An Energy Model for Recognizing the Histone Modification Signal Levels in lncRNA Promoter Regions of HCC Based on Molecular Structure

Menglan Li , Yingli Chen * , Qianzhong Li * , Pengyu Du , Dimeng Zhang , Yuanyuan Zhao

Posted Date: 20 April 2026

doi: 10.20944/preprints202604.1389.v1

Keywords: hepatocellular carcinoma; long non-coding RNA; histone modification; statistical physics model; interaction energy

 doi:10.20944/preprints202604.1389.v1

Article

An Energy Model for Recognizing the Histone Modification Signal Levels in lncRNA Promoter Regions of HCC Based on Molecular Structure

Menglan Li [1], Yingli Chen [1,2,*], Qianzhong Li [1,2,*], Pengyu Du [1], Dimeng Zhang [1] and Yuanyuan Zhao [1]

[1] Inner Mongolia Autonomous Region Key Laboratory of Biophysics and Bioinformatics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China

[2] State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Inner Mongolia University, Hohhot 010021, China

* Correspondence: stchenyl@imu.edu.cn (Y.C.); qzli@imu.edu.cn (Q.L.); Tel.: +86-471-499-2914

Abstract

In hepatocellular carcinoma (HCC), aberrant histone modifications are linked to the dysregulation of long non-coding RNA (lncRNA) expression. Although existing computational models can accurately predict some associations, they lack deep physical interpretability. We constructed an energy model based on the physical principle that energy determines molecular structure. Total DNA segment energy was calculated by summing adjacent trinucleotide interaction energies and applied to analyze 11 key histone modifications in HCC, specifically within lncRNA promoter regions where modification signals were increased or decreased. Finally, ten-fold cross-validation revealed that significant energy differences between sequences with increased and decreased histone signals enable excellent classification performance. These results indicted a strong correlation between the total energy of local DNA structures and histone modification signal. Furthermore, introducing longer k-mers led to computational redundancy without a consistent improvement, confirming that the trinucleotide model most effectively acquires the local DNA structural changes associated with histone modification levels. Our model can effectively distinguish DNA sequences associated with different histone modification levels from a physical energy perspective. This model serves as an interpretable tool for epigenetic research while providing a new understanding a new perspective for understanding the dysregulation of lncRNA expression in HCC.

Keywords: hepatocellular carcinoma; long non-coding RNA; histone modification; statistical physics model; interaction energy

1. Introduction

Hepatocellular carcinoma (HCC) is one of the leading causes of cancer death in the world, and its development is a complex process involving dysregulated gene expression.[1,2] More and more evidence has shown that long non-coding RNA (lncRNA) are aberrantly expressed in HCC, participating in key oncogenic processes including proliferation, invasion, metastasis, and drug resistance[3–5]. Recently, research on lncRNAs in cancer has rapidly expanded into new areas such as neutrophil extracellular traps (NETs) and various modes of programmed cell death like ferroptosis. These advancements provide significant clinical value for enhancing prognosis assessment and biomarker development [6,7]. In addition to these approaches, the dysregulated expression of lncRNA in HCC is closely associated with epigenetic modifications in their promoter regions, in which histone modifications play a particularly critical role [8,9]. Therefore, analyzing the changes of abnormal histone modifications in the promoter regions of lncRNA is crucial for finding the underlying mechanisms of HCC pathogenesis [10].

 doi:10.20944/preprints202604.1389.v1

In HCC, the patterns of histone modifications are often significantly different from normal hepatocytes [11,12]. A number of studies have demonstrated these alterations in multiple levels. First, in the protein-coding gene level, it has been found that signals of activating modifications such as H3K4me3 and H3K27ac are aberrantly enhanced in the promoter regions of oncogenes like MYC and CCND1, while repressive modifications like H3K27me3 and H3K9me3 are related to the silencing of tumor suppressor genes [13–15]. Second, in the non-coding RNA level, similar modification patterns regulate the expression levels of oncogenic lncRNA (e.g., MALAT1) and tumor-suppressive lncRNA (e.g., DAW), directly affecting tumor proliferation and metastasis [16,17]. Additionally, recent research has highlighted the coregulatory effects of multiple histone modifications on key ferroptosis-related genes, emphasizing the complexity of the epigenetic [18]. Furthermore, recent studies have revealed that aberrant histone acetylation in HCC can enhance the expression of specific m6A-related lncRNA, which act as signaling molecules to further cooperate with key pathways such as Wnt/ β-catenin, jointly influencing immune response and therapeutic resistance [19,20]. Finally, in clinical applications, these abnormal modification patterns have been verified as key signals for assessing HCC subtypes and predicting patient prognosis and diagnosis [21–23]. In summary, these findings highlight the great potential of histone modifications and lncRNAs as biomarkers and potential therapeutic targets for cancer [24,25].

Although current experimental techniques can effectively reveal the distribution of histone modifications or observe their effects on chromatin structure and gene expression [26–28], these methods still face significant limitations. On the technical level, approaches such as Chromatin Immunoprecipitation Sequencing (Chip-seq) and enzyme inhibitor treatments are often costly and time-consuming, while High-throughput Chromatin Conformation Capture (Hi-C) technology struggles to resolve the fine local structure of DNA over short distances. More importantly, on the mechanistic level, most of these experimental methods can only provide descriptive correlations and fail to delve into the physical level of molecular interaction [29].

To overcome the limitations of purely experimental approaches, a series of predictive algorithms have been developed in the field of bioinformatics [30–37]. For example, machine learning and deep learning models, such as those based on CNN or Transformer architectures, integrate multi-omics data including genomic sequences, chromatin accessibility, methylation, and histone modifications. These models have demonstrated good potential in identifying genome-wide histone modification regions, inferring the effects of histone modification signals on lncRNA expression, and evaluating the prognosis of HCC patients based on multiple epigenetic modification data.

However, a key issue exists with these mainstream computational methods. They tend to focus on using complex parameters to analyze the correlation between sequences and epigenetic modifications and their effects on gene expression, most of these works were based on analyzing the association between sequence features and epigenetic modifications, with no regard for the influence and interpretation of DNA's physical properties, such as its structure. This lack of a physical perspective results in a critical limitation: while these models can accurately predict the impact, they fail to fundamentally explain the underlying physical reasons. This limitation highlights the necessity of constructing interpretable analytical models based on physical principles [38,39].

Therefore, from a physical perspective, identifying and understanding the histone modification signal levels in lncRNA promoter regions is crucial for studying the molecular origins of cancer development [40,41]. Based on the physical principle that molecular structure is determined by its energy, a statistical physics energy model using the interaction energy between adjacent dinucleotides has been proposed. This model has been successfully applied to the recognition of prokaryotic promoters, confirming an intrinsic correlation between the interaction energy of DNA segments and their biological functions [42].

To better understanding the cooperative effects of adjacent bases on local DNA structure, this study aimed to develop an improved trinucleotide energy model by integrating bioinformatics and statistical physics. In this model, the total energy of a promoter region is computed as the sum of interaction energies between adjacent trinucleotides. Applying this model, we used the data of 11 key

 doi:10.20944/preprints202604.1389.v1

histone modifications in HCC to calculate the total energy of local structures in the promoter regions of differentially expressed lncRNA where modification signals were significantly increased or decreased. Excellent results were obtained, providing a new theoretical tool for understanding the changes in epigenetic signals from a physical perspective.

