Electronic Research Archive

http://www.aimspress.com/journal/ERA

ERA, 31(9): 5442–5456.
DOI: 10.3934/era.2023276
Received: 27 May 2023
Revised: 18 July 2023
Accepted: 23 July 2023
Published: 03 August 2023

Research article

MTNA: A deep learning based predictor for identifying multiple types of N-terminal protein acetylated sites

Yongbing Chen[1], Wenyuan Qin[1], Tong Liu[1], Ruikun Li[1], Fei He[1,*], Ye Han[2,*], Zhiqiang Ma[3,*] and Zilin Ren[4,*]

[1] School of Information Science and Technology, Northeast Normal University, Changchun 130117, China
[2] School of Information Technology, Jilin Agricultural University, Changchun 130118, China
[3] Department of Computer Science, College of Humanities & Sciences of Northeast Normal University, Changchun 130119, China
[4] Changchun Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Changchun 130122, China

* **Correspondence:** Email: hef740@nenu.edu.cn, yeh@jlau.edu.cn, mazq@nenu.edu.cn, zilin.ren@outlook.com.

Abstract: N-terminal acetylation is a specific protein modification that occurs only at the N-terminus but plays a significant role in protein stability, folding, subcellular localization and protein-protein interactions. Computational methods enable finding N-terminal acetylated sites from large-scale proteins efficiently. However, limited by the number of the labeled proteins, existing tools only focus on certain subtypes of N-terminal acetylated sites on frequently detected amino acids. For example, NetAcet focuses on alanine, glycine, serine and threonine only, and N-Ace predicts on alanine, glycine, methionine, serine and threonine. With the growth of experimental N-terminal acetylated site data, it is observed that N-terminal protein acetylation occurs on nearly ten types of amino acids. To facilitate comprehensive analysis, we have developed MTNA (Multiple Types of N-terminal Acetylation), a deep learning network capable of accurately predicting N-terminal protein acetylation sites for various amino acids at the N-terminus. MTNA not only outperforms existing tools but also has the capability to identify rare types of N-terminal protein acetylated sites occurring on less studied amino acids.

Keywords: protein translational modification; protein acetylation; N-terminal acetylated sites; deep learning

1. Introduction

Protein acetylation is one of the most common types of protein post-translational modification (PTM) in eukaryotes [1], and it is involved in the regulation of many cellular processes [2]. Acetylation can be categorized into two main types based on different substrates: amino-terminal acetylation (N^{α}-acetylation) [3] and lysine acetylation (N^{ε}-lysine acetylation) [4].

N^{α}-acetylation is an irreversible co-translational modification that targets the acetyl groups to the α-amino group of amino-terminal residues. Such modification occurs in more than 80% of human proteins [5] and plays a significant role in several biological processes, such as protein stability [6], folding [7], subcellular localization [8] and protein-protein interactions (PPIs) [9]. In contrast, N^{ε}-lysine acetylation is a reversible post-translational modification of protein whereby the acetyl group is transferred to the E-amino group of lysine residues to neutralize the positive charge of the lysine, which directly affects the electrostatic state of the modified protein [10]. It regularizes gene expression by modifying core histone tails by histone acetyl-transferases (HATS) or histone deacetylases (HDACs). A growing number of studies reveal that many diseases, including developmental disabilities [11] and cancers [12], are associated with these two types of N-terminal acetylation. Therefore, the identification of N-terminal protein acetylated sites is vital for understanding the function of proteins in cells.

Up until now, various biological approaches have been employed to accurately identify protein N-terminal acetylated sites. These include the radioanalytical method [13], chromatin immunoprecipitation (ChIP) [14] and mass spectrometry [15]. However, these wet laboratory experiments suffer from limitations such as labor-intensive procedures, high failure rates and costliness. With the accumulation of known protein N-terminal acetylated sites, computational prediction methods were developed to overcome these problems.

Most of the existing computational tools put more effort into identifying N^{ε}-lysine protein acetylated sites but less on N^{α}-acetylated sites. However, these lysine-specific tools cannot be generalized to predict N^{α}-acetylated sites [16]. Only a few tools are specifically designed to target all types of acetylated sites, such as NetAcet [17] and NT-AcPredictor [18].

NetAcet proposed a neural network to predict whether the first three positions of a protein are subject to N-terminal acetylated modifications. NT-AcPredictor investigated the first five residues to predict their possible N-terminal acetylated sites. However, those tools suffered from the shallow perspective fields (considering only the top 3–5 residues) and the single input modality (individually taking sequence or physiochemical properties in featurization) in modeling. These limitations will lead to incomplete and biased representations of proteins for identifying their N-terminal acetylated sites.

This study presents a predictor named MTNA for predicting Multiple Types of N-terminal Acetylation. We designed a deep ensemble architecture to learn deep representations from two modalities, including sequences and physiochemical properties. Two types of deep graphics are integrated by a self-attention strategy, which drives our model to pay more attention to the informative components in merged representation generation. Our model outputs the acetylated sites at the first 30 positions from N-terminal with such merged representations. In addition, we analyzed the effectiveness and informative distribution of learned representations via saliency map [19] and UMAP (uniform manifold approximation and projection) [20] to interpret our model, which may provide new insights into the learned patterns related to N-terminal protein acetylated sites.

