

Prediction of Lysine Acetylation Sites in Escherichia Coli Using DenseNet and Attention Mechanisms

Leyun Xing, Busheng Li

School of Information Engineering, Jingdezhen Ceramic University, Jingdezhen, Jiangxi, CO 333403, China

Abstract: Lysine acetylation is an important post-translational modification that plays a critical role in regulating protein conformational stability, catalytic activity, and cellular metabolism. Owing to the cost and turnaround time of mass spectrometry-based assays, developing efficient and generalizable computational predictors is of practical significance. Using publicly available Escherichia coli protein sequences, we design a lightweight deep learning model composed of a DenseNet convolutional backbone and a CBAM attention module to identify acetylation sites. Each input peptide is encoded by a 23-channel representation that concatenates 20-amino-acid one-hot vectors with three physicochemical indicators (polarity, hydrophobicity, and charge). During training, predicted probabilities are calibrated on the validation set via temperature scaling, and a decision threshold is selected under a specificity constraint of $Sp \approx 0.900$; the calibrated temperature and threshold are then applied to the test set for evaluation. On the E. coli test set, the proposed method achieves $Acc = 0.645$, $Sn = 0.363$, $Sp = 0.901$, $Pre = 0.802$, $F1 = 0.528$, and $AUC = 0.762$. These results indicate that, under a high-specificity operating regime, our model attains overall discrimination comparable to a strong baseline without relying on transfer learning, while delivering higher precision and stability at the selected operating point.

Keywords: DenseNet CNN; CBAM Attention; Temperature Scaling; Deep Learning.

1. Introduction

Lysine acetylation (Kac) is a prevalent post-translational modification (PTM) that participates in cellular metabolism, signal transduction, and stress responses. By altering the charge state of lysine residues and affecting protein conformation, acetylation modulates protein stability and molecular interactions, thereby exerting broad regulatory effects. Although mass-spectrometry (MS)-based assays can identify acetylation sites with high confidence, they are costly and time-consuming, limiting routine large-scale profiling. Consequently, sequence-based computational predictors are of practical importance.

Early studies often relied on traditional machine-learning models such as support vector machines (SVMs) with hand-crafted features. With the advent of deep learning, end-to-end architectures have shown superior performance by learning informative representations directly from sequences. MDDeep-Ace (Liu et al., 2025) employs multi-domain adaptation to enable cross-species transfer, but its training pipeline is relatively complex and computationally demanding [1]. In contrast, species-specific and lightweight models are attractive for efficient deployment and for interpreting organism-specific sequence patterns.

In this work, we focus on Escherichia coli and propose a compact deep model that combines a DenseNet backbone with the Convolutional Block Attention Module (CBAM)[2]. DenseNet captures local sequence regularities through feature reuse across layers, while CBAM emphasizes discriminative signals via channel- and spatial-attention. Inputs are encoded by concatenating 20-amino-acid one-hot vectors with three physicochemical indicators (polarity, hydrophobicity, and charge), yielding 23 channels. During training, we perform temperature scaling on the validation set to calibrate predicted probabilities and select a decision threshold under a specificity constraint of $Sp \approx 0.900$, which is then carried forward to the test set for evaluation. Empirically, under this

high-specificity operating regime, the proposed method achieves strong precision and robustness on the E. coli dataset, offering an efficient and reproducible solution for species-specific lysine acetylation site prediction.

2. Data and Methods

2.1. Data Source and Sample Construction

The Escherichia coli (E. coli) protein acetylation dataset used in this study was obtained from the publicly available multi-species lysine acetylation database released by the MDDeep-Ace project. The original data were curated through rigorous experimental validation and sequence alignment, and include positive samples (acetylated sites identified by mass spectrometry) as well as sequences of non-acetylated sites[3]. To ensure reliable model training, we first performed sequence standardization: fragments containing non-standard amino-acid symbols or unknown characters were removed, and all sequences were unified to a length of 31 amino-acid residues by extracting a fixed window centered on the target lysine or padding when necessary.

For data partitioning, to guarantee independence during training and fairness in evaluation, we adopted stratified sampling with a fixed random seed and divided the entire dataset into training, validation, and test sets, which were used for model fitting, hyperparameter tuning, and final performance assessment, respectively. The proportion of positive and negative samples was kept consistent across splits to avoid the impact of class imbalance on model performance. The resulting dataset provides a comprehensive representation of E. coli protein acetylation and supplies high-quality inputs for subsequent model training.

