Briefings in Bioinformatics, 2026, 27, bbag339
https://doi.org/10.1093/bib/bbag339
Published: 24 June 2026
Problem Solving Protocol

OXFORD

scDeepAPA: a deep learning framework for single-cell alternative polyadenylation identification

Jialu Liang[1,‡], Qing Wang[1,‡], Sen Guo[1], Wei Zhang[2], Mingyi Xie[3,4,5], Qianqian Song[1,*]

[1]Department of Health Outcomes and Biomedical Informatics, University of Florida, 1889 Museum Rd, Suite 7000, Gainesville, FL 32611, United States
[2]Department of Computer Science, University of Central Florida, 4328 Scorpius St. Building 116, Room 246, Orlando, FL 32816, United States
[3]Department of Biochemistry and Molecular Biology, University of Florida, 1200 Newell Drive, Gainesville, FL 32610, United States
[4]UF Health Cancer Center, University of Florida, Gainesville, FL 32610, United States
[5]UF Genetics Institute, University of Florida, 2033 Mowry Road, Gainesville, FL 32610, United States

[*]Corresponding author. Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL 32611, USA. E-mail: qsong1@ufl.edu
[‡]The authors contributed equally to this work.

Abstract

Alternative polyadenylation (APA) is a widespread post-transcriptional regulatory mechanism that diversifies transcript isoforms and modulates mRNA stability, localization, and translation. Although single-cell RNA sequencing (scRNA-seq) provides an unprecedented opportunity to study cell-type-specific APA dynamics, existing computational tools are largely designed for bulk RNA-seq data or rely heavily on gene annotations, limiting their applicability to single-cell contexts. Here, we present scDeepAPA, a deep learning framework specifically optimized for scRNA-seq data to enable accurate polyadenylation site (PAS) detection, isoform quantification, and functional interpretation of APA events at single-cell resolution. Trained on high-confidence annotations from PolyASite v3.0, scDeepAPA integrates convolutional feature extraction with Mamba-based state-space modeling and bidirectional LSTM layers to capture both long-range and local sequence dependencies. Comprehensive benchmarking against five state-of-the-art PAS prediction models demonstrates that scDeepAPA consistently achieves superior performance across accuracy, F1 score, and area under the receiver operating characteristic metrics in both human and mouse datasets. Applying scDeepAPA to Alzheimer's disease mouse brain data revealed widespread, cell-type-specific APA remodeling across immune and glial populations, including shifts toward proximal PAS usage and 3′ UTR shortening. In KRAS-mutant small cell lung cancer, scDeepAPA uncovered global proximal PAS activation and tumor-specific intronic polyadenylation events. Notably, several intronic APA events generated truncated transcripts encoding predicted neoantigenic peptides with strong major histocompatibility complex class I binding affinity, supported by structural modeling and tumor-specific expression patterns. By enabling accurate PAS identification and quantitative APA profiling, scDeepAPA facilitates in-depth downstream analyses of regulatory mechanisms and immunogenic consequences in single-cell transcriptomics, advancing the understanding of post-transcriptional regulation in neurodegeneration and cancer.

Keywords single-cell RNA sequencing, alternative polyadenylation (APA), polyadenylation site (PAS) prediction, deep learning, mamba architecture, bidirectional LSTM, post-transcriptional regulation, neoantigen discovery, immunogenomics

Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized transcriptomics by enabling high-resolution gene expression profiling at the resolution of individual cells. This technique allows the simultaneous processing of thousands of cells, offering a comprehensive view of cellular heterogeneity within complex tissues that bulk RNA-seq approaches fail to capture [1, 2]. By isolating and analyzing each cell independently, scRNA-seq uncovers distinct transcriptional states, rare cell populations, and developmental trajectories. A variety of scRNA-seq platforms have been developed to achieve scalable single-cell transcriptome profiling, including 10× Chromium [3], Drop-seq [4], inDrop [5], Microwell-seq [6], Smart-seq2 [7], sci-RNA-seq [8], and other droplet based [9] or microwell based [10] technologies. Despite differences in experimental implementation, most widely adopted scRNA-seq protocols rely on oligo(dT)-based priming to capture polyadenylated RNA molecules. By leveraging poly(A)-primed capture strategies together with single-cell resolution, scRNA-seq protocols are cost-effective and technically streamlined, while retaining sensitivity to biologically meaningful post-transcriptional processes. As a result, they are particularly well-suited for studying

Received: March 4, 2026. **Revised:** April 30, 2026. **Accepted:** May 30, 2026

cleavage and polyadenylation (CPA), a fundamental regulatory mechanism governing transcript termination and isoform diversity.

CPA enables alternative polyadenylation (APA), encompassing both 3′ UTR APA and intronic polyadenylation (IPA), thereby generating transcript isoforms with variable 3′ untranslated regions (3′ UTRs) or prematurely terminated coding sequences. These isoforms can differ in mRNA stability, subcellular localization, and translational efficiency. Dysregulated APA has been implicated in a variety of diseases, including cancer, immune disorders, and neurodegenerative conditions [11]. Beyond altering transcript length, APA can modulate cis-regulatory elements, such as microRNA (miRNA) [12] and RNA-binding protein (RBP) binding sites [13], thereby influencing post-transcriptional gene regulation. For example, *INSIG1* undergoes APA-mediated 3′ UTR shortening, which removes miRNA-binding sites, such as those for miR-122, thereby alleviating post-transcriptional repression and promoting oncogenic expression [14]. RBPs like HuD (*ELAVL4*) also regulate APA by interacting with the polyadenylation machinery in cancer samples [15]. Moreover, APA can generate tumor-specific isoforms encoding neoantigenic peptides, such as those derived from HCC1428, which enhance tumor immune evasion by altering major histocompatibility complex class I (MHC-I) presentation [16]. These findings highlight the mechanistic complexity and biological significance of APA, particularly in disease settings.

The inherent features of scRNA-seq make it an ideal platform for uncovering cell-type-specific APA dynamics and their regulatory consequences. However, computational methods for APA analysis remain limited. Existing approaches for APA analysis can be broadly categorized into annotation-based methods such as QAPA [17] and deep learning-based methods (e.g. SCAPTURE [18], SANPolyA [19], DeepGSR [20], DeepPolyA [21], and DeeReCT [22]). Annotation-based methods rely on existing gene annotations and read coverage to infer APA events, but they often miss unannotated or cell-type-specific poly(A) sites, particularly in single-cell data. In contrast, deep learning models have improved PAS prediction by learning sequence-level features from large annotated datasets. For example, DeepGSR [20] and DeepPolyA [21] use CNN or CNN-RNN architectures to classify PAS, but they are primarily trained on bulk RNA-seq data. DeeReCT incorporates both sequence and structural features for better prediction accuracy, while SANPolyA [19] leverages self-attention for modeling long-range dependencies, yet both remain limited by bulk-derived training data. SCAPTURE [18] is among the few models designed for 3′-tag scRNA-seq and supports cell-type specific APA analysis, but still relies heavily on existing annotations and lacks integration with functional regulatory analyses. In addition, several computational methods [23–26], such as scAPA [27], have been specifically developed for APA analysis in single-cell RNA-seq data. These methods enable the identification and quantification of APA events at cell-type or single-cell resolution, typically relying on peak-calling strategies, statistical modeling, or the adaptation of bulk RNA-seq frameworks to sparse single-cell data. While these approaches have significantly advanced the study of APA dynamics in heterogeneous cell populations, many of them focus on peak-level inference or APA usage estimation, rather than precise nucleotide-level PAS localization within candidate regions. While these models offer strong predictive power, their limited adaptability to single-cell contexts and lack of downstream regulatory insights highlight the need for more comprehensive, scRNA-seq-aware approaches.

To overcome these limitations, we developed scDeepAPA, a deep learning model specifically trained on cleavage-enriched 3′ tag-based data for the accurate detection and characterization of APA sites in scRNA-seq data. scDeepAPA leverages the recent release of PolyASite v3.0 [28], which catalogs over 18 million high-confidence PASs across multiple species, providing a significantly enhanced resource for model training and APA discovery. scDeepAPA employs a hybrid architecture that integrates Mamba layers [29] for efficient long-range dependency modeling with bidirectional LSTM layers [30] to capture the local sequential context within polyadenylation sequences. The model takes as input fixed-length genomic sequences centered on putative polyadenylation sites (PASs) and outputs the probability of cleavage activity. It is trained on high-confidence PAS annotations from the PolyASite v3.0 database and is optimized for high-resolution modeling of cleavage activity in single-cell transcriptomic data. scDeepAPA enables the precise identification, quantification, and regulatory interpretation of APA dynamics across individual cells and cell types. A list of abbreviations used in this study is provided in Supplementary Table S1. The model is publicly available on GitHub (https://github.com/QSong-github/scDeepAPA).

