Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prioritization of Androgenetic Alopecia-Associated Lactylation Sites through Multi-Omics Integration and the HybridKla Deep Learning Framework

Yuan Zhu
Tongji University https://orcid.org/0009-0009-3618-2794
Pei Li
Tongji University
Chika Hasegawa
Tongji University
Seungyeon Lee
Tongji University
Chenlong He
Tongji University
Ming Yin
Tongji University
Shengming Wu
Shanghai East Hospital
Zhe Tao
Tongji University
Yilong Wang
yilongwang@tongji.edu.cn

Tongji University
Yufei Li
yufeili@tongji.edu.cn

Tongji University

Research Article

Keywords: androgenetic alopecia, lysine lactylation, ZBTB7B, multi-omics integration, deep learning, HybridKla.

Posted Date: October 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11394744/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

The contribution of lysine lactylation to androgenetic alopecia (AGA) remains poorly characterized. Here, we integrated scalp single-cell transcriptomics, proteomics, lactylomics, and HybridKla-based deep learning to prioritize AGA-associated lactylation candidates. Single-cell data from three frontal–occipital donor pairs provided a T-cell-related transcriptional context, while proteomic and lactylomic profiling of two donor pairs identified 1,804 lactylation sites on 551 proteins. ZBTB7B, a transcription factor involved in T-cell lineage regulation, showed opposing regional trends in total-protein and modified-peptide abundance: frontal-to-occipital ratios were 1.806 for total ZBTB7B, 0.523 for K206-associated lactylated peptides, and 0.587 for K212-associated lactylated peptides. These directions were consistent across both donors, although neither site passed multiple-testing correction. To support candidate prioritization, we adapted HybridKla using 46,403 reference site records with protein-grouped five-fold partitioning for base-model evaluation. Five-score fusion yielded reported receiver operating characteristic and precision–recall areas under the curve of 0.8281 and 0.7984, respectively. Classification-head adaptation increased mean recall from 78.09% to 90.75% on 346 confirmed positive sites from 176 held-out proteins; this positive-only evaluation assessed detection sensitivity rather than overall classification performance. Both K206 and K212 were included among predicted-positive ZBTB7B sites and supported by mass-spectrometric localization records. Together, these findings prioritize ZBTB7B K206 and K212 as candidate sites for investigating lactylation-associated regulation in AGA. The integrated workflow connects tissue-level observations with residue-specific evidence, providing a focused basis for examining post-translational regulation within the follicular immune microenvironment.

Introduction

Androgenetic alopecia (AGA) is characterized by progressive hair follicle miniaturization, and its molecular basis needs to be understood within the context of the follicle and its surrounding tissue microenvironment. Recent single-cell and spatial transcriptomic studies have revealed altered follicular cell states, progenitor-cell fate, and intercellular communication in AGA, highlighting the involvement of multiple interacting cell populations in impaired hair growth. (Li, Yang et al. 2026) However, transcriptional information does not fully capture protein abundance or post-translational modification states. Integrating cellular transcriptional features with tissue-level protein modifications therefore provides an opportunity to identify candidate regulatory molecules associated with AGA.

Lysine lactylation (Kla) offers an entry point for investigating the relationship between metabolic states and protein regulation. The discovery of histone lactylation established a connection between lactate-associated metabolism, chromatin modification, and gene expression. (Zhang, Tang et al. 2019)However, this finding does not directly establish the functions of individual lactylation sites on non-histone proteins. Investigating lactylation in AGA requires identifying proteins and sites supported by traceable modification evidence and determining whether changes in modified-peptide abundance accompany changes in total-protein abundance. Joint proteomic and lactylomic analysis can help

distinguish these two levels of variation and guide candidate prioritization, although modified-peptide abundance cannot itself be interpreted as site occupancy.

ZBTB7B, also known as ThPOK, is a transcription factor involved in T-cell lineage determination, with an experimentally established role in CD4 versus CD8 T-cell lineage commitment. (He, He et al. 2005) This biological background makes ZBTB7B a candidate for connecting protein regulation with T-cell-associated transcriptional features. In AGA, integrating its cellular expression context, tissue-level protein abundance, and lactylation-site records could generate hypotheses concerning its potential regulatory roles. Nevertheless, modifications detected in whole scalp tissue cannot be directly assigned to T cells or interpreted as evidence of altered T-cell function or hair follicle growth. Establishing site-level evidence is therefore a necessary initial step toward developing testable functional hypotheses.

Protein language models provide a methodological basis for computational prioritization of candidate modification sites. Integrating multiple protein language models has enabled single-sequence structure prediction, demonstrating the utility of sequence representations for structural inference. (Jing, Wu et al. 2024) Incorporating biophysical dynamics into model training has also improved mutation-effect prediction in selected settings, suggesting the complementary potential of different information sources. (Hou, Zhao et al. 2026) For post-translational modifications, protein language model representations have been combined with atom-level structural features to predict several classes of lysine modification sites. (Luo, Zhu et al. 2026) Specifically for lactylation, the publicly available HybridKla framework combines ESM2, a long short-term memory network, and physicochemical and compositional descriptors to support multi-feature prediction. (Ning, Qin et al. 2025) However, disease-specific applications require careful consideration of data partitioning, evaluation independence, and experimental evidence. In particular, the ability to recognize predictive sequence patterns must be distinguished from evidence of a candidate site's biological role in disease.

Here, we integrated single-cell transcriptomic, proteomic, and lactylomic information from paired frontal and occipital scalp samples to prioritize candidate lactylation sites associated with AGA. We first examined tissue-associated functional patterns and transcriptional changes in a candidate T-cell population and compared regional changes in ZBTB7B total-protein and lactylated-peptide abundance. We then adapted and evaluated the HybridKla framework, comparing ESM2 representations and model-fusion strategies and assessing positive-site recall on previously unseen proteins following laboratory-data adaptation. (Ning, Qin et al. 2025) Finally, we combined predictions with modified-peptide records, site-localization probabilities, and predicted structural context to prioritize ZBTB7B K206 and K212 for subsequent investigation. This study aims to establish a candidate-site prioritization workflow linking tissue-level multi-omic observations with sequence-based prediction and to provide molecular leads for investigating lactylation in AGA, rather than to establish a causal disease mechanism.

Materials and Methods

Study design and datasets

This exploratory study used within-donor comparisons between frontal (FH) and occipital (OC) scalp. The single-cell transcriptomic dataset comprised six samples from three donors. The proteomic and lactylomic datasets comprised four samples from two donors, designated FH_A, OC_A, FH_B, and OC_B. Occipital scalp served as the within-donor regional comparator. Each omics dataset was analyzed separately, followed by integration with sequence predictions, modified-peptide records, and localization information to prioritize candidate lactylation sites on ZBTB7B.

Single-cell data organization and target-cell extraction

Analyses used raw UMI count matrices, quality-controlled cell barcodes, cluster metadata, and UMAP coordinates. Existing clusters and dimensionality-reduction results were retained. C21 cells, provisionally annotated as T cells, were extracted from cluster metadata and matched by barcode to the sample-specific count matrices in Matrix Market format. Target-cell numbers and aggregated UMI counts were checked for each sample. The resulting population comprised 650 cells and was used for donor-paired analyses. C0 and C1 from the existing T-cell subclustering results were used for functional enrichment displays without assigning additional, independently unverified T-cell subtype identities.

Donor-paired pseudobulk analysis

Raw gene-level UMI counts were summed across C21 cells within each donor–region combination to generate a six-column pseudobulk matrix. Counts per million (CPM) were calculated by dividing each gene count by the total target-cell UMI count in the corresponding sample and multiplying by 10^6, followed by log2(CPM + 1) transformation. Genes with a total raw count ≥ 10 and CPM ≥ 1 in at least three samples were retained, yielding 10,217 genes for statistical testing.

Within-donor differences were calculated as FH minus OC on the log2(CPM + 1) scale. The arithmetic mean of the three paired differences was used as the effect estimate. Regional differences were evaluated using two-sided paired t-tests, with Benjamini–Hochberg (BH) correction across all tested genes. Directional consistency across donors and the range of mean differences after leaving out one donor at a time were recorded as descriptive assessments of stability. The figure label “mean paired log2FC” denotes the mean paired difference in log2(CPM + 1). All six samples were retained, including the MRJ occipital sample containing five target cells.

Expression-module scoring and functional enrichment

Three exploratory expression modules were defined. The helper-associated module comprised ZBTB7B, BCL11B, CBFB, GATA3, AHR, BATF, IL13, IL26, and TNFRSF18; the quiescence/memory module comprised KLF2, NR4A1, CCR7, SELL, and IL7R; and the cytotoxic module comprised CD8A, CCL5, CCL4, NKG7, GNLY, GZMB, GZMK, and PRF1. For detected genes within each module, log2(CPM + 1) values were standardized gene-wise across the six samples. Sample-level scores were calculated as the arithmetic mean of the module gene z scores. FH and OC scores were compared using two-sided paired

t-tests, with nominal P values reported. Selected-gene heatmaps used row-standardized expression values.

Tissue-proteomic GO analysis used a hypergeometric test with identified proteins carrying the relevant annotations as the background. Functional displays for C0 and C1 used existing GO results for regional differential-expression analyses, retaining the GeneRatio, Count, and p.adjust fields. Terms were selected for relevance to the study question. Separate color scales were used for the cell groups, and cell-level enrichment results were interpreted separately from donor-paired pseudobulk tests.

Protein extraction, digestion, and lactylated-peptide enrichment

Tissue was ground in liquid nitrogen and lysed in buffer containing 8 M urea, 1 mM PMSF, and 2 mM EDTA. Samples were sonicated on ice for 5 min and centrifuged at 15,000 × g for 10 min at 4 °C. Protein concentrations were measured using a BCA assay. For total proteomics, 100 µg protein was reduced with 5 mM DTT for 45 min at 37 °C and alkylated with 11 mM iodoacetamide for 15 min at room temperature in the dark. Samples were diluted with 25 mM ammonium bicarbonate and digested overnight with trypsin. Peptides were acidified and desalted using C18 material before mass spectrometry.

For lactylomics, a urea-based lysis system supplemented with deacylase inhibitors was used. Protein aliquots of 4 mg were reduced with 10 mM DTT, alkylated with 50 mM iodoacetamide, precipitated with precooled acetone, and digested overnight with trypsin and Lys-C at 37 °C. Peptides were dissolved in NETN buffer and incubated overnight at 4 °C with anti-lysine-lactylation antibody-conjugated beads. Following washing with NETN buffer and water, peptides were eluted three times with 0.2% trifluoroacetic acid. Eluates were pooled and desalted using C18 material.

Liquid chromatography–mass spectrometry and database searches

Samples were analyzed using a Vanquish Neo UHPLC coupled to an Orbitrap Astral mass spectrometer in positive-ion data-independent acquisition (DIA) mode. Total-proteome analysis used a trap-and-analytical-column configuration with a 150 µm × 15 cm, 2 µm C18 analytical column at 55 °C. The injection amount was 200 ng, the flow rate was 2.5 µL/min, the effective gradient was 6.9 min, and the total run time was 8 min. MS1 spectra were acquired over m/z 380–980 at a resolution of 240,000 at m/z 200. DIA acquisition used 299 windows with a 2 Th isolation width and an HCD collision energy of 25%.