2.2. Feature Encoding

To comprehensively characterize the physicochemical and compositional properties of protein sequences, each amino-acid residue was encoded with a multidimensional feature

vector. Specifically, every residue was represented by a 23-dimensional vector: the first 20 dimensions correspond to a one-hot encoding of the standard amino acids, capturing sequence composition; the remaining 3 dimensions encode physicochemical attributes—polarity, hydrophobicity, and charge—to reflect residue chemistry and environmental adaptability. This tripartite property scheme follows commonly used classifications of protein physicochemical features and has been shown in multiple studies to effectively distinguish modified from unmodified residues.

For sequence construction, a fixed-length window of 31 amino acids centered on the target lysine was extracted to incorporate upstream and downstream context. After feature encoding, each window yields a tensor of shape $(23 \times 31 \times 1)$ as input to the convolutional neural network. For sequences shorter than the required window, zero-padding was applied to maintain a uniform dimensionality. This encoding strategy balances biological interpretability with learnability, providing a rich and stable input representation for subsequent convolutional and attention modules.

2.3. Model Architecture

The input is a $(23 \times 31 \times 1)$ window tensor. It is first processed by a 3×3 convolutional stem, and then fed into a DenseNet backbone. The backbone comprises three Dense Blocks, each containing four 3×3 convolutional layers with a growth rate of 12, interleaved with two Transition layers (a 1×1 convolution followed by $(2,1)$ average pooling) that downsample only along the window (channel-height) dimension; the sequence-length axis is preserved to maintain upstream-downstream context. Next, a CBAM (Convolutional Block Attention Module) is applied to reweight feature saliency along both channel and spatial dimensions. The resulting feature map is converted “to-sequence” by applying mean and max pooling across the window dimension and concatenating the two projections to form a length-wise sequence. Finally, the representation is flattened and passed through a two-layer fully connected head to produce binary class probabilities via a softmax output. The overall architecture is illustrated in Figure 1.

2.4. Model Training and Probability Calibration

The model was trained using the Adam optimizer with categorical cross-entropy as the loss function. To mitigate potential class imbalance, class weights were set to $\{0: 1.0, 1: 1.33\}$. During training, the validation AUC served as the primary performance monitor, combined with early stopping to prevent overfitting; model weights were saved at the epoch achieving the best validation performance[4].

After training, we applied temperature scaling for probability calibration on the validation set. The temperature parameter TTT was determined by minimizing the negative log-likelihood (NLL) of the calibrated probabilities on validation data, thereby improving agreement between predicted probabilities and the empirical outcome distribution. Based on the calibrated ROC curve, we then selected a decision threshold under the constraint of specificity $Sp \geq 0.900$, choosing the operating point closest to the target specificity. The resulting TTT and threshold were fixed and transferred to the test set for out-of-sample evaluation, providing a balanced trade-off between reliability (calibration) and sensitivity at a high-specificity operating regime.

3. Experiments and Results

3.1. Data Statistics

This study uses an *E. coli* dataset comprising 17,362 samples with strict class balance (8,681 positives and 8,681 negatives; 50.0%/50.0%). The data are split at approximately 8:1:1 into training, validation, and test sets: 13,889 for training (6,931 positives; 6,958 negatives; positive rate 49.9%), 1,736 for validation (874 positives; 862 negatives; 50.3%), and 1,737 for testing (876 positives; 861 negatives; 50.4%). The class proportions across splits closely match the overall distribution, indicating near-stratified partitioning and supporting unbiased performance estimation on the validation and test sets. Table 1 summarizes the dataset statistics for *E. coli*, including class balance and split sizes.