Materials and methods

Training data construction

To construct the training dataset for scDeepAPA, we first extracted annotated PAS positions from the PolyASite v3.0 database [28], which served as true PAS (positive) samples. For each annotated PAS, a 200-bp positive sequence was generated by extracting 100 nucleotides upstream and 100 nucleotides downstream of the site. To balance the dataset, an equal number of negative samples representing non-PAS regions were included. Negative sequences were randomly sampled from intergenic and non-PAS genomic regions and were explicitly filtered to ensure no overlap with any known PAS annotations. To preserve genome-wide representativeness, negative samples were proportionally distributed across chromosomes according to chromosome length. Together, these positive (true PAS) and negative (non-PAS) sequences constituted the training dataset used to train the scDeepAPA model.

Model overview

The structure of scDeepAPA is illustrated in Fig. 1. The model is composed of two functionally distinct but complementary modules: (i) a feature embedding module, which encodes raw nucleotide sequences into enriched local representations, and (ii) a sequence analysis module, which models long-range and contextual dependencies required for accurate PAS recognition. This modular design reflects the biological characteristics of PAS signals, which are defined by both short local motifs (e.g. polyadenylation signals) and a broader sequence context extending hundreds of nucleotides.

Feature embedding module

Given that previous methods encode the initial sequence as a simple one-hot vector, we employ the “Token2Embedding” method [31], widely used in large language models, to enhance the richness of the sequence representation. Specifically, a unique token is assigned to each of the four nucleotide bases, “A,” “G,” “C,” and “T,” as well as to the unknown element “N,” forming a dictionary {A: 0, G: 1, C: 2, T: 3, N: 4}. Assuming the input sequence is $S \in R^{200}$, we perform initial embedding on the tokenized sequence S through the

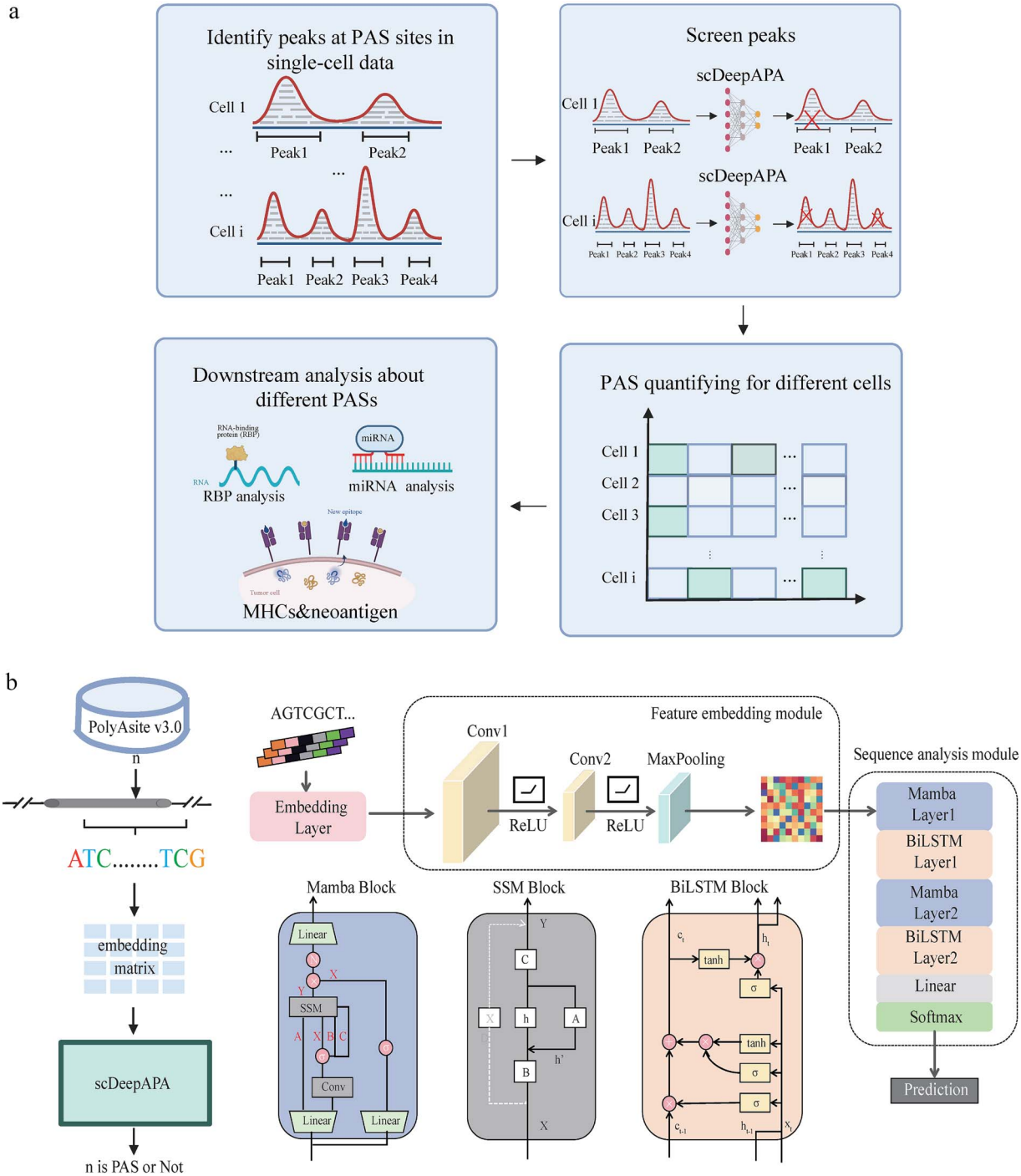

Figure 1 Overview of the scDeepAPA framework for single-cell PAS discovery and regulatory analysis. (a) Schematic illustration of the scDeepAPA framework for accurate single-cell PAS detection, quantification, and downstream regulatory analyses. Top left: Candidate PAS peaks are initially detected as poly(A) enriched 3′ end signals from scRNA-seq data across diverse cell types using genomic analysis tools. Top right: scDeepAPA applies a deep learning approach to evaluate candidate peaks and selects high-confidence PAS sites for downstream analyses. Bottom right: Reads mapped to high-confidence PAS peaks are quantified to calculate RE values across different cellular populations, enabling detailed exploration of APA dynamics. Bottom left: PAS RE values are integrated with RBP expression, miRNA targeting profiles, and neoantigen predictions to systematically dissect regulatory signals associated with APA. (b) Overview of the scDeepAPA model architecture. The model integrates convolutional layers for local sequence feature extraction with Mamba and BiLSTM modules to capture long-range dependencies and bidirectional sequence context, enabling accurate PAS prediction.

embedding layer to obtain $S_0 \in R^{200\times5}$. Then we use two layers of 1D convolution neural network [19] and the ReLU activation function for feature extraction. We set the input channel of the first convolution layer to 5, the output channel to 128, the kernel size to 12, and the padding size to 5. The parameters of the second convolutional network are in_channels = 128, out_channels = 64, kernel_size = 6, and padding_size = 2. The formulas are:

$$S_1 = \mathbf{ReLU}\left(\mathbf{Conv}_1\left(S_0\right)\right), \tag{1}$$

$$S_2 = \mathbf{ReLU}\left(\mathbf{Conv}_2\left(S_1\right)\right), \tag{2}$$

where $S_1 \in R^{128\times199}$ is the output of the first layer of convolutional neural network (CNN) and the input of the second layer of CNN and $S_2 \in R^{64\times198}$ is the output of the second layer of CNN. Then we set a max pooling layer with kernel size of 4 and stride of 4. The pooling layer is represented as follows:

$$S_3 = \mathbf{MaxPool}\left(S_2\right), \tag{3}$$

where $S_3 \in R^{64\times49}$ is the output of the max pooling layer.

Sequence analysis module

Mainstream sequence analysis models include LSTM and Transformer. However, LSTM has limited contextual perception, and while the Transformer model offers a robust structure, it does not perform well with sparse vocabularies (such as a nucleotide vocabulary of only five tokens). To address these limitations, we incorporated Mamba blocks, which employ state-space modeling (SSM) to efficiently capture long-range dependencies across genomic sequences. Specifically, we designed the sequence analysis module by cross-stacking two layers each of the Mamba and BiLSTM models, with the output of each layer feeding into the input of the next.

$$S_4 = \mathbf{Mamba}_1\left(S_3^T\right), \tag{4}$$

$$S_5 = \mathbf{BiLSTM}_1\left(S_4\right), \tag{5}$$

$$S_6 = \mathbf{Mamba}_2\left(S_5\right), \tag{6}$$

$$S_7 = \mathbf{BiLSTM}_2\left(S_6\right), \tag{7}$$

where $S_3^T \in R^{49\times64}$ is the transposed output of the pooling layer serving as the input to the $\mathbf{Mamba}_1(\cdot)$ block. Both $\mathbf{Mamba}_1(\cdot)$ and $\mathbf{Mamba}_2(\cdot)$ have $d_{\text{model\_output}}$=64, yielding outputs $S_4 \in R^{49\times64}$ and $S_6 \in R^{49\times64}$ respectively. $\mathbf{BiLSTM}_1(\cdot)$ has input size 64 and hidden size 32 per direction, producing $S_5 \in R^{49\times64}$. $\mathbf{BiLSTM}_2(\cdot)$ has input size 64 and hidden size 128 per direction, producing $S_7 \in R^{49\times256}$.

