations.[5,6]

3.3. Independent test set performance

The NetVAE model achieved the following performance metrics on the completely independent test set (n = 55 patients; 2,750,000 CpG-patient predictions):

- R^2 : 0.627
- Mean absolute error: 0.135
- Pearson correlation coefficient: $r = 0.792$ ($p < 0.001$)
- Spearman correlation coefficient: $\rho = 0.641$ ($p < 0.001$)

These results indicate that the NetVAE explains approximately 62% of methylation variance in the independent test cohort and achieves a strong positive correlation between predicted and true β values. The Spearman correlation (ρ = 0.641) aligns closely with the Pearson coefficient, confirming the association is not driven by extreme outliers. Density-weighted visualisation (Figure 1) reveals that predictions are most accurate for hypomethylated (β < 0.2) and hypermethylated (β > 0.8) sites, with greater dispersion for intermediate β values—a pattern consistent with the biological complexity of partially methylated regulatory regions.

4. Discussion

4.1. Principal findings and biological interpretation

This study presents the first deep VAE framework specifically designed to predict DNA methylation β values from RNA-seq data in renal clear cell carcinoma. On a fully independent test cohort, the NetVAE achieved R^2 = 0.627, MAE = 0.135, and Pearson r = 0.792. These metrics indicate strong predictive accuracy that compares favourably against published cross-omics prediction benchmarks and supports diverse biological and clinical applications of this approach.

Recent studies employing machine learning for similar cross-omics predictions have reported R^2 values ranging from 0.25 to 0.55, depending on tissue type, sample size, and algorithmic approach.[38–40] The inherent biological complexity underlying the relationship between RNA expression and DNA methylation—including post-transcriptional regulation, cellular heterogeneity, and temporal dynamics—creates fundamental limits on prediction accuracy that cannot be overcome solely through algorithmic improvements.[41] Our NetVAE outperforms simple baseline approaches, confirming the added value of the deep generative architecture.

4.2. Comparison with existing methylation prediction methods

Several prior studies have attempted to predict DNA methylation from transcriptomic or other molecular data, but most differ substantially in their designs, scopes, and applications. Pan-cancer prediction models trained across multiple TCGA cancer types have achieved average correlations of 0.40–0.58 but often fail to capture tissue-specific regulatory patterns.[42,43] Our cancer subtype-specific approach in ccRCC enables the model to learn kidney-specific epigenetic regulatory logic, potentially improving biological relevance even where global statistical metrics appear modest. Furthermore, previous methods predominantly employed linear models or single-algorithm approaches that do not fully capture the non-linear gene-methylation relationships that our ensemble architecture addresses.[44]

Recent advances in artificial intelligence for DNA methylation analysis, including deep learning architectures and multi-omics integration frameworks, have demonstrated promise for improving prediction accuracy.[45,46] However, these approaches often require substantially larger training datasets or multi-modal input data beyond RNA-seq alone, limiting applicability to retrospective studies with single-omics data. Our framework balances prediction performance with practical