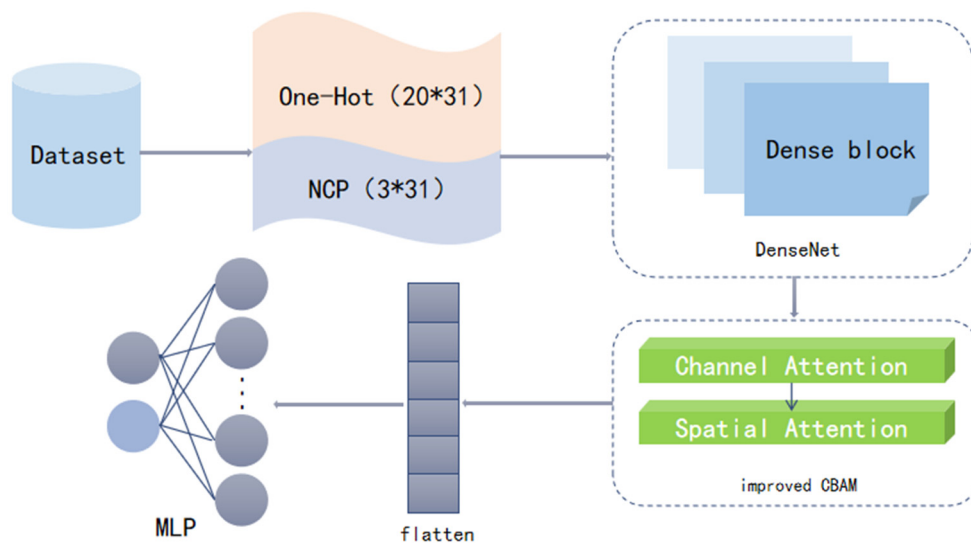


Figure 1. Schematic of the overall model architecture.

Table 1. Dataset statistics for E. coli.

Dataset	Samples	Positives	Negative
Train	13889	6931	6958
Validation	1736	874	862
Test	1737	876	861
Total	17362	8681	8681

3.2. Results and Analysis

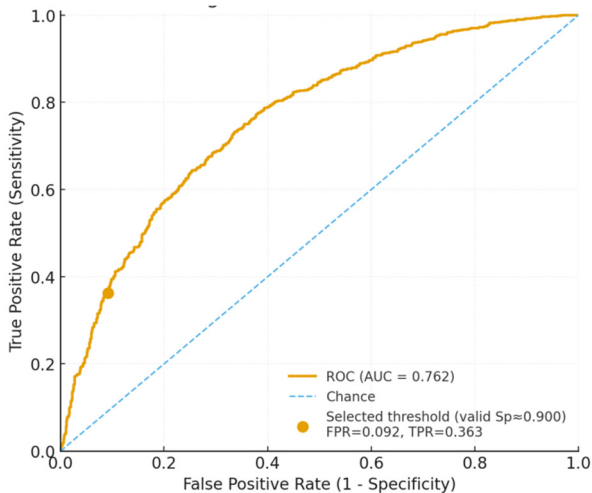
On the E. coli test set—using the validation-calibrated threshold with $Sp \approx 0.90$ transferred for evaluation—the proposed DenseNet+CBAM achieves virtually the same overall discrimination as MDDeep-Ace (AUC: 0.762 vs 0.763). At matched sensitivity ($Sn = 0.363$) and comparable specificity (0.901 vs 0.900), DenseNet+CBAM attains higher accuracy (0.645 vs 0.629; +1.6 pp), precision (0.802 vs 0.787; +1.5 pp), and F1 (0.528 vs 0.497; +3.1 pp). This indicates that,

under a fixed false-positive-rate constraint, our method produces fewer false positives for the same recall, thereby improving practical discriminative quality. Notably, the near-identical AUCs imply similar ranking performance across thresholds; the observed differences arise primarily from the precision – recall trade-off at the selected operating point. As reported in Table 2, at $Sp \approx 0.900$ our method attains $Acc = 0.645$, $Pre = 0.802$, and $F1 = 0.528$. The corresponding ROC curve is shown in Figure 2.

Table 2. Test-set performance of the proposed model on E. coli

model	Acc	Sn	Sp	Pre	F1	AUC
DenseNet+CBAM	0.645	0.363	0.901	0.802	0.528	0.762
MDDeep-Ace	0.629	0.363	0.900	0.787	0.497	0.763

The ROC curve on the E. coli test set yields $AUC = 0.762$, indicating mid-to-strong discrimination across operating thresholds. The decision threshold was chosen on the validation set after temperature scaling (optimal $T \approx 1.4T$) under a specificity constraint ($Sp \approx 0.900$) and then transferred to the test set. At this operating point, the test set attains $FPR = 0.092$ ($Sp \approx 0.908$) and $TPR = 0.363$ ($Sn \approx 0.363$), consistent with Table II ($Pre = 0.802$, $Acc = 0.645$, $F1 = 0.528$). The curve stays well above the chance line in the low-FPR region (≤ 0.2), indicating useful sensitivity under strict false-positive control, and exhibits a stable concave shape, suggesting robust ranking performance over a broad range of thresholds.

