BMC Biology

RESEARCH

Open Access

PBertKla: a protein large language model for predicting human lysine lactylation sites

Hongyan Lai[1†], Diyu Luo[1†], Mi Yang[2], Tao Zhu[1], Huan Yang[3], Xinwei Luo[4], Yijie Wei[4], Sijia Xie[4], Feitong Hong[4], Kunxian Shu[1*], Fuying Dao[5*] and Hui Ding[4*]

Abstract

Background Lactylation is a newly discovered type of post-translational modification, primarily occurring on lysine (K) residues of both histones and non-histones to exert diverse effects on target proteins. Research has shown that lysine lactylation (Kla) modification is ubiquitous in different cells and participates in the determination of cell function and fate, as well as in the initiation and progression of various diseases. Precise identification of Kla sites is fundamental for elucidating their biological functions and uncovering their application potential.

Results Here, we proposed a novel human Kla site predictor (named PBertKla) through curating a reliable benchmark dataset with proper sample length and sequence identity threshold to train a protein large language model with optimal hyperparameters. Extensive experimental results consistently demonstrated that our model possessed robust human Kla site prediction ability, achieving an AUC (area under receiver operating characteristic curve) value of over 0.880 on the independent validation data. Feature visualization analysis further validated the effectiveness of in feature learning and representation from Kla sequences. Moreover, we benchmarked PBertKla against other cutting-edge models on an independent testing dataset from different sources, highlighting its superiority and transferability.

Conclusions All results indicated that PBertKla excelled as an automatic predictor of human Kla sites, and it would advance the investigation of lactylation modifications and their significance in health and disease.

Keywords Protein large language model, BERT, Transformer, Lysine lactylation site, Human

†Hongyan Lai and Diyu Luo contributed equally to this work.

*Correspondence:
Kunxian Shu
shukx@cqupt.edu.cn
Fuying Dao
fuying.dao@ntu.edu.sg
Hui Ding
hding@uestc.edu.cn
[1] Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2] Clinical Hospital of Chengdu Brain Science Institute, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[3] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, Zhejiang 324000, China
[4] Center for Informational Biology, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China
[5] School of Biological Sciences, Nanyang Technological University, Singapore 639798, Singapore

Background

Protein post-translational modifications (PTMs) refer to the enzyme-catalyzed covalent modification of amino acid residues after protein synthesis. PTMs do not alter the original amino acid sequence but regulate the physicochemical properties, structure, and stability of the proteins, thereby diversifying their functions [1]. At present, a wide variety of PTMs have been identified and well-studied, including phosphorylation, acetylation, ubiquitination, methylation, and glycosylation [2–4]. Different PTMs regulate various physiological and biochemical processes, influencing growth, development, and disease progression [5–7].

Lysine lactylation (Kla), as a novel type of PTM [8], is catalyzed by lactyltransferases (e.g., HAT p300) [9] and delactylases (e.g., class I histone deacetylases, HDAC1-3) [10] in the lactate metabolism process. It dynamically regulates cellular metabolism, cell signaling, neuronal excitation, immune homeostasis, and the maintenance of cardiovascular and cerebrovascular functions. Studies have shown that Kla is closely associated with metabolic disease [11], neurodegenerative diseases [12], cardiovascular diseases [13], and cancers [14–17], exhibiting functional heterogeneity. In many cancers, Kla demonstrates tumor-promoting properties. For instance, Zhou et al. [18] found that upregulated histone Kla levels enhance liver metastasis of colorectal cancer. Conversely, Kla can also exert tumor-suppressive effects. Longhitano et al. [19] reported that lactate increases the levels of lactate-induced histone H3K18la modification, significantly reducing the proliferation and migration of uveal melanoma cells.

Therefore, accurately identifying Kla sites is crucial for understanding their biological functions and unraveling the mechanism of Kla in various diseases. Liquid chromatography–mass spectrometry (LC–MS/MS) is a powerful technique for high-throughput and high-precision identification of Kla sites. However, its application is constrained by high sequencing costs and time consumption. To fill this gap, a series of computational models have been developed [20]. For example, Jiang et al. [21] proposed an ensemble deep learning (DL) model, FSL-Kla, to identify Kla sites, achieving an AUC value of 0.889. Additionally, Lv et al. [22] first leveraged LC–MS/MS generated global lactylome profile to construct an attention-based Kla site prediction tool, DeepKla, which achieved an AUC value of 0.972 on independent datasets, demonstrating robustness and transferability. Based on the same dataset, Guan et al. [23] developed DeepKlapred to identify Kla sites in rice via multi-view feature fusion. Similarly, tools such as Auto-Kla [24] and A/EBFF-Kla [25] were developed based on recently published Kla profiles to accurately predict human Kla sites. Despite significant progress in Kla site identification, current models still have room for improvement in generalizability and accuracy.

In this study, we introduced a novel tool, PBertKla, based on a protein large language model for precise identification of human Kla sites. We first curated a reliable human Kla site benchmark dataset with appropriate sample length and similarity threshold. Furthermore, we fine-tuned ProteinBERT, a deep language model pretrained on large-scale protein sequences and Gene Ontology (GO) annotation knowledges [26], on our Kla site dataset using optimal hyperparameter to build PBertKla. Comprehensive experimental results demonstrated that PBertKla achieved consistently high predictive performance (AUC value exceeding 0.880) across both training and independent validation data. Feature visualization analysis confirmed the effectiveness of PBertKla's large-model strategy for Kla site identification. To further validate its performance, we compared PBertKla with other state-of-the-art models, highlighting its superiority across independent validation datasets from diverse sources. We believe that PBertKla will facilitate the overcoming of obstacles in identifying Kla sites, thereby advancing the in-depth exploration of its pivotal role in cellular function regulation.

Results

Curation of human Kla benchmark dataset

Given that the bias of amino acids specific to their positions relative to Kla site, it was of considerable importance to select an appropriate window size for flanking sequences around Kla sites and to characterize Kla samples well. On the other hand, an unsuitable sequence similarity threshold of CD-HIT would compromise both sample distribution and model performance by either failing to eliminate redundant sequences (high threshold) or reducing sequence diversity (low threshold). To create a reliable and practical benchmark dataset and a strong predictive model for effectively identifying human Kla sites, it required determining the proper sample length and sequence similarity of human Kla peptide samples, which was implemented based on a preliminary analysis of their impact on prediction performance.

This thorough analysis experiment involved grid search over sample length ranging from 33 to 59 (with an increment of 2 amino acids) and sequence similarity threshold ranging from 30 to 70% (with a 10% increment). We analyzed the predictive performance of each dataset with specific sample length and sequence similarity using ProteinBERT model with default parameters. The prediction AUC values obtained on independent validation datasets were summarized in Fig. 1 and Additional file 1: Table S1. AUC values of all datasets fluctuated between 0.765 and