Specifically, for the $\mathbf{Mamba}_1(\cdot)$ block, given input $S_3^T \in R^{49\times64}$, a single linear projection maps it to a combined representation:

$$P = S_3^T \cdot w_{in} + b_{in}, P \in R^{49\times386}, \tag{8}$$

where $w_{in} \in R^{64\times386}$ is the weight matrix and $b_{in} \in R^{49\times386}$ is bias. Note here the $d_{in} = 2d_{\text{inner}} + 2d_{\text{state}} + n_{\text{heads}} = 386$, with $d_{\text{inner}} = 128$ (expansion factor 2), $d_{\text{state}} = 64$, and $n_{\text{heads}} = 2$. P is then split along the last dimension into three components:

$$[\mathbf{z}, \mathbf{xBC}, \mathbf{dt}] = \mathbf{split}(P), \tag{9}$$

where $\mathbf{z} \in R^{49\times128}$ serves as a gating signal, $\mathbf{xBC} \in R^{49\times256}$ carries the combined state projection, and $\mathbf{dt} \in R^{49\times2}$ is the time-step scaling factor processed through softplus activation. **xBC** is passed through a depthwise Conv1d layer (kernel_size = 4) followed by SiLU activation:

$$\mathbf{xBC}' = \mathbf{SiLU}\left(\mathbf{Conv1d}\left(\mathbf{xBC}\right)\right), \mathbf{xBC}' \in R^{49\times256}, \tag{10}$$

$$\mathbf{SiLU}(s) = s \cdot \frac{1}{1+e^{-s}}, \tag{11}$$

xBC′ is then split into three components:

$$[x, B, C] = \mathbf{split}\left(\mathbf{xBC}'\right), \tag{12}$$

where $\mathbf{x} \in R^{49\times128}$ is the input signal, $\mathbf{B} \in R^{49\times64}$ is the input projection matrix, and $\mathbf{C} \in R^{49\times64}$ is the output projection matrix of the SSM. x is further reshaped to $\mathbf{x} \in R^{49\times2\times64}$ to accommodate the multi-head structure with $n_{\text{heads}} = 2$ and head dimension is 64. The SSM ($h_0 = 0 \in R^{2\times64\times64}$) update process in timestep t is:

$$h_t = A \cdot h_{t-1} + B \cdot x_t, \tag{13}$$

$$h_t = \mathbf{exp}\left(\mathbf{dt} \cdot A\right) \cdot h_{t-1} + \mathbf{dt} \cdot B \cdot \left(\mathbf{dt} \cdot x_t\right), \tag{14}$$

$$y_t = C \cdot h_t + D \cdot x_t, \tag{15}$$

where $h_t \in R^{2\times64\times64}$ is the hidden state with shape (n_{heads}, head dimension, d_{state}), $A \in R^2$ is the state transition parameter (learnable) in log-parameterized form, $D \in R^2$ is a learnable skip connection parameter, and $y_t \in R^{49\times2\times64}$ is the raw SSM output. y_t is reshaped to $\in R^{49\times128}$ and gated with z through gated root mean square layer normalization (RMSNorm):

$$\mathbf{y}' = y_t \cdot \mathbf{SiLU}(z), \ \ \mathbf{y}' \in R^{49\times128}, \tag{16}$$

$$\mathbf{y}'' = \mathbf{y}' \cdot \frac{\mathbf{w}}{\sqrt{\mathbf{mean}\left(y'^2\right) + \varepsilon}}, \quad \mathbf{y}'' \in R^{49\times128}, \tag{17}$$

where $\mathbf{w}$ is a learnable weight and ε is a small constant for numerical stability. Finally, a linear layer maps $\mathbf{y}''$ back to d_model:

$$S_4 = \mathbf{y}'' \cdot w_{\text{out}} + b_{\text{out}}, \ \ S_4 \in R^{49\times64}, \tag{18}$$

After passing through the cross-stacked Mamba and BiLSTM layers, the final output is $S_7 \in R^{49\times256}$. We take the feature at the last sequence position:

$$S_T = S_7\left[-1, :\right], \ \ S_T \in R^{1\times256}, \tag{19}$$

and pass it through a fully connected layer followed by a softmax function to obtain the prediction:

$$\mathbf{logits} = \mathbf{Softmax}\left(\mathbf{FC}\left(S_T\right)\right), \quad \mathbf{logits} \in R^{1\times2}. \tag{20}$$

We use BCEWithLogitsLoss as the loss function:

$$\mathbf{BCEWithLogitsLoss}\left(\mathbf{logits}, \mathbf{gt}\right) = -\left(\mathbf{gt} \cdot \log\left(\frac{1}{1+e^{-\mathbf{logits}}}\right) + \left(1-\mathbf{gt}\right) \cdot \log\left(1 - \frac{1}{1+e^{-\mathbf{logits}}}\right)\right), \tag{21}$$

where $\mathbf{gt}$ is the ground-truth binary label.

The training process is conducted on a single A100 from NVIDIA DGX A100 640GB System. The batch size is set to 300, the learning rate is 0.0001, and the optimizer is AdamW.

Performance evaluation and benchmarking

We performed a systematic benchmark evaluation of scDeepAPA with five existing polyA site prediction tools (SANPolyA [19], DeeReCT [22], DeepPolyA [21], DeepGSR [20], and SCAPTURE [18]) using experimentally validated PASs from PolyASite v3.0 as ground truth. A 10-fold cross-validation strategy was adopted, and model performance was assessed using five standard metrics: Accuracy, F1-score, area under the receiver operating characteristic (AUROC), Precision, and Recall.

The datasets were constructed from experimentally validated PAS-centered sequences derived from PolyASite v3.0 annotations and were randomly split into training and test sets at an 80%/20% ratio. The consistency of chromosome-level distributions across training and validation splits is further evaluated in Supplementary Note S1 (Supplementary Fig. S1 and Supplementary Table S2). Detailed information on dataset composition and sample sizes is provided in the Zenodo repository associated with this study (Zenodo under the DOI 10.5281/zenodo.15066940).

To ensure a fair and consistent comparison, all methods were evaluated using a unified input representation, in which each candidate PAS was encoded using a fixed-length sequence window spanning ±100 bp around the site. This standardized input design enables direct comparison of predictive performance across different models. Among the benchmarked methods, SANPolyA, DeeReCT, DeepPolyA, and our model natively adopt a fixed-length sequence input, and therefore required no architectural modification under this setting. DeepGSR originally proposes two alternative convolutional architectures, including a 1D-CNN and a 2D-CNN. In this study, we employed the 1D-CNN implementation of DeepGSR, which operates on linear genomic sequences and is directly compatible with the unified input format. For SCAPTURE, we retained the original DeepPASS model architecture during benchmarking and applied it to the same standardized ±100 bp sequence inputs, thereby ensuring comparability while preserving the original model design. All models were also retrained using the same datasets, preprocessing procedures, and cross-validation splits, detailed implementation is provided in Supplementary Note S2.

Identification of 3′ UTR regions

To define candidate regions for PAS identification, we focused primarily on 3′ UTRs and intronic regions, which are known to be enriched for APA events. Based on GENCODE annotations for human and mouse genes, and following the strategy proposed by Miura *et al.* [32], we used the gencode_regions tool (https://github.com/saketkc/gencode_regions) to extract standardized coordinates for annotated 3′ UTRs and introns. In cases where multiple 3′ UTRs were annotated for a single gene (e.g. across different transcripts), we retained all of them to comprehensively cover potential polyadenylation signals and avoid missing genuine PASs.

PAS detection

To detect PASs, we processed the raw scRNA-seq data using the Cell Ranger pipeline to generate alignment files (BAM) and gene expression matrices [3]. From the BAM files, we extracted valid reads that contained both Unique Molecular Identifier (UB) and cell barcode (CB) tags. Polymerase chain reaction (PCR) duplicates were removed using UMI-tools to ensure that only reads derived from unique molecular events were retained. The deduplicated reads were then used for subsequent peak calling [33].