Lactylome analysis used a 150 µm × 15 cm, 2 µm C18 column at 55 °C, a flow rate of 800 nL/min, and a 30-min gradient. Mobile phase A was water containing 0.1% formic acid, and mobile phase B was 80% acetonitrile containing 0.1% formic acid. The MS1 range and resolution were as above; MS2 acquisition used a 3 Th isolation width and an HCD collision energy of 25%.

Total-proteome data were analyzed using DIA-NN 1.8.1 in library-free mode with match-between-runs enabled. Lactylome data were analyzed using Spectronaut 19. The search database was the human

UniProt FASTA dated February 24, 2025, containing 83,385 sequences. Lactylome searches specified strict trypsin specificity and up to three missed cleavages. Cysteine carbamidomethylation was a fixed modification, whereas lysine lactylation and methionine oxidation were variable modifications, with up to three variable modifications permitted. Identification filtering used FDR < 1% and a site-localization probability ≥ 0.75.

Protein and modification-site quantification

Total-protein and lactylation-site abundances were median-normalized within each sample, separately for the two datasets. FH/OC fold change (FC) was defined as the ratio of arithmetic mean relative abundances. P values and BH-adjusted values in the quantification tables were retained for statistical displays. Exploratory screening used nominal $P \leq 0.05$ together with $FC \geq 1.5$ or ≤ 0.6667. Lines connecting samples indicated donor pairing and did not redefine the statistical tests underlying the existing P values.

Site-level FC values were matched to the corresponding total-protein FC values by protein accession and residue position. Total-protein log2FC was plotted against site-level log2FC to describe their relationship. This comparison characterized abundance-change directions and did not estimate absolute modification occupancy. Sample distributions were displayed using the existing lactylome PCA coordinates and explained-variance values. Sequence-context displays used MEME/MoMo-related outputs spanning six residues on either side of the modified lysine and retained the original positional preference scores.

Reference sequences and protein-grouped cross-validation

The lactylation prediction workflow was adapted from HybridKla. The initial reference dataset contained 47,004 records, with 23,502 positives and 23,502 reference negatives. Protein identifiers were recovered by matching peptides to UniProt sequences. Excluding 601 unresolved records yielded 46,403 records, comprising 23,502 positives and 22,901 reference negatives across 6,789 protein groups. Each site was represented by a 51-amino-acid window containing the central lysine and 25 residues on either side, with padding at sequence boundaries. Reference-negative labels were retained from the source data and were not equated with experimentally confirmed absence of modification.

Protein-grouped five-fold partitioning was performed using StratifiedGroupKFold with random seed 42. All base models shared the same partitions. Within each outer training set, 10% of proteins were reserved for internal validation and model selection. Sites from the same protein group were not included in both training and testing within a fold.

ESM2 representation comparison and base models

The pretrained esm2_t30_150M_UR50D model was used, comprising 30 Transformer layers with 640-dimensional residue embeddings. Following addition of CLS and EOS tokens, the central lysine embedding was extracted at zero-based token index 26. Mean pooling included only valid residues, excluding special tokens and padded positions. Five representations were compared: CLS alone (E0),

central lysine (E1), masked mean pooling (E2), CLS concatenated with central lysine (E3), and central lysine concatenated with masked mean pooling (E4). Single and concatenated representations were 640- and 1,280-dimensional, respectively, and were derived from a single model forward pass.

The classification head comprised a linear projection to 512 units, LayerNorm, ReLU, dropout of 0.3, and a two-class output layer followed by softmax. Both the ESM2 backbone and classification head were fine-tuned during representation comparison. Training used a batch size of 32, an initial learning rate of 5×10^{-5}, weight decay of 0.05, cross-entropy loss, cosine learning-rate scheduling, and 150 warmup steps, for a maximum of 40 epochs. Checkpoints were selected by validation ROC-AUC, with an early-stopping patience of six epochs and a minimum improvement of 0.0001.

Additional base models comprised LSTM and classifiers based on AAINDEX, CKSAAP, OBC, ACF, and PSEAAC, representing sequence order, physicochemical properties, spaced residue-pair composition, position-specific encoding, autocorrelation, and composition/sequence-order information, respectively. Base-model out-of-fold predictions were generated using the shared protein-grouped partitions.

Multi-model score fusion and performance assessment

The 7D MetaModel received positive-class scores from ESM2-E4, LSTM, and the five handcrafted-feature models. The 5D MetaModel received scores from ESM2-E4, LSTM, AAINDEX, CKSAAP, and OBC. Existing fusion-evaluation outputs were used to compare ROC-AUC, PR-AUC, and threshold-dependent metrics at 0.5: Matthews correlation coefficient (MCC), sensitivity, specificity, F1 score, and balanced accuracy, defined as the mean of sensitivity and specificity. ESM2 representation comparisons reported means and standard deviations across five folds, whereas fusion-model comparisons used pooled-prediction metrics. Between-model differences were described numerically without significance testing.

Laboratory-data adaptation and held-out testing

After deduplication by UniProt identifier and residue position, the adaptation dataset contained 3,528 records from 444 proteins, comprising 1,196 confirmed positives and 2,332 unlabeled lysines. Sites from proteins represented in the reference dataset were used for adaptation training and validation. The training set contained 680 positives and 1,328 unlabeled sites from 214 proteins. The validation set contained 170 positives and 331 unlabeled sites from 54 proteins, with no protein overlap between these sets. The remaining 176 proteins, absent from the reference dataset, were held out in their entirety. Their 346 confirmed positive sites were used for fixed test evaluation; their 673 unlabeled lysines were excluded from adaptation, validation, and this positive-only test.

Starting from each of the five E4 fold models, the ESM2 backbone was frozen and only the classification head was fine-tuned. Adaptation used AdamW with a learning rate of 5×10^{-6}, weight decay of 0.01, batch size of 16, gradient clipping at 1.0, a maximum of eight epochs, and an early-stopping patience of three. Checkpoints were selected using validation weighted cross-entropy. Confirmed positives received a loss weight of 2.0, whereas unlabeled sites were assigned class 0 with a weight of 0.2 as low-weight

background. Held-out proteins were not used for training or model selection. Recall was calculated separately for each of the five models on the same 346 positives at a threshold of 0.5, and the arithmetic mean was used for the before-and-after comparison.

Five-model score-space clustering

Scores from ESM2-E4, LSTM, AAINDEX, CKSAAP, and OBC formed a five-dimensional vector for each record, producing a 46,403 × 5 matrix. Each dimension was mean-centered and standardized by its standard deviation. K-means clustering was performed in the full five-dimensional space, and the K = 2 solution was displayed. PCA projected standardized scores into three dimensions for visualization only. The three scatterplots used the same 15,000 displayed records and coordinates, colored by cluster assignment, reference class, or MetaModel score. Cluster sizes, positive proportions, and mean MetaModel scores were summarized across all 46,403 records. This analysis described prediction-score patterns rather than independent biological subtypes.

ZBTB7B site mapping and structural visualization

The 539-amino-acid canonical UniProt O15156 sequence was used for consistent residue numbering and verification of candidate lysines and modified-peptide positions. A binary site map was generated from the candidate set reported by the original 5D model at a threshold of 0.5. Peptides and localization probabilities for K206 and K212 were obtained from the PTM result table. Four records were labeled Run 1–Run 4 in export order.

The UniProt-linked AlphaFold model AF-O15156-F1-model_v6 was retrieved on September 27, 2026. After verification of chain A and the target residues, Cα coordinates were used to display the full-length trace, highlight K206 and K212, and extract their pLDDT values. Structural visualization described residue positions and local prediction confidence; no lactyl-group modeling or modified-versus-unmodified structural simulation was performed.

Statistical analysis

Donors were the biological replicates for human omics analyses, whereas site windows were the record units for machine-learning analyses. Gene-level paired comparisons used two-sided tests, with BH-FDR < 0.05 as the multiple-testing significance threshold. Module scores were reported with nominal P values, and proteomic and lactylomic displays included existing nominal P values and BH-adjusted values. Identification FDR, site-localization probability, differential-analysis FDR, and model prediction scores were interpreted separately. Score-space clustering and model-performance comparisons were summarized descriptively.

Results

Tissue-level functional profiling and paired transcriptomic analysis establish a T-cell-related context for candidate prioritization

We analyzed six paired frontal and occipital scalp single-cell transcriptomic samples from three donors and four paired scalp proteomic and lactylomic samples from two donors. Tissue-level functional profiles and cell-associated transcriptional patterns were integrated to guide candidate site prioritization (Fig. 1A). Selected Gene Ontology (GO) terms associated with differentially abundant tissue proteins included intermediate filament organization, keratinization, epithelial cell differentiation, and the classical and alternative complement pathways. Intermediate filament organization and the classical complement pathway each included 11 proteins, whereas keratinization included nine proteins and epithelial cell differentiation and the alternative complement pathway each included six proteins (Fig. 1B). These annotations describe epithelial and immune-related tissue features; the numbers or proportions of annotated proteins do not measure pathway activation.

The original single-cell atlas contained 23 clusters, including C21, a broad candidate T-cell population provisionally annotated as T cells in the existing atlas (Fig. 1C). This population comprised 650 cells. Cell counts were 34, 141, 5, 106, 241, and 123 in the occipital and frontal samples from donors WJR, MRJ, and FLW, respectively (Fig. 1D). The MRJ occipital sample contained only five target cells, indicating substantial sampling uncertainty. In the original T-cell subclustering analysis, regional differential-expression results for C0 were associated with T-cell activation, lymphocyte differentiation, and T-cell receptor signaling. C1 results included T-cell receptor signaling and STAT regulation, together with cornification-related terms (Fig. 1E). These findings were treated as exploratory functional clues rather than independent confirmation of cell identity or regional immune activation.

To retain donors as the biological replicates, raw unique molecular identifier (UMI) counts were aggregated within the target population for paired pseudobulk analysis. Of 17,434 detected genes, 10,217 passed expression filtering; none met a Benjamini–Hochberg false discovery rate (BH-FDR) threshold of 0.05. A selected-gene heatmap illustrated variation in helper-associated, quiescence/memory-associated, and cytotoxic transcriptional patterns (Fig. 1F). The helper-associated module was higher in frontal scalp in all three donors, with a mean paired score difference of 1.094 and a nominal P value of 0.0754. The quiescence/memory module changed in the opposite direction in all three donors, with a mean difference of −0.902 (P = 0.220). The cytotoxic module showed inconsistent directions across donors, with a mean difference of −0.425 (P = 0.638; Fig. 1G). None of these module comparisons reached nominal statistical significance.

ZBTB7B showed a mean paired difference in log2(CPM + 1) of 0.534, with higher expression in frontal scalp in two of three donors. This difference was not statistically significant (P = 0.221; BH-FDR = 0.789; Fig. 1H). Collectively, these analyses establish a tissue and T-cell-related context for hypothesis generation, but do not demonstrate significant ZBTB7B transcriptional upregulation or increased T-cell infiltration.

Total ZBTB7B and its K206- and K212-associated lactylation signals show opposite regional trends

Integrated profiling of total proteins and enriched lactylated peptides in two frontal–occipital sample pairs identified 1,629 lactylated peptides, 1,804 modification sites, and 551 modified proteins (Fig. 2A,

B). A sequence-context analysis spanning six residues on either side of the modified lysine showed higher reported positional preference scores for residues including glycine, proline, and lysine at selected positions (Fig. 2C). This analysis describes the sequence context of the identified lactylome, rather than differential residue enrichment between scalp regions.