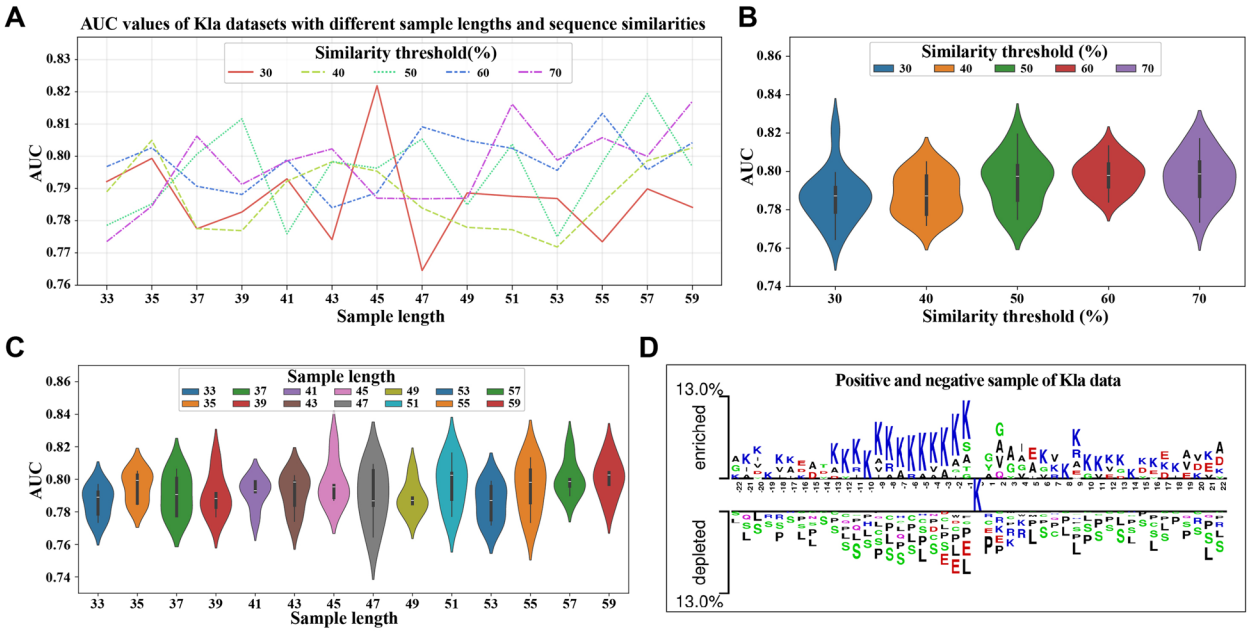

Fig. 1 The determination of optimal sample length and sequence similarity threshold for defining human Kla benchmark datasets. **A** The line chart showing each AUC value of each dataset with specific sample length and sequence similarity, in which x axis is sample length, different colored lines correspond to different sequence similarity thresholds. **B**–**C** Two violin plots visualizing the AUC values of datasets generated based on different sequence similarity thresholds of CD-HIT (corresponding to different colors) and different sample lengths (corresponding to different colors), respectively. **D** Sequence characteristics of Kla and non-Kla sites in the training data

0.822, indicating a marginal difference of approximately 6%. Thereinto, datasets with sample length of 45 yielded AUCs ranging from 0.787 to 0.822, and AUCs of datasets with 30% sequence similarity ranged from 0.765 to 0.822. Although the precise reasons for this variability remain unclear, it is evident that ill-suited data resulted in less stable model performance. To optimize computation efficiency and facilitate comparison, the dataset with sample length of 45 and 30% sequence similarity, including 9640 Kla and non-Kla peptide samples and demonstrating superior performance over all other datasets, was thus determined as the human Kla benchmark dataset used for all subsequent analyses in this study, unless otherwise specified.

To intuitively examine the amino acid preferences around lactylated sites, we utilized Two-Sample-Logo [27] to generate sequence maps of fragments surrounding both Kla and non-Kla sites, thereby comparing the amino acid distributions between positive and negative samples. The analysis focused on 45-residue fragments centered on Kla sites (positions − 22 to + 22 relative to the central lysine). As shown in Fig. 1D, the analysis revealed an enrichment of lysine (K) and alanine (A) in both the upstream and downstream regions of lactylation sites, while glycine (G) exhibited a downstream-specific preference. These findings indicate specific amino acid preferences around lactylation sites, supporting the interpretation of these regions as contextually meaningful ("semantic") sequences.

Determination of PBertKla predictor

In view of parameter tuning, especially the learning rate and batch size hyperparameters, in DL models is critical for achieving efficient convergence, stable training, and well generalization performance [28]. Generally, to balance training speed and model accuracy, hyperparameter fine-tuning could be effectively conducted by grid search. Here, to improve the predictive performance of the pretrained protein language model for human lysine lactylation (Kla) site identification, we first conducted systematic hyperparameter optimization through a grid search evaluating 36 combinations of learning rates (six values) and batch sizes (six values). Furthermore, to systematically evaluate the impact of sequence length parameter on Kla identification task, we performed a comprehensive analysis by varying sequence length and assessing model's predictive capability on the validation data. The specific search space of three parameters was enumerated in Additional file 1: Table S2. The Acc, MCC, and AUC evaluation metrics were employed to determine the optimal parameter settings.

The human Kla site prediction performance of different parameters on the validation data was demonstrated in Fig. 2. For different learning rates and batch sizes,

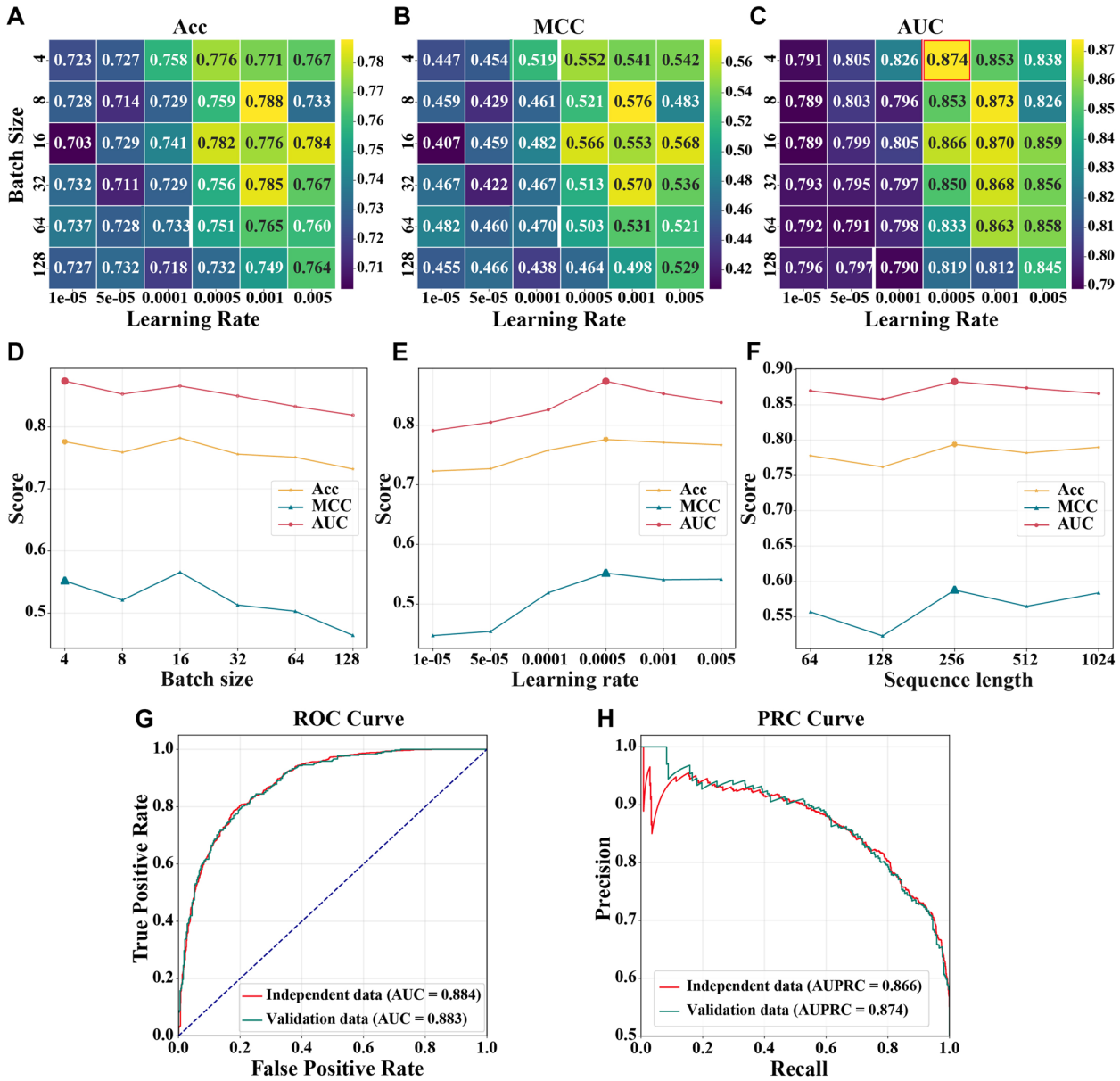

Fig. 2 The parameter optimization of the pretrained protein language model for accurately predicting human Kla sites. **A**–**C** The heatmap of learning rate (x-axis) and batch size (y-axis) optimization in terms of Acc, MCC, and AUC metrics, respectively. Brighter yellow blocks correspond to higher metric values in the heatmap. **D**–**F** representing the performance of different learning rates, batch sizes, and sequence length. Yellow, green, and red lines indicate Acc, MCC, and AUC metrics. **G** ROC (receiver operating characteristic) curve and **H** PRC (precision-recall) curve plotting the overall prediction performance of PBertKla with the optimal parameter setting on the validation data (green line) and independent validation data (red line), respectively