2. Materials and methods

2.1. Data collection

We retrieved the N-terminal acetylated proteins from the Swiss-prot [21] database using the keyword "N?-acety". Since such search may return some acetylation enzymes, we filter them out according to their annotations with 64 keywords such as "transferase", "hydrolase", "kinase", etc. To keep consistent with the training scale in the existing tools, we built our training set and test set with the N-terminal acetylated proteins before Jan 2016, which is the approximate time the tools were released. Furthermore, to avoid overestimation, we removed the proteins from a training set with 30% similarity to the test set using CD-HIT [22]. After processing, 3477 N-terminal acetylation modification sequences were left as our training data with 794 protein sequences as test data.

Our work involves seven types of N-terminal acetylation we collected from the Swiss-prot database: N-acetylalanine (short in A), N-acetylcysteine (short in C), N-acetylglycine (short in G), N-acetylmethionine (short in M), N-acetylserine (short in S), N-acetylthreonine (short in T), N-acetylvaline (short in V). Table 1 shows the numbers of each type of N-terminal acetylated site.

Table 1. The number of each type of N-terminal acetylated site.

Acetylated types	Training set	Test set
A	1233	293
C	15	45
G	54	34
M	1049	336
S	905	130
T	207	34
V	25	32

2.2. Data preprocessing and analysis

Biological experiments show that N-terminal acetylation occurs on one protein's first 30 amino acids. Hence, we will identify the acetylated sites on proteins' first 30 amino acids [1]. Existing tools consider removing the first residue, methionine (M), because N-terminal acetylation occurs at the N-terminus, and the start codon translates into methionine in the first position of each sequence. However, recent analysis has shown that acetylation can indeed occur on Methionine as the first residue. Therefore, we have decided to keep methionine in our study.

To exclude irrelevant amino acids far away from N-terminal, we cut out 60 amino acids from the N-terminal of proteins per sample. Compared to the existing works [17,18] that employed a window size of five or seven, our selection can conserve the context from protein sequences. For better demonstrating the potential sequential patterns on the fragments, we analyzed the first 60 amino acids of all N-terminal acetylated proteins from the training set using the WebLogo 3 website [23]. WebLogo is a sequence analysis tool that is used to visualize and analyze protein or DNA sequences. It allows users to generate a graphical representation of sequence logos which displays the conservation and variability of amino acids or nucleotides at each position in a sequence alignment. The x-axis identifies the position of the first 60 bits of the sequence, and the y-axis in WebLogo represents the height or frequency of occurrence of each amino acid or nucleotide at a particular position. The higher the symbol on the y-axis, the more frequently that specific residue is found at that position in the sequence alignment.

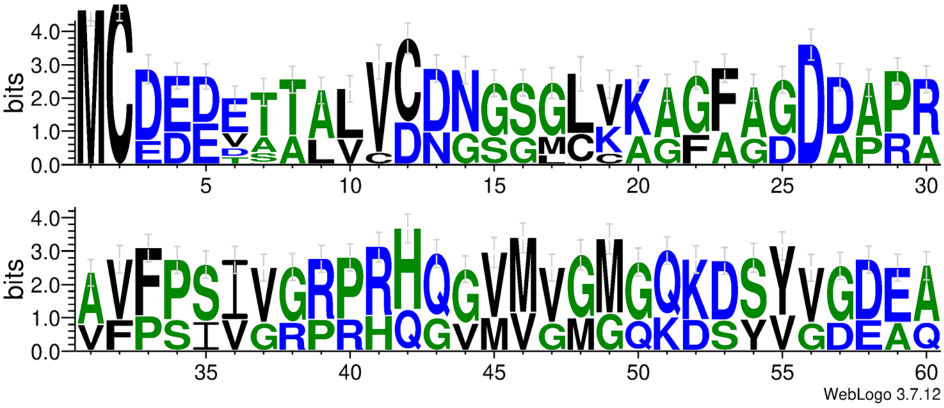

Figure 1. WebLogo graph of N-acetylcysteine.

Taking N-acetylcysteine as an example, Figure 1 shows the first residue, methionine (M), serves as the start codon of the sequence, while the second residue, cysteine (C), exhibits a high level of conservation. The 11th amino acid enriches valine (V), and the 26th amino acid fixes to aspartic acid (D). These positional preferences are clues for N-terminal protein acetylated sites. Meanwhile, we also observed that the signals led to the scattering of acetylated sites at all 60 positions of N-terminal proteins. It inspired us to consider not only the short-range but also the long-range dependency of residual compositions to identify protein N-terminal acetylated sites in this work.

To evaluate our proposed method fairly, we used a 5-fold cross-validation strategy to split the collected data [24]. In 5-fold cross-validation, the training data was randomly divided into 5 non-overlapping and equal-sized subsets. Of the five subsets, one subset was assigned for validating the model, and the remaining four subsets were treated as training data. The training and validation processes were repeated five times to calculate the average unbiased validation results.

2.3. Data encoding

To convert the protein fragments into a computational format, we encoded each fragment into two types of code: one-hot vector and physicochemical properties.

2.3.1. One-hot coding

According to Figure 1, some underlying sequential patterns are located at protein around acetylated sites. To encode the fragment sequences, one-hot vector was employed in this work [18]. In one-hot coding, each 60 amino acid fragment is represented as a 60*20 2-dimensional (2D) matrix, which consists of a 20-dimensional binary vector with a one in the index corresponding to the amino acid in the protein sequence. In the encoding scheme, every protein fragment will be mapped to an exclusive and sparse matrix, quantifying amino acids and maintaining their relative positions.