Principal component analysis (PCA) of lactylation quantification data attributed 46.52% and 30.94% of the variance to PC1 and PC2, respectively. The four samples showed region-associated differences in their distribution, while the two occipital samples were dispersed along PC2, indicating that both regional variation and sample heterogeneity warrant consideration (Fig. 2D). Applying exploratory thresholds of nominal $P \leq 0.05$ and a frontal-to-occipital abundance ratio ≥ 1.5 or ≤ 0.6667 identified 31 increased and 26 decreased sites; 1,747 sites did not meet these combined criteria (Fig. 2E). These counts reflect nominal screening thresholds, not significance after multiple-testing correction.

When site-level changes were mapped against their corresponding total-protein changes, ZBTB7B K206 and K212 fell in the region representing higher total protein but lower modified-peptide abundance (Fig. 2F). The frontal-to-occipital ratio for total ZBTB7B was 1.806, with a reported P value of 0.2184 and BH-FDR of 0.801 (Fig. 2G). The corresponding ratios were 0.523 for K206 and 0.587 for K212, with reported P values of 0.04638 and 0.001054 and BH-FDR values of 0.750 and 0.127, respectively (Fig. 2H, I). Both donors showed the same directional pattern: higher total ZBTB7B and lower signals for the two modified sites in frontal scalp.

This discordance supports prioritizing K206 and K212 for further investigation. However, neither site passed BH-FDR correction, and comparison of separately quantified protein and modified-peptide abundances does not directly establish a reduction in modification occupancy. Subsequent analyses therefore examined sequence-based prediction and site-localization evidence for these candidates.

Center-lysine and local-context representations yield higher reported discrimination than CLS-only ESM2 features

Using an adaptation of the HybridKla framework, we predicted lysine lactylation from 51-amino-acid, lysine-centered sequence windows and compared ESM2 representations and score-fusion strategies (Fig. 3A). The initial reference dataset contained 47,004 records, comprising 23,502 positive and 23,502 reference-negative records. Excluding 601 records for which protein identifiers could not be recovered yielded 46,403 site records: 23,502 positives and 22,901 reference negatives across 6,789 protein groups. All base models used the same protein-grouped five-fold partition, preventing records from the same protein group from appearing in both training and testing within a fold.

Five ESM2 representations were evaluated: CLS alone (E0), the central lysine embedding (E1), masked mean pooling over valid residues (E2), concatenated CLS and central-lysine embeddings (E3), and concatenated central-lysine and masked-mean embeddings (E4). Training loss generally decreased across epochs, whereas validation loss initially decreased and subsequently fluctuated or increased to

varying degrees (Fig. 3B–F). Thus, continued reduction in training loss did not indicate continued improvement in generalization.

Mean five-fold ROC-AUC increased from 0.8078 for E0 to 0.8180 for E1 and 0.8194 for E2, reaching 0.8222 for both E3 and E4 (Fig. 3G). Corresponding mean PR-AUC values were 0.7756, 0.7885, 0.7888, 0.7936, and 0.7923 (Fig. 3H). E3 had a slightly higher PR-AUC than E4; E4 was therefore not uniformly superior. ROC-AUC values calculated from pooled out-of-fold predictions were 0.8219 for E3 and 0.8215 for E4. No statistical comparison established a difference between these closely performing representations.

Within this dataset and evaluation setting, representations incorporating the target residue or valid sequence window produced higher reported metrics than the CLS-only representation, with modest further gains from concatenation. E4 was selected for subsequent fusion because it explicitly combines target-site and local-context information while providing a favorable balance of threshold-dependent metrics.

Five-score fusion provides modest numerical gains, while laboratory adaptation increases positive-site recall

In the reported reference-set evaluation, the 5D MetaModel achieved a ROC-AUC of 0.8281 and PR-AUC of 0.7984, exceeding the corresponding reported values for ESM2-E4, LSTM, and the three handcrafted-feature base models shown in Fig. 4A and B. ESM2-E4 achieved ROC-AUC and PR-AUC values of 0.8215 and 0.7907, whereas LSTM achieved 0.8117 and 0.7782. Removing ACF and PSEAAC score inputs from the 7D MetaModel changed ROC-AUC from 0.8264 to 0.8281 and PR-AUC from 0.7957 to 0.7984 (Fig. 4C). These are modest numerical differences rather than demonstrated statistically significant improvements.

At a fixed decision threshold of 0.5, the 5D MetaModel achieved a Matthews correlation coefficient (MCC) of 0.5193, sensitivity of 0.8392, specificity of 0.6722, F1 score of 0.7776, and balanced accuracy of 0.7557. Although sensitivity and some summary metrics improved relative to ESM2-E4, specificity decreased from 0.6776 to 0.6722 and was also slightly lower than that of the 7D model (Fig. 4D). Score fusion therefore involved a sensitivity–specificity trade-off rather than improvement across every metric.

To assess adaptation to laboratory-derived data, we fine-tuned the E4 classification head while freezing the ESM2 backbone. Before and after adaptation, the same 346 confirmed positive sites from 176 previously unseen proteins were evaluated. Mean recall across the five separately tested E4 fold models increased from 0.7809 to 0.9075, an increase of 12.66 percentage points (Fig. 4E, F). This value represents the mean of five individual-model recall estimates, not the recall of a probability-averaged ensemble.

In the adaptation validation set, the proportion of unlabeled lysines predicted as positive also increased, from 0.4290 to 0.6423, indicating a more positive prediction tendency. Because the held-out evaluation

included only confirmed positive sites, these findings support improved positive-site detection in unseen proteins from the same laboratory source, but cannot establish improved specificity, ROC-AUC, or overall classification performance. Moreover, out-of-fold base-model scores alone do not establish independent evaluation of the MetaModel; interpretation of fusion performance remains contingent on verification of the complete meta-level cross-fitting procedure.

Five-model score-space clustering reveals high- and low-score patterns associated with reference labels

Each site was represented by five prediction scores from ESM2-E4, LSTM, AAINDEX, CKSAAP, and OBC, producing a 46,403 × 5 matrix. Following dimension-wise standardization, K-means clustering was performed in the full five-dimensional space, and PCA was used only for three-dimensional visualization. The same displayed records and coordinates were colored by cluster assignment, reference class, or MetaModel prediction probability (Fig. 5A–C). PC1, PC2, and PC3 explained 58.2%, 15.4%, and 12.7% of the variance, respectively, or approximately 86% in total. The plots display 15,000 records, whereas summary statistics include all 46,403 records.

In the displayed K = 2 solution, Cluster 1 contained 20,036 records, of which 22.05% were reference positives. Cluster 2 contained 26,367 records, with a positive proportion of 72.38% (Fig. 5D). Mean MetaModel probabilities were 0.2318 and 0.7293, respectively (Fig. 5E). Cluster 2 therefore represented a higher-score pattern with a greater positive-label proportion, whereas Cluster 1 represented a relatively lower-score pattern. Both clusters contained positive and reference-negative records, indicating incomplete class separation.

These results describe the relationship between model score patterns and reference labels. Although K-means did not directly use class labels, its inputs were outputs of supervised models. The clustering is therefore not validation independent of supervised learning and does not identify distinct biological lactylation subtypes.

Sequence prediction and mass-spectrometric localization support prioritization of ZBTB7B K206 and K212

The canonical ZBTB7B sequence, UniProt O15156, contains 539 amino acids and 18 lysines. K206 and K212 lie between the illustrated N-terminal BTB/POZ domain and C-terminal C2H2 zinc-finger region (Fig. 6A). At a threshold of 0.5, the original 5D model report identified nine predicted-positive positions: K60, K61, K206, K212, K353, K360, K418, K441, and K450. Both omics-derived candidates were included (Fig. 6B). This panel shows thresholded candidate positions, not equal probabilities across sites.

In the exported mass-spectrometry records, K206 was assigned to the first lysine of KAFLQTK, corresponding to residue 206 of the full-length sequence. K212 was assigned to the sixth residue of AFLQTKGAR, corresponding to full-length residue 212 (Fig. 6C). These peptide schematics illustrate sequence mapping and do not constitute additional MS/MS fragment-spectrum evidence. Across four

exported run records, localization probabilities for K206 were 0.8960, 0.9999, 1.0000, and 0.9996; probabilities for K212 were 1.0000 in all four records (Fig. 6E). These scores support modification-site assignment in the original search, but do not quantify the significance of regional abundance differences.

In the associated AlphaFold model, the predicted local distance difference test (pLDDT) scores at K206 and K212 were 32.28 and 25.09, respectively (Fig. 6D). Both positions therefore have very low local structural confidence. The model was used to illustrate their locations within the full-length protein, not to infer solvent exposure, precise contacts, binding interfaces, or lactylation-induced conformational changes.

Together, the tissue-level abundance patterns, sequence predictions, and localization records nominate K206 and K212 for targeted validation. Their cellular origins, absolute modification occupancies, and effects on ZBTB7B transcriptional activity or hair-follicle growth remain unresolved.

Discussion

Multi-omics integration provides site-level leads for investigating AGA mechanisms

This study integrated scalp single-cell transcriptomics, proteomics, lactylomics, and HybridKla-based sequence prediction to prioritize ZBTB7B K206 and K212 for investigation in androgenetic alopecia (AGA). The evidence layers served complementary purposes: tissue-level functional analyses established an epithelial and immune-related context; transcriptional analysis of a candidate T-cell population provided cellular clues; protein and modified-peptide quantification captured regional trends at different molecular levels; and sequence prediction and localization records focused the investigation on specific residues. The principal contribution is to extend attention beyond expression-associated candidate molecules toward site-specific regulation, providing defined targets for examining relationships between protein functional states and the follicular microenvironment.

This site-oriented approach complements research on the tissue environment of AGA. A recent study combining single-cell and spatial transcriptomics with functional experiments showed that aberrant connective tissue sheath contraction can impair hair-follicle growth through PIEZO1-associated mechanotransduction, highlighting the pathophysiological importance of perifollicular tissue states. (Li, Yang et al. 2026) Our study approaches the problem from the perspective of post-translational regulation by linking the immune-associated transcription factor ZBTB7B to candidate lactylation sites. The resulting question extends beyond which genes or proteins change in AGA-associated tissue to whether modification of specific residues contributes to altered protein function. This adds a protein-modification dimension to the investigation of AGA mechanisms.

ZBTB7B provides a biological entry point for linking lactylation to immune regulation

ZBTB7B, also known as ThPOK, has an established role in CD4-versus-CD8 T-cell lineage commitment, providing a clear rationale for studying its post-translational regulation. (He, He et al. 2005) In the present study, functional analyses of the candidate T-cell population involved T-cell activation, lymphocyte differentiation, and antigen-receptor-related processes, while the helper-associated module showed a consistent frontal increase across three donors. These observations provide context for investigating ZBTB7B regulation. However, neither ZBTB7B expression nor module scores reached statistical significance in donor-paired analyses, and the inclusion of ZBTB7B in the helper-associated module means that these observations are not independent confirmations.

Combining this transcriptional context with tissue lactylomic data yields a more focused working hypothesis: the AGA-associated tissue environment may accompany changes in site-specific ZBTB7B modification, which, if present in T cells, could influence local immune states through altered transcriptional regulation. The value of this hypothesis is that it converts a broad proposed relationship between lactylation and immunity into a question involving a defined protein, specific residues, and a candidate cellular context. Nevertheless, whole-scalp modification signals and candidate T-cell transcriptional signals represent different observational levels. Establishing cellular origin is therefore a necessary link between these evidence types rather than a conclusion that follows directly from their integration.