**Figure 2.** Test-set ROC curve.

4. Discussion

This study employs DenseNet as the convolutional backbone, leveraging cross-layer feature reuse to enhance the learning of local sequence patterns. Incorporating CBAM further amplifies salient responses at both the channel and spatial levels with negligible parameter overhead, thereby highlighting discriminative cues within a fixed 31-aa window. Validation-set temperature scaling (optimal $T \approx 1.4T$) improves confidence calibration, enabling a more stable and interpretable threshold transfer under a high-specificity constraint ($Sp \approx 0.900$)[5].

On the E. coli test set, the proposed method achieves $AUC = 0.762$, which is nearly identical to MDDeep-Ace (0.763), indicating comparable ranking ability across thresholds. However, at the operating point with matched sensitivity ($Sn = 0.363$) and comparable specificity ($Sp \approx 0.90$), our DenseNet+CBAM model attains higher precision (0.802 vs 0.787) and F1 (0.528 vs 0.497), along with a slightly higher accuracy (0.645 vs 0.629). These results suggest that, in applications requiring low false-positive rates, the proposed approach yields fewer false positives at the same recall, thereby offering improved practical utility.

5. Conclusion

This study addresses lysine acetylation site prediction with a lightweight model that combines a DenseNet backbone and CBAM attention, using 20-AA one-hot features augmented by three physicochemical indicators (polarity, hydrophobicity, charge). We calibrate predicted probabilities on the validation set via temperature scaling and select a decision threshold under a specificity constraint of $Sp \approx 0.900$, which is then

transferred to the test set for evaluation. On the E. coli dataset, the proposed method attains AUC = 0.762, yielding overall discrimination comparable to the strong baseline MDDeep-Ace. At matched sensitivity ($S_n = 0.363$) and comparable specificity, our model achieves Acc/Pre/F1 = 0.645/0.802/0.528, outperforming the baseline (0.629/ 0.787/ 0.497). These results indicate that, in low-false-positive scenarios, the proposed approach delivers higher discriminative quality and more stable confidence at the same recall.

The advantages of this work lie in its simplicity, reproducibility, and deployment friendliness. Moreover, the validation-calibrated, specificity-constrained evaluation protocol enhances comparability and transferability across data splits. Current limitations include a relatively simple feature representation, the absence of cross-species pretrained embeddings, and the fixed window that may underrepresent long-range dependencies. Future work will explore incorporating protein pretrained embeddings (e.g., ESM), cost-sensitive training, and multi-scale receptive fields, as well as small-scale ensembles and more robust threshold selection strategies, aiming to improve recall without sacrificing high specificity.

Acknowledgments

This work was supported in part by the Jingdezhen Ceramic University Degree and Graduate Education Teaching Reform Research Project titled “Focusing on ‘Digital

Intelligence’: Innovative and Practical Graduate Teaching Model Integration of Industry and Education — A Case Study of the Big Data Science and Application Program.”

References

- [1] Liu Y, Ye C, Lin C, Mao K, Zhu M. MDDeep-Ace: species-specific acetylation site prediction based on multi-domain adaptation[J]. *PeerJ*, 2025, 13: e19649.
- [2] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional Block Attention Module[C]//Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. *Computer Vision – ECCV 2018*. Cham: Springer, 2018: 3–19.
- [3] Yu, K., Zhang, Q., Liu, Z., Du, Y., Gao, X., Zhao, Q., Cheng, H., Li, X., & Liu, Z.-X. (2020). Deep-PLA: deep learning based prediction of reversible HAT/HDAC-specific lysine acetylation. *Briefings in Bioinformatics*, 21(5), 1798–1805.
- [4] Lingkuan Meng, Xingjian Chen, Ke Cheng, Nanjun Chen, Zetian Zheng, Fuzhou Wang, Hongyan Sun, Ka-Chun Wong, TransPTM: a transformer-based model for non-histone acetylation site prediction, *Briefings in Bioinformatics*, Volume 25, Issue 3, May 2024, bbae219.
- [5] Wu, M., Yang, Y., Wang, H. et al. A deep learning method to more accurately recall known lysine acetylation sites. *BMC Bioinformatics* 20, 49 (2019).
- [6] Liu, Y.; Wang, Q.; Xi, J. DeepDA-Ace: A Novel Domain Adaptation Method for Species-Specific Acetylation Site Prediction. *Mathematics* 2022, 10, 2364.