the values of Acc, MCC, and AUC metrics (as shown in Fig. 2A–C) were 70.3 ~ 78.8%, 0.407 ~ 0.576, and 0.789 ~ 0.874, respectively. The optimization of these two hyperparameters would achieve about 9% increase in Acc and AUC, as well as an ~ 17% improvement in MCC. As illustrated in Fig. 2D, there was a gradual decline trend in model's performance with the batch size increasing from 4 to 128. Referring to Fig. 2E, we observed an obvious positive trend in both Acc and MCC as well as AUC with learning rate less than 0.005, and a downward trend with learning rate more than 0.005. Regarding sequence length (Fig. 2F), analysis results revealed that Acc, MCC, and AUC all improved as sequence lengths approaching 256, but began to decline for sequences longer than

256. Finally, the optimal learning rate, batch size, and sequence length were set to 0.0005, 4, and 256, respectively, corresponding to the model with the highest AUC value of 0.883. Through the above exploration and analysis, the best parameter configuration was determined to construct PBertKla model for identifying human Kla sites. On the independent validation data, PBertKla achieved a prediction accuracy of 80.3%, sensitivity of 78.7%, specificity of 82.0%, and MCC, AUC, and AUPRC values of 0.607, 0.884, and 0.866, respectively (Fig. 2G–H).

Feature visualization analysis of PBertKla training process

To intuitively illustrate the feature learning process on human Kla sequence samples and investigate the classification ability of PBertKla, we further visualized the gradually changing feature maps of the training dataset in a 2-dimensional (2D) space using Uniform Manifold Approximation and Projection (UMAP) [29]. Figure 3 shows the sample distribution in each feature space during the training process of PBertKla, in which each point represented each sample, Kla sites were labeled with red color while non-Kla sites with blue color.

As seen from Fig. 3A, all samples were mixed, indicating that model could not distinguish human Kla from non-Kla samples based on the original features generated from embedding layer. When mapping the data to the low-dimensional space on the basis of the output features of six transformer-like blocks, these two types of samples become more dispersed (Fig. 3B–G). After merging all local and global representations through a concatenation layer, a relatively obvious boundary appeared between Kla and non-Kla samples, with Kla samples predominantly distributed on the right and non-Kla samples on the left (Fig. 3H). This suggested that the key features determining PBertKla's classification performance were likely encoded in this layer, highlighting the effectiveness of hybrid DL architectures. Overall, these feature visualization observations further illustrated the capacity of PBertKla model in capturing potential information for the identification of human Kla sites.

Performance of manual feature engineering on human Kla representation

For comparison, this section delved into the exploration of Kla sample representation with manual feature engineering. As reported by literature [30], the fusion scheme of seven classical feature encodings could characterize different types of PTM sites well. These comprehensive features would be taken into consideration to represent human Kla sites. Specifically, we extracted the sequence information of our human Kla benchmark dataset using position weight amino acid composition (PWAAC), amino acid relative position composition (AARPC), composition of k-space amino acid pairs (CKSAAP), composition of physical and chemical properties (CPCP),

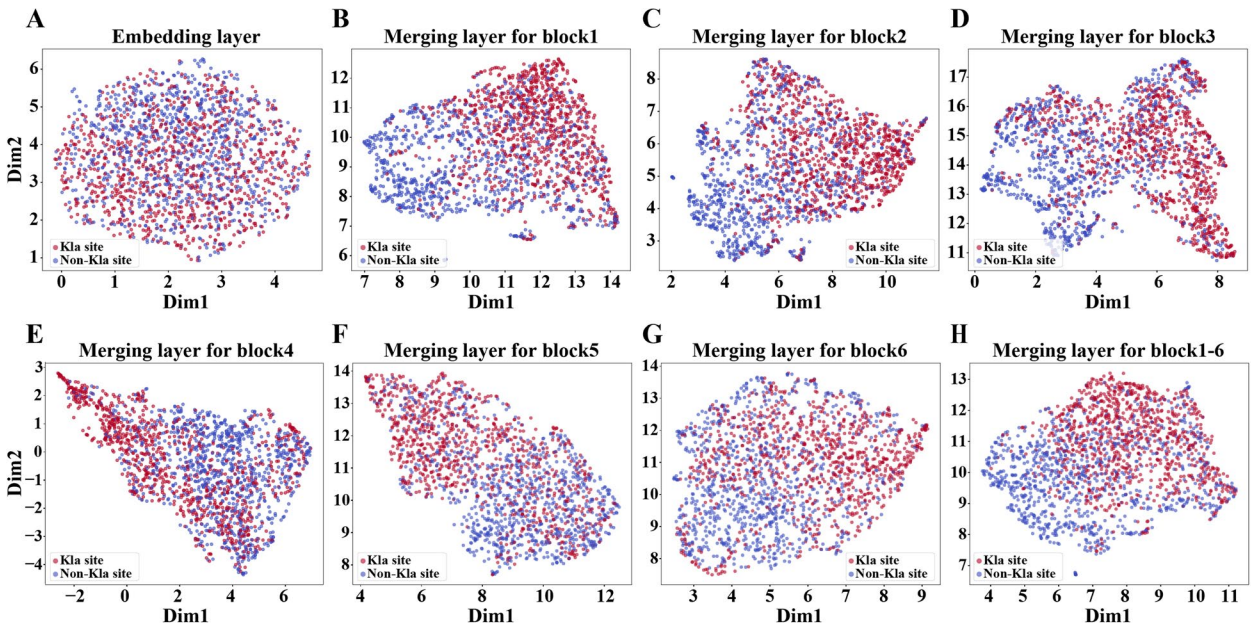

Fig. 3 The UMAP-based visualization of human Kla and non-Kla site sample distribution in 2D feature space. **A** The original feature map for input sequences originating from embedding transformation, **B**–**G** the feature maps originating from six transformer-like blocks, **H** the feature map originating from the integration of all blocks. Red and blue dots denoting human Kla and non-Kla site samples, respectively

encoding based on grouped weight (EBGW), composition, transition and distribution (CTD), and cone feature space (CFS) feature descriptors. Meanwhile, a simple end-to-end manner was adopted to integrate these features for multi-information fusion. As a result, 3080-dimensional feature vectors were produced for encoding Kla and non-Kla peptide samples. In this work, we applied both the base ("_L1") and stacking ("_L2") modes of PyTorch and fastai v1 neural network (NN) models to distinguish human Kla from non-Kla sites.