To identify candidate PAS peaks, we utilized the makeTagDirectory and findPeaks tools from the HOMER suite, with parameters specifically optimized for single-cell sequencing data (peak size = 50 bp, fragment length = 100 bp, strand-specific mode enabled, and minimum distance between peaks = 1 bp) [34]. Due to the high variability and noise inherent in single-cell data, especially for low-abundance transcripts, an initial large number of candidate peaks were detected, including many potential false positives. To improve robustness against sparsity in single-cell data, peak detection was performed using an aggregated signal across cells rather than relying on individual cells alone.

To distinguish true PASs from artifacts such as internal priming events in A-rich regions, we applied our deep learning model scDeepAPA to evaluate the local sequence context surrounding each candidate peak to predict its authenticity. Only peaks predicted as label 1 by scDeepAPA were retained as high-confidence PASs and used for downstream analysis. Finally, for 3′ UTR-associated APA analysis, high-confidence PASs located within annotated 3′ UTRs were further categorized into proximal and distal sites. The PAS closest to the midpoint of the 3′ UTR was designated as the proximal site (cUTR), while downstream PASs were classified as distal(aUTR). IPA events shown in results were identified from intronic regions annotated in the reference genome. For introns exhibiting detectable 3′ end cleavage signals, scDeepAPA was applied to model the local polyadenylation landscape and to pinpoint the most probable cleavage site, which was subsequently defined as the IPA site for downstream analyses, as described in the Methods section. For high-confidence PASs located in intronic regions, the sites were retained but not further categorized into proximal or distal classes.

PAS usage quantification

To analyze APA dynamics in single cells, we quantified the usage of PAS by calculating the read enrichment (RE) value for each PAS. Following a method inspired by APAlyzer [35] but adapted to our customized PAS annotation, we defined RE as the log-ratio of read counts between the distal (aUTR) and proximal (cUTR) regions:

$$\mathrm{RE} = \log_2\left(\frac{\mathrm{RD}_{\mathrm{aUTR}} + 1}{\mathrm{RD}_{\mathrm{cUTR}} + 1}\right)$$

where $\mathrm{RD}_{\mathrm{aUTR}}$ and $\mathrm{RD}_{\mathrm{cUTR}}$ represent the read counts mapped to the distal and proximal regions, respectively.

For cell type annotation, single-cell expression matrices were first normalized using Seurat [36]. In Alzheimer's disease (AD) samples, cell type identities were assigned using SingleR [37], whereas in lung cancer samples, scType was used to classify cells based on canonical marker gene expression [31]. These annotations served as the foundation for downstream analyses of cell type specific PAS usage.

To accurately quantify RE at PAS regions, we first used featureCounts to assign raw sequencing reads to updated PAS annotation intervals. Based on these assignments, we then applied UMI-tools to count the number of reads at both the proximal (cUTR) and distal (aUTR) PAS regions for each individual cell. RE values are computed

at the single-cell level, downstream APA analyses are primarily presented at the cell-type or population level to ensure robustness against sparsity and dropout effects inherent to scRNA-seq data.

APA-centered regulatory analysis

To identify differentially expressed genes (DEGs) and APA usage events between biological conditions (e.g. sample groups or cell types), we first preprocessed both the RNA expression matrix and the regional expression (RE) matrix. For gene expression analysis, raw count matrices from two groups were merged and normalized using standard single-cell RNA-seq workflows. Differential expression analysis was performed using the FindMarkers function from the Seurat R package. Genes with an adjusted *P*-value < .05 were considered significantly differentially expressed and were subsequently used for downstream regulatory analyses.

For APA dynamics, we first constructed the RE matrix, where each entry represents the relative expression level of a PAS by calculating RE values based on the normalized read densities in the proximal and distal regions surrounding each annotated PAS regions. Specifically, RE values were derived by quantifying the normalized number of reads mapped to defined upstream (proximal) and downstream (distal) regions adjacent to each candidate PAS, as previously described. Using this RE matrix, we performed differential analysis with the FindMarkers function from the Seurat R package, comparing RE values between different biological conditions (e.g. tumor versus normal, or cell type comparisons). PAS regions exhibiting statistically significant changes in RE values (adjusted *P*-value < .05) were identified and classified as differentially regulated APA sites. These significantly altered APA regions were subsequently used for integrative regulatory analyses.

To investigate the potential role of RBPs in APA regulation, we performed correlation analysis between RBP expression levels and RE values in both human and mouse datasets. RBP–target gene annotations for human samples were obtained from the starBase v2.0 database [38], and annotations for mouse samples were sourced from the EuRBPDB database [39]. RBP-target gene pairs were extracted and matched with normalized RNA expression matrices and RE matrices. For each RBP-target gene pair, Pearson correlation coefficients were calculated between the expression level of the RBP and the RE value of the target. Only significantly correlated pairs, defined as those with *P*-values < .05, were retained for further analysis.

At the miRNA level, we annotated predicted miRNA-binding sites near candidate PAS regions using data from the TargetScan database for human [40]. The genomic coordinates of the predicted miRNA target sites were converted to the appropriate reference genome build using the UCSC liftOver tool [41]. Among these annotations, only miRNA-binding sites located within the distal region downstream of a defined PAS were labeled as "miRNA related." This classification is based on the biological rationale that if a proximal PAS is preferentially utilized due to APA regulation, the downstream distal region would not be transcribed, thereby preventing the corresponding miRNA from binding to its target site and exerting regulatory effects. Therefore, only miRNA-binding sites located in distal regions were considered potentially influenced by APA usage shifts. Subsequent correlation analysis was performed between RE values and gene expression levels for these miRNA target genes, excluding RE values equal to zero. As in the RBP analysis, only significantly correlated gene-RE pairs ($P < .05$ and $|r| > 0.5$) were retained for further regulatory inference.

Finally, significantly correlated RBP-RE pairs, significantly correlated miRNA-RE pairs, DEGs, and differentially regulated APA sites were each categorized into separate sets. We then performed intersection analysis across these four regulatory layers to identify overlapping genes and APA events potentially influenced by multiple regulatory mechanisms. Visualization of these complex overlaps was carried out using ComplexUpset [42], a package specifically designed for intuitive and detailed depiction of multi-set intersections.

Identification of APA-derived neoantigens

Tumor cells of interest were first defined based on cell-type annotations, focusing on pulmonary alveolar type II (AT2)-like malignant cells. scRNA-seq reads corresponding to these cells were extracted from tumor samples according to their cell barcodes and converted from BAM to FASTQ format using bamtofastq. Reads originating from the same sample were then combined at the read-processing stage to generate a sample-level dataset with sufficient coverage for downstream analyses. To enable accurate MHC-I typing, reads mapping to chromosome 6, which harbors the HLA loci, were extracted using samtools. Low-quality and unmapped reads were removed by alignment to the reference genome with bwa mem, followed by additional filtering steps. The resulting high-quality FASTQ files were subsequently used for HLA genotyping with OptiType, allowing reliable inference of sample-specific MHC-I alleles. For comparison with normal lung tissue, AT2 and AT1 epithelial cells were selected from control samples to match the epithelial lineage of the tumor population. Reads from these normal epithelial cells were processed using the same workflow to ensure consistency in downstream IPA and neoantigen analyses.

Following HLA genotyping, the processed FASTQ files were aligned to the hg38 human reference genome using HISAT2 with default parameters. The resulting BAM files were sorted and indexed with samtools and used as input for subsequent IPA event prediction. Rather than relying on predefined IPA annotations, we systematically scanned both KRAS samples and Control samples for RE signals within annotated intronic regions. Exon–intron boundaries were defined according to RefSeq gene annotations. Intronic regions exhibiting localized read accumulation were identified and processed using the same peak detection and filtering strategy described in the PAS detection module, enabling precise extraction of candidate cleavage peak regions. These intronic peak regions were then subjected to scDeepAPA-based prediction to identify high-confidence PASs independently in KRAS and Control samples. To ensure tumor specificity, IPA-associated PAS detected in both KRAS and Control samples were excluded from downstream analyses. Novel transcript isoforms were reconstructed based on the retained KRAS-specific IPA events, with particular emphasis on identifying the first in-frame stop codon upstream of each IPA site. Transcripts containing such premature stop codons were classified as truncated protein-coding isoforms.

Finally, coding sequences were extracted and translated into amino acid sequences using gffread, and peptides of 8–11 amino acids containing at least one intron-derived residue were selected to match the typical binding affinity range of MHC-I molecules. The binding affinities of the candidate peptides to the inferred MHC-I alleles were predicted using NetMHCpan v4.0, and peptides with predicted percentile ranks below 2% were retained as candidate neoantigens [43]. To minimize the occurrence of false positives, peptides matching sequences present in the UniProt human reference proteome and RefSeq database were excluded. Through this rigorous workflow, a

final set of tumor-specific IPA-derived neoantigens with high predicted MHC-I binding affinity was identified. To visualize the interaction between predicted neoantigens and MHC molecules, we used AlphaFold2-Multimer to model the binding structures using the amino acid sequences of each neoantigen and its corresponding MHC-I allele [44]. The predicted complex structures were then visualized using PyMOL, allowing us to assess structural feasibility and potential binding interfaces [45].