2.3.2. Physicochemical properties coding

Some studies indicate that there are strong connections between some physicochemical properties of amino acids and various protein translational modifications [25]. To introduce this potential modality to N-terminal acetylated site prediction, we encoded each protein fragment by all the physicochemical properties from the AAindex [26] and looked for the most critical physicochemical properties information by observing the gradient when training the model. Finally, we chose the top 25% of the physicochemical features with the highest contribution during training as inputs for the model, totaling 531 dimensions.

2.4. The architecture of MTNA

We initially designed an LSTM (long short-term memory) network based on one-hot encoding due to the conservation of the sequence. However, as mentioned earlier, we preserved the first residue, methionine, in the sequence. However, due to the primary ability of LSTM networks to capture long-range information, the sequence-based subnetwork was unable to predict whether N-acetylmethionine would occur. Therefore, we introduced physicochemical properties as additional encoding to enhance the input information. Additionally, we specifically designed a CNN model to capture the short-range information within the input sequence. Finally, we designed two different subnets for two input modalities to better obtain the residual descriptions from the sequence fragments. The architecture of MTNA is shown in Figure 2.

As shown in Figure 2, the one-hot network consists of two layers of LSTM and one LayerNorm layer. Subsequently, a linear layer is applied to reduce the dimensionality. The physicochemical property subnetwork comprises four layers of CNN (convolutional neural network) followed by one linear layer. After each CNN layer, regularization is performed on each batch. Similarly, a linear layer is applied at the end for dimensionality reduction. The results from the two subnets are inputted into the ensemble network and then undergo dimensionality reduction via two linear layers to generate the prediction results.

The PTM site is related to its structure [18], so the long-range information of the protein fragment needs to be considered to detect the structural motifs that may be sparse in the sequences. We chose the LSTM layer to process the one-hot vector to extract the long-range information as follows.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.2}$$

$$\tilde{C} = \tanh\,(W_c \cdot [h_{t-1}, x_t] + b_i) \tag{2.3}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C} \tag{2.4}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{2.5}$$

$$h_t = o_t * \tanh\,(C_t) \tag{2.6}$$

where x_t is the current amino acid on the sequence, and h_{t-1} is the hidden representation of its previous neighboring amino acid. LSTM holds the amino acid upstream information in cell C_t by multiplying x_t and h_{t-1} with a learnable weight matrix W_t, called short-term memory. The LSTM achieves long-term memory in a cell C_t by multiplying x_t and h_{t-1} with another learnable weight matrix W_f. With input amino acids from the sequence accumulated, the cell state C_{t-1} will overlay the updated information from the current amino acid by f_t and further restrain some historical information. Hence, the important long-range feature of input protein fragments will be enhanced by such “memory” and “forget” operations.