These models were learned on the training dataset and were assessed using the independent validation dataset. Performance evaluation results were presented in Fig. 4 and Additional file 1: Table S3. Based on the seven handcrafted features, NN models could identify human Kla sites with Acc of 72.0 ~ 79.6%, MCC of 0.445 ~ 0.595, and AUC of 0.808 ~ 0.882. As illustrated in Fig. 4, the stacking NN models performed better than base NN models, the Acc, MCC, and AUC values of NeuralNetFastAI_L2 (green) were higher than that of NeuralNetFastAI_L1 (yellow), and the same phenomenon was observed between NeuralNetTorch_L2 (red) and NeuralNetTorch_L1 (blue). The PyTorch models were more effective than fastai v1 models in identifying human Kla sites, and NeuralNetTorch_L2 performed the best and achieved the highest accuracy, MCC, and AUC of 79.6%, 0.595, and 0.882, respectively. Overall, the multi-view feature fusion strategy for distinguishing human Kla sites from non-Kla sites achieved maximum values of 79.6% Acc, 0.595 MCC, and 0.882 AUC, which underperformed compared to PBertKla's results of 80.3% Acc, 0.607 MCC, and 0.884 AUC (as shown in Fig. 4A–B). All analysis results indicated that this protein language model incorporating local and global representations constituted an effective approach for capturing Kla site-specific information.

Comparison of PBertKla with existing methods

For further assessing the effectiveness of our proposed PBertKla, we conducted a comparative analysis against state-of-the-art Kla prediction methods. Due to

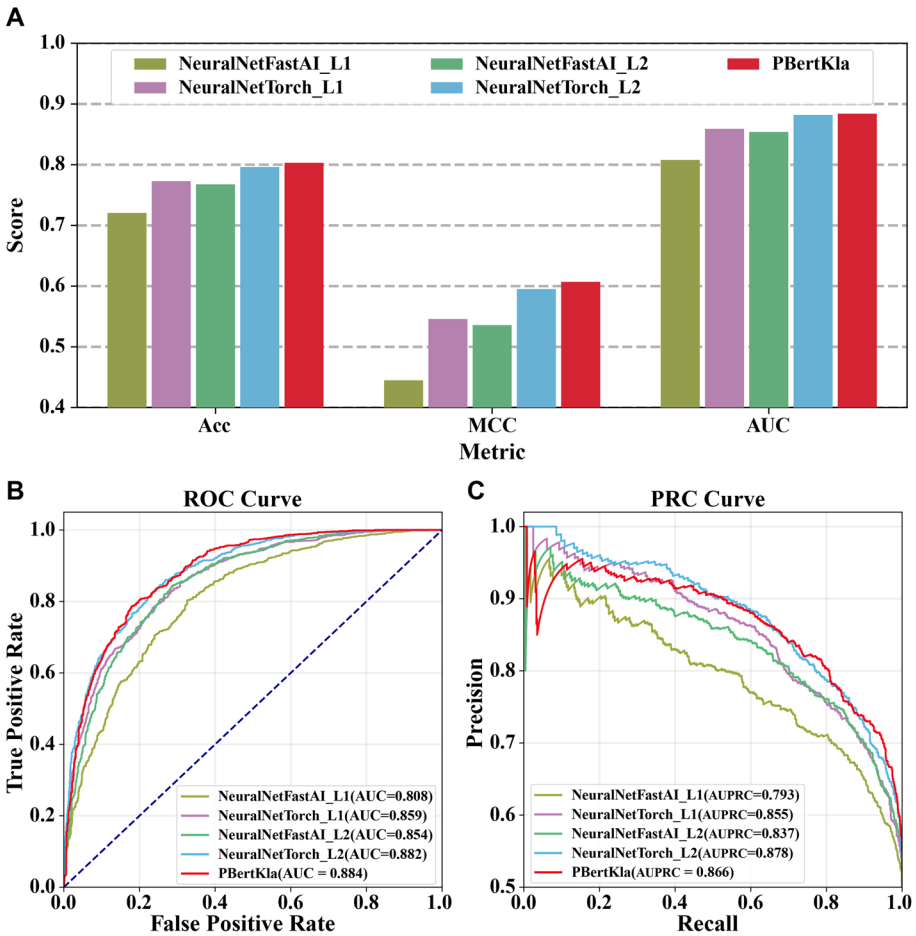

Fig. 4 Performance evaluation of handcrafted features for representing human Kla sites based on neural network models. **A** The Acc, MCC, and AUC metrics, as well as **B** ROC and **C** PRC curves of the four neural network algorithms and our proposed PBertKla method (in red color)

challenges in accessing or implementing FSL-Kla and A/EBFF-Kla, our comparison was limited to DeepKla [22], Auto-Kla [24], and DeepKlapred [23]. To ensure equitable comparison of algorithmic performance characteristics, the adoption of standardized training and validation protocols across all models on the same curated benchmark datasets was methodologically imperative. Thus, we learned DeepKla, AutoKla, and DeepKlapred models on our human Kla training data to predict the independent data. Independent validation results (Fig. 5 and Additional file 1: Table S4) showed that the human Kla site prediction accuracy, MCC, and AUC values of DeepKla were 76.1%, 0.533, and 0.853, those of AutoKla and DeepKlapred were 72.5%, 0.486, 0.837 and 70.9%, 0.417, 0.767. PBertKla demonstrated consistent superiority over DeepKla, AutoKla, and DeepKlapred models across most evaluation metrics, achieving relative improvements of approximately 4–9% in accuracy, 7–19% in MCC, 3–12% in AUC, and 2–11% in AUPRC, respectively (as depicted in Fig. 5A, C–D). Overall, our introduced PBertKla demonstrated an outstanding performance than other tools in predicting human Kla sites.

To evaluate the robustness and generalization ability of a model, independent testing datasets were usually utilized to confirm the prediction performance. Consequently, we established another independent testing dataset covering 2488 human Kla peptides derived from OSCC (oral squamous cell carcinoma) and NA (normal adjacent) tissues for further assessing the predictive capability of the above tools. As expected, the superior performance of PBertKla was sustained into the testing phase, and an apparent discrepancy in Kla identification accuracy presented in Fig. 5B. The PBertKla (in red color) accurately identified 2164 Kla sites, achieving the highest Acc value of 90.6% (2255/2488). Meanwhile, DeepKla (in green color), AutoKla (in yellow color), and DeepKlapred (in purple color) underperformed with accuracy of 79.1% (1968/2488), 65.5% (1608/2455), and 86.5% (2151/2488), respectively. All consistent results highlighted the PBertKla's excellent prediction capability and robust transferability, and further affirmed the architectural superiority