The discovery of histone lactylation established a connection between lactate-associated metabolism and chromatin-dependent transcription. (Zhang, Tang et al. 2019) ZBTB7B is a non-histone transcription factor, and whether its lactylation affects chromatin occupancy, cofactor interactions, or downstream transcriptional programs represents a molecule-specific extension of that broader concept. Accordingly, the present findings define a candidate route for investigating relationships among the metabolic environment, transcription-factor modification, and immune regulation, rather than an established epigenetic causal mechanism.

Domain organization and chromatin-regulatory mechanisms guide investigation of ZBTB7B site function

Human canonical ZBTB7B comprises 539 residues. The UniProt O15156 record archived in this project annotates a BTB domain at residues 34–115 and four zinc fingers at 346–368, 374–396, 402–424, and 430–454, with the fourth annotated as atypical. K206 and K212 lie between, rather than within, these domain regions. Their potential functions should therefore be considered in terms of regulation from the intervening region, rather than assumed direct DNA contacts or established BTB interaction interfaces. Low local pLDDT values also do not establish a specific flexible conformation, solvent exposure, or allosteric mechanism.

Mechanistically, Gao and colleagues demonstrated that the ThPOK BTB domain recruits the nucleosome remodeling and deacetylase complex NuRD to support repression of Runx3 and CD8-lineage genes. NuRD recruitment and maintenance of Cd4 expression were not equivalent functional branches. (Gao, Zamisch et al. 2022) This distinction motivates separate measurement of protein abundance and

regulatory activity. For the present candidates, relevant questions are whether modification affects NuRD association or target-gene occupancy, rather than whether total abundance alone implies stronger or weaker function. Because K206/K212 are outside the BTB domain, any influence on recruitment remains a hypothesis of indirect regulation.

Previous acetylation and ubiquitination studies motivate investigation of modification crosstalk

ZBTB7B already has an experimentally investigated post-translational regulatory context. Zhang and colleagues reported p300-mediated ThPOK acetylation at K210, K216, and K339, using the numbering in their study, and competition with ubiquitination that increased protein stability. Lysine substitutions also altered stability and repression of CD8-associated expression. (Zhang, Zhang et al. 2010) Particularly relevant here, human UniProt O15156 annotates K206, K212, and K335 as EP300-associated acetylation sites based on similarity to mouse Q64321, not direct experiments in human AGA tissue. This provides a prior modification context while requiring clear separation of species numbering, annotation, and direct evidence.

The resulting question is whether alternative modification states at candidate lysines affect ZBTB7B turnover or transcriptional activity. Competition among lactylation, acetylation, and ubiquitination was not demonstrated here, and p300-dependent acetylation does not establish that p300 lactylates these residues. Nonetheless, paired assessment of the modifications, protein half-life, ubiquitination, and complex association would distinguish stability-related from activity-related effects. Importantly, higher total protein alongside lower lactylated-peptide signals cannot be directly explained by a simple model in which lactylation stabilizes ZBTB7B.

A lung adenocarcinoma study additionally reported NEDD4-dependent, K450-associated ubiquitination and proteasomal degradation of ZBTB7B, affecting downstream ADPGK transcription and glycolysis. (Liu, Huang et al. 2026) This is a separate example connecting ZBTB7B modification, turnover, and metabolic output; its different site and disease setting preclude treating it as evidence for the same pathway in AGA. The novelty here is therefore not the general proposition that ZBTB7B can be modified, but the placement of AGA-associated K206/K212 lactylation candidates within its established regulatory context. A limited literature search likewise does not justify claiming the first discovery of ZBTB7B lactylation.

Discordant protein and modification trends highlight the value of site-specific investigation

A notable observation was the opposing regional trend in total ZBTB7B and its K206- and K212-associated lactylated peptides: both donors showed higher total protein but lower modified-peptide signals in frontal scalp. This pattern indicates that total-protein abundance alone does not describe the observed modification changes and makes site-specific regulation a relevant question for follow-up. Examining total proteomic and modification data together helps avoid attributing every change in a modified-peptide signal to a corresponding change in protein abundance.

Traceable modified peptides, high localization probabilities, and sequence-model support provide a basis for targeted investigation of K206 and K212. These evidence types address different questions: localization probabilities support residue assignment within a peptide, quantitative measurements describe abundance relationships across samples, and prediction scores reflect learned sequence patterns. Their combination makes candidate selection more explicit, but the evidence types are not interchangeable. In particular, neither site's regional difference survived BH correction, and separately normalized total-protein and modified-peptide signals do not directly quantify absolute modification occupancy. Tissue composition, peptide detectability, and enrichment efficiency could also contribute to the observed pattern. The findings therefore support prioritization of these sites rather than a definitive estimate of their modification changes or functional direction.

Deep learning-assisted prediction extends the evidence-integration approach to lactylation research

A methodological feature of this study is the incorporation of computational prediction into disease-oriented candidate selection rather than its presentation solely as a separate benchmarking exercise. Within the HybridKla framework, ESM2 supplied learned sequence-context representations, LSTM modeled sequential information, and physicochemical and compositional descriptors provided additional sequence representations that were combined through score fusion. (Ning, Qin et al. 2025) This arrangement allowed existing sequence and omics data to be examined around a shared candidate question, connecting modification identification, molecular prioritization, and experimental design. The contribution lies in this disease-oriented adaptation and integration, rather than a claim to have newly invented the existing HybridKla framework.

Representation comparisons provided practical evidence about how to use the protein language model for this task. Central-lysine or valid-window representations yielded higher reported metrics than CLS alone, supporting attention to the target residue and its local context. The similar performance of E3 and E4 favors task-specific evaluation of representations rather than a universal preference for one concatenation strategy. The 5D model retained fewer score inputs while obtaining modest numerical gains over the 7D model. Thus, adding feature types did not necessarily improve performance in this setting; selecting an appropriate combination may be more useful than simply expanding the input set.

This direction is consistent with broader applications of protein language models. Integration of language models has supported single-sequence structure prediction, while representations trained on biophysical dynamics have informed mutation-effect prediction. (Jing, Wu et al. 2024, Hou, Zhao et al. 2026) Combining language-model representations with structural context has also been applied to lysine PTM-site prediction, illustrating the potential of integrating different information levels. (Luo, Zhu et al. 2026)The present study applied this computationally assisted approach to AGA-related lactylation candidate selection, but neither three-dimensional coordinates nor molecular dynamics were training inputs. The method is therefore appropriately described as deep learning-assisted sequence prediction, rather than direct simulation of lactylation or protein dynamics.

In terms of research strategy, the study strengthens the connection between omics observations, computational screening, site-level evidence integration, and functional-experiment design. Potential benefits include narrowing the candidates requiring priority validation, extracting more information from existing datasets, and focusing subsequent experiments on defined residues. Although reductions in experimental cost or time were not measured, this prioritization strategy may help concentrate resources without relying exclusively on a single differential-abundance threshold or undirected testing of individual sites.

The methodological novelty is best framed as question-driven integration across scales: cellular and tissue data establish the AGA-related context, site-to-protein comparisons identify modification patterns of interest, and deep learning plus peptide-localization evidence establishes experimental priorities. This workflow connects why a protein is selected, why particular residues are prioritized, and which functions should be tested. It thereby advances AGA investigation from descriptions of expression differences toward testable questions about protein turnover, regulatory-complex assembly, and chromatin regulation. Its contribution is a more precise molecular focus and validation sequence, not computational prediction as a substitute for mechanism.

Data adaptation and score-space analysis clarify the practical interpretation of model outputs

Laboratory adaptation made the prediction workflow more directly responsive to a specific data source. With the ESM2 backbone frozen, classification-head adaptation increased mean recall on the same held-out positive sites from 0.7809 to 0.9075. This suggests that existing sequence representations can be adjusted through a limited set of parameters to increase coverage of positive candidates. Such high-recall screening is relevant to discovery-oriented studies. However, the positive-prediction rate among unlabeled lysines also increased, and the adapted fusion model did not achieve higher ROC-AUC or PR-AUC on the original reference set. The benefit is therefore improved detection of laboratory-derived positive sites, rather than a general improvement in classification ability.

Five-model score-space analysis provided an additional description of how predictions were organized. The higher-score cluster had a greater reference-positive proportion and mean MetaModel score, whereas the lower-score cluster showed the opposite pattern. This offers a visualization for examining agreement and disagreement among model outputs, but does not explain which sequence features caused each prediction. Because clustering used supervised-model outputs, its correspondence with reference labels and fusion scores is not independent validation and does not identify naturally occurring lactylation subtypes. Biological interpretation would require association with measured modification states, cellular origins, and functional outcomes.

Study boundaries and the path from candidate sites to mechanistic validation

The main limitations of this exploratory integration concern biological replication and the connections between evidence layers. The single-cell dataset contained three donors, including one sample with very few target cells, whereas proteomic and lactylomic analyses included only two pairs. Donor-level

pseudobulk aggregation helps avoid treating individual cells as biological replicates but cannot compensate for a small number of donors. Methodological studies have emphasized the importance of accounting for variation between biological replicates in single-cell differential-expression analysis. (Squair, Gautier et al. 2021) Cornification-related signals in the candidate population also warrant reassessment of annotation and cellular purity. Ambient RNA is a recognized confounder in single-cell data, although its specific contribution here remains to be assessed. (Young and Behjati 2020) In addition, the absence of corresponding samples from healthy individuals means that frontal–occipital differences may include normal anatomical variation.

For the computational analyses, protein grouping reduces within-protein overlap across training and test sets, but homologous sequences, duplicate windows, and overlap between ZBTB7B candidate windows and model-development data require separate examination. Fusion training and model selection must also remain isolated from testing; research on cross-validation has demonstrated optimistic bias when selection and performance estimation are not properly separated. (Varma and Simon 2006) The reported fusion results therefore warrant confirmation through a documented cross-fitting workflow and independent evaluation. Out-of-fold base-model predictions alone do not establish that the complete pipeline is free of information leakage.

Mechanistic follow-up can proceed in stages around K206 and K212. Synthetic modified peptides and targeted mass spectrometry in an expanded paired cohort would first verify the sites while quantifying the corresponding total protein. Sorted-cell analyses could establish cellular origin, followed by experiments testing effects on ZBTB7B localization, interactions, and transcriptional activity, and ultimately by hair-follicle-relevant functional assays. Both positions have low local confidence in the AlphaFold model, making the structure suitable for positional illustration rather than identification of interfaces or conformational mechanisms. Site-directed mutations should be interpreted alongside modification-specific measurements and rescue experiments to distinguish effects of lysine substitution from those of altered lactylation.

Overall, this study adds a molecular perspective centered on candidate ZBTB7B lactylation sites to AGA research and demonstrates how deep learning can connect multi-omics observations to experimentally tractable questions. Its value lies both in prioritizing K206 and K212 and in establishing an evidence-integration route from tissue context to site-specific functional investigation, providing a foundation for examining relationships among post-translational regulation, immune states, and the follicular microenvironment.