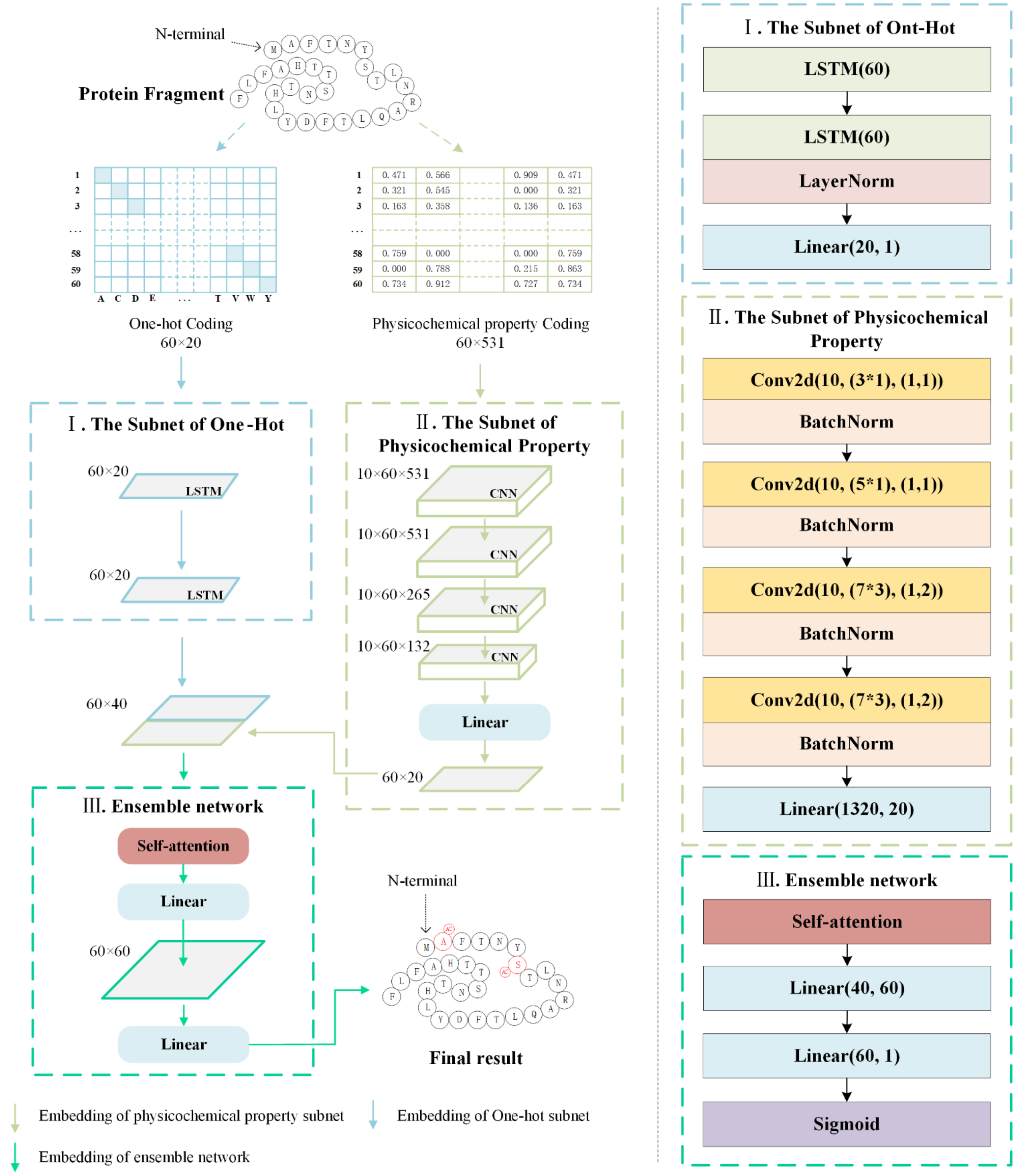

Figure 2. The architecture of MTNA.

On the other side, local sequential information is also contributed to protein N-terminal acetylation. We designed some CNN layers to extract the short-range information from the protein fragments for the modality of physiochemical properties.

$$z(u,v)=\sum_{i=-\infty}^{\infty}\sum_{i=-\infty}^{\infty}x_{i,j}\cdot k_{u-i,v-i} \tag{2.7}$$

X is the physicochemical properties of protein fragments, and K is the convolution kernel. In such a CNN layer, X is convoluted with its neighbors inside the learnable kernel with size of $i\times j$, resulting in a position-aware local latent representation Z. After four CNN layers in MTNA, the physicochemical properties at each position output a 19-dimensional embedding Z as the local-range descriptor.

After concatenating the embedding of the LSTM and CNN subnet outputs, we put them into a self-attention [27] layer to enhance the informative components across two modalities.

$$Attention(Q,K,V)=softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.8}$$

Q, K and V represent query, key and value by multiplying the input merged embedding with three learnable matrixes. Q multiplied with K^T calculates their inter-relation, which helps reveal the implicit importance of each component from the input embedding. Then, multiplication of the importance with V will lead to an enhanced embedding, allowing MTNA to pay more attention to the informative positions. This enhanced embedding will be used in making prediction results at the final linear layer.

2.5. Training strategy

N-terminal protein acetylated site prediction is a typical imbalanced classification. In this work, we employed focal loss [28] to control the data imbalance and the interference from some hard-to-identify samples. Focal loss is defined as follows:

$$FL(p_t)=-\partial_t(1-p_t)^{\gamma}\log(p_t) \tag{2.9}$$

$$p_t=\begin{cases} p & if\ \ y=1 \\ p-1 & otherwise \end{cases} \tag{2.10}$$

where p_t are the logits of positive samples. γ is the sample weight and set to 0.033 according to the distribution between positives and negatives in the training set. δ_t is the modulation factor to scale the contribution from complex cases and set to 2 based on [28].

For training MTNA, we chose the Adam optimizer with an initial learning rate of 0.001 and a decay rate of 0.8. We also arranged the early stopping strategy checking at each epoch. The training process will stop if the validation loss cannot improve within 15 epochs. Once a model comes to converging, we can input any test sequences to the trained model to identify N-terminal protein acetylated sites.

Due to the imbalanced distribution of positives and negatives, we evaluated our method and the existing tools via metrics indicating the performance from the positive and negative perspectives, such as sensitivity (SN), specificity (SP), Matthews correlation coefficient (MCC) and F1-score [29]. At the same time, we performed a comparison with other tools using the average precision (AP) and area under curve (AUC) metrics, which derive from the precision-recall curve (PR curve) and receiver operating characteristic curve (ROC curve) [30]. The PR curve is a curve of precision versus recall (sensitivity) for all possible cut-offs for a test. ROC is a plot between true sensitivity and false 1–specificity for all potential cut-offs for a test. AP is the area under the PR curve, and AUC is the area under the ROC. These two metrics reflect the overall performance compared to other evaluation metrics. Hence, we employ them in the comparisons between MTNA and other tools.

3. Results and discussion

3.1. Performance of MTNA

Table 2 lists the averages and the variances of performance for each type of N-terminal protein acetylated sites from 5-fold cross-validation.

Table 2. Performance of MTNA with 5-fold cross-validation.

Acetylated type	SN	SP	MCC	F1-Score	AP	AUC
N-acetylalanine (A)	0.921 ± 0.02	0.998 ± 0.00	0.908 ± 0.01	0.922 ± 0.01	0.874 ± 0.05	0.986 ± 0.00
N-acetylcysteine (C)	0.789 ± 0.19	0.999 ± 0.00	0.868 ± 0.11	0.949 ± 0.05	0.960 ± 0.04	0.999 ± 0.00
N-acetylglycine (G)	0.922 ± 0.05	0.998 ± 0.00	0.888 ± 0.07	0.920 ± 0.05	0.860 ± 0.08	0.982 ± 0.00
N-acetylmethionine (M)	0.729 ± 0.04	0.995 ± 0.00	0.710 ± 0.03	0.734 ± 0.03	0.775 ± 0.03	0.983 ± 0.01
N-acetylserine (S)	0.871 ± 0.01	0.997 ± 0.00	0.857 ± 0.01	0.868 ± 0.01	0.747 ± 0.07	0.980 ± 0.01
N-acetylthreonine (T)	0.782 ± 0.13	0.995 ± 0.00	0.772 ± 0.06	0.801 ± 0.07	0.756 ± 0.09	0.979 ± 0.02
N-acetylvaline (V)	0.641 ± 0.28	0.991 ± 0.00	0.593 ± 0.22	0.761 ± 0.14	0.698 ± 0.28	0.986 ± 0.01
ALL	0.816 ± 0.03	0.997 ± 0.00	0.812 ± 0.02	0.820 ± 0.01	0.821 ± 0.02	0.985 ± 0.01