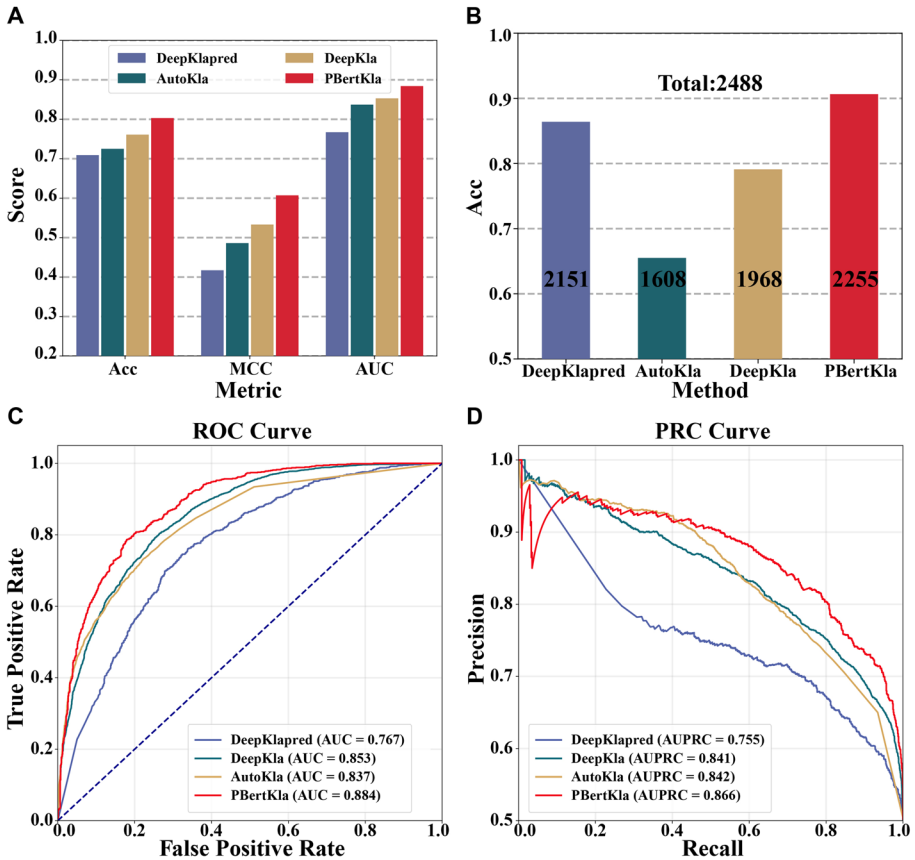

Fig. 5 Performance comparison between PBertKla and existing tools in identifying human Kla sites. **A**, **C**–**D** The prediction Acc, MCC, and AUC metrics, as well as ROC and PRC curves of PBertKla (in red color), DeepKla (in green color), AutoKla (in yellow color), and DeepKlapred (in purple color) on the independent validation dataset. **B** The human Kla site identification accuracy of each tool on the independent testing dataset

of the pretrained protein large language model in predicting human Kla sites.

Discussion

Lysine lactylation (Kla), an important type of protein PTMs, plays a key regulatory role in cellular metabolism, inflammatory response, and the onset and progression of various diseases. More and more researches pay attention to explore the application potential of Kla modifications as biomarkers for disease prognosis [31], therapy [32], drug-effect monitoring [33], and so on. However, the accurate identification of lactylated proteins and exact localization of lactylated amino acid residues are essential for revealing the molecular mechanisms and functional diversity of Kla modifications. Currently, aside from time-consuming and labor-intensive experimental methods, only a handful of computational tools are available for the convenient and rapid detection of human Kla sites.

Given the current scarcity of effective models for automatically identifying human Kla sites, we proposed a novel human Kla site prediction approach in this work. A reliable benchmark dataset is essential for developing an efficient and stable classifier, serving as the foundation that significantly enhances the overall performance of the model. Therefore, we firstly conducted a preliminary analysis to determine the optimal and representative human Kla benchmark dataset with appropriate sample length of 45 and sequence identity of 30%. Secondly, we employed a pretrained protein language model to learn from the above-mentioned human Kla peptide sequence data, resulting in the development of the PBertKla model for identifying human Kla sites. This model enabled the capture of both local (character level) and global (whole sequence level) representations of human Kla peptides by incorporating the classic BERT and Transformer modules. It exhibited impressive and robust prediction performance on both independent validation and testing datasets, achieving accuracies of about 80.3% and 90.6%, respectively. Moreover, the comparative experiment findings demonstrated that PBertKla exceeded the performance of current Kla predictors in identifying human Kla sites. Owning to the limitation of data and calculation, the potential for our proposed method to identify Kla sites in other species remains unexplored in this research.

Furthermore, the following aspects on automatically predicting human Kla site warrant further investigation. First, with the growing popularity of MS/LC–MS technology and the significantly increasing focus on lactylation, there is a rapid expansion of human lactylation data. To fully leverage these and information, a broad range of data from databases and published literature can be systematically collected in the future to construct a more comprehensive and standardized dataset of human lactylation sites. Second, to date, no research has thoroughly examined the effectiveness of different traditional machine learning (ML) algorithms and conventional feature engineering based on amino acid sequences in distinguishing human lactylation from non-lactylation sites. The potential of ML-based human Kla site predictors merit additional exploration. Third, it would be beneficial to investigate the development of multi-scale deep protein language learning models for the interpretable prediction of Kla sites. This model should incorporate protein structural data, evolutionary conservation information, and other relevant features into advanced deep learning approaches [34, 35], such as the recently published ESM3 model [36]. By integrating these diverse data sources, the model would provide a more comprehensive understanding of the biological context and improve the accuracy and interpretability of Kla site prediction. Lastly, to enhance accessibility for researchers in this field, we propose developing an advanced online platform. This platform would offer tools for data exploration, facilitate easy downloads, and provide human (or even potentially across various species) lactylation modification prediction services while also supporting comprehensive analysis of prediction results from different models.

In addition, extending studies beyond human data should be performed to analyze lactylation in different species, which could reveal conserved mechanisms and broaden our understanding of its evolutionary significance. It would be meaningful to investigate the crosstalk between lactylation and other modifications (e.g., phosphorylation) to uncover potential synergistic or antagonistic effects on protein function and gain a more comprehensive understanding of protein regulation.

Conclusions

In this study, we developed PBertKla, a novel deep learning-based predictor for accurately identifying human lysine lactylation (Kla) sites. By curating a reliable benchmark dataset and leveraging a pretrained protein language model, PBertKla effectively captures both local and global sequence representations through BERT and Transformer modules. Our model demonstrated outstanding predictive performance, achieving AUC and AUPRC values of 0.884 and 0.866, respectively, on the independent validation data. Comparative analyses further confirmed its superior accuracy, robustness, and generalization ability over existing Kla site predictors. The source code of PBertKla is publicly available to offer a powerful approach for advancing lactylation research and its potential implications in disease mechanisms and therapeutic strategies.

The development of PBertKla provides a useful computational tool to complement experimental strategies for lactylation site detection, facilitating further exploration of lactylation modifications in health and disease. Despite its strong performance, there remain areas for future improvement. Expanding the dataset with newly identified lactylation sites, incorporating structural and evolutionary information, and exploring multi-scale deep learning architectures could further enhance predictive power. Additionally, integrating Kla prediction tools into a user-friendly online platform would improve accessibility for researchers, enabling broader applications across multiple species and fostering deeper insights into the biological significance of lactylation.

Methods

Benchmark dataset

The primary raw data of this work was collected from the literature [37], including 9275 Kla sites 2437 proteins identified by LC–MS/MS experiments on 110 tumor and adjacent liver samples from 52 hepatitis B virus-related hepatocellular carcinoma (HCC) patients. The sequences of all human proteins were downloaded from the Uniprot (Universal Protein Resource) database [38]. According to the location information of Kla sites and the corresponding annotated protein sequences, we obtained the positive samples, namely Kla peptides with length of $2n + 1$ and lactylated lysine residue located at the center. The Kla sites with less than n upstream or downstream residues were not taken into consideration in this work. Similarly, the negative samples, with center of normal lysine residue, were extracted from all proteins without Kla sites. Then, the CD-HIT, a widely used bioinformatics program for clustering and comparing large number of protein or nucleic acid sequences at different sequence identity levels [39, 40], was utilized to remove redundant sequences and avoid model overfitting. As most proteins do not undergo lactylation modification, a considerable number of negative samples were obtained. To balance the positive and negative samples, the equal amount of nonredundant negative samples was randomly selected out to be non-Kla dataset.

To build a reliable and practical benchmark dataset of human Kla site, we conducted a preliminary analysis for determining the optimal sample length and the optimal sequence identity threshold of CD-HIT. As a result, the dataset with sample length of 45 and sequence identity of 30% was selected as the final benchmark dataset, consisting of 4820 Kla and 4820 non-Kla sequence samples. Notably, this study divided the human Kla benchmark dataset into a training subset for learning the prediction models, comprising 3856 positives and 3856 negatives, and a validation subset for independently testing the prediction performance, containing 964 positives and 964 negatives (Table 1). This strategy guarantees a consistent and comparative evaluation of the models' effectiveness. For further validating the human Kla identification ability of models, a smaller LC–MS/MS experiment dataset, including 2488 Kla sites from oral squamous cell carcinoma (OSCC) tissues and normal adjacent tissues (NATs) [41], was preprocessed to be independent testing data.