Conclusion

By integrating scalp single-cell transcriptomics, proteomics, lactylomics, and HybridKla-based sequence prediction, this study prioritized ZBTB7B K206 and K212 as candidate lactylation sites for investigation in androgenetic alopecia (AGA). Paired scalp samples showed opposing regional trends in total ZBTB7B abundance and the lactylated peptides associated with these sites, highlighting the value of

investigating site-specific regulation beyond changes in protein abundance. Together with the T-cell-related transcriptional context, these findings connect the established immunoregulatory role of ZBTB7B with its potential regulation by lactylation, providing a defined molecular entry point for exploring relationships among protein modifications, immune states, and the follicular microenvironment in AGA. The integration of multi-omics evidence with deep learning predictions also provides a candidate-prioritization strategy that advances AGA research from tissue-level associations toward site-specific functional investigation.

Declarations

Code availability

The code used in this study is available on GitHub at https://github.com/lipei2026/kla_2026

Data availability

The single-cell RNA sequencing data analyzed in this study are available in the NCBI Sequence Read Archive under BioProject accession PRJNA1252963. The mass spectrometry proteomics and lysine lactylation proteomics data generated in this study are available in iProX under subproject accession IPX0020147001.

Acknowledgements

This work was supported by Fund of the Ministry of Education of the People's Republic of China (grant number 231103242232720; 2408271251) as well as Clinical Research Fund of Shanghai East Hospital(grant number SEHHH-2021□KJ□-0669-KJB-485; SEHHH-2023□KJ□-0143-KYB-111)

Patients and Ethical Approval

This study involved 5 male patients diagnosed with AGA who underwent follicular unit transplantation at the Department of Plastic Surgery, Shanghai East Hospital, Tongji University. Written informed consent was obtained from all participants before sample collection. The study protocol was approved by the Ethics Committee of Shanghai East Hospital, Tongji University. All procedures adhered to institutional ethical guidelines.

Declaration of Interest Statement

The authors declare no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Author contributions

Y.Z. drafted the manuscript, integrated the project, and organized the figures. **P.L.** trained the deep learning models and deposited the code in the public repository. **Chika Hasegawa** collected and preprocessed the samples. **S.L.** and **Z.T.** prepared and formatted the figures. **Chenlong He** and **M.Y.** contributed to data curation and analysis. **S.W.** contributed to interpretation of the findings and manuscript revision. **Y.W.** contributed to study supervision and critical review of the manuscript. **Y.L.** led and supervised the study. **Y.W.** and **Y.L.** served as corresponding authors.

References

1. Gao Y, Zamisch M, Vacchio M, Chopp L, Ciucci T, Paine EL, Lyons GC, Nie J, Xiao Q, Zvezdova E, Love PE, Vinson CR, Jenkins LM, Bosselut R (2022) NuRD complex recruitment to Thpok mediates CD4(+) T cell lineage differentiation. Sci Immunol 7(72):eabn5917
2. He X, He X, Dave VP, Zhang Y, Hua X, Nicolas E, Xu W, Roe BA, Kappes DJ (2005) The zinc finger transcription factor Th-POK regulates CD4 versus CD8 T-cell lineage commitment. Nature 433(7028):826–833
3. Hou C, Zhao H, Shen Y (2026) Protein language models trained on biophysical dynamics inform mutation effects. Proc Natl Acad Sci U S A 123(4):e2530466123
4. Jing X, Wu F, Luo X, Xu J (2024) Single-sequence protein structure prediction by integrating protein language models. Proc Natl Acad Sci U S A 121(13):e2308788121
5. Li G, Yang L, Duan S, Chen M, Zhang Y, Liu F, Wang Y, Li J, Xu S, Wu Z, Wang M, Wang B, Zhao Z, Shi W, Lei M, Xie H, Tang Y, Deng Z, Li J (2026) Single-cell transcriptomics reveals hair growth retardation mediated by aberrant connective tissue sheath contraction in male androgenetic alopecia. Nat Commun 17(1)
6. Liu FP, Huang T, Yang HY, Zha JH, Xu C, Qiu CP, Wang X, Lu Z, Wang T (2026) Inhibition of ZBTB7B-mediated ADPGK transcription by NEDD4 impedes glycolysis and progression of lung adenocarcinoma. Oncogenesis 15(1)
7. Luo M, Zhu X, Bai C, Warshel A, Chen L (2026) Mining lysine post-translational modification sites by integrating protein language model representations with structural context. Proc Natl Acad Sci U S A 123(7):e2529141123
8. Ning W, Qin F, Zhou Z, Yang H, Li C, Guo Y (2025) HybridKla: a hybrid deep learning framework for lactylation site prediction. Brief Bioinform 26(4)
9. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, Hudelle R, Qaiser T, Matson KJE, Barraud Q, Levine AJ, La Manno G, Skinnider MA, Courtine G (2021) Confronting false discoveries in single-cell differential expression. Nat Commun 12(1):5692
10. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7:91
11. Young MD, Behjati S (2020) SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience 9(12)

12. Zhang D, Tang Z, Huang H, Zhou G, Cui C, Weng Y, Liu W, Kim S, Lee S, Perez-Neut M, Ding J, Czyz D, Hu R, Ye Z, He M, Zheng YG, Shuman HA, Dai L, Ren B, Roeder RG, Becker L, Zhao Y (2019) Metabolic regulation of gene expression by histone lactylation. Nature 574(7779):575–580
13. Zhang M, Zhang J, Rui J, Liu X (2010) p300-mediated acetylation stabilizes the Th-inducing POK factor. J Immunol 185(7):3960–3969

Figures

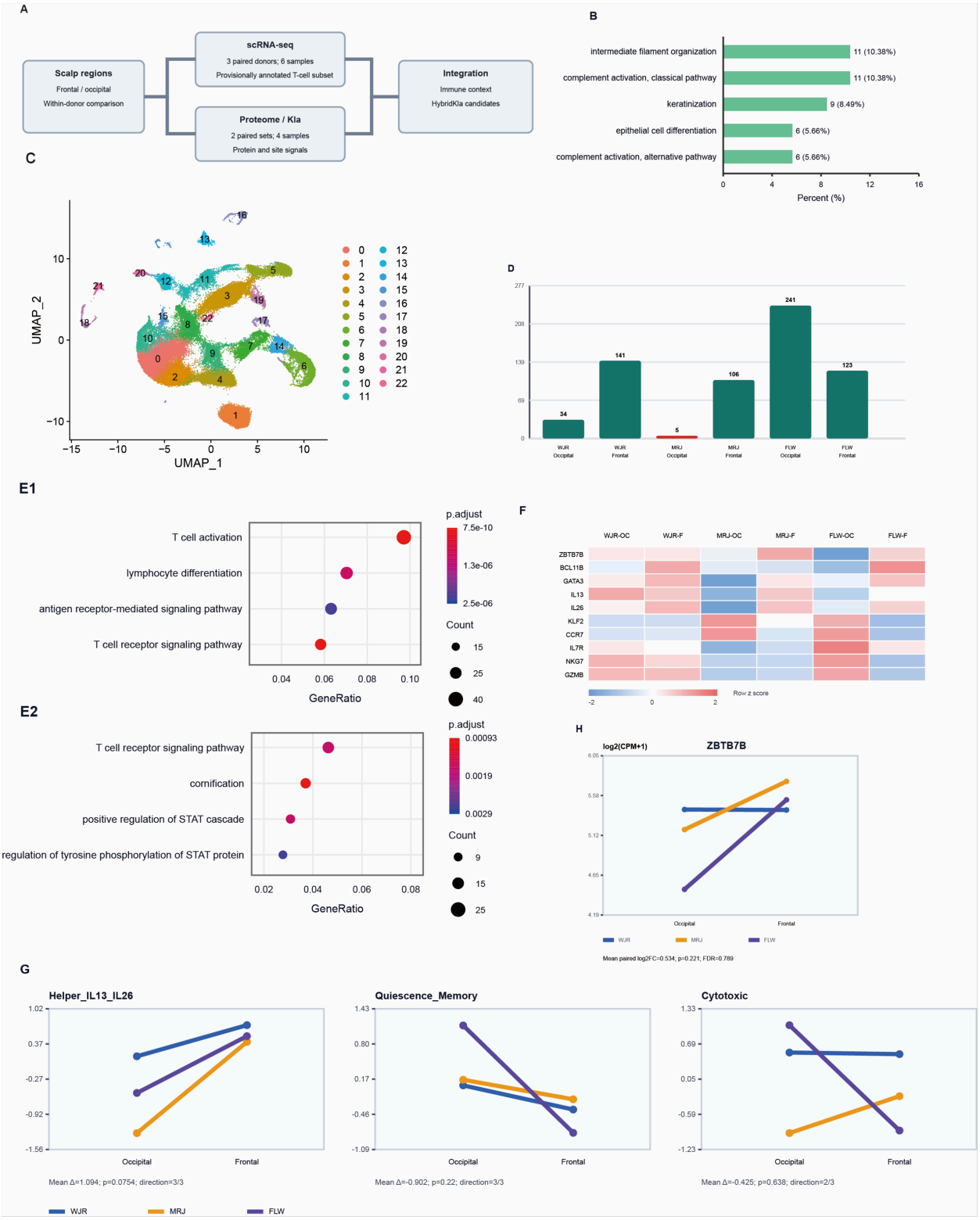


Figure 1

Tissue proteomic features and T-cell-associated transcriptional profiles in paired scalp samples from androgenetic alopecia.

(A) A paired multi-omics design integrates tissue-level changes with T-cell-associated transcriptional features. Single-cell RNA sequencing, tissue proteomics and lysine lactylation profiling of paired frontal

and occipital scalp samples were integrated to inform subsequent candidate-site prioritization using HybridKla. The single-cell cohort comprised six samples from three donors, whereas the proteomic and lactylomic datasets comprised four samples from two paired sample sets. Donor overlap between cohorts was not assumed, and occipital samples served as within-donor regional comparators.

(B) Differentially abundant tissue proteins are associated with complement-related processes and epithelial structure and differentiation. Five selected biological process terms from the original GO enrichment analysis are shown. Bar lengths indicate the percentage of differential proteins assigned to each term relative to the biological process differential-protein set used in the original analysis (denominator, 106). Labels indicate protein counts and percentages. Green denotes the GO category rather than the direction of protein changes or pathway activation.

(C) The complete single-cell atlas provides the context for identifying the candidate T-cell population. UMAP visualization includes 68,310 cells, with colors and numbers indicating the 23 clusters in the original analysis. Subsequent T-cell-associated analyses used the provisionally annotated C21 population, comprising 650 cells, whose cellular identity requires further verification.

(D) Captured candidate T-cell numbers vary substantially among samples, with particularly limited representation in the MRJ occipital sample. Bars show C21 cell counts for each donor and scalp region. Red indicates a sample containing fewer than 20 cells; the MRJ occipital sample contained only five cells. These counts reflect captured cells rather than a direct measure of tissue infiltration.

(E1) Selected enriched terms in C0 involve T-cell activation, lymphocyte differentiation and antigen receptor-associated signaling. Four terms selected from the original C0 differential-gene GO enrichment analysis are shown: T-cell activation, lymphocyte differentiation, antigen receptor-mediated signaling and T-cell receptor signaling.

(E2) C1 exhibits TCR/STAT regulatory terms alongside a cornification-related signal. Four terms selected from the original C1 differential-gene GO enrichment analysis are shown. The cornification term warrants caution when interpreting the cellular annotation of this population. In E1 and E2, GeneRatio indicates the proportion of input genes assigned to each term, bubble size indicates the associated gene count, and color indicates the adjusted P value. The plots use separate color scales and derive from the original cell-level analyses rather than donor-paired pseudobulk enrichment analyses.