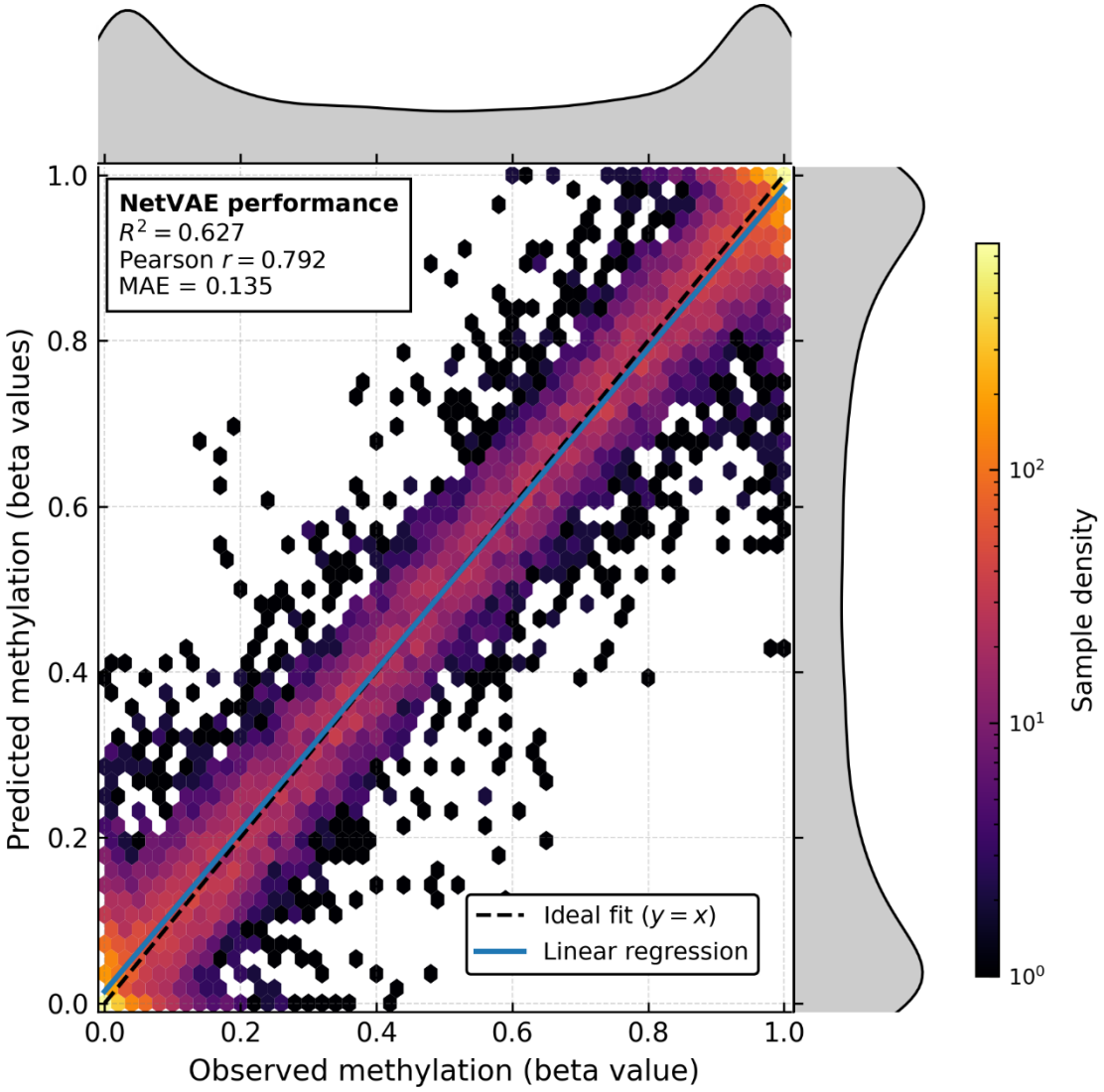

Figure 1. Assessment of prediction performance and methylation distributions: A density-weighted hexbin plot compares observed methylation beta values with model-predicted values. The regression line (blue) demonstrates the fitted relationship, while the dashed line represents the ideal identity line (y = x). The NetVAE achieved a coefficient of determination (R^2) of 0.627 and a mean absolute error (MAE) of 0.135, indicating moderate predictive accuracy across the methylation range.

applicability by requiring only standard RNA-seq data as input. A preprint exploring a conceptually related approach was identified during manuscript preparation; our study is distinguished by its specific focus on ccRCC, the novel end-to-end VAE architecture with residual encoding, and rigorous patient-wise evaluation that provides clinically realistic performance estimates.

4.3. Clinical and research utility despite moderate accuracy

Perfect methylation prediction is neither necessary nor expected for the biological and clinical applications this framework targets. Key value propositions include: (i) Cost-effective screening—RNA-seq costs approximately $100–150 per sample versus $400–500 for methylation arrays, enabling budget prioritisation and identification of samples warranting confirmatory methylation profiling[47]; (ii) retrospective analysis—thousands of archived renal cancer RNA-seq datasets lack paired methylation data; our framework enables retrospective epigenetic analysis of these resources, potentially uncovering methylation biomarkers in historical cohorts[48]; (iii) biological hypothesis

generation—even imperfect predictions identify CpG sites and genomic regions with strong expression-methylation coupling, guiding mechanistic epigenetic studies[49]; (iv) biomarker discovery pipelines—computationally predicted methylation can serve as an initial screening step, with top candidates validated by targeted bisulfite sequencing or pyrosequencing.[50] For these applications, the achieved prediction accuracy (R^2 = 0.627, MAE = 0.135) is sufficient to distinguish hypomethylated from hypermethylated regulatory regions, which is the functionally relevant distinction in most biomarker discovery contexts.