2. Results

2.1. Analysis of the lncRNA Differential Expression Profile

To identify lncRNA with significant expression changes in HCC, we analyzed total RNA-seq data from the normal hepatocyte and HCC cell lines. From the annotation file, we first extracted a total of 13,880 lncRNA. Subsequently, by applying the thresholds of an adjusted P-value < 0.05 and $|\log_2 FC| > 1$, we identified 930 significantly differentially expressed lncRNAs (DElncRNAs). Among these, 407 lncRNAs were significantly up-regulated and 523 were significantly down-regulated in the HCC cell line (Figure 1).

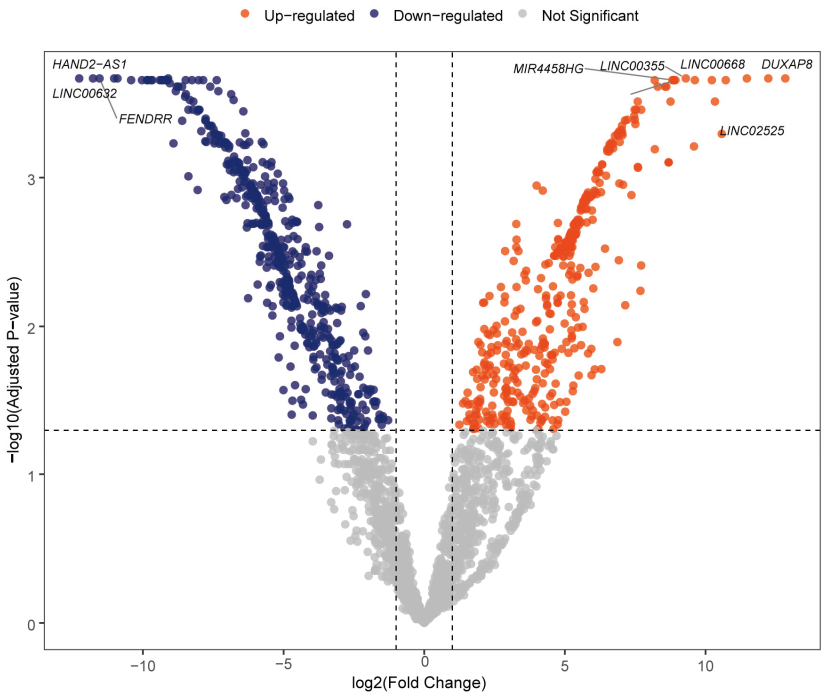

Figure 1. Volcano plot of differentially expressed lncRNAs (DElncRNAs). The x-axis displays the ($|\log_2$ (Fold Change), and the y-axis represents the statistical significance as -log10(Adjusted P-value). The horizontal dashed line indicates the significance threshold (Adjusted P-value = 0.05), while the vertical dashed lines represent the fold change threshold ($|\log_2 FC| = 1$). Compared to the normal group, points colored in orange represent significantly up-regulated lncRNAs in the tumor group, points in dark blue represent significantly down-regulated lncRNAs, and points in gray represent non-significantly expressed lncRNAs.

2.2. Construction and Screening of Model Input Datasets

Following the identification of DElncRNAs promoter regions, we analyzed the signals of 11 histone modifications within these regions and identified regions where these signals were significantly altered.

Finally, using the eight histone modifications with sufficient sample sizes, we constructed the positive sets (sequences from regions with significantly enhanced signals in DElncRNAs promoters)

 doi:10.20944/preprints202604.1389.v1

and negative sets (sequences from regions with significantly weakened signals). The sample counts for each dataset are summarized in Table 1, and these sequences were used for the subsequent training and validation of the energy model.

Table 1. Statistics of the positive and negative sample sizes used for energy model training and evaluation.

Histone Modification	*Positive Samples*	*Negative Samples*	*Total Samples*
H2AFZ	*284*	*2122*	*2406*
H3K27ac	*1318*	*1795*	*3113*
H3K4me1	*1293*	*745*	*2038*
H3K4me2	*951*	*1601*	*2552*
H3K4me3	*2307*	*652*	*2959*
H3K79me2	*1149*	*304*	*1453*
H3K9ac	*998*	*1189*	*2187*
H4K20me1	*156*	*177*	*333*

2.3. Energy Values and Energy Differences in Positive and Negative Sample Sets

To visually analyze how the energy model distinguishes between lncRNA promoters with increased (positive set) and decreased (negative set) modification levels in HCC, we plotted the distribution of energy values and the energy difference (ΔE) for the eight histone modifications. The analysis was performed on the test sets in a 10-fold cross-validation.

Energy analysis revealed a consistent pattern across all eight histone modifications: sequences from regions with increased signals exhibited significantly lower energy values than those from regions with decreased signals. Consequently, the ΔE values for the positive set were consistently negative, which means the model can effectively distinguish between the two states.

We illustrate these findings using four key modifications as examples: H3K27ac, H3K4me1, H3K4me3, and H3K9ac. To validate the model's effectiveness, we first analyzed the energy distribution features for H3K27ac (Figure 2).

The energy difference (ΔE) distribution plots (Figure 2a,c) showed that points representing increased samples (orange) were predominantly clustered in the negative region of the ΔE axis,whereas decreased samples (blue) were mainly distributed in the positive region. This trend was further confirmed in the energy plots (Figure 2b,d), where the energy curve for the positive model ($E_{positive}$, orange) was globally lower than that of the negative model ($E_{negative}$, blue) for sequences with lower ΔE values, clearly demonstrating the model's discriminative ability.

Similarly, other key modifications such as H3K4me1 (Figure 3), H3K4me3 (Figure 4), and H3K9ac (Figure 5) exhibited highly similar patterns in their energy and energy difference distributions (see Supplementary Figures S1–S9 for other modifications).

To further investigate the structural basis of the energy model, we compared the performance of models based on dinucleotide and trinucleotide features. A comparison of the energy difference plots (Figure 2a vs. Figure 2c) revealed that the trinucleotide model had a stronger discriminative power. Specifically, its ΔE distribution range (approx. -600 to 600) was much distinct than that of the dinucleotide model (approx. -300 to 300), suggesting that trinucleotide interaction energies can capture finer-scale differences in local DNA structure. Furthermore, the energy profile plots (Figure 2b vs. Figure 2d) also showed that the degree of separation between the two energy ($E_{positive}$ and

 doi:10.20944/preprints202604.1389.v1

$E_{negative}$) was significantly greater in the trinucleotide model, further confirming its superior performance.