The protein large language model

In this work, we mainly made full use of ProteinBERT, a deep language model specifically designed for proteins [26, 42], to learn on Kla and non-Kla sequence data for automatically and accurately predicting Kla sites (Fig. 6). In this section, we would briefly describe the proposal of PBertKla, including the pretraining process on a huge number of protein sequences and relevant annotation knowledge, the critical fine-tuning process on human Kla benchmark dataset, as well as the architecture and methodology.

Encoding of protein/peptide sequence and annotation

During pretraining and fine-tuning, protein and peptide sequences were encoded using 26 distinct integer tokens to represent 20 standard amino acids, selenocysteine (U), an undefined amino acid (X), another amino acid (OTHER), and three special tokens (START, END, PAD). Thereinto, START and END tokens marked the beginning and end of each sequence, while PAD token was used to extend shorter sequences to match the minibatch length. START and END tokens also aided in modeling proteins longer than the chosen length by selecting a random subsequence, excluding at least one end. The absence of the START or END token indicated that only part of the sequence was provided. Meanwhile, GO annotations were encoded as a binary vector of 8943 elements, with zeros everywhere except for entries corresponding to the protein's GO terms. In scenarios of lacking GO annotation data, such as during fine-tuning or evaluation on human Kla benchmarks, it was set as a zero vector.

Table 1 The benchmark dataset of human lysine lactylation site

Dataset	Sample type	Sample number
Training data (80%)	Positive	3856
	Negative	3856
Independent data (20%)	Positive	964
	Negative	964

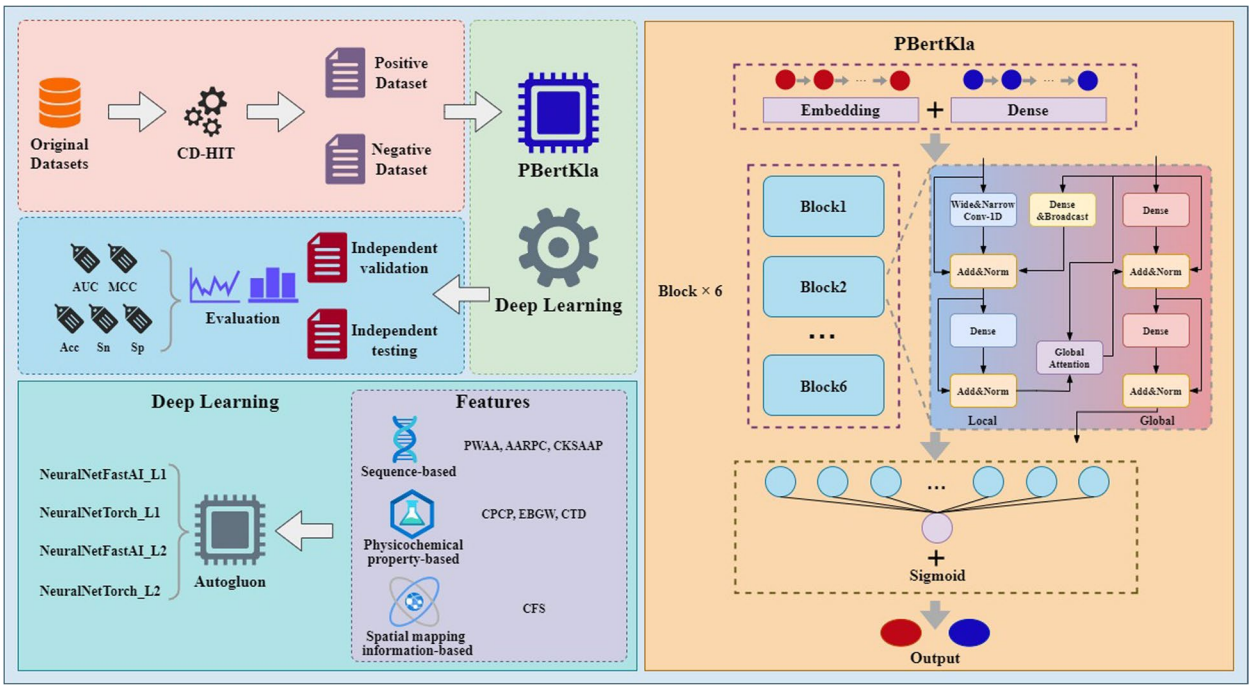

Fig. 6 Schematic diagram of the proposed PBertKla method for identifying human Kla sites

The learning of PBertKla on human Kla benchmark dataset

To learn comprehensive protein representations, ProteinBERT [26], a DL framework modeled after the Bidirectional Encoder Representations from Transformers (BERT) architecture [43], was pretrained on approximately 106 M UniRef90 protein amino acid sequences along with their associated GO annotations. We directly obtained the pretrained protein language model framework, implemented in TensorFlow's Keras [44]. The ProteinBERT framework was subsequently fine-tuned on our curated human Kla benchmark dataset to learn PBertKla.

The fine-tuning protocol of PBertKla involved three stages: in the initial phase, all layers of the pretrained model were kept frozen, with training restricted solely to a newly incorporated fully connected layer for a maximum of 40 epochs. Subsequently, all layers were unfrozen, and the model underwent further training for up to an additional 40 epochs. In the final step, the model was trained for one last epoch using a larger sequence length (setting to 1024). In addition, hyperparameter optimization was conducted on the learning rate, batch size, and sequence length (i.e., the number of tokens encoding the input sequences), the search space of which was enumerated in Additional file 1: Table S2. We trained the model across different parameter settings and assessed its performance on the validation data. The best-performing parameter configuration was selected to finalize the human Kla prediction model, namely PBertKla. It should be noted that throughout the fine-tuning and performance evaluation phases, no GO annotation data was utilized, namely the GO-annotation input remained a fixed all-zero vector. The entire procedure described above was executed on a NVIDIA RTX 3090 GPU.

Framework overview of PBertKla

Referring to the pretrained protein large language model (ProteinBERT) [26], when fine-tuning on the Kla benchmark dataset, the model receives two types of input information: (1) peptide sequences, represented as amino acid token sequences, (2) GO annotations, encoded as fixed-size all-zero vectors (as lack of GO annotation data). Architecturally, it features two quasi-parallel subnetworks: one dedicated to local representations and the other to global representations (Fig. 6). The local representations are three-dimensional tensors with the shape $B \times L \times d_{local}$, where B represents the batch size, L is the sequence length within a mini-batch, and d_{local} is the number of channels for the local representations (setting to 128). The global representations are two-dimensional tensors of shape $B \times d_{global}$ (setting to 512). In the model's initial layers, an embedding layer with d_{local} output features transforms input sequences into a local-representation 3D tensor. This embedding layer is applied uniformly across each position. Meanwhile, a fully connected layer with d_{global} output features converts input annotations into a global-representation 2D tensor.

The model's core architecture comprises six transformer-inspired blocks designed to process the local and global representations generated by the initial layers. Each block refines the local representations using 1D convolutional layers, which integrate both narrow and wide convolutional layers to capture contextual dependencies at varying scales—ranging from local to distant sequence positions. These convolutional operations are followed by a fully connected layer. Simultaneously, the global representations are updated through two fully connected layers within each block. Every hidden layer in the model, including both fully connected and convolutional layers, employs GELU (Gaussian Error Linear Unit) as the activation function.

In each block, the local and global representations interact exclusively through broadcast fully connected layers (from global to local) and global attention layers (from local to global). The broadcast layers are fully connected layers that convert the global representation's d_{global} features into the local representation's d_{local} features, then duplicate this representation across all L sequence positions.

When PBertKla was trained on labeled human Kla datasets, a new fully connected layer was added to its output, connected to the local representations. This involved concatenating all the local normalization layers (two per block) with the output sequence, which were then passed into the final fully connected layer. This layer included a dropout layer and a sigmoid activation function for binary classification labels. Overall, the innovative integration of local and global representations, combined with dual-input pretraining strategy, this robust framework would achieve efficient and stable learning outcome on Kla peptide sequences. By synergistically integrating local residue-level patterns with global functional embeddings through a dual-modality pretraining strategy, this architecture would achieve robust generalization capabilities, enabling efficient and stable learning on Kla peptide sequence prediction tasks.