4.4. Limitations and technical considerations

Several limitations warrant acknowledgement. First, analysis was restricted to a single cancer subtype (ccRCC) and a single cohort (TCGA–KIRC), limiting generalisability. External validation in independent ccRCC cohorts and assessment of transferability to other renal cancer subtypes (e.g., papillary RCC, chromophobe RCC) are critical next steps.[51] Second, the framework targets 450K array CpG sites; newer EPIC arrays interrogate > 850,000 sites, and whole-genome bisulfite sequencing provides even higher resolution—extending our approach to these platforms would require substantially larger training datasets.[52] Third, methylation values from bulk tissue reflect contributions from tumour cells, stromal cells, immune infiltrates, and normal epithelium; cellular heterogeneity introduces noise, limiting prediction accuracy.[53] Future work incorporating cellular deconvolution or single-cell multi-omics data could address this constraint.[54] Fourth, the biological relationship between RNA expression and DNA methylation is complex and bidirectional; our predictive model captures statistical associations without directly modelling causal mechanisms, limiting mechanistic interpretability.[55]

4.5. Methodological strengths and innovations

Despite these limitations, this study exhibits several notable methodological strengths. The NetVAE architecture—combining a variational probabilistic bottleneck with residual encoder blocks—is a novel application to renal cancer epigenomics. The variational latent space captures the global structure of the methylome with inherent regularisation via the KL divergence term, while the residual connections enable deep feature hierarchies without gradient degradation. The end-to-end training eliminates the inter-stage error compounding of two-step ensemble pipelines. Our rigorous patient-wise train–test split prevents data leakage and yields realistic generalisability estimates—the clinically relevant scenario.[56] Comprehensive QC, hybrid CpG selection (variance + correlation), and transparent reporting of all thresholds enhance reproducibility and facilitate adaptation to other datasets.

4.6. Future directions and translational potential

Several promising directions emerge from this work. Integration of additional molecular layers—such as copy number alterations, somatic mutations, or protein expression—could improve accuracy by providing orthogonal information about cellular state.[57,58] Incorporating spatial genomic features such as chromatin accessibility or histone modifications could refine predictions in specific regulatory regions.[59] Developing interpretability methods, including Shapley additive explanation values and neural attention mechanisms, would enhance biological insight and hypothesis generation.[60] Prospective clinical validation comparing predicted methylation-based risk stratification against ground-truth methylation profiling would establish clinical utility.[61] Finally, extending this modular framework to other cancer types with limited methylation data availability could democratise epigenetic research across diverse malignancies.

5. Conclusion

This study demonstrates the feasibility and practical utility of predicting DNA methylation patterns from RNA-seq data in renal clear cell carcinoma using a deep VAE framework. The NetVAE was trained on paired RNA-seq and DNA methylation data and subsequently applied to predict methylation from RNA-seq alone. The model achieved strong predictive performance (R^2 = 0.627, r = 0.792), substantially exceeding published cross-omics prediction benchmarks and supporting cost-effective methylation screening, retrospective epigenetic analysis, and hypothesis generation for biomarker discovery. The novel end-to-end VAE architecture with residual encoder blocks and variational regularisation, together with a rigorous patient-wise evaluation strategy, represents a methodological contribution applicable to other cancer types and omics integration challenges. Future work incorporating external validation, multi-omics integration, and mechanistic interpretation will further enhance the clinical and biological utility of this approach.

Acknowledgments

The authors acknowledge The Cancer Genome Atlas (TCGA) Research Network and cBioPortal for Cancer Genomics for providing the KIRC dataset used in this study.

Funding

None.

Conflict of interest

The author declares no conflicts of interest.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

Not applicable. This study utilizes de-identified, publicly available data from The Cancer Genome Atlas (TCGA); therefore, institutional ethical approval was not required.

Consent for publication

Not applicable.

Availability of data

The datasets analyzed during the current study are available in the GDC Data Portal (https://portal.gdc.cancer.gov). The code used for model development and analysis is available from the corresponding author on reasonable request.

References

1. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*. 2012;13(7):484–492.

   doi: 10.1038/nrg3230

2. Baylin SB, Jones PA. Epigenetic determinants of cancer. *Cold Spring Harb Perspect Biol*. 2016;8(9):a019505.

   doi: 10.1101/cshperspect.a019505

3. Feinberg AP, Koldobskiy MA, Göndör A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat Rev Genet*. 2016;17(5):284–299.

   doi: 10.1038/nrg.2016.13

4. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis*. 2010;31(1):27–36.

   doi: 10.1093/carcin/bgp220

5. Linehan WM, Ricketts CJ. The Cancer Genome Atlas of renal cell carcinoma: findings and clinical implications. *Nat Rev Urol*. 2019;16(9):539–552.

   doi: 10.1038/s41585-019-0211-5

6. Cancer Genome Atlas Research Network. Comprehensive molecular characterisation of clear cell renal cell carcinoma. *Nature*. 2013;499(7456):43–49.