As shown in Table 2, MTNA achieved good results in all involved types of N-terminal acetylated sites. Such excellence comes from our architecture’s multiple modalities and feature learning ability. MTNA works with two input modalities, including sequences and physiochemical properties, further designing two subnets to learn their deep representations. The separate subsets detect global and local information using LSTM layers and CNN layers. These deep representations get a weighted combination in the following self-attention layer. Such design gives the power of comprehensively describing the residues at protein N-terminal and leads to predicting the N-terminal acetylated sites accurately.

3.2. Performance of MTNA

To further indicate the superiority of MTNA, we compared MTNA with the currently available N-terminal acetylated tools, NetAcet [17] and NT-AcPredictor [18], with the independent set. Since these benchmark tools provide evaluation scores, AP and AUC were chosen as objective measures to evaluate the models. The comparative results on an independent test dataset are presented in Tables 3 and 4, respectively.

3.2.1. Comparison of MTNA and NetAcet

NetAcet only supports identifying N-acetylalanine, N-acetylglycine, N-acetylserine and N-acetylthreonine, and it just looks into the first three amino acids from a protein fragment. Meanwhile, MTNA can predict all types of N-terminal acetylated sites at the first 30 positions. We calculated the metrics based on the predictions of the first three amino acids for N-acetylalanine, N-acetylglycine, N-acetylserine and N-acetylthreonine on the independent test set.

As shown in Table 3, NetAcet cannot detect any N-acetylglycine sites from the independent set, while MTNA can identify most of them according to its 0.981 AUC. Among the other three types, MTNA also performed better than NetAcet.

NetAcet develops a fully connected neural network to combine all input physiochemical properties but fails to introduce sequential and positional information. Meanwhile, MTNA implicitly extracts sequential features along with LSTM-based operation position by position. In addition, physicochemical properties also complement in MTNA. Hence, MTNA overall beats NetAcet on the acetylated types of its interest.

Table 3. Comparison of MTNA and NetAcet with independent test set.

Acetylated type	Metric	NetAcet	MTNA
N-acetylalanine (A)	AP	0.716	**0.880**
	AUC	0.766	**0.998**
N-acetylglycine (G)	AP	-	**0.879**
	AUC	-	**0.981**
N-acetylserine (S)	AP	0.589	**0.775**
	AUC	0.651	**0.983**
N-acetylthreonine (T)	AP	0.504	**0.811**
	AUC	0.641	**0.996**

3.2.2. Comparison of MTNA and NT-AcPredictor

NT-AcPredictor predicts the type of N-terminal acetylation on the fragments by a rule-based decision tree approach. Since it can only calculate acetylation sites for the first 10 residues, we only use the first 10 residues to compute the metrics when comparing.

As shown in Table 4, MTNA gained higher APs and AUCs on most N-terminal acetylated types except for N-acetylmethionine. It is observed that NT-AcPredictor performed poorly on N-acetylthreonine and N-acetylvaline. This is because the developers did not consider these two types in their rule summary. Hence, the decision tree is insensitive to the two types. MTNA can naturally cover all types of N-terminal acetylated sites involved in the training data as it is a data-driven predictor and does not require any feature engineering. This advantage will bring MTNA better generalization. On N-acetylmethionine, NT-AcPredictor performed perfectly, but MTNA failed in a few cases in AP and AUC. For such types with explicit patterns, a rule-based model can easily achieve better results than data-driven approaches. However, our deep architecture can still get superior performance in most cases.

Table 4. Comparison of MTNA and NT-AcPredictor with independent test set.

Acetylated types	Metric	NetAcet	MTNA
N-acetylalanine (A)	AP	**0.883**	0.874
	AUC	0.940	**0.986**
N-acetylcysteine (C)	AP	0.764	**0.960**
	AUC	0.878	**0.999**
N-acetylglycine (G)	AP	0.426	**0.860**
	AUC	0.703	**0.982**
N-acetylmethionine (M)	AP	**1.000**	0.775
	AUC	**1.000**	0.983
N-acetylserine (S)	AP	0.698	**0.747**
	AUC	0.844	**0.980**
N-acetylthreonine (T)	AP	0.117	**0.756**
	AUC	0.543	**0.979**
N-acetylvaline (V)	AP	0.121	**0.698**
	AUC	0.545	**0.986**

3.3. Model interpretation

In order to whiten the black box of our deep architecture, we attempted to analyze the effectiveness and contributions of its intermediate embedding.

3.3.1. Correlation analysis of two subnets

To check the independence between the embedding from the one-hot subnet and physicochemical property subnet, we calculated their Pearson correlation coefficients (PCCs), and the density curves of their PCCs are shown in Figure 3.