(F) Selected immune-associated genes exhibit donor- and region-dependent expression heterogeneity. Pseudobulk expression of 10 genes across the six samples is shown. Raw UMI counts from C21 cells were aggregated within each sample, transformed to $\log_2(\text{CPM}+1)$, and standardized for each gene across samples. Colors indicate row-wise z scores, with the color scale clipped at −2 and 2. F and OC denote frontal and occipital samples, respectively.

(G) Helper-associated and quiescence/memory modules show directionally consistent increases and decreases, respectively, whereas cytotoxic changes vary among donors. Frontal scores were higher for

the helper-associated module and lower for the quiescence/memory module in all three donors, while cytotoxic scores showed heterogeneous changes. None of the three comparisons reached nominal statistical significance. Module scores were calculated by averaging the across-sample standardized $\log_2$(CPM+1) expression values of constituent genes. Matching colors connect paired samples from the same donor. Mean Δ denotes the mean frontal-minus-occipital difference, and “direction” denotes the number of donors showing the majority direction of change divided by the total donor number. Each module uses its own y-axis range. The helper-associated module includes ZBTB7B and therefore does not provide evidence independent of H.

(H) ZBTB7B shows a positive mean paired expression difference. Paired pseudobulk expression is shown for the three donors. The mean paired difference in $\log_2$(CPM+1) was 0.534 (P=0.221; FDR=0.789).

Kla, lysine lactylation; GO, Gene Ontology; UMAP, uniform manifold approximation and projection; UMI, unique molecular identifier; CPM, counts per million; FDR, false discovery rate.

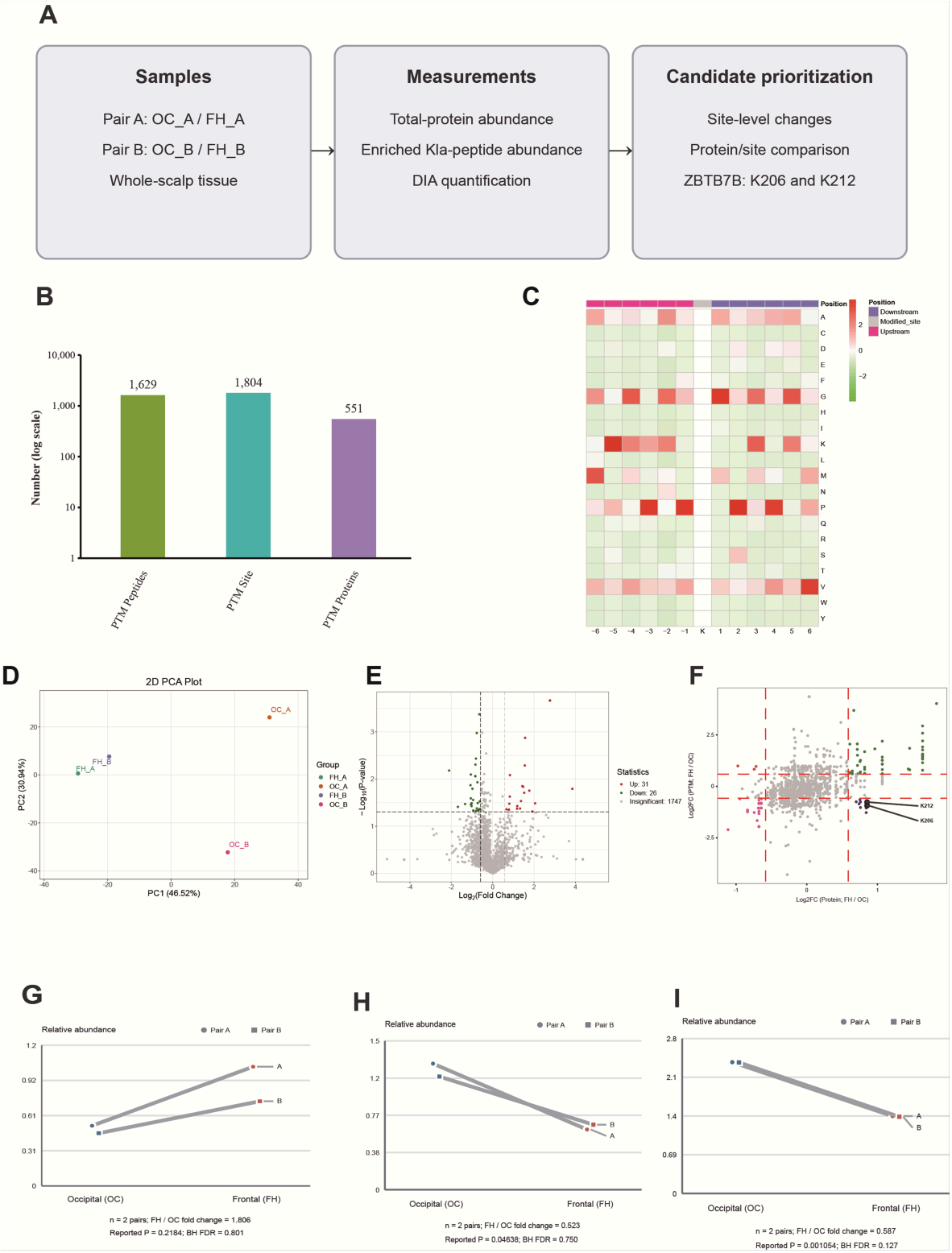


Figure 2

ZBTB7B protein abundance is higher, whereas the abundances of its K206 and K212-lactylated peptides are lower, in frontal scalp tissue.

(A) Overview of paired proteomic and lactylomic profiling and candidate-site prioritization. Paired frontal (FH) and occipital (OC) scalp tissues from two donors were analyzed by data-independent acquisition

(DIA) to quantify total proteins and enriched lactylated peptides. Site-level changes were integrated with corresponding protein-abundance changes to prioritize candidate lactylation sites on ZBTB7B.

(B) Lactylomic profiling identified 1,804 modification sites across 551 proteins. Bars indicate the numbers of identified lactylated peptides (1,629), lactylation sites (1,804), and lactylated proteins (551), with a logarithmic y-axis. These counts describe identification coverage rather than differential lactylation.

(C) Sequences flanking lactylation sites exhibit position-dependent amino acid patterns. The heatmap summarizes amino acid features at six residues upstream and downstream of all identified lactylation sites. The central K denotes the modified lysine; negative and positive positions indicate upstream and downstream residues, respectively. Colors represent relative positional scores in the original analysis output, with red and green indicating higher and lower scores, respectively. The top annotation distinguishes upstream residues, the modified site, and downstream residues. This analysis describes global sequence features rather than differential motifs between FH and OC.

(D) Principal component analysis summarizes intersample variation in the lactylomic profiles. Each point represents one scalp tissue sample. PC1 and PC2 explain 46.52% and 30.94% of the variance, respectively. The distribution describes overall sample similarity and individual variation and does not constitute statistical evidence of group separation.

(E) Exploratory differential analysis identified 31 increased and 26 decreased lactylation sites. The x-axis shows $\log_2$ fold change (FH/OC), and the y-axis shows $-\log_{10}$ nominal P value. Sites were screened using nominal $P < 0.05$ and fold change > 1.5 or $< 1/1.5$. Red, green, and gray indicate increased sites, decreased sites, and sites not meeting these thresholds, respectively. These classifications were not based on significance after multiple-testing correction.

(F) The two candidate ZBTB7B lactylation sites show abundance changes opposite to those of the corresponding total protein. The x-axis shows total-protein $\log_2$ fold change (FH/OC), and the y-axis shows lactylated-peptide $\log_2$ fold change (FH/OC). Each point represents a modification site paired with its corresponding protein-level change. The two annotated points represent ZBTB7B K206 and K212, not two distinct proteins. Dashed lines indicate the boundaries corresponding to a 1.5-fold change and its reciprocal.

(G) ZBTB7B protein abundance is higher in FH than in OC in both donors. The FH/OC fold change is 1.806, with a reported P value of 0.2184 and a Benjamini–Hochberg-adjusted false discovery rate (BH FDR) of 0.801.

(H) ZBTB7B K206-lactylated peptide abundance is lower in FH than in OC in both donors. The FH/OC fold change is 0.523, with a reported P value of 0.04638 and a BH FDR of 0.750.

(I) ZBTB7B K212-lactylated peptide abundance is lower in FH than in OC in both donors. The FH/OC fold change is 0.587, with a reported P value of 0.001054 and a BH FDR of 0.127.

In (G–I), each point represents one tissue sample; circles and squares denote donor pairs A and B, respectively. Lines connect matched OC and FH samples from the same donor (n = 2 biological pairs). Panels use independent y-axis scales, and abundance values should not be compared directly across panels. Neither ZBTB7B protein abundance nor either candidate site met BH FDR < 0.05; these findings are exploratory. Lactylated-peptide abundance does not measure site-specific lactylation occupancy, and measurements from whole scalp tissue do not establish T-cell-specific changes.

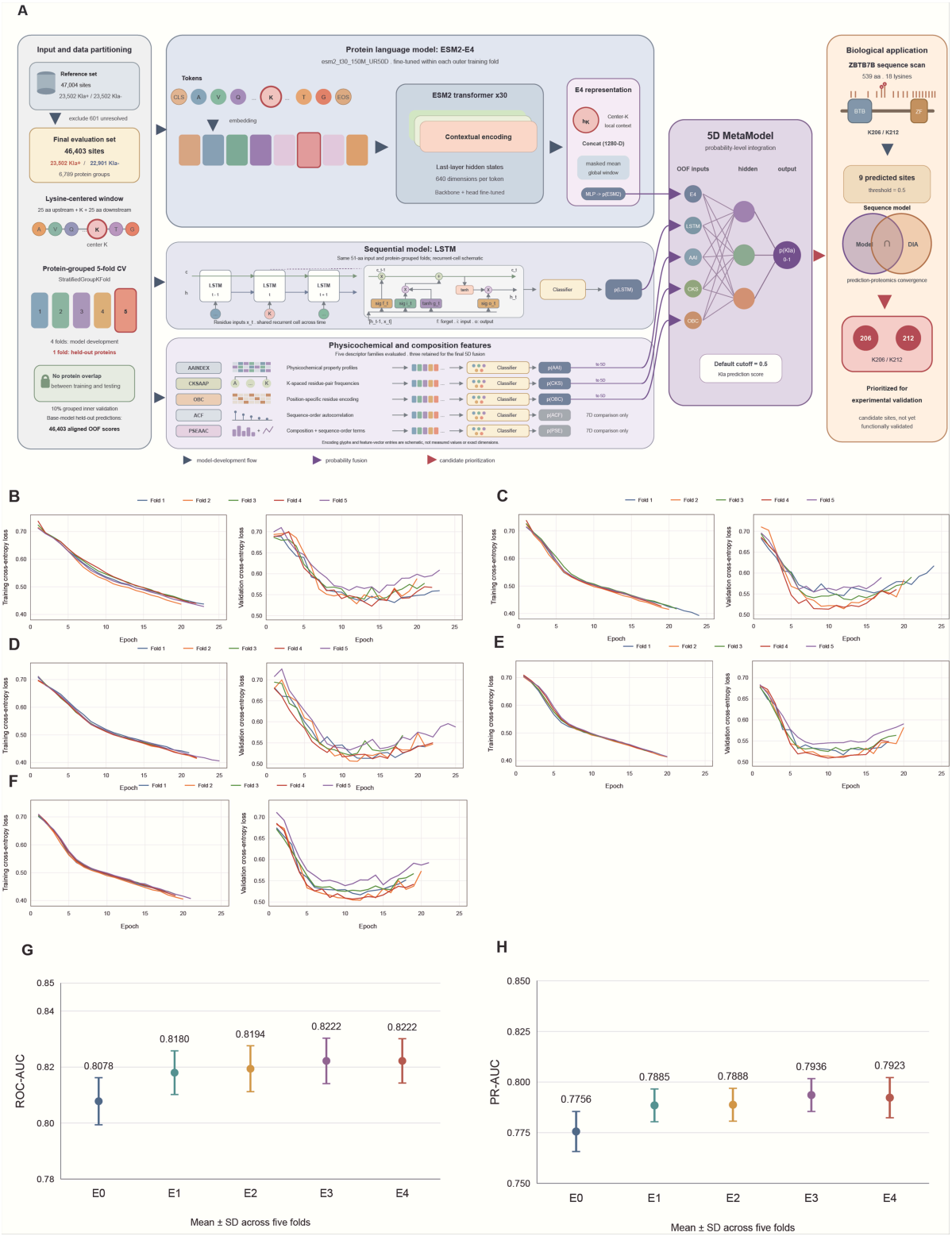


Figure 3

Joint representations of the central lysine and sequence context improve mean ESM2 discrimination of lysine lactylation sites.