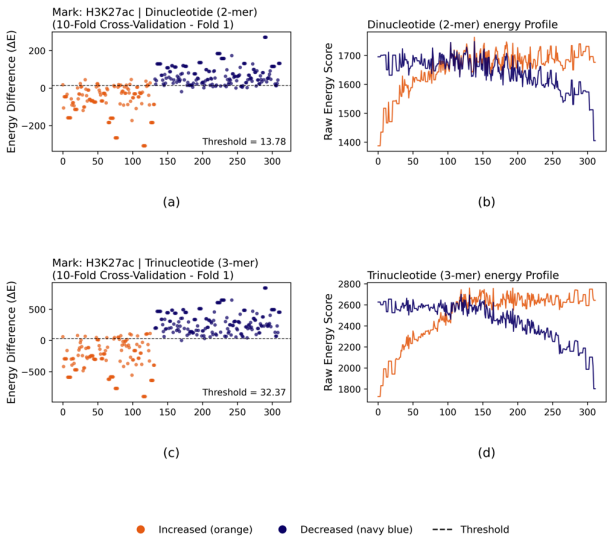

Figure 2. Comparison of dinucleotide and trinucleotide models for discriminating H3K27ac signals. (a) and (b) represent the results of the dinucleotide model for a representative 10-fold cross-validation fold. (c) and (d) represent the results of the trinucleotide model for a representative 10-fold cross-validation fold. (a) and (c) show the distribution of the energy difference (ΔE) for each sequence in the test sets, where the horizontal dashed line indicates the classification threshold. (b) and (d) display the score curves of the positive ($E_{positive}$) and negative ($E_{negative}$) energies for the test sequences sorted by their ΔE values, providing the energy profile.

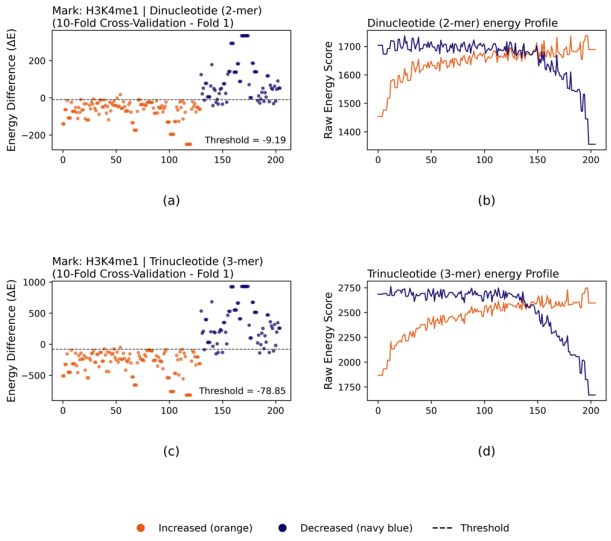

Figure 3. Comparison of dinucleotide and trinucleotide models for discriminating H3K4me1 signals. (a) and (b) represent the results of the dinucleotide model for a representative 10-fold cross-validation fold. (c) and (d) represent the results of the trinucleotide model for a representative 10-fold cross-validation fold. (a) and (c) show the distribution of the energy difference (ΔE) for each sequence in the test sets, where the horizontal dashed line indicates the classification threshold. (b) and (d) display the score curves of the positive ($E_{positive}$) and negative ($E_{negative}$) energies for the test sequences sorted by their ΔE values, providing the energy profile.

 doi:10.20944/preprints202604.1389.v1

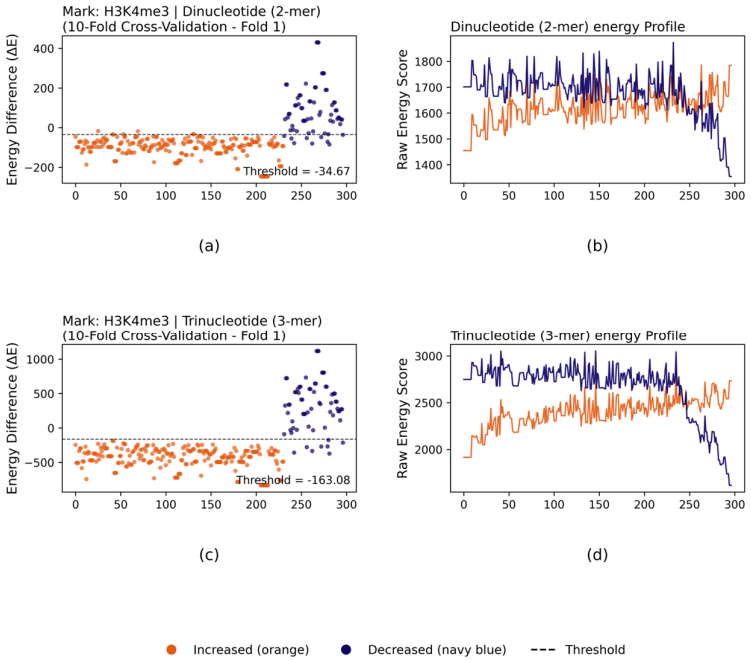

Figure 4. Comparison of dinucleotide and trinucleotide models for discriminating H3K4me3 signals. (a) and (c) represent the results of the dinucleotide model for a representative 10-fold cross-validation fold. (a) shows the distribution of the energy difference (ΔE) for each sequence in the test sets, where the horizontal dashed line indicates the classification threshold. (c) displays the score curves of the positive ($E_{positive}$) and negative ($E_{negative}$) energies for the test sequences sorted by their ΔE values. (b) and (d) show the corresponding results for the trinucleotide model, with (b) representing the ΔE distribution and (d) showing the raw energy separation profile.

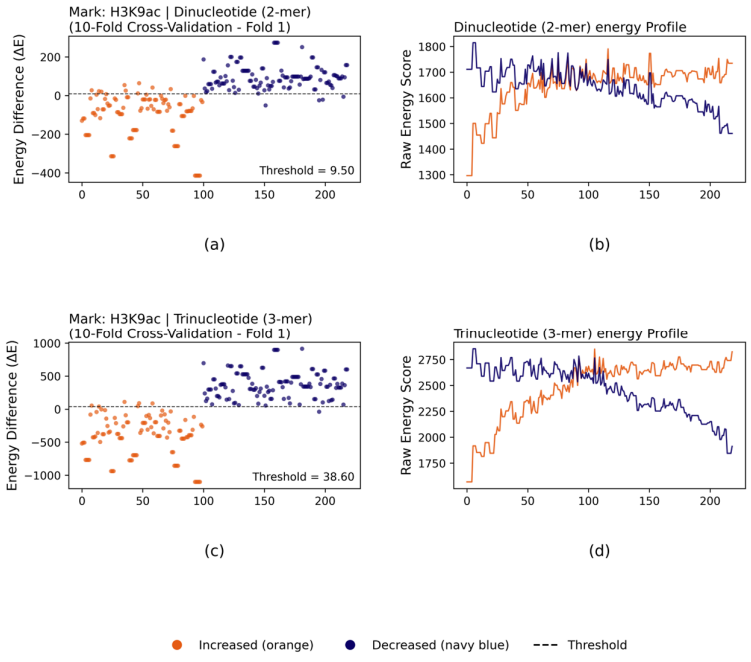

Figure 5. Comparison of dinucleotide and trinucleotide models for discriminating H3K9ac signals. (a) and (b) represent the results of the dinucleotide model for a representative 10-fold cross-validation fold. (c) and (d) represent the results of the trinucleotide model for a representative 10-fold cross-validation fold. (a) and (c) show

 doi:10.20944/preprints202604.1389.v1

the distribution of the energy difference (ΔE) for each sequence in the test sets, where the horizontal dashed line indicates the classification threshold. (b) and (d) display the score curves of the positive ($E_{positive}$) and negative ($E_{negative}$) energies for the test sequences sorted by their ΔE values, providing the energy profile.