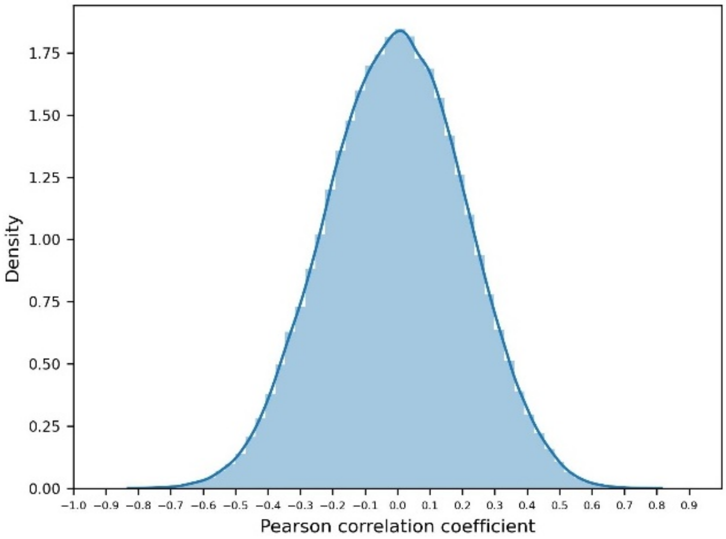

Figure 3. The density curves of PCC between the embedding from one-hot subnet and physicochemical property subnet.

In Figure 3, the x-axis represents the PCCs between two embeddings from two subsets on all training protein fragments, and the y-axis denotes the density of each PCC bin. The low PCCs between these two embeddings indicate that the input modalities describe a protein fragment from different perspectives. This suggests that the two subnets can generate complementary embeddings, which can be utilized in ensemble learning at the subsequent self-attention layer. The informative embeddings lay a solid foundation for precisely identifying N-terminal protein acetylated sites.

3.3.2. Positional embedding analysis

In this work, we employed saliency map at the last layer of MTNA to demonstrate the positional patterns of N-terminal acetylated sites. The salience map [19] is a visualization technique based on the gradient to show the contributions from input signals to outputs.

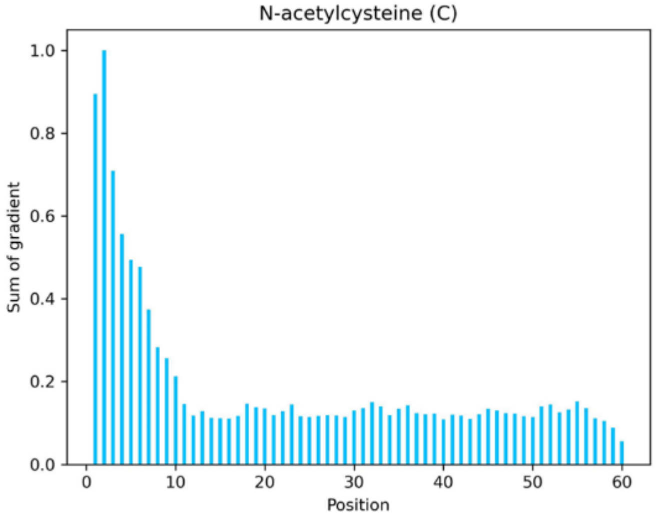

Figure 4. The contribution of each location to model results.

Figure 4 indicates the contribution of the amino acid at each position to model output. For N-terminal acetylated site prediction, the first 10 amino acids play a more important role than other positions. This implies that N-terminal acetylated site prediction is a particular PTM compared to lysine acetylated sites, since its major signals concentrate at the N-terminal. That might be why the existing tools investigate the first 3–5 positions. However, the latter positions also provide useful information in our model, resulting in better performance than the tools only considering shallow perspective fields.

3.3.3. UMAP analysis for embedding from different layers

To observe the distinctive ability of the intermediate embedding from each hidden layer, UMAP [20] is adopted to reduce the dimensions of these embeddings to a 2-dimensional vector. Thereby, the discrimination of these embeddings can be observed from the scatter plots on the 2-dimensional vectors.