(A) Schematic overview of the HybridKla algorithm for lysine lactylation site prediction and its application to candidate-site prioritization in ZBTB7B.

(B) The CLS baseline shows decreasing training loss and a late increase in validation loss. Five-fold training and inner-validation cross-entropy losses are shown for E0, which uses the CLS token representation.

(C) The central-lysine model shows decreasing training loss with fold-dependent validation trajectories. Training and validation losses are shown for E1, which uses the contextualized representation of the central lysine.

(D) The window-average model shows initial decreases followed by fluctuations and increases in validation loss. Training and validation trajectories are shown for E2, which averages valid-residue representations while excluding CLS, EOS, and padding positions.

(E) Joint CLS and central-lysine representations show late divergence between training and validation losses. Training and validation trajectories are shown for E3, which concatenates CLS and central-lysine representations into a 1,280-dimensional representation.

(F) Joint central-lysine and window-average representations show decreasing training loss and a late increase in validation loss. Training and validation trajectories are shown for E4, which concatenates the central-lysine representation with the masked mean and is used for downstream 5D fusion.

In B–F, the left and right plots show training and inner-validation losses, respectively. Blue, orange, green, red, and purple indicate folds 1–5. Curves connect the recorded epochs without smoothing, imputation, or extrapolation. Model checkpoints were selected using inner-validation ROC-AUC rather than minimum validation loss.

(G) E3 and E4 achieve the same highest mean ROC-AUC, numerically exceeding the CLS baseline. Five-fold mean ROC-AUC values for E0–E4 are 0.8078, 0.8180, 0.8194, 0.8222, and 0.8222, respectively.

(H) E3 achieves the highest mean PR-AUC, closely followed by E4, with both exceeding the CLS baseline. Five-fold mean PR-AUC values for E0–E4 are 0.7756, 0.7885, 0.7888, 0.7936, and 0.7923, respectively.

OOF, out-of-fold; DIA, data-independent acquisition; Kla, lysine lactylation.

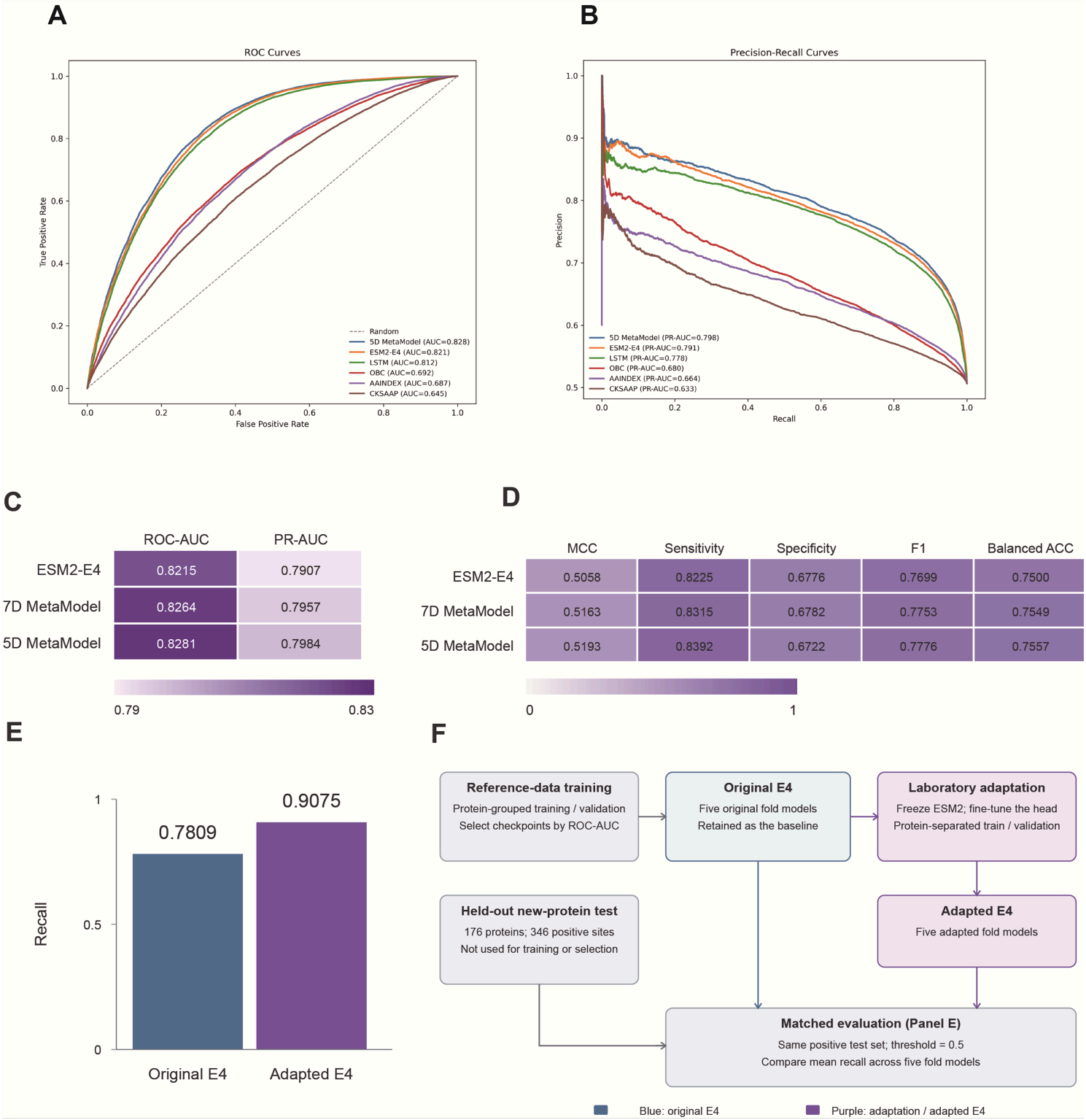


Figure 4

Five-model fusion yields higher lactylation-site discrimination, and laboratory-data adaptation increases E4 recall for positive sites on previously unseen proteins.

(A) The 5D MetaModel achieves a higher ROC-AUC than the individual base models compared. Receiver operating characteristic (ROC) curves compare the lactylation-site discrimination of the 5D MetaModel

with that of ESM2-E4, LSTM, AAINDEX, CKSAAP, and OBC. The 5D MetaModel integrates the predicted probabilities from these five base models and achieves a ROC-AUC of 0.8281.

(B) The 5D MetaModel achieves a higher PR-AUC than the individual base models compared. Precision–recall (PR) curves illustrate the relationship between precision and recall across prediction thresholds for the same models. The 5D MetaModel achieves a PR-AUC of 0.7984.

(C) Five-dimensional fusion yields slightly higher ROC-AUC and PR-AUC values than seven-dimensional fusion and ESM2-E4. The heatmap compares ESM2-E4, the 7D MetaModel, and the 5D MetaModel. The 7D MetaModel additionally incorporates predictions from the ACF and PSEAAC base models. ROC-AUC values are 0.8215, 0.8264, and 0.8281, and PR-AUC values are 0.7907, 0.7957, and 0.7984, respectively. A shared color scale spanning 0.79–0.83 highlights the small numerical differences.

(D) At a fixed threshold, five-dimensional fusion yields higher sensitivity and several composite metrics, with slightly lower specificity. Matthews correlation coefficient (MCC), sensitivity, specificity, F1 score, and balanced accuracy are compared at a prediction threshold of 0.5. The corresponding values for the 5D MetaModel are 0.5193, 0.8392, 0.6722, 0.7776, and 0.7557, respectively, illustrating a trade-off between positive-site detection and specificity.

(E) Laboratory-data adaptation increases mean E4 recall on held-out positive sites from previously unseen proteins from 0.7809 to 0.9075. The five original and five adapted E4 fold models were evaluated individually on the same test set, comprising 346 confirmed lactylation sites from 176 proteins absent from the reference dataset. A prediction threshold of 0.5 was used throughout. Bar heights represent the arithmetic mean of recall across the five fold models, rather than recall calculated from an ensemble of their predictions. Blue and purple indicate original and adapted E4 models, respectively.

(F) Schematic overview of E4 training, laboratory-data adaptation, and matched evaluation. Reference-data training and validation use protein-grouped partitions, with original checkpoints selected according to validation ROC-AUC. Laboratory-data adaptation freezes the ESM2 backbone and fine-tunes only the classification head, with no protein overlap between the adaptation training and validation sets. The held-out new-protein test set is excluded from training and model selection and is used to compare positive-site recall before and after adaptation. Blue denotes original E4, whereas purple denotes the adaptation process and adapted E4.

Panels (A–D) present performance point estimates from the reference-dataset analysis; numerical differences do not establish statistically significant improvements. The test set in (E) contains confirmed positive sites only and therefore supports evaluation of positive-site recall, but not specificity, ROC-AUC, or overall classification performance.