Other deep learning models

To evaluate PBertKla, we also trained and tested several other DL models on our human Kla benchmark dataset. We firstly employed ProtBERT [45], another BERT-based large protein language model, to predict human Kla sites. Moreover, the research of Lv et al. [30] indicated that the feature engineering based on amino acid composition information, physicochemical property information, and spatial mapping information was effective to represent PTM sites. Therefore, we encoded human Kla and non-Kla samples using the seven feature descriptors proposed by them, including PWAAC, AARPC, CKSAAP, CPCP, EBGW, CTD, and CFS. These features were then fused to feed into DL models to predict human Kla sites. Here, we utilized the program package provided by Lv et al. [30] to perform the aforementioned feature extraction. The AutoGluon, an open source software with automated machine learning (AutoML) capabilities, was developed by Amazon team [46]. They provided the convenient and effective implementation for a series of ML and DL models, including PyTorch neural networks and fastai v1 neural networks, which were employed in this work to classify human Kla sites.

Additionally, we evaluated three existing Kla prediction tools (including DeepKla, AutoKla, and DeepKlapred) to predict human Kla sites. All of these tools were developed based on deep learning (DL) architectures. The architecture of DeepKla consisted of four closely connected subnetworks, including a word embedding layer, convolutional neural network (CNN), bidirectional gated recurrent units (BiGRU), and attention mechanism layer [22]. DeepKlapred integrated six biochemical/evolutionary sequence descriptors into a BiGRU-Transformer framework, with a cross-attention fusion mechanism dynamically bridging sequence embeddings and descriptor features to capture their synergistic interactions [23]. The Auto-Kla model was mainly composed of adaptive embedding module, Transformer encoder module, and multi-layer perceptron classifier module [24]. Independent validation datasets were accordingly tested on them for performance comparison among all models.

Performance evaluation

The performance of each Kla prediction model is evaluated using following common metrics, including accuracy (Acc), sensitivity (Sn), specificity (Sp), and Matthews correlation coefficient (MCC) [47–51]. The formulas for computing these metrics as follows:

$$\begin{cases} Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \\ Sensitivity = \frac{TP}{TP+FN} \\ Specificity = \frac{TN}{TN+FP} \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \quad (1)$$

where TP (true positive) and FP (false positive) are the numbers of sequences correctly and incorrectly predicted as Kla samples, respectively. TN (true negative) and FN (false negative) are the numbers of sequences correctly and incorrectly classified as non-Kla samples. ROC (receiver operating characteristic) curves and PR (precision-recall) curves are also employed to visually represent the overall predictive performance, and the areas under ROC curve (AUC) and PR curve (AUPRC) are further calculated to quantify performance. All indicator values range from [0, 1], higher values indicating better prediction performance.

Abbreviations

PTM	Post-translational modification
Kla	Lysine lactylation
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
LC–MS/MS	Liquid chromatography–mass spectrometry
DL	Deep learning
GO	Gene Ontology
UMAP	Uniform Manifold Approximation and Projection
NN	Neural network
ML	Machine learning
OSCC	Oral squamous cell carcinoma
Uniprot	Universal Protein Resource
PWAAC	Position weight amino acid composition
AARPC	Amino acid relative position composition
CKSAAP	Composition of k-space amino acid pairs
CPCP	Composition of physical and chemical properties
EBGW	Encoding based on grouped weight
CTD	Composition, transition and distribution
CFS	Cone feature space
Acc	Accuracy
Sn	Sensitivity
Sp	Specificity
MCC	Matthews correlation coefficient
AUC	Area under receiver operating characteristic (ROC) curve
AUPRC	Area under precision-recall (PRC) curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-025-02202-1.

Additional file 1: PBertKla: a protein large language model for predicting human lysine lactylation sites, Tables S1–S4. Table S1 AUC values of datasets across different sample lengths and sequence similarities. Table S2 The search space for the three parameters of the protein large language model. Table S3 Performance of feature descriptors based on neural networks. Table S4 Performance comparison of PBertKla with other deep learning models as well as existing tools.

Acknowledgements
We are grateful to the anonymous reviewers for their insightful and constructive comments, which will be instrumental in improving the presentation of this paper.

Authors' contributions
H.L. and H.D. designed this study, and prepared the initial and final drafts of the manuscript. D.L. collected and analyzed the data, performed the research. M.Y., T.Z., X.L., Y.W., S.X., and F.H. prepared the methodology. H.Y., K.S., and F.D. critiqued drafts, and approved the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding
The work was supported by the National Natural Science Foundation of China (62271120), Natural Science Foundation of Chongqing (CSTB2024NSCQ-MSX1241), Science and Technology Department of Sichuan Province (2024ZYD0039), Health Commission of Sichuan Province (24CXTD11), Sichuan Medical Association (S23012), Chengdu Science and Technology Bureau (2022-YF05-01867-SN), and Health Commission of Chengdu (2024141).

All data generated or analysed during this study are included in this published article, its supplementary information files and publicly available repositories. The human lysine lactylation sites were collected from the National Omics Data Encyclopedia (NODE, https://www.biosino.org/node) database under accession number OEP002852 (Yang Z, Yan C, Ma J, Peng P, Ren X, Cai S, et al. Lactylome analysis suggests lactylation-dependent mechanisms of metabolic adaptation in hepatocellular carcinoma) [37] and the Proteomics Identification Database (PRIDE, https://www.ebi.ac.uk/pride/) via PXD047535 (Jing F, Zhu L, Zhang J, Zhou X, Bai J, Li X, et al. Multi-omics reveals lactylation-driven regulatory mechanisms promoting tumor progression in oral squamous cell carcinoma) [41]. The core codes for implementing ProteinBERT [26], DeepKla [22], DeepKlapred [23], and AutoKla [24] models referred to https://github.com/nadavbra/protein_bert, https://github.com/linDing-group/DeepKla, https://github.com/GGCL7/DeepKlapred, and https://github.com/tubic/Auto-Kla. The data and source code supporting PBertKla are available in the following repositories: https://github.com/laihongyan/PBertKla and https://doi.org/10.5281/zenodo.15107500.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 16 January 2025 Accepted: 31 March 2025
Published online: 07 April 2025