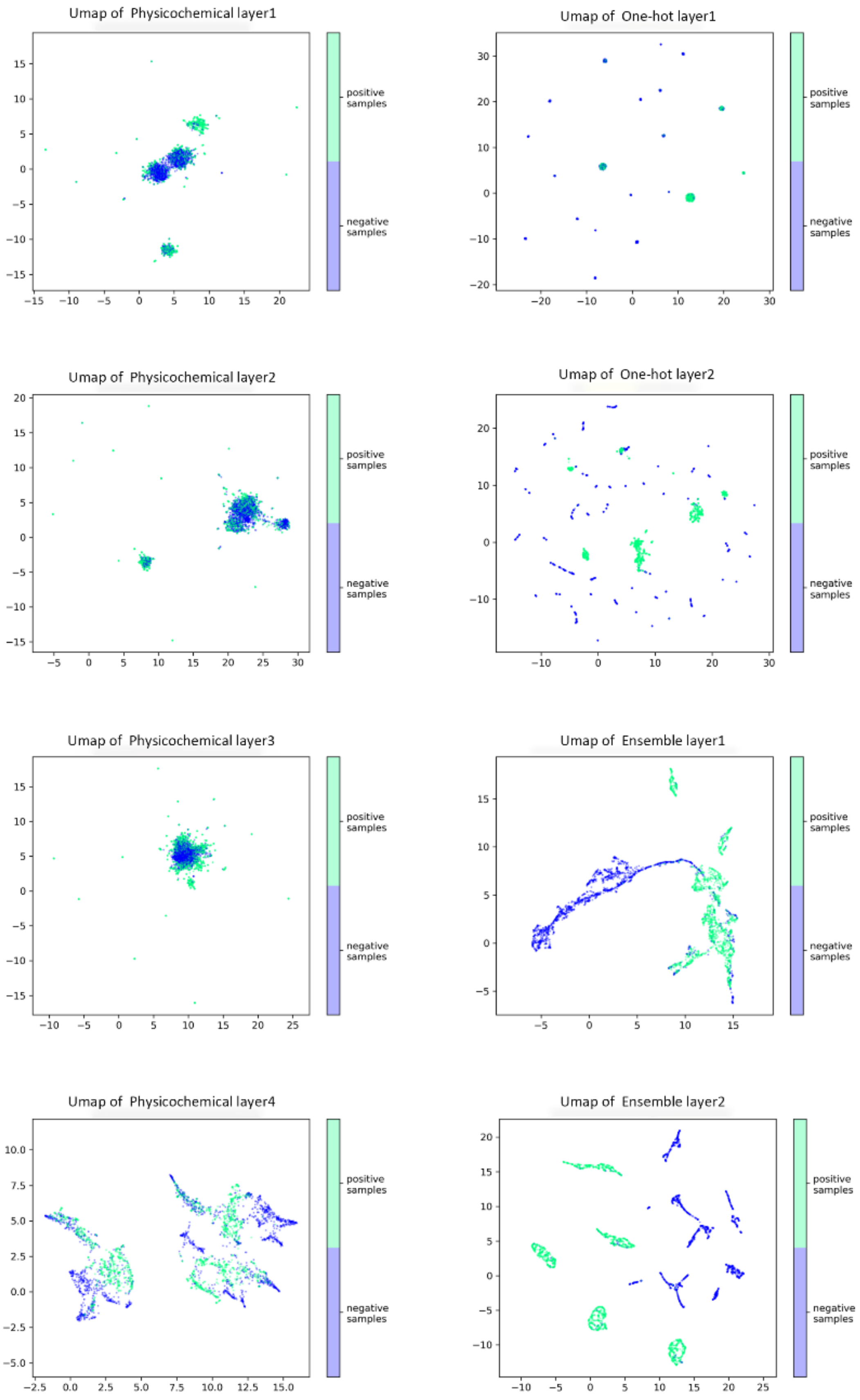

Figure 5. The scatter plots of 2D embeddings from different layers by UMAP.

The plots from the dimension-reduced vectors of each layer's embedding are shown as Figure 5, where the green dots represent the acetylated amino acids, and the blue dots are the non-acetylated ones. It can be seen that the positives and negatives are gradually separated along with the deeper CNN layers in the physicochemical property subnet. Similarly, after two LSTM layers, the embeddings of positives and negatives are also gradually separated, and the positives are also getting concentrated. These figures indicate that with deeper layers the generative embeddings obtain a better distinctive ability for identifying N-terminal protein acetylated sites.

4. Conclusions

In this study, we have designed a deep ensemble learning architecture called MTNA that performs better N-terminal protein acetylated site prediction. We compared MTNA with two existing N-terminal protein acetylated site prediction tools with an independent set. The results show that our predictor has stronger robustness and generalization than other tools. Further experiments of model interpretation exhibited that our input modalities contain rich information for describing protein fragments, the positional preference for N-terminal acetylated sites and the distinguishability of generative embeddings from our architecture. Our study improved the performance of N-terminal acetylated site prediction and provided new insights into deep learning network design for this field. Through continuous experimentation, we found that when retaining the first residue of the input sequence, methionine, the network fails to recognize N-acetylmethionine sites without the introduction of additional features. However, this issue can be effectively addressed by employing an ensemble learning strategy. In future experiments, we intend to incorporate protein language pre-training models to extract high-dimensional features from the sequences, aiming to address this issue more effectively. Our source code is available at https://github.com/Chenyb939/N-terminal-Acetylation.

Use of AI tools declaration

The authors declare they have not used artificial intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Jilin Provincial Natural Science Foundation under Grant No. 20200201288JC, No. 20210101174JC and No. 20200201158JC and the National Natural Science Foundation of China under Grant No. 42171450.

Conflict of interest

All authors declare no conflict of interest.

References

1. B. Polevoda, F. Sherman, N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins, *J. Mol. Biol.*, **325** (2003), 595–622. https://doi.org/10.1016/S0022-2836(02)01269-X