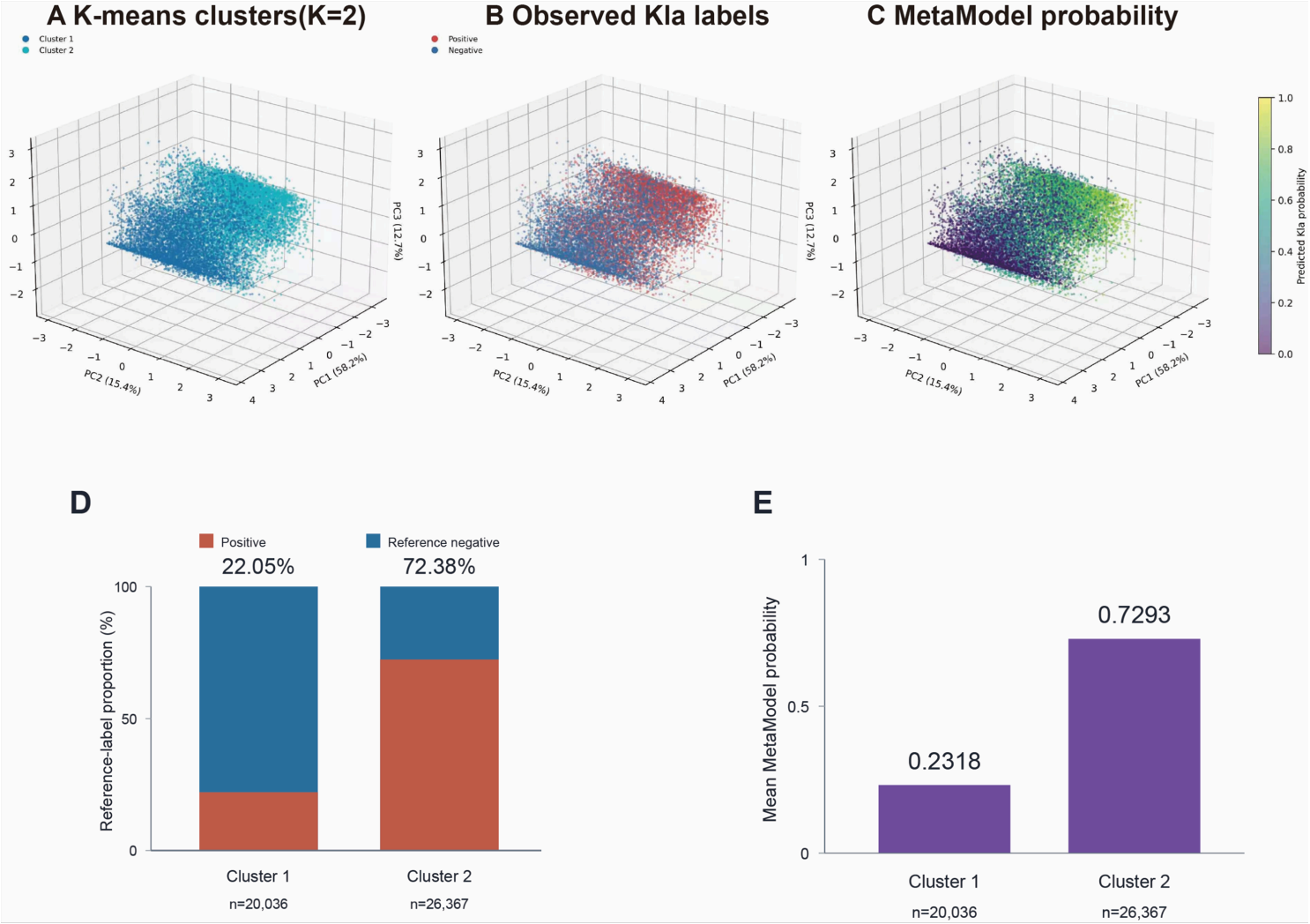


Figure 5

Two clusters in the five-model prediction-score space differ in reference-positive proportions and MetaModel prediction probabilities.

(A) K-means partitions samples into two distinct prediction-score patterns. Each sample is represented by the out-of-fold (OOF) prediction probabilities from five base models: ESM2-E4, LSTM, AAINDEX, CKSAAP, and OBC. After standardization of the five dimensions, K-means clustering (K = 2) was performed in the full five-dimensional space, followed by projection into three dimensions using principal component analysis (PCA). Blue and turquoise indicate Cluster 1 and Cluster 2, respectively. Reference-positive and reference-negative labels were not directly used in the clustering step.

(B) Reference-positive and reference-negative samples show different but substantially overlapping distributions in score space. The same samples and PCA coordinates as in (A) are colored by reference labels. Red indicates lactylation-positive samples, and blue indicates reference-negative samples. This view illustrates the correspondence between prediction-score clusters and reference labels.

(C) MetaModel outputs form a low-to-high prediction-probability gradient within the same score space. The coordinates shown in (A, B) are colored continuously according to the lactylation probabilities predicted by the 5D MetaModel. Purple, green, and yellow indicate low, intermediate, and high predicted probabilities, respectively, on a scale from 0 to 1.

(D) Cluster 2 has a higher reference-positive proportion than Cluster 1. Stacked bars summarize reference-label composition across all samples in each cluster. Cluster 1 contains 20,036 samples, of which 22.05% are reference-positive; Cluster 2 contains 26,367 samples, of which 72.38% are reference-positive. Red and blue denote reference-positive and reference-negative samples, respectively. Values above the bars indicate positive proportions.

(E) Cluster 2 has a higher mean MetaModel prediction probability than Cluster 1. Mean predicted probabilities are 0.2318 and 0.7293 for Clusters 1 and 2, respectively, as indicated above the bars. This comparison describes concordance between the base-model score patterns and the fusion-model outputs and does not constitute independent validation.

Panels (A–C) display the same subset of 15,000 samples from a total of 46,403, whereas (D, E) summarize all 46,403 samples. PC1, PC2, and PC3 explain 58.2%, 15.4%, and 12.7% of the variance, respectively. Clustering was performed in the standardized five-dimensional score space; PCA was used only for visualization. Reference-negative labels do not imply experimentally confirmed absence of lactylation. These clusters reflect similarities in supervised-model prediction scores and should not be interpreted as distinct biological lactylation subtypes. The differences shown are descriptive, with no inferential statistical tests presented.

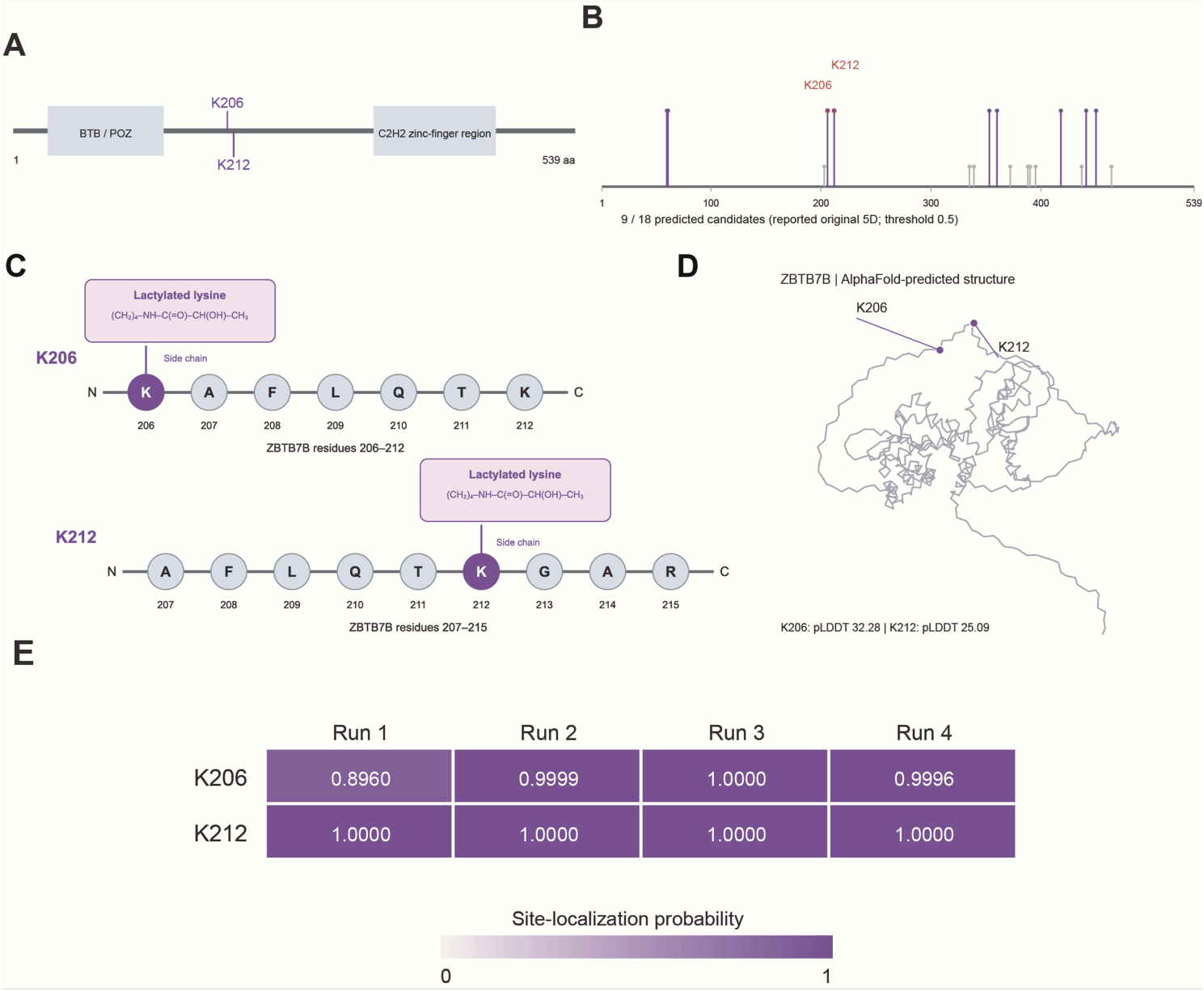


Figure 6

Model predictions and mass-spectrometric identification records support prioritizing ZBTB7B K206 and K212 as candidate lactylation sites for further validation.

(A) K206 and K212 lie between the BTB/POZ domain and the C2H2 zinc-finger region of ZBTB7B. The schematic shows the domain organization of the canonical human ZBTB7B isoform (UniProt: O15156; 539 amino acids) and the positions of the two candidate sites. Purple markers indicate K206 and K212. Domain boundaries are approximate.

(B) The reported five-dimensional fusion model predicts nine candidate lactylation sites, including K206 and K212. The plot shows the positions of all 18 lysine residues in ZBTB7B and their reported binary classifications by the original 5D MetaModel at a prediction threshold of 0.5. Purple indicates predicted candidate sites, light blue-gray indicates sites below the threshold, and red outlines highlight K206 and

K212 recorded by mass spectrometry. Marker heights distinguish classification categories rather than prediction probabilities; markers at adjacent positions may overlap.

(C) Two modified peptides in the mass-spectrometric identification records map to ZBTB7B K206 and K212. The upper peptide, KAFLQTK, corresponds to residues 206–212 and carries the modification at K206. The lower peptide, AFLQTKGAR, corresponds to residues 207–215 and carries the modification at K212. Circles represent amino acid residues, with their positions in the full-length protein indicated below. N and C denote the peptide amino and carboxyl termini, respectively. Purple highlights the target lysines, and the insets illustrate the lactylated lysine side-chain linkage. These peptide-chain schematics are based on identification records and are neither experimentally determined structures nor annotated MS/MS spectra.

(D) Both K206 and K212 occur in low-confidence regions of the AlphaFold model. The Cα trace of the AlphaFold-predicted ZBTB7B structure (AF-O15156-F1, version 6) is shown, with the candidate sites highlighted in purple. Local prediction confidence, expressed as pLDDT, is 32.28 for K206 and 25.09 for K212. Accordingly, this view illustrates site positions within the predicted model but does not establish their actual local conformations, solvent accessibility, or functional roles. Lactyl groups were not incorporated into the model.

(E) Mass-spectrometric analysis assigns high modification-localization probabilities to the two candidate sites. The heatmap displays site-localization probabilities across four mass-spectrometry runs. Values for K206 are 0.8960, 0.9999, 1.0000, and 0.9996, whereas all four values for K212 are displayed as 1.0000, rounded to four decimal places. Increasing color intensity indicates increasing localization probability on a scale from 0 to 1. Run numbering follows the original export order and does not denote four independent donors.

Concordance between model predictions and mass-spectrometric records supports candidate prioritization rather than independent functional validation. Site-localization probability does not measure lactylation abundance or modification occupancy.