2.4. Performance Comparison of Trinucleotide and Dinucleotide Models

We next systematically evaluated the model's performance using a 10-fold cross-validation procedure. This analysis confirmed that the model could effectively distinguish between lncRNA promoters with increased and decreased histone modification levels in HCC. For all eight modifications tested, the energy difference distributions between the positive and negative sets were significantly different.

The dinucleotide-based model performed well across all eight histone modifications, achieving MCC > 0.92 (Table 2). However, the trinucleotide-based energy model consistently outperformed the dinucleotide model for every modification (Table 2, Figure 6). With the trinucleotide model, the average MCC increased to over 0.96. These results suggest that considering longer-range adjacent base dependencies captures additional structural information relevant to the changes in histone modification levels.

Table 2. Predictive performance of the dinucleotide and trinucleotide energy models in distinguishing increased versus decreased histone modification states in hepatocellular carcinoma (10-fold cross-validation).

Mark	*Model*	*Sn(%)*	*Sp(%)*	*Ac(%)*	*MCC*	*auROC*
H2AFZ	*2-mer*	*90.14*	*99.53*	*98.42*	*0.92*	*1.00*
H2AFZ	*3-mer*	*92.60*	*100.00*	*99.13*	*0.96*	*1.00*
H3K27ac	*2-mer*	*88.09*	*92.42*	*90.59*	*0.81*	*0.97*
H3K27ac	*3-mer*	*92.71*	*97.38*	*95.41*	*0.91*	*0.99*
H3K4me1	*2-mer*	*95.44*	*83.63*	*91.12*	*0.81*	*0.97*
H3K4me1	*3-mer*	*97.29*	*88.45*	*94.06*	*0.87*	*0.99*
H3K4me2	*2-mer*	*87.06*	*90.63*	*89.30*	*0.78*	*0.97*
H3K4me2	*3-mer*	*88.43*	*97.44*	*94.08*	*0.87*	*0.99*
H3K4me3	*2-mer*	*98.83*	*89.27*	*96.72*	*0.90*	*0.99*
H3K4me3	*3-mer*	*99.96*	*92.49*	*98.31*	*0.95*	*0.99*
H3K79me2	*2-mer*	*98.96*	*84.51*	*95.94*	*0.87*	*0.98*
H3K79me2	*3-mer*	*99.39*	*86.82*	*96.76*	*0.90*	*0.99*
H3K9ac	*2-mer*	*93.29*	*91.00*	*92.04*	*0.84*	*0.98*
H3K9ac	*3-mer*	*95.49*	*96.21*	*95.88*	*0.92*	*1.00*
H4K20me1	*2-mer*	*91.58*	*92.58*	*92.10*	*0.84*	*0.99*
H4K20me1	*3-mer*	*84.58*	*100.00*	*92.77*	*0.86*	*0.99*

 doi:10.20944/preprints202604.1389.v1

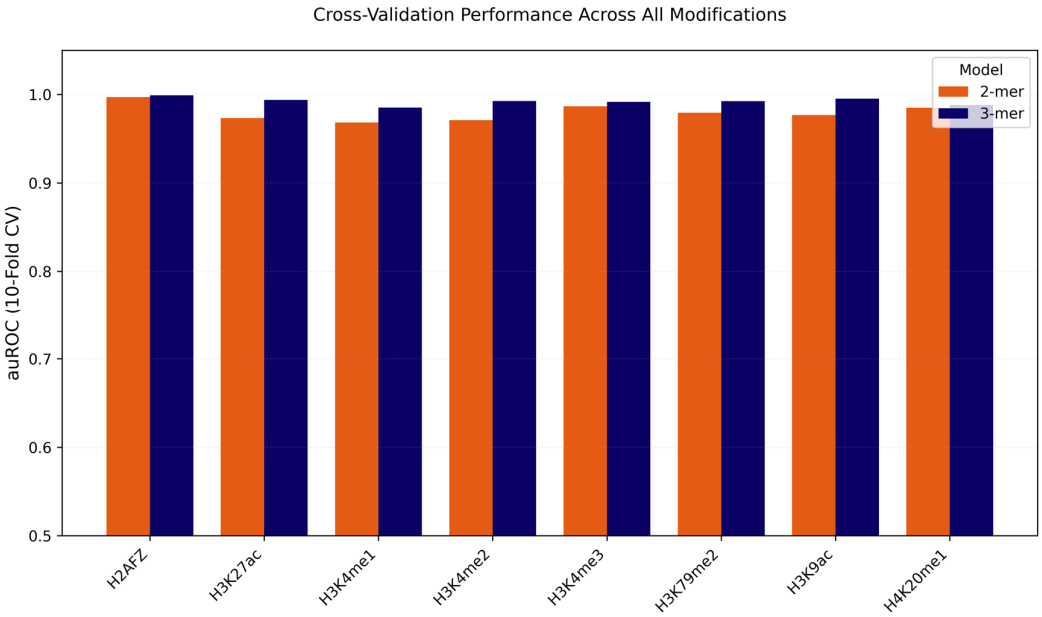

Figure 6. Comparison of auROC results between dinucleotide and trinucleotide energy models based on 10-fold cross-validation. Area Under the Curve (AUC) values for the two models across 8 histone modifications. The blue bars indicate the AUC values for the trinucleotide model, and the orange bars indicate the AUC values for the dinucleotide model.

3. Discussion

The results of this study demonstrate that the local structural energy of DNA sequences can effectively distinguish between lncRNA promoter regions with increased and decreased histone modification signals in HCC.

In this study, we constructed and validated a statistical physics model based on trinucleotide interaction energies to investigate the physical basis of histone modification signal changes associated with lncRNA expression in HCC. The results confirm that the local interaction energy of DNA is an effective physical parameter for discriminating between lncRNA promoter regions with increased and decreased modification signals.

In contrast to many deep learning approaches, which require extensive parameters, our method calculates total energy based on foundational principles of statistical physics and information theory. This integration of bioinformatics and physics provides a novel and interpretable perspective on the intrinsic relationship between epigenetics and the physical properties of DNA. Our findings also show that the trinucleotide model is better at distinguishing the two groups than the dinucleotide model. It is likely that the DNA structural changes are influenced by the synergistic effects of adjacent bases. A dinucleotide interaction model may not adequately capture these local structural features, which are dependent on a longer sequence context. By incorporating information from more extensive neighboring interactions, the trinucleotide model more sensitively detects local DNA structural alterations associated with histone modifications, thereby more accurately reflecting the energy differences between distinct modification levels.