2. C. Yi, M. Ma, L. Ran, J. Zheng, J. Tong, J. Zhu, et al., Function and molecular mechanism of acetylation in autophagy regulation, *Science*, **336** (2012), 474–477. https://doi.org/10.1126/science.1216990
3. B. Polevoda, F. Sherman, The diversity of acetylated proteins, *Genome Biol.*, **3** (2002), 1–6. https://doi.org/10.1186/gb-2002-3-5-reviews0006
4. X. J. Yang, The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases, *Nucleic Acids Res.*, **32** (2004), 959–976. https://doi.org/10.1093/nar/gkh252
5. T. Arnesen, P. Van Damme, B. Polevoda, K. Helsens, R. Evjenth, N. Colaert, et al., Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans, *Proc. Natl. Acad. Sci.*, **106** (2009), 8157–8162. https://doi.org/10.1073/pnas.0901931106
6. C. S. Hwang, A. Shemorry, A. Varshavsky, N-Terminal acetylation of cellular proteins creates specific degradation signals, *Science*, **327** (2010), 973–977. https://doi.org/10.1126/science.1183147
7. A. J. Trexler, E. Rhoades, N-terminal acetylation is critical for forming α-helical oligomer of α-synuclein, *Protein Sci.*, **21** (2012), 601–605. https://doi.org/10.1002/pro.2056
8. R. Behnia, B. Panic, J. R. C. Whyte, S. Munro, Targeting of the Arf-like GTPase Arl3p to the Golgi requires N-terminal acetylation and the membrane protein Sys1p, *Nat. Cell Biol.*, **6** (2004), 405–413. https://doi.org/10.1038/ncb1120
9. D. C. Scott, J. K. Monda, E. J. Bennett, J. W. Harper, B. A. Schulman, N-Terminal acetylation acts as an avidity enhancer within an interconnected multiprotein complex, *Science*, **334** (2011), 674–678. https://doi.org/10.1126/science.1209307
10. T. Y. Lee, J. B. K. Hsu, F. M. Lin, W. C. Chang, P. C. Hsu, H. D. Huang, N-Ace: Using solvent accessibility and physicochemical properties to identify protein N-acetylation sites, *J. Comput. Chem.*, **31** (2010), 2759–2771. https://doi.org/10.1002/jcc.21569
11. A. F. Rope, K. Wang, R. Evjenth, J. Xing, J. J. Johnston, J. J. Swensen, et al., Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency, *Am. J. Hum. Genet.*, **89** (2011), 345. https://doi.org/10.1016/j.ajhg.2011.07.008
12. T. V. Kalvik, T. Arnesen, Protein N-terminal acetyltransferases in cancer, *Oncogene*, **32** (2013), 269–276. https://doi.org/10.1038/onc.2012.82
13. D. J. Welsch, G. L. Nelsestuen, Amino-terminal alanine functions in a calcium-specific process essential for membrane binding by prothrombin fragment 1, *Biochemistry*, **27** (1988), 4939–4945. https://doi.org/10.1021/bi00413a052
14. D. Umlauf, Y. Goto, R. Feil, Site-specific analysis of histone methylation and acetylation, *Epigenet. Protoc.*, **287** (2004), 99–120. https://doi.org/10.1385/1-59259-828-5:099
15. K. F. Medzihradszky, In-solution digestion of proteins for mass spectrometry, *Methods Enzymol.*,**405** (2005), 50–65. https://doi.org/10.1016/S0076-6879(05)05003-2
16. C. Xia, Y. Tao, M. Li, T. Che, J. Qu, Protein acetylation and deacetylation: An important regulatory modification in gene transcription, *Exp. Ther. Med.*, **20** (2020), 2923–2940. https://doi.org/10.3892/etm.2020.9073
17. L. Kiemer, J. D. Bendtsen, N. Blom, NetAcet: Prediction of N-terminal acetylation sites, *Bioinformatics*, **21** (2005), 1269–1270. https://doi.org/10.1093/bioinformatics/bti130

18. K. D. Yamada, S. Omori, H. Nishi, M. Miyagi, Identification of the sequence determinants of protein N-terminal acetylation through a decision tree approach, *BMC Bioinf.*, **18** (2017), 289. https://doi.org/10.1186/s12859-017-1699-4
19. K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, *arXiv preprint*, (2013), arXiv:1312.6034. https://doi.org/10.48550/arXiv.1312.6034
20. L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, *arXiv preprint*, (2018), arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426
21. The UniProt Consortium, UniProt: the universal protein knowledgebase in 2023, *Nucleic Acids Res.*, **51** (2023), D523–D531. https://doi.org/10.1093/nar/gkac1052
22. Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT suite: A web server for clustering and comparing biological sequences, *Bioinformatics*, **26** (2010), 680–682. https://doi.org/10.1093/bioinformatics/btq003
23. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: A sequence logo generator, *Genome Res.*, **14** (2004), 1188–1190. https://doi.org/10.1101/gr.849004
24. J. Zhang, H. Chai, S. Guo, H. Guo, Y. Li, High-throughput identification of mammalian secreted proteins using species-specific scheme and application to human proteome, *Molecules*, **23** (2018), 1448. https://doi.org/10.3390/molecules23061448
25. P. Radivojac, V. Vacic, C. Haynes, R. R. Cocklin, A. Mohan, J. W. Heyen, et al., Identification, analysis, and prediction of protein ubiquitination sites, *Proteins*, **78** (2010), 365–380. https://doi.org/10.1002/prot.22555
26. S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, *Nucleic Acids Res.*, **36** (2007), D202–D205. https://doi.org/10.1093/nar/gkm998
27. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *2017 Advances in Neural Information Processing Systems*, (2017), 1–11.
28. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár. Focal loss for dense object detection, in *2017 International Conference on Computer Vision (ICCV)*, IEEE, (2017), 2980–2988.
29. J. Zhang, Y. Zhang, Z. Ma, In silico prediction of human secretory proteins in plasma based on discrete firefly optimization and application to cancer biomarkers identification, *Front. Genet.*, **10** (2019), 542. https://doi.org/10.3389/fgene.2019.00542
30. T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PloS One*, **10** (2015), e0118432. https://doi.org/110.1371/journal.pone.0118432