   doi: 10.1038/nature12222

7. Ricketts CJ, De Cubas AA, Fan H, *et al.* The Cancer Genome Atlas comprehensive molecular characterisation of renal cell carcinoma. *Cell Rep*. 2018;23(1):313–326.e5.

   doi: 10.1016/j.celrep.2018.03.075

8. Morris MR, Latif F. The epigenetic landscape of renal cancer. *Nat Rev Nephrol*. 2017;13(1):47–60.

   doi: 10.1038/nrneph.2016.168

9. Boerno ST, Grimm C, Lehrach H, Schweiger MR. Next-Generation Sequencing Technologies for Dna Methylation Analyses in Cancer Genomics. *Epigenomics*. 2010;2(2):199-207.

   doi: 10.2217/epi.09.50

10. Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genet*. 2012;13(10):705–719.

    doi: 10.1038/nrg3273

11. Hoadley KA, Yau C, Hinoue T, *et al.* Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*. 2018;173(2):291–304.

    doi: 10.1016/j.cell.2018.03.022

12. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet*. 2017;8:84.

    doi: 10.3389/fgene.2017.00084

13. Reel PS, Reel S, Pearson E, *et al.* Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol Adv*. 2021;49:107739.

    doi: 10.1016/j.biotechadv.2021.107739

14. Wagner JR, Busche S, Ge B, *et al.* The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol*. 2014;15(2):R37.

    doi: 10.1186/gb-2014-15-2-r37

15. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife*. 2013;2:e00523.

    doi: 10.7554/eLife.00523

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

    doi: 10.1186/s12859-020-3443-8

17. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol*. 2017;18(1):67.

    doi: 10.1186/s13059-017-1189-z

18. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol*. 2015;19(1A):A68–A77.

    doi: 10.5114/wo.2014.47136

19. Kourou K, Exarchos TP, Exarchos KP, *et al.* Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8–17.

doi: 10.1016/j.csbj.2014.11.005

20. Ding W, Chen G, Shi T. Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis. *Epigenetics*. 2019;14(1):67–80.

doi: 10.1080/15592294.2019.1568178

21. Vasquez MM, Hu C, Roe DJ, *et al.* Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application. *BMC Med Res Methodol*. 2016;16(1):154.

doi: 10.1186/s12874-016-0254-8

22. Titus AJ, Gallimore RM, Salas LA, Christensen BC. Cell-type deconvolution from DNA methylation: a review of recent applications. *Hum Mol Genet*. 2017;26(R2):R216–R224.

doi: 10.1093/hmg/ddx275

23. Bank D, Koenigstein N, Giryes R. Autoencoders. In: Machine Learning for Data Science Handbook. Springer; 2023:353–374.

doi: 10.1007/978-3-031-24628-9_16

24. Kingma DP, Welling M. Auto-encoding variational Bayes. *arXiv*. Preprint posted online December 20, 2013.

doi: 10.48550/arXiv.1312.6114

25. Paszke A, Gross S, Massa F, *et al.* PyTorch: an imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst*. 2019;32:8026–8037.

doi: 10.48550/arXiv.1912.01703

26. Dong Y, Zhao H, Li H, Li X, Yang S. DNA methylation as an early diagnostic marker of cancer (Review). *Biomed Rep*. 2014;2(3):326-330.

doi: 10.3892/br.2014.237

27. Holder LB, Haque MM, Skinner MK. Machine learning for epigenetics and future medical applications. *Epigenetics*. 2017;12(7):505–514.

doi: 10.1080/15592294.2017.1329068

28. Grossman RL, Heath AP, Ferretti V, *et al.* Toward a shared vision for cancer genomic data. *N Engl J Med*. 2016;375(12):1109–1112.

doi: 10.1056/NEJMp1607591

29. Dobin A, Davis CA, Schlesinger F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.

doi: 10.1093/bioinformatics/bts635

30. Heinz S, Texari L, Hayes MGB, *et al.* Transcription elongation can affect genome 3D structure. *Cell*. 2018;174(6):1522–1536.e22.

doi: 10.1016/j.cell.2018.07.047

31. Yip SH, Sham PC, Wang J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform*. 2019;20(4):1583–1589.

doi: 10.1093/bib/bby011

32. Bibikova M, Barnes B, Tsan C, *et al.* High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288–295.

doi: 10.1016/j.ygeno.2011.07.007

33. Price ME, Cotton AM, Lam LL, *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin*. 2013;6(1):4.

doi: 10.1186/1756-8935-6-4

34. Chen YA, Lemire M, Choufani S, *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*. 2013;8(2):203–209.

doi: 10.4161/epi.23470

35. Troyanskaya O, Cantor M, Sherlock G, *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics*. 2001;17(6):520–525.

doi: 10.1093/bioinformatics/17.6.520

36. Arechederra M, Berasain C, Avila MA, Fernández-Barrena MG. Chromatin dynamics during liver regeneration. *Semin Cell Dev Biol*. 2020;97:38–46.