In summary, the results of this study show that a statistical physics model based on trinucleotide interaction energies can effectively identify DNA sequences associated with different histone modification levels. This model not only provides a new theoretical tool for understanding epigenetic regulation in HCC from a physical perspective but also demonstrates that the local structural energy of DNA could serve as a potential biomarker, offering new directions for future cancer diagnosis and treatment.

 doi:10.20944/preprints202604.1389.v1

4. Materials and Methods

4.1. Data Acquisition and Pre-Processing

The raw count data from RNA-sequencing (RNA-seq) and the Chip-seq data for 11 key histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K79me2, H4K20me1, H3K9ac, H3K27ac, and H2AFZ) for the hepatocellular carcinoma cell line (HepG2) and normal hepatocytes were downloaded from the ENCODE project (ENCODE; https://www.encodeproject.org/). The genome annotation and lncRNA coordinates (GRCh38 version, Release 49) were sourced from the GENCODE database (GENCODE; https://www.gencodegenes.org/human/). All genomic coordinates were standardized to the hg38 reference genome.

4.2. Differential Expression Analysis of lncRNAs

To identify differentially expressed lncRNAs in the HCC cell line, we performed differential expression analysis on the RNA-seq data using the limma package [43]. LncRNAs were considered as significantly differentially expressed if they met the criteria of an adjusted P-value (FDR) < 0.05 and $|\log_2 FC| > 1$.

4.3. Identification of Differential Histone Modification Regions

To accurately identify genomic regions with differential histone modification signals between the HCC cell line (HepG2) and the normal hepatocyte cell line (Hepa), we employed a signal intensity-based differential analysis method. For each histone modification, we first integrated all peaks identified by MACS2 in both cell lines to construct a unified set of candidate regions [44]. Subsequently, for each region in this set, we calculated the mean signal intensity of histone modification in these regions for all biological replicates from the bigWig files of the HepG2 and Hepa cell lines. The $\log_2 FC$ of signal intensity for each region i was calculated using the following formula:

$$\log_2 FC = \log_2\left(\frac{Signal_{HepG2,i} + \beta}{Signal_{Hepa,i} + \beta}\right) \quad (1)$$

where Signal (HepG2, i) and Signal (Hepa, i) represent the average signal intensities of region i in the HepG2 and Hepa cell lines, respectively. β is a pseudocount, set to 1×10^{-6}, to avoid division by zero. Finally, regions with significantly differential histone modifications were selected based on a threshold of $|\log_2 FC| > 1.0$.

4.4. Identification of Differential Histone Modification Regions Within lncRNA Promoters

To focus on key functional regions directly associated with transcriptional regulation, the core promoter regions for each differentially expressed lncRNA (DElncRNAs) was defined as the interval spanning from 1500 bp upstream to 500 bp downstream of its transcription start site (TSS). Subsequently, to identify the differential histone modification peaks located within these promoters, only those peaks that spatially overlapped with the defined promoter regions were retained for further analysis.

4.5. Construction of DNA Sequence Fragments

To prepare a standardized input dataset for the energy model, we extracted DNA sequence fragments of a uniform length (600 bp) from the differential histone modification regions within the lncRNA promoters identified previously. For narrow peaks, the summit position (point of maximum signal) was selected as the center. For broad peaks, we calculated and used the geometric center of

 doi:10.20944/preprints202604.1389.v1

the peak region. The final 600 bp DNA fragments were obtained by extending 300 bp upstream and 300 bp downstream from each defined anchor point (i.e., summit or geometric center).

A 600 bp window was selected because it appropriately covers a complete nucleosome (~147 bp) and its adjacent linker DNA (20-90 bp). This length represents a typical spatial context for the interplay of histone modifications and is better to capture the features of local DNA structure.

4.6. Physical Principles of the Energy Model

Based on the fundamental biological principle that "structure determines function," we supposed that changes in histone modification levels directly influence local DNA structure, thereby affecting gene transcription. Compared to normal hepatocytes, the histone modification signals in promoter regions were significantly increased or decreased in HCC, due to their different biological functions, they should have different structures. Therefore, it was proposed that the interaction energies should be different for distinct histone modification levels. As the total energy is the sum of interaction energies of these units across all sites, the total local structural energy of the DNA must be different between these two groups of regions.

4.7. Construction of Model

To prepare for the energy model analysis, we first constructed an independent input dataset for each histone modification. The sequences were partitioned into two classes based on their signal intensity changes in HCC cells: (1) Positive set, consisting of sequences regions with significantly enhanced histone modification signals; and (2) Negative set, consisting of sequences from regions with significantly weakened signals.

Subsequently, all datasets were filtered based on sample size to ensure statistical significance for robust model training. We found that the number of differential Peaks for H3K27me3, H3K9me3, and H3K36me3 within lncRNA promoter regions was too low to support robust model training. Therefore, these three modifications were excluded. The final energy model analysis was conducted on the remaining eight modifications: H2AFZ, H3K27ac, H3K9ac, H3K4me1, H3K4me2, H3K4me3, H3K79me2, and H4K20me1.

4.8. Position-Correlation Probability Matrix (PCPM) and Pseudocount

To estimate the interaction energies defined in our model, we first calculated the positional probabilities for all trinucleotides from the sequence data. This was accomplished by constructing a Position-Correlation Probability Matrix (PCPM).

In this framework, each lncRNA promoter fragment of length L (L= 600 in this study) was treated as a chain of L-2 (i.e., 598) overlapping trinucleotides. For a given histone modification state ζ ($\zeta \in (\{pos, neg\})$, the probability ($p^i_{j\zeta}$) of a specific trinucleotide j (where $j \in$ {AAA, AAC, ..., TTT}) occurring at site i was calculated using a pseudocount method to improve the estimate, as follows [42]:

$$p^i_{j\zeta} = \frac{n^i_{j\zeta} + b}{N_\zeta + B} \tag{2}$$

where $p^i_{j\zeta}$ is the estimated probability of trinucleotide j at position i for the sequence set in state ζ; $n^i_{j\zeta}$ is the observed count of trinucleotide j at position i for the sequence set ζ; N_ζ is the total number of sequences in the sample set ζ; and b and B are the pseudocount parameters calculated as follows:

$$B = \sqrt{N_\zeta}, \tag{3}$$

$$b = p_0 \sqrt{N_\zeta} \tag{4}$$

 doi:10.20944/preprints202604.1389.v1

where p_0 represents the average background frequency of a trinucleotide.

4.9. Construction of the Energy Model

The physical properties of each local site were described by its local probability $p_{j\zeta}^{i}$, and a local partition function, $z_{i\zeta}$ the local partition function for site i was defined by according to the principles of statistical physics, as:

$$z_{i\zeta} = \sum_{allj}^{64} e^{-\beta\varepsilon_{j\zeta}^{i}} \tag{5}$$

where $\varepsilon_{j\zeta}^{i}$ is the interaction energy of the physical unit (trinucleotide j) at site j, and $\beta = 1/kT$, with k being the Boltzmann constant and T the absolute temperature.