References

1. Lee JM, Hammarén HM, Savitski MM, Baek SH. Control of protein stability by post-translational modifications. Nat Commun. 2023;14(1):201. https://doi.org/10.1038/s41467-023-35795-8.
2. Dao F, Xie X, Zhang H, Guan Z, Wu C, Su W, et al. PlantEMS: a comprehensive database of epigenetic modification sites across multiple plant species. Plant Commun. 2024:101228.
3. Lin H. Artificial intelligence with great potential in medical informatics: a brief review. Medinformatics. 2024;1(1):2–9.
4. Chen Y, Gao Q, Wang D, Zou X, Li X, Ji J, et al. An overview of research advances in oncology regarding the transcription factor ATF4. Curr Drug Targets. 2025;26(1):59–72.
5. Lin Y, Lin P, Lu Y, Zheng J, Zheng Y, Huang X, et al. Post-translational modifications of RNA-modifying proteins in cellular dynamics and disease progression. Adv Sci (Weinh). 2024;11(44):e2406318.
6. Huang Z, Guo X, Qin J, Gao L, Ju F, Zhao C, et al. Accurate RNA velocity estimation based on multibatch network reveals complex lineage in batch scRNA-seq data. BMC Biol. 2024;22(1):290.
7. Zhu H, Hao H, Yu L. Identification of microbe-disease signed associations via multi-scale variational graph autoencoder based on signed message propagation. BMC Biol. 2024;22(1):172.
8. Zhang D, Tang Z, Huang H, Zhou G, Cui C, Weng Y, et al. Metabolic regulation of gene expression by histone lactylation. Nature. 2019;574(7779):575–80.
9. Dong M, Zhang Y, Chen M, Tan Y, Min J, He X, et al. ASF1A-dependent P300-mediated histone H3 lysine 18 lactylation promotes atherosclerosis by regulating EndMT. Acta Pharm Sin B. 2024;14(7):3027–48.
10. Moreno-Yruela C, Zhang D, Wei W, Bæk M, Liu W, Gao J, et al. Class I histone deacetylases (HDAC1–3) are histone lysine delactylases. Sci Adv. 2022;8(3):eabi6696.
11. Wu X, Xu M, Geng M, Chen S, Little PJ, Xu S, et al. Targeting protein modifications in metabolic diseases: molecular mechanisms and targeted therapies. Signal Transduct Target Ther. 2023;8(1):220.
12. Wang X, Liu Q, Yu H-T, Xie J-Z, Zhao J-N, Fang Z-T, et al. A positive feedback inhibition of isocitrate dehydrogenase 3β on paired-box gene 6 promotes Alzheimer-like pathology. Signal Transduct Target Ther. 2024;9(1):105.
13. Ouyang J, Wang H, Huang J. The role of lactate in cardiovascular diseases. Cell Commun Signal. 2023;21(1):317.
14. Yang Y, Luo N, Gong Z, Zhou W, Ku Y, Chen Y. Lactate and lysine lactylation of histone regulate transcription in cancer. Heliyon. 2024;10(21):e38426.
15. Li H, Sun L, Gao P, Hu H. Lactylation in cancer: current understanding and challenges. Cancer Cell. 2024;42(11):1803–7.

16. Manayalan B, Basith S, Shin TH, Wei L, Lee G. mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics. 2019;35(16):2757–65.
17. Wei L, He W, Malik A, Su R, Cui L, Manavalan B. Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. Brief Bioinform. 2021;22(4):bbaa275.
18. Zhou J, Xu W, Wu Y, Wang M, Zhang N, Wang L, et al. GPR37 promotes colorectal cancer liver metastases by enhancing the glycolysis and histone lactylation via Hippo pathway. Oncogene. 2023;42(45):3319–30.
19. Longhitano L, Giallongo S, Orlando L, Broggi G, Longo A, Russo A, et al. Lactate rewrites the metabolic reprogramming of uveal melanoma cells and induces quiescence phenotype. Int J Mol Sci. 2022;24(1):24.
20. Liu M, Li C, Chen R, Cao D, Zeng X. Geometric deep learning for drug discovery. Expert Syst Appl. 2024;240:122498.
21. Jiang P, Ning W, Shi Y, Liu C, Mo S, Zhou H, et al. FSL-Kla: a few-shot learning-based multi-feature hybrid system for lactylation site prediction. Comput Struct Biotechnol J. 2021;19:4497–509.
22. Lv H, Dao F-Y, Lin H. DeepKla: an attention mechanism-based deep neural network for protein lysine lactylation site prediction. Imeta. 2022;1(1):e11.
23. Guan J, Xie P, Dong D, Liu Q, Zhao Z, Guo Y, et al. DeepKlapred: a deep learning framework for identifying protein lysine lactylation sites via multi-view feature fusion. Int J Biol Macromol. 2024;283(Pt 3):137668.
24. Lai F-L, Gao F. Auto-Kla: a novel web server to discriminate lysine lactylation sites using automated machine learning. Brief Bioinform. 2023;24(2):bbad070.
25. Yang Y-H, Yang J-T, Liu J-F. Lactylation prediction models based on protein sequence and structural feature fusion. Brief Bioinform. 2024;25(2):bbad539.
26. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022;38(8):2102–10.
27. Vacic V, Iakoucheva LM, Radivojac P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics. 2006;22(12):1536–7.
28. Liu Y, Zhou Z, Cao X, Cao D, Zeng X. Effective drug-target affinity prediction via generative active learning. Inf Sci. 2024;679:121135.
29. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint. 2020;arXiv:1802.03426.
30. Lv H, Zhang Y, Wang J-S, Yuan S-S, Sun Z-J, Dao F-Y, et al. iRice-MS: an integrated XGBoost model for detecting multitype post-translational modification sites in rice. Brief Bioinform. 2022;23(1):bbab486.
31. Li F, Si W, Xia L, Yin D, Wei T, Tao M, et al. Positive feedback regulation between glycolysis and histone lactylation drives oncogenesis in pancreatic ductal adenocarcinoma. Mol Cancer. 2024;23(1):90.
32. Zhang N, Zhang Y, Xu J, Wang P, Wu B, Lu S, et al. α-Myosin heavy chain lactylation maintains sarcomeric structure and function and alleviates the development of heart failure. Cell Res. 2023;33(9):679–98.
33. Chen Y, Wu J, Zhai L, Zhang T, Yin H, Gao H, et al. Metabolic regulation of homologous recombination repair by MRE11 lactylation. Cell. 2024;187(2):294–311.
34. Abdelkader GA, Kim JD. Advances in protein-ligand binding affinity prediction via deep learning: a comprehensive study of datasets, data preprocessing techniques, and model architectures. Curr Drug Targets. 2024;25(15):1041–65.
35. Jiang Y, Wang R, Feng J, Jin J, Liang S, Li Z, et al. Explainable deep hypergraph learning modeling the peptide secondary structure prediction. Adv Sci (Weinh). 2023;10(11):e2206151.
36. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, et al. Simulating 500 million years of evolution with a language model. Science. 2025;387(6736):850–8.
37. Yang Z, Yan C, Ma J, Peng P, Ren X, Cai S, et al. Lactylome analysis suggests lactylation-dependent mechanisms of metabolic adaptation in hepatocellular carcinoma. Nat Metab. 2023;5(1):61–79.
38. The UC. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523–31.
39. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
40. Meng Q, Guo F, Tang J. Improved structure-related prediction for insufficient homologous proteins using MSA enhancement and pre-trained language model. Brief Bioinform. 2023;24(4):bbad217.
41. Jing F, Zhu L, Zhang J, Zhou X, Bai J, Li X, et al. Multi-omics reveals lactylation-driven regulatory mechanisms promoting tumor progression in oral squamous cell carcinoma. Genome Biol. 2024;25(1):272.
42. Zulfiqar H, Guo Z, Ahmad RM, Ahmed Z, Cai P, Chen X, et al. Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings. Front Med (Lausanne). 2023;10:1291352.
43. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint. 2019;arXiv:1810.04805.
44. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. Proceedings of the 12th USENIX conference on operating systems design and implementation; Savannah, GA, USA: USENIX Association; 2016. p. 265–83.
45. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2022;44(10):7112–27.
46. Erickson N, Mueller J, Shirkov A, Zhang H, Larroy P, Li M, et al. Autogluon-tabular: robust and accurate automl for structured data. arXiv preprint. 2020;arXiv:2003.06505.
47. Zou X, Ren L, Cai P, Zhang Y, Ding H, Deng K, et al. Accurately identifying hemagglutinin using sequence information and machine learning methods. Front Med (Lausanne). 2023;10:1281880.
48. Wang Y, Zhai Y, Ding Y, Zou Q. SBSM-Pro: support bio-sequence machine for proteins. SCIENCE CHINA Inf Sci. 2024;67(11):212106.
49. Ai C, Yang H, Liu X, Dong R, Ding Y, Guo F. MTMol-GPT: de novo multi-target molecular generation with transformer-based generative adversarial imitation learning. Plos Comput Biol. 2024;20(6):e1012229.
50. Liu B, Gao X, Zhang H. BioSeq-Analysis 2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res. 2019;47(20):e127.
51. Yan K, Lv H, Guo Y, Peng W, Liu B. sAMPpred-GAT: prediction of antimicrobial peptide by graph attention network and predicted peptide structure. Bioinformatics. 2023;39(1):btac715.

Publisher's Note