Results

Overview of the scDeepAPA framework

To accurately detect and quantify APA events in single-cell RNA-seq data, we developed scDeepAPA, a deep learning-based framework that integrates peak detection, PAS classification, isoform quantification, and downstream functional exploration. As illustrated in Fig. 1a, scDeepAPA begins by identifying read enriched peaks within 3′ UTR from single-cell RNA-seq data at the individual cell level. These candidate peaks are then screened by scDeepAPA to predict whether each represents a true PAS. Sites identified as true PASs are further quantified across cells by computing RE values [35] within 3′ UTR regions upstream and downstream of the PAS, resulting in a PAS-by-cell expression matrix that captures APA usage. This output supports a range of downstream functional analyses, including association studies with post-transcriptional regulatory elements such as miRNAs and RBPs, as well as neoantigen detection through MHC-binding prediction. These analyses facilitate the interpretation of APA events in both regulatory and immunogenomic contexts.

Figure 1b details the architecture of the scDeepAPA model. Input sequences centered around candidate PAS sites (±100 nt) are first embedded and passed through a feature embedding module, consisting of two convolutional layers followed by max pooling to extract local sequence features. These embeddings are then fed into a sequence analysis module consisting of stacked Mamba and BiLSTM blocks. The Mamba blocks leverage SSMs to capture long-range dependencies over kilobase-scale sequence contexts, complemented by BiLSTM blocks that model bidirectional contextual information by processing sequences in both forward and reverse directions. The output is passed through a final linear-softmax layer for binary classification of PAS versus non-PAS. The model is trained using labeled genomic sequences from PolyASite v3.0, which provides 1 750 661 high-confidence PASs in mouse and 18 432 135 in human, and is tailorly designed for 3′ tag-based scRNA-seq data. The detailed procedures for PAS screening, RE values [35], and regulatory annotation are organized into a streamlined workflow, as shown in Supplementary Fig. S2. This illustrates the modularity and comprehensiveness of the scDeepAPA framework, which supports systematic investigation of APA events and regulation across diverse cell types and pathological conditions.

Performance comparison of scDeepAPA and existing poly(A) site prediction tools

To rigorously assess the predictive capability of scDeepAPA, we conducted a comprehensive benchmarking analysis using high-confidence PAS annotations curated from the PolyASite v3.0 database. We compared scDeepAPA with five representative deep learning-based PAS prediction tools: SCAPTURE, SANPolyA, DeeReCT, DeepPolyA, and DeepGSR. To ensure fairness, all models were re-implemented or retrained under a unified input format using the same training and testing datasets derived from PAS annotations in the PolyASite v3.0 database, including PASs from human and mouse genomes. Further details on data preprocessing and model configuration are provided in the Methods section.

As shown in Fig. 2a and detailed in Supplementary Table S3, scDeepAPA outperformed all baseline models across five standard evaluation metrics: accuracy, F1 score, AUROC, precision, and recall. In the human dataset, scDeepAPA achieved the highest F1 score of 0.839 ± 0.0012, substantially surpassing DeeReCT (0.7400 ± 0.0014) and DeepPolyA (0.797 ± 0.0023), with over 4% improvement in F1, reflecting its balanced precision (0.8723 ± 0.0075) and recall (0.8089 ± 0.0076). SANPolyA and DeepGSR showed competitive but lower performance, indicating that scDeepAPA maintains superior classification fidelity even when benchmarked against attention- and CNN-based architectures. In Fig. 2b, in the mouse dataset, scDeepAPA similarly outperformed all baseline models across the same five evaluation metrics. Specifically, scDeepAPA achieved the highest F1 score of 0.849 ± 0.0017, markedly exceeding DeeReCT (0.790 ± 0.0016) and DeepPolyA (0.794 ± 0.0018), representing an improvement of ∼4%–5% in F1 score. This performance gain was supported by a favorable balance between precision (0.868 ± 0.0098) and recall (0.830 ± 0.0116). SANPolyA and DeepGSR again demonstrated competitive yet lower performance, indicating that scDeepAPA maintains robust and consistent classification accuracy across species.

To further evaluate classification effectiveness, we analyzed the area under the ROC curve (AUC values) for each model across human and mouse datasets (Fig. 2c). scDeepAPA achieved an AUC value of 0.8856 ± 0.0030, outperforming SANPolyA (0.8222 ± 0.0026) and DeepGSR (0.8260 ± 0.0036), indicating improved ability to distinguish true PASs from background sequences with high sensitivity and specificity. Similarly, in the mouse dataset (Fig. 2d), scDeepAPA continued to demonstrate superior classification performance, achieving an AUROC of 0.90 ± 0.0041, which was consistently higher than SANPolyA (0.83 ± 0.0019), DeepGSR (0.82 ± 0.0027), DeeReCT (0.82 ± 0.0031), and DeepPolyA (0.82 ± 0.0035). This result indicates that scDeepAPA maintains robust discrimination capability across species, effectively distinguishing true PASs from background sequences with high sensitivity and specificity. All models were also retrained using identical datasets, preprocessing procedures, and cross-validation splits. The results are shown in Supplementary Tables S4 and S5 and Supplementary Fig. S3. Collectively, these results demonstrate that scDeepAPA not only offers state-of-the-art predictive performance across diverse metrics but also generalizes robustly across species and experimental settings, establishing it as a reliable tool for PAS detection in single-cell transcriptomics. We further evaluated model generalization using independent external datasets. scDeepAPA achieves top performance across metrics while maintaining a balanced precision-recall trade-off. Details are provided in Supplementary Note S3 (Supplementary Fig. S4; Supplementary Tables S6 and S7). In addition, the architectural contribution of individual components is further evaluated through ablation analysis, as described in Supplementary Note S4. The corresponding results are presented in Supplementary Fig. S5 and Supplementary Tables S8 and S9.

In addition, we performed several complementary analyses to further validate the robustness and reliability of our evaluation. These include a direct comparison with a representative single-cell APA baseline under shared peak regions to assess PAS localization accuracy (Supplementary Note S5; Supplementary Figs. S6 and S7;

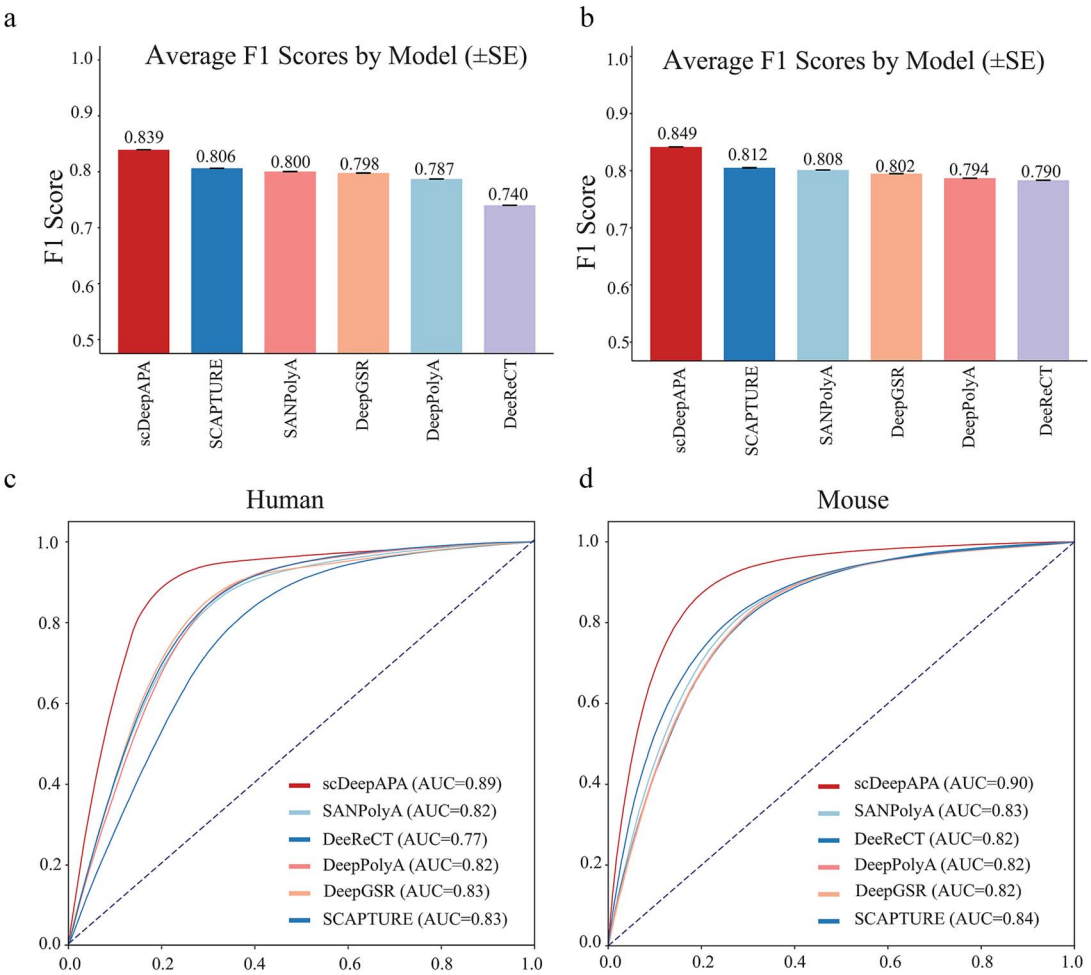

Figure 2 Performance evaluation of the scDeepAPA model. (a) Comparison of predictive performance between scDeepAPA and existing PAS prediction models on human datasets, evaluated using F1 scores. Values represent the mean F1 score across 10-fold cross-validation, with error bars indicating SD. (b) Comparison of predictive performance between scDeepAPA and existing PAS prediction models on mouse datasets, evaluated using F1 scores. Values represent the mean F1 score across 10-fold cross-validation, with error bars indicating SD. (c) ROC curves illustrating model performance on human PAS datasets. AUROC values indicate the sensitivity and specificity of each method. (d) ROC curves illustrating model performance on mouse PAS datasets. AUROC values indicate the sensitivity and specificity of each method.

Supplementary Table S10), and a controlled analysis to disentangle the contributions of upstream peak-calling and downstream PAS prediction (Supplementary Note S6; Supplementary Figs. S8 and S9; Supplementary Table S11). Together, these analyses provide additional support for the robustness, fairness, and methodological validity of the proposed framework.

Investigation of APA alteration in Alzheimer's disease mouse models

To investigate APA alterations in AD, we applied scDeepAPA to the scRNA-seq data from the brain tissues of AD and wild-type (WT) mice [46]. PAS usage was quantified using RE values [35], enabling characterization of APA events at single-cell resolution across immune-related and glial cell populations. UMAP-based clustering guided by canonical cell-type markers identified major cell populations, including microglia, monocytes, B cells, and T cells (Fig. 3a).

Differential PAS usage between AD and WT samples was evident across several immune-related and neural development-associated genes. In Fig. 3b and c, we visualized the read coverage across AD and WT samples within specific cell types, including *Cul3* (microglia), *Itgb7* (B cells), *Tmem263* (monocytes), *Lsp1* (other), and *Ybx1* (T cells), all exhibiting marked differences in PAS usage between AD and WT conditions. Specifically, *Itgb7* displayed increased distal PAS usage in AD B-cells, whereas WT samples showed a preference for proximal sites. *Tmem263* favored distal PAS usage in WT monocytes. Conversely, *Ybx1* and *Cul3* exhibited a prominent shift toward proximal PAS usage in AD, indicative of 3′ UTR shortening and potential loss of post-transcriptional regulatory elements. These genes have established roles in RNA processing, cellular signaling, and immune modulation [47–49], and their APA activities suggests functional implications in the AD brain microenvironment. We next assessed APA usage shifts globally by computing the RE values across immune-related cell types (Fig. 3d). Comparison of AD and WT samples revealed cell type-specific APA alterations. Across cell types in WT and AD, proximal PAS usage exhibited pronounced cell type–specific patterns. Among all examined populations, microglia consistently showed the most left-shifted cumulative distributions of RE, indicating a globally higher preference for proximal PAS usage relative to other immune cell types. In contrast, monocytes, T cells, and B cells displayed comparatively right-shifted

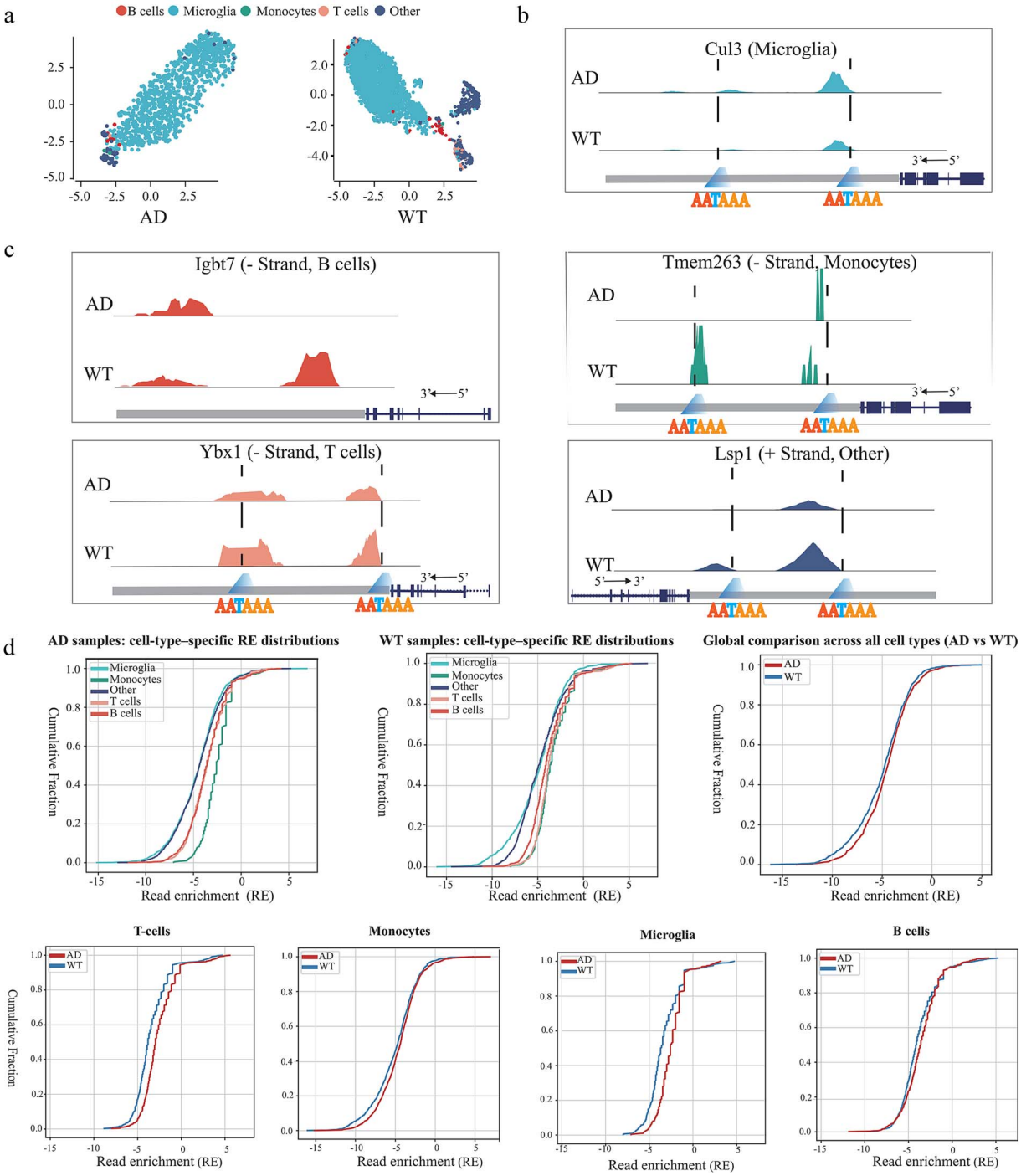

Figure 3 scDeepAPA reveals distinct APA dynamics and associated regulatory mechanisms in AD. (a) UMAP plot of single-cell transcriptomes from AD and WT mouse brain samples. Cells are clustered and colored according to five major cell types. (B-C) Read coverage plots illustrating cell-type-specific differential RE values across AD and WT mouse brain samples. Representative APA events are shown for *Cul3* (microglia), *Igh7* (B cells), *Tmem263* (monocytes), *Ybx1* (T cells), and *Lspl* (other cells). For each gene, 3′ end read coverage is displayed separately for AD and WT samples, highlighting shifts in RE values. Corresponding transcript annotations are shown below each plot. (d) Cumulative distribution functions showing the distribution of RE for PAS events located within 3′ UTRs. The top row, from left to right, shows RE distributions across all cell types for AD, WT, and their overall comparison.The bottom row shows RE distributions for T cells, microglia, monocytes, and B cells, from left to right. AD samples are shown for comparison in each distribution plot.

distributions, consistent with increased distal PAS usage. Within the same cell type, WT and AD samples exhibited clear and systematic distributional differences. When all cell types were pooled, the overall cumulative distribution revealed a modest but consistent rightward shift of AD relative to WT, indicating a global tendency toward more distal PAS usage at the aggregate level [50]. We further analyze WT and AD RE trendency within individual cell types. In microglia, which exhibited the most proximal-biased PAS usage across cell types, AD samples showed a clear distributional left shift relative to WT, indicating altered proximal PAS selection despite the overall proximal preference of this cell type. In monocytes, AD and WT distributions were also separated, with AD displaying a relative shift toward distal PAS usage compared with WT. Similar WT–AD differences were observed in T cells and B cells, although the magnitude of the shift varied between cell types, suggesting heterogeneous APA remodeling across immune populations. Overall, AD samples exhibited a modest but consistent increase in distal PAS usage relative to WT across the analyzed dataset.

Analysis of APA dysregulation in KRAS-mutant small cell lung cancer

To explore APA activties in KRAS-mutant small-cell lung cancer (SCLC), we applied scDeepAPA to the single-cell RNA-seq data from tumor and adjacent normal lung tissues [51, 52] derived from KRAS-mutatnt SCLC cohorts. As shown in Fig. 4a, UMAP clustering identified major lung-resident cell populations in both tumor and normal samples, including alveolar macrophages, pulmonary alveolar types II cells, ciliated cells, and immune system cells. Pulmonary alveolar types II cells in the tumor samples were enriched with tumor cell specific marker expressions.

To further investigate the regulatory landscape of APA events in KRAS-mutant lung cancer, we conducted a multi-layer intersection analysis of APA events that exhibited significant changes in RE values within tumor cells. This analysis incorporated three regulatory layers: DEGs, predicted miRNA target sites, and RBP-binding regions. DEGs were defined as genes showing significant expression differences between KRAS-mutant tumor and control samples (adjusted *P*-value $< .05$). miRNA target sites and RBP-binding regions were identified within APA-associated regions spanning between proximal and distal PASs, capturing regulatory elements potentially affected by APA shifts (details in the Methods section: APA-Centered Regulatory Analysis). As shown in Fig. 4b, the majority of APA events overlapped with at least one regulatory category, with the most prominent overlap observed between APA dynamics and gene expression changes in tumor cells. Additionally, a subset of APA sites was found to intersect with both miRNA and RBP features, suggesting potential multilayered regulatory influences. These findings indicate that in the context of lung tumor cells, APA acts not only as an independent co-transcriptional regulatory mechanism but may also cooperate with other regulatory factors to modulate transcript expression. This highlights APA's potential involvement in the transcriptional regulatory associated with tumor progression.

Prominent APA alterations were observed in KRAS-mutant lung cancer tissues compared to normal lung tissues (Fig. 4c). For pulmonary alveolar type II cells, both *GNL3L* and *TCF12* exhibited distinct APA activities. *GNL3L* displayed enhanced distal PAS usage in tumor tissue, while *TCF12* showed a clear shift toward proximal PAS usage and 3′ UTR shortening in tumor tissue. These two genes are known to be involved in cancer-related pathways, where *GNL3L* contributes to oncogenic activities through the NF-κB pathway [53] whereas *TCF12* has been implicated in promoting tumor progression via TGF-β signaling [54]. For ciliated cells, *C1ORF112* demonstrated increased proximal PAS usage in tumor tissue. *C1ORF112* has been identified as a potential biomarker across multiple tumor types, indicating its broad relevance in cancer biology [55]. For immune cells, *SCYL3* showed a notable switch in PAS usage, from predominantly proximal in normal lung tissues to distal usage in tumor tissues, indicating pontential APA-mediated regulatory changes in immune cell function. Specifically for pulmonary alveolar type II cells, Fig. 4d further presents the RE-based APA usage of *GNL3L*, *TCF12*, and *SOD3*. All three genes exhibited statistically significant shifts in PAS usage between tumor and normal tissues, indicating altered 3′UTR isoform preferences in tumors, indicating specific APA activities associated with tumor progression.

Next, we performed cumulative distribution analysis of proximal polyadenylation usage (RE) across major lung cell populations (Fig. 4e). KRAS (across cell types) summarizes the global RE distribution of KRAS-mutant tumor cells across all cell types, providing an overview of proximal PAS usage at the whole-tumor level. Control (across cell types) panel shows the corresponding global RE distribution in normal control samples, serving as a baseline for comparison. Direct comparison of pooled distributions in Overall Comparison of KRAS versus Control revealed a pronounced leftward shift in KRAS-mutant samples relative to controls, indicating a global increase in proximal PAS usage associated with oncogenic KRAS activation. Cell type-specific comparisons further illustrated this trend. In Pulmonary alveolar type II cells, KRAS-mutant samples exhibited a strong shift toward proximal PAS usage compared with controls, consistent with marked 3′ UTR shortening in this epithelial population. Similar but cell type-dependent leftward shifts were observed in Alveolar macrophages, Ciliated cells, Immune system cells, and Other cells, although the magnitude of the effect varied across lineages. Overall, KRAS-mutant cancer samples displayed a global shift in APA patterns, particularly favoring proximal PAS usage, suggesting that KRAS mutations may broadly affect 3′ UTR selection and mRNA stability regulation across diverse cell types.

Prediction of potential APA-derived neoantigen candidates

APA, particularly those occur in intronic regions, can generate truncated transcripts that encode novel C-terminal peptide sequences not present in canonical protein isoforms. In cancer, these aberrant isoforms represent a potential source of tumor-specific neoantigens that may be presented by MHC-I molecules and recognized by the immune system. Having identified APA activation in KRAS-mutant lung cancer using scDeepAPA, we next sought to evaluate whether certain APA events could give rise to putative neoantigenic peptides with the capacity for MHC-I presentation.

To enable neoantigen prediction and peptide–MHC modeling, reliable human leukocyte antigen (HLA) genotyping was first required. Although scRNA-seq data are not optimized for HLA typing, we assessed the feasibility and quality of MHC-I inference by visualizing read alignment patterns across representative HLA alleles inferred using OptiType [56]. Specifically, we examined six alleles, A*02:07:01, A*11:01:01, B*35:01:01, B*40:02:01, C*04:01:01:01, and C*04:11, which together provided broad allelic coverage. As shown in Fig. 5a, all selected alleles exhibited strong read support, low mismatch rates, and high mapping confidence, suggesting that HLA genotyping is feasible

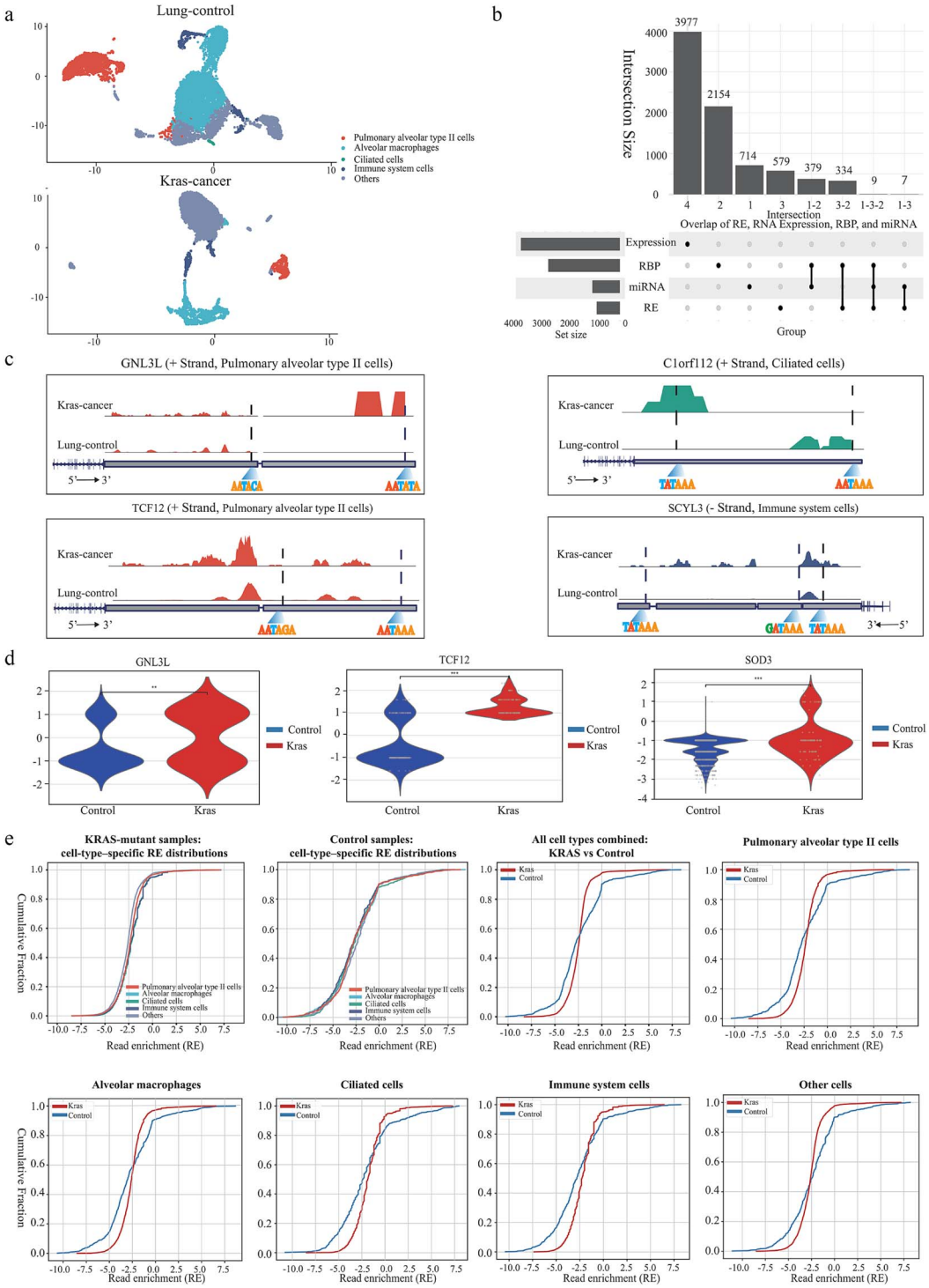

Figure 4 scDeepAPA reveals cell-type-specific APA alterations in KRAS-mutant SCLC. (a) UMAP plot of single-cell transcriptomes from normal lung (top) and KRAS-mutant SCLC tumor (bottom) samples. Cells are clustered and colored by five major cell types. (b) UpSet plot integrating differential RE values, RBP expression, miRNA targeting, and gene expression changes. Intersection analysis identifies candidate genes subject to coordinated APA-related regulatory modulation specific to KRAS-driven tumorigenesis. (c) Read coverage plots illustrating cell-type-specific differential PAS usage across KRAS-mutant lung tumor and matched normal lung samples. APA events are shown for *GNL3L* and *TCF12* (pulmonary alveolar type II cells), *C1orf112* (ciliated cells), and *SCYL3* (immune cells). For each gene, read coverage at the 3′ end is displayed for both KRAS-mutant and normal conditions, highlighting shifts in PAS usage. Annotated transcript structures and predicted polyadenylation signals (AAUAAA motifs) are shown below each plot. (d) Violin plots showing RE for *GNL3L*, *TCF12*, and *SOD3* across lung normal cells and KRAS-mutant tumor cells. Significant shifts in APA usage are observed between non-mutant and KRAS-mutant conditions (Wilcoxon test; $^{***}P < .001$, $^{**}P < .01$, $^{*}P < .05$). (e) CDFs showing the distribution of RE for PAS usage within 3′ UTRs. The top row shows RE distributions across all cell types for KRAS, Control, and their overall comparison, from left to right. The bottom row shows RE distributions for pulmonary alveolar type II cells, Alveolar macrophages, ciliated cells, immune system cells, and other cells, from left to right. KRAS and control samples are displayed for comparison in each distribution plot.

under the current data, although with inherent limitations due to the 3′-end bias of scRNA-seq. Among these, A11:01:01, B35:01:01, and B*40:02:01 were prioritized for downstream analysis based on both reliable sequence coverage and high predicted binding affinity to candidate APA-derived peptides. The full set of inferred HLA genotypes is summarized in Supplementary Table S12.

Building on this foundation, we evaluated the immunogenic potential of APA-derived peptides by selecting representative candidates with strong predicted MHC-I binding affinity (%Rank < 0.1; Supplementary Table S13) and modeling peptide–MHC class I complexes using AlphaFold-Multimer [44]. As shown in Fig. 5b, modeled complexes included peptides derived from *AKAP10* and *SRSF7* bound to HLA-A11:01, MARCHF9 bound to HLA-B35:01, and *HEXD* bound to HLA-B40:02. Structural predictions revealed canonical MHC-I binding geometries, with stable peptide anchoring and favorable interactions within the peptide-binding groove. Notably, the AKAP10 peptide formed clear side-chain interactions with the B and F pockets of HLA-A11:01, while MARCHF9 and HEXD peptides exhibited tight spatial packing within their respective MHC alleles. Although the SRSF7-derived peptide adopted a slightly more superficial configuration, it maintained stable anchoring without steric clashes. Together, these results indicate that those APA-derived peptides not only satisfy sequence-based neoantigen prediction criteria but also adopt structurally favorable conformations for antigen presentation.

Finally, we sought transcript-level evidence supporting the expression of these predicted neoantigens. scDeepAPA coverage analysis revealed tumor-specific activation of IPA events in *HEXD*, *MARCHF9*, *AKAP10*, and *DAPK1*, characterized by sharp read accumulation near intronic PASs in KRAS-mutant tumors but not in matched normal tissues (Fig. 5c). Strand-specific read distributions further matched the expected orientation of APA-induced premature transcription termination: *HEXD* and *MARCHF9* (forward strand) showed proximal enrichment on the left side of the plots, whereas *AKAP10* and *DAPK1* (reverse strand) exhibited proximal peaks on the right. The absence of corresponding signals in normal samples supports the tumor specificity of these truncated isoforms. Collectively, these findings indicate that scDeepAPA can facilitate the identification of APA-derived transcript variants that may give rise to tumor-specific peptide sequences with predicted MHC-I binding potential, supporting its potential application in neoantigen candidate discovery and design. However, we acknowledge that scRNA-seq data, particularly 3′-end sequencing platforms, provide limited coverage of highly polymorphic HLA regions. Therefore, these inferred HLA types should be interpreted as approximate estimates rather than definitive genotypes.

Discussion

In this study, we present scDeepAPA, a deep learning framework specifically designed for accurate PAS detection and APA analysis from scRNA-seq data. Trained on high-confidence PAS annotations from PolyASite v3.0, scDeepAPA enables robust, high-resolution inference of PAS usage at single-cell resolution. Comprehensive benchmarking across human and mouse datasets demonstrates that scDeepAPA consistently outperforms existing PAS prediction methods, exhibiting strong generalization capacity and compatibility with both bulk and single-cell sequencing formats. The competitive performance of scDeepAPA is driven by its hybrid model architecture, which integrates Mamba-based SSM to capture long-range sequence dependencies with bidirectional LSTM layers that encode local contextual features. This combination allows the model to accurately recognize both canonical and non-canonical PAS signals, addressing a major challenge in polyadenylation analysis. Together, these architectural innovations underpin scDeepAPA's high prediction accuracy, robustness across species and platforms, and biological relevance in PAS identification.

Beyond improved predictive performance, scDeepAPA enables the systematic detection and quantification of APA events at single-cell resolution, a level of detail not achievable with bulk RNA-seq-based approaches. scDeepAPA supports cell type-specific analyses and integration with post-transcriptional regulatory features such as RBP expression and miRNA targeting. In AD mouse models, scDeepAPA revealed widespread APA remodeling across immune-related cell populations, characterized by increased proximal PAS usage and global 3′UTR shortening, with T-cells and microglia showing the most pronounced changes. These APA shifts were largely independent of differential gene expression and predicted miRNA or RBP regulation, indicating that APA remodeling represents an additional isoform-level regulatory layer during AD progression. In KRAS-mutant SCLC, scDeepAPA uncovered a global shift toward proximal PAS usage in tumor cells, with genes such as *TCF12* and *GNL3L* exhibiting switching between proximal and distal PASs. Extending beyond canonical APA, we identified IPA events that generate truncated transcripts encoding isoform-specific peptides predicted to bind MHC-I molecules, including *HEXD*, *MARCHF9*, *AKAP10*, and *SRSF7*. Structural modeling and transcript-level coverage analyses further supported tumor-specific expression and peptide–MHC binding, establishing a direct link between aberrant APA, particularly IPA, and the diversification of the tumor immunopeptidome, with potential HLA allele-dependent effects.

Despite these advances, several limitations should be acknowledged. Our disease-focused analyses were performed using publicly available datasets with modest sample sizes, which may not fully capture the heterogeneity of APA regulation across broader patient populations. In addition, while scDeepAPA effectively quantifies PAS usage through peak-centered signal integration, the current framework does not yet achieve base-pair-level resolution of cleavage sites, limiting the precise localization of polyadenylation events. Future work will focus on refining PAS mapping to base-level