Based on the assumption that the probability of a specific unit occurring at this site follows the Boltzmann distribution, we have:

$$p_{j\zeta}^{i} = \frac{1}{z_{i\zeta}} e^{-\beta\varepsilon_{j\zeta}^{i}} \tag{6}$$

By rearranging the terms in the Boltzmann distribution equation, the interaction energy of the physical unit at site i can be derived as:

$$\varepsilon_{j\zeta}^{i} = -\frac{1}{\beta}(\ln p_{j\zeta}^{i} + \ln z_{i\zeta}) \tag{7}$$

where $\varepsilon_{j\zeta}^{i}$ is the interaction energy for trinucleotide j at position i under state ζ. This equation allows for the direct calculation of interaction energies from the previously computed probabilities ($p_{j\zeta}^{i}$) and partition functions ($z_{j\zeta}^{i}$).

To account for the continuous and holistic nature of the physical properties along the DNA fragment, we defined the total probability P_ζ, and the total partition function Z_ζ, for the entire fragment. These global parameters were calculated as the product of their respective local parameters across all sites.

$$P_\zeta = \prod_{i=1}^{L-2} p_{j\zeta}^{i} \tag{8}$$

$$Z_\zeta = \prod_{i=1}^{L-2} z_{j\zeta}^{i} \tag{9}$$

Here, the total probability P_ζ is the joint probability of observing the specific sequence of trinucleotides ($j_1, j_2, \ldots, j_{L-2}$) that constitute the fragment. The total partition function Z_ζ aggregates the statistical weights from all $L-2$ sites.

According to the principles of statistical mechanics, the total energy of the local DNA structure E_ζ, is related to the total probability P_ζ, and the total partition function Z_ζ by the following equation:

$$E_\zeta = -kT(\ln P_\zeta + \ln Z_\zeta) \tag{10}$$

By substituting Equations (8) and (9) into Equation (10), an expression for the total energy as a sum of local contributions was obtained:

$$E_\zeta = -kT\sum_{i=1}^{L-2}(\ln p_{j\zeta}^{i} + \ln z_{i\zeta}) \tag{11}$$

Equation (11) demonstrates that the total energy of the fragment is equivalent to the sum of the interaction energies of all overlapping physical units:

$$E_\zeta = \sum \varepsilon_{j\zeta}^{i} \tag{12}$$

 doi:10.20944/preprints202604.1389.v1

4.10. Discriminant Function and Classification Strategy

To distinguish he different histone modification signal levels, the energy difference ΔE between the positive and negative sets was defined as the discriminant score:

$$\Delta E = E_{pos} - E_{neg} \tag{13}$$

Substituting Equation (11) into Equation (13), the final energy discriminant function was derived:

$$\Delta E = -kT\left[\sum_{i=1}^{L-2}\left(\ln p_{pos}^{i} - \ln p_{neg}^{i}\right) + \sum_{i=1}^{L-2}\left(\ln z_{i,pos} - \ln z_{i,neg}\right)\right] \tag{14}$$

This expression can be rearranged as:

$$\Delta E = -kT\sum_{i=1}^{L-2}(\ln\frac{p_{pos}^{i}}{p_{neg}^{i}} + C) \tag{15}$$

$$C = -kT\sum_{i=1}^{L-2}(\ln z_{i,pos} - \ln z_{i,neg}) \tag{16}$$

The term C is a constant determined by the local physical environments under different modification levels and serves as the decision threshold for the energy model. In practice, if a query region satisfies $\Delta E > C$, the DNA fragment is classified as a region with significantly increased histone modification signals; otherwise, it is classified as a region with significantly decreased signals.

4.11. Performance Evaluation of the Model

To comprehensively evaluate the effectiveness and reliability of our energy model in discriminating between differential histone modification regions, we employed four key metrics: Sensitivity (S_n), Specificity (S_p), Accuracy (Acc), and the Matthews Correlation Coefficient (Mcc). The definitions and formulas for these metrics are as follows:

Sensitivity (S_n): Represents the model's ability to correctly identify regions with significantly increased histone modification signals (the positive set).

$$S_n = \frac{TP}{TP + FN} \tag{17}$$

Specificity (S_p): Represents the model's ability to correctly identify regions with significantly decreased histone modification signals (the negative set).

$$S_p = \frac{TN}{TN + FP} \tag{18}$$

Accuracy (Acc): Reflects the overall proportion of samples correctly classified in both the positive and negative sets.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

Matthews Correlation Coefficient (Mcc): A comprehensive metric for classification quality that is robust to class imbalance. The Mcc value ranges from -1 to +1, where a value closer to +1 indicates a stronger ability to distinguish between the physical features.

$$Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{20}$$

In the above formulas, TP (True Positives) is the number of correctly identified positive samples; TN (True Negatives) is the number of correctly identified negative samples; FP (False Positives) is the number of negative samples incorrectly classified as positive; and FN (False Negatives) is the number of positive samples incorrectly classified as negative.

 doi:10.20944/preprints202604.1389.v1

4.12. Receiver Operating Characteristic (ROC) Curve and AUC Value

To further evaluate the model's classification performance, we performed an ROC curve analysis. The ROC curve was plotted with the True Positive Rate (S_n) on the y-axis against the False Positive Rate (1-S_p) on the x-axis. The Area Under the Curve (AUC) was calculated as a key parameter to quantify the performance of the energy model. The AUC value ranges from 0.5 to 1.0, where a larger value indicates a higher accuracy of the model in distinguishing the total energy differences in local DNA structure caused by different modification signal levels.

5. Conclusions

The results of this study indicate that the local structural energy features of DNA sequences can effectively distinguish between regions with significantly enhanced and weakened histone modification signals in the promoters of lncRNAs in HCC. Furthermore, the energy model based on trinucleotide interactions showed a much better ability of distinguish compared to the dinucleotide model. This suggests that longer-range nucleotide interactions contain richer structural information, making it to better detect the DNA physical features linked to histone modification signals. In summary, this research provides a novel computational perspective and a model tool for understanding the states of histone modifications linked to lncRNA expression from a physical energy standpoint.

Author Contributions: Conceptualization, Y.C., Q.L. ; methodology, M.L., Y.C., Q.L., P.D., D.Z. and Y.Z.; formal analysis, M.L.; investigation, M.L., D.Z. and Y.Z.; data curation, M.L.; writing—original draft preparation, M.L.; writing—review and editing, M.L., Y.C. and Q.L.; visualization, M.L.; project administration, M.L.; supervision, Y.C. and Q.L.; funding acquisition, Y.C., Q.L.. All authors have read and agreed to the published version of the manuscript.

Funding information: This work was supported by the National Natural Science Foundation of China [Nos. 62361047 and 32160216].

Institutional Review Board Statement: Not applicable. Informed Consent Statement: Not applicable.

Data Availability Statement: All the data for this study were obtained from publicly available databases such as ENCODE (https://www.encodeproject.org/, accessed on 6 August 2025).

Acknowledgments: All authors sincerely appreciate the sharing of data from ENCODE and GENCODE databases.

Conflicts of Interest: The authors declare no conflicts of interest.

Supplementary Materials: The following supporting information can be downloaded at: Preprints.org.

Abbreviations

lncRNA	long noncoding RNA
HCC	hepatocellular carcinoma
AML	acute myeloid leukemia
LUAD	lung adenocarcinoma
NETs	neutrophil extracellular traps
DElncRNAs	differentially expressed lncRNAs.

 doi:10.20944/preprints202604.1389.v1

RNA-seq	RNA sequencing
ChIP-seq	Chromatin Immunoprecipitation sequencing
TSS	transcription start site
ΔE	energy difference
PCPM	Position-Correlation Probability Matrix
ROC	receiver operating characteristic
AUC	area under the curve

References

1. Wolinska, E.; Skrzypczak, M. Epigenetic Changes Affecting the Development of Hepatocellular Carcinoma. *Cancers* **2021**, *13*, 4237. https://doi.org/10.3390/cancers13164237
2. Bueloni, B.; Garcia Fernandez de Barrena, M.; Avila, M.A.; Bayo, J.; Mazzolini, G. Epigenetic mechanisms involved in hepatocellular carcinoma development and progression. *eGastroenterology* **2025**, *3*, e100186. https://doi.org/10.1136/egastro-2024-100186
3. Mehra, M.; Chauhan, R. Long Noncoding RNAs as a Key Player in Hepatocellular Carcinoma. *Biomark. Cancer* **2017**, *9*, 1179299X17737301. https://doi.org/10.1177/1179299X17737301
4. Verma, S.; Sahu, B.D.; Mugale, M.N. Role of lncRNAs in hepatocellular carcinoma. *Life Sci.* **2023**, *325*, 121751. https://doi.org/10.1016/j.lfs.2023.121751
5. Liang, W.; Zhao, Y.; Meng, Q.; Jiang, W.; Deng, S.; Xue, J. The role of long non-coding RNA in hepatocellular carcinoma. *Aging* **2024**, *16*, 4052–4073. https://doi.org/10.18632/aging.205523
6. Liu, S.H.; Chen, Y.L.; Li, Q.Z.; Fan, Z.Y.; Li, M.L.; Du, P.Y. A prognostic model for acute myeloid leukemia based on ferroptosis-related lncRNA and immune infiltration analysis. *Biophys. Rep.***2024**, *10*, 240029. https://doi.org/10.52601/bpr.2024.240029
7. Zhao, Y.Y.; Chen, Y.L.; Li, Q.Z. Neutrophils extracellular traps-related lncRNAs as potential biomarkers in lung adenocarcinoma. *Biophys. Rep.* **2025**, *11*, 250058. https://doi.org/10.52601/bpr.2025.250058
8. Shi, Z.; Jin, S.; Liu, X.; Jiang, M.; Fang, Y.; Khadaroo, P.A.; Lin, H.; Fan, X. The epigenetic regulatory network of long noncoding RNAs in hepatocellular carcinoma. *Genes Dis.* **2025**, *12*, 101534. https://doi.org/10.1016/j.gendis.2025.101534
9. Du, P.; Chen, Y.; Li, Q.; Gai, Z.; Bai, H.; Zhang, L.; Liu, Y.; Cao, Y.; Zhai, Y.; Jin, W. CancerMHL: the database of integrating key DNA methylation, histone modifications and lncRNAs in cancer. *Database* **2024**, *2024*, baae029. https://doi.org/10.1093/database/baae029
10. Hong, Y.; Zhang, Y.; Zhao, H.; Chen, H.; Yu, Q.Q.; Cui, H. The roles of lncRNA functions and regulatory mechanisms in the diagnosis and treatment of hepatocellular carcinoma. *Front. Cell Dev. Biol.* **2022**, *10*, 1051306. https://doi.org/10.3389/fcell.2022.1051306
11. Liu, M.; Jiang, L.; Guan, X.Y. The genetic and epigenetic alterations in human hepatocellular carcinoma: a recent update. *Protein Cell* **2014**, *5*, 673–691. https://doi.org/10.1007/s13238-014-0065-9
12. Liu, Y.X.; Song, J.L.; Li, X.M.; Lin, H.; Cao, Y.N. Identification of target genes co-regulated by four key histone modifications of five key regions in hepatocellular carcinoma. *Methods* **2024**, *231*, 165–177. https://doi.org/10.1016/j.ymeth.2024.09.017
13. Liu, Y.X.; Li, Q.Z.; Cao, Y.N.; Zhang, L.Q. Identification of key genes and important histone modifications in hepatocellular carcinoma. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 2657–2669. https://doi.org/10.1016/j.csbj.2020.09.013
14. Rajan, P.K.; Udoh, U.A.; Sanabria, J.D.; Manne, U.; Khuder, S.A.; Shafi, M.A.; Moore, R.; Gupta, S.; Smith, J.E.; Jones, R.L.; et al. The role of histone acetylation-methylation-mediated apoptotic gene regulation in hepatocellular carcinoma. *Int. J. Mol. Sci.* **2020**, *21*, 8894. https://doi.org/10.3390/ijms21238894

 doi:10.20944/preprints202604.1389.v1

15. Hung, S.Y.; Lin, H.H.; Yeh, K.T.; Chang, J.G. Histone-modifying genes as biomarkers in hepatocellular carcinoma. *Int. J. Clin. Exp. Pathol.* **2014**, *7*, 2496–2507.
16. Wang, D.; Chen, F.; Zeng, T.; Tang, Q.; Chen, B.; Chen, L.; Dong, Y.; Li, X. Comprehensive biological function analysis of lncRNAs in hepatocellular carcinoma. *Genes Dis.* **2021**, *8*, 58–70. https://doi.org/10.1016/j.gendis.2019.12.013
17. Li, H.; Mu, H.; Xiao, Y.; Zhao, Z.; Cui, X.; Wu, D. Comprehensive analysis of histone modifications in hepatocellular carcinoma reveals different subtypes and key prognostic models. *J. Oncol.* **2022**, *2022*, 5961603. https://doi.org/10.1155/2022/5961603
18. Qi, Y.C.; Bai, H.; Hu, S.L.; Li, S.J.; Li, Q.Z. Coregulatory effects of multiple histone modifications in key ferroptosis-related genes for lung adenocarcinoma. *Epigenomics* **2024**,*16*, 348–359. https://doi.org/10.2217/epi-2023-0403
19. Liang, W.; Shi, C.; Hong, W.; Li, P.; Zhou, X.; Fu, W.; Lin, L.; Zhang, J. Super-enhancer-driven lncRNA-DAW promotes liver cancer cell proliferation through activation of Wnt/β-catenin pathway. *Mol. Ther. Nucleic Acids* **2021**, *26*, 1351–1363. https://doi.org/10.1016/j.omtn.2021.10.028
20. Su, Y.Y.; Ji, F.Y.; Zheng, Z.J.; Peng, J.Y.; Xie, J.J. The role of super-enhancer-driven lncRNAs in cancer. *Comput. Struct. Biotechnol. J.* **2025**, *27*, 3897–3907. https://doi.org/10.1016/j.csbj.2025.09.011
21. Cheng, C.C.Y.; Cheung, M.F.; Lee, A.Y.; Wu, Q.; Chow, S.H.C.; Ang, J.Y.J.; Medina, I.R.; Lo, G.; Wu, H.; Yang, W.; et al. Multi-omic analysis of hepatocellular carcinoma reveals aberrant cis-regulatory changes and dysregulated retrotransposons with prognostic potential. *Commun. Biol.* **2025**, *8*, 245. https://doi.org/10.1038/s42003-025-09154-w
22. Wei, Y.; Zeng, W.; Wan, G.; Zhan, J.; Song, J.; Xian, S.; Sun, Z.; Cao, J.; Gi, W.; Yang, J.; et al. Enhancer-associated H3K27ac marks define distinct molecular subtypes and therapeutic vulnerabilities in hepatocellular carcinoma. *Nat. Commun.* **2020**, *11*, 1–15. https://doi.org/10.1038/s41467-020-15607-x
23. Li, J.; Huang, S.; Yuan, F.; Yang, X.; Xie, Q. Prognostic and immunotherapeutic response prediction in hepatocellular carcinoma: role of non-histone acetylation/deacetylation scoring. *Discov. Oncol.* **2025**, *16*, 333. https://doi.org/10.1007/s12672-025-03339-9
24. Yao, W.; Hu, X.; Wang, X. Crossing epigenetic frontiers: the intersection of novel histone modifications and diseases. Signal Transduct. *Target. Ther*. **2024**, 9, 232. https://doi.org/10.1038/s41392-024-01918-w
25. Abdelmonem, B.H.; Kamal, L.T.; Wardy, L.W.; Ragheb, M.; Hanna, M.M.; Elsharkawy, M.; Abdelnaser, A. Non-coding RNAs: emerging biomarkers and therapeutic targets in cancer and inflammatory diseases. *Front. Oncol*. **2025**, 15, 1534862. https://doi.org/10.3389/fonc.2025.1534862
26. Donovan, B.T.; Luo, Y.; Meng, Z.; Poirier, M.G. The nucleosome unwrapping free energy landscape defines distinct regions of transcription factor accessibility and kinetics. *Nucleic Acids Res*. **2023**, 51, 1139 – 1155. https://doi.org/10.1093/nar/gkac1267
27. Jung, J.; Werner, M.S. The histone code at a crossroads: history, context, and new approaches. *Trends Genet.***2026**,*42*, 126–136. https://doi.org/10.1016/j.tig.2025.09.003
28. Abbasova, L.; Urbanaviciute, P.; Hu, D.; Ismail, J.N.; Schilder, B.M.; Nott, A.; Skene, N.G.; Marzi, S.J. CUT&Tag recovers up to half of ENCODE ChIP-seq histone acetylation peaks. *Nat. Commun*. **2025**, 16, 2993. https://doi.org/10.1038/s41467-025-58137-2
29. Liu, R.; Xu, R.; Yan, S.H.; Li, P.Y.; Jia, C.T.; Sun, H.Q.; Sheng, K.W.; Wang, Y.J.; Zhang, Q.; Guo, J.; Xin, X.Z.; Li, X.L.; Guo, D.H. Hi-C, a chromatin 3D structure technique advancing the functional genomics of immune cells. *Front. Genet.* **2024**, *15*, 1377238. https://doi.org/10.3389/fgene.2024.1377238
30. Tahir, M.; Norouzi, M.; Khan, S.S.; Davie, J.R.; Yamanaka, S.; Ashraf, A. Artificial intelligence and deep learning algorithms for epigenetic sequence analysis: A review for epigeneticists and AI experts. *Comput. Biol. Med.* **2024**, *183*, 109302. https://doi.org/10.1016/j.compbiomed.2024.109302
31. Yin, Q.; Wu, M.; Liu, Q.; Lv, H.; Jiang, R. DeepHistone: a deep learning approach to predicting histone modifications. *BMC Genomics* **2019**, *20* (Suppl 2), 193. https://doi.org/10.1186/s12864-019-5489-4
32. Chen, Y.; Xie, M.; Wen, J. Predicting gene expression from histone modifications with self-attention based neural networks and transfer learning. *Front. Genet.* **2022**, *13*, 1081842. https://doi.org/10.3389/fgene.2022.1081842

 doi:10.20944/preprints202604.1389.v1

33. Suita, Y.; Bright, H., Jr.; Pu, Y.; Toruner, M.D.; Idehen, J.; Tapinos, N.; Singh, R.; Goudarzi, S.; Wang, X.; Zhang, L.; et al. Machine learning on multiple epigenetic features reveals H3K27Ac as a driver of gene expression prediction across patients with glioblastoma. *PLoS Comput. Biol.* **2025**, *21*, e1012272. https://doi.org/10.1371/journal.pcbi.1012272
34. Zheng, S.; Thakkar, N.; Harris, H.L.; Liu, S.; Zhang, M.; Gerstein, M.; Aiden, E.L.; Rowley, M.J.; Noble, W.S.; Gürsoy, G.; et al. Predicting A/B compartments from histone modifications using deep learning. *iScience* **2024**, *27*, 109570. https://doi.org/10.1016/j.isci.2024.109570
35. Yin, Q.; Wu, M.; Liu, Q.; Lv, H.; Jiang, R. Accurate and highly interpretable prediction of gene expression from histone modifications. *BMC Bioinformatics* **2022**, *23*, 151. https://doi.org/10.1186/s12859-022-04651-5
36. Singh, R.; Lanchantin, J.; Sekhon, A.; Qi, Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* **2016**, *32*, i639–i648. https://doi.org/10.1093/bioinformatics/btw427
37. Eraslan, G.; Avsec, Ž.; Gagneur, J.; Theis, F.J. Deep learning: new computational modelling techniques for genomics.*Nat. Rev. Genet.* **2019**, *20*, 337–357. https://doi.org/10.1038/s41576-019-0122-6
38. Sievers, A.; Sauer, L.; Bisch, M.; Sprengel, J.; Hausmann, M.; Hildenbrand, G. Moderation of structural DNA properties by coupled dinucleotide contents in eukaryotes. *Genes* **2023**, *14*, 755. https://doi.org/10.3390/genes14030755
39. Sievers, A.; Hausmann, M.; Hildenbrand, G. Repeats influence structural DNA properties around functional annotations associated with 3D organization and transcription. *Genes* **2025**, *16*, 1082. https://doi.org/10.3390/genes16091082
40. Murthy, S.; Dey, U.; Olymon, K.; Abbas, E.; Yella, V.R.; Kumar, A. Discerning the role of DNA sequence, shape, and flexibility in recognition by drosophila transcription factors. *ACS Chem. Biol.* **2024**, *19*, 1533–1543. https://doi.org/10.1021/acschembio.4c00202
41. Boev, N.B.; Gerstein, M.B.; Kumar, S. DNA shape and epigenomics distinguish the mechanistic origin of human genomic structural variations. *Nucleic Acids Res.* **2025**, 53, gkaf1325. https://doi.org/10.1093/nar/gkaf1325
42. Chen, Y.L.; Guo, D.H.; Li, Q.Z. An energy model for recognizing the prokaryotic promoters based on molecular structure. *Genomics* **2020**, *112*, 2072–2079. https://doi.org/10.1016/j.ygeno.2019.12.001
43. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. https://doi.org/10.1093/nar/gkv007
44. Zhang, Y.; Liu, T.; Meyer, C.A.; Eeckhoute, J.; Johnson, D.S.; Bernstein, B.E.; Nusbaum, C.; Myers, R.M.; Brown, M.; Li, W.; et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **2008**, *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137