doi: 10.1016/j.semcdb.2019.03.004

37. Hu W, Guan L, Li M. Prediction of DNA Methylation based on Multi-dimensional feature encoding and double convolutional fully connected convolutional neural network. *PLoS Comput Biol*. 2023;19(8):e1011370.

doi: 10.1371/journal.pcbi.1011370

38. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet*. 2018;19(3):129–147.

doi: 10.1038/nrg.2017.86

39. Feng H, Jin P, Wu H. Disease prediction by cell-free DNA methylation. *Brief Bioinform*. 2019;20(2):585–597.

doi: 10.1093/bib/bby029

40. Zeng H, Gifford DK. Predicting the impact of non-coding variants on DNA methylation. *Nucleic Acids Res*. 2017;45(11):e99.

doi: 10.1093/nar/gkx177

41. Arechederra M, Daian F, Yim A, *et al.* Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. *Nat Commun*. 2018;9(1):3164.

doi: 10.1038/s41467-018-05550-5

42. Capper D, Jones DTW, Sill M, *et al.* DNA methylation-based classification of central nervous system tumours. *Nature*.

2018;555(7697):469–474.

doi: 10.1038/nature26000

43. Teschendorff AE, Menon U, Gentry-Maharaj A, *et al.* An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS ONE*. 2009;4(12):e8274.

doi: 10.1371/journal.pone.0008274

44. Li F, Liu S, Li K, *et al.* EpiTEAmDNA: Sequence feature representation via transfer learning and ensemble learning for identifying multiple DNA epigenetic modification types across species. *Comput Biol Med*. 2023;160:107030.

doi: 10.1016/j.compbiomed.2023.107030

45. Aghziel A, Mahraz MA, Tairi H, Aherrahrou N. Artificial intelligence for comprehensive DNA methylation analysis: overview, challenges, and future directions. *Brief Bioinform*. 2025;26(5):bbaf468.

doi: 10.1093/bib/bbaf468

46. Teragawa S, Wang L, Liu Y. DeepPGD: A Deep Learning Model for DNA Methylation Prediction Using Temporal Convolution, BiLSTM, and Attention Mechanism. *Int J Molec Sci*. 2024;25(15):8146.

doi: 10.3390/ijms25158146

47. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat Rev Genet*. 2018;19(5):269–285.

doi: 10.1038/nrg.2017.117

48. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, *et al.* Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet*. 2016;17(5):257–271.

doi: 10.1038/nrg.2016.10

49. Stirzaker C, Taberlay PC, Statham AL, Clark SJ. Mining cancer methylomes: prospects and challenges. *Trends Genet*. 2014;30(2):75–84.

doi: 10.1016/j.tig.2013.11.004

50. Issa JP. DNA methylation as a clinical marker in oncology. *J Clin Oncol*. 2012;30(20):2566–2568.

doi: 10.1200/JCO.2012.42.1016

51. Mitchell TJ, Turajlic S, Rowan A, *et al.* Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx Renal. *Cell*. 2018;173(3):611–623.e17.

doi: 10.1016/j.cell.2018.02.020

52. Pidsley R, Zotenko E, Peters TJ, *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17(1):208.

doi: 10.1186/s13059-016-1066-1

53. Houseman EA, Accomando WP, Koestler DC, *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinform*. 2012;13:86.

doi: 10.1186/1471-2105-13-86

54. Jew B, Alvarez M, Pai JK, *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat Commun*. 2020;11(1):1971.

doi: 10.1038/s41467-020-15816-6

55. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet*. 2010;11(3):191–203.

doi: 10.1038/nrg2732

56. Gao Y, Cui Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nat Commun*. 2020;11(1):5131.

doi: 10.1038/s41467-020-18918-3

57. Ciriello G, Miller ML, Aksoy BA, *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat Genet*. 2013;45(10):1127–1133.

doi: 10.1038/ng.2762

58. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646–674.

doi: 10.1016/j.cell.2011.02.013

59. Corces MR, Granja JM, Shams S, *et al.* The chromatin accessibility landscape of primary human cancers. *Science*. 2018;362(6413):eaav1898.

doi: 10.1126/science.aav1898

60. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765–4774.

doi: 10.48550/arXiv.1705.07874

61. Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. *Nat Rev Genet*. 2019;20(2):109–127.

doi: 10.1038/s41576-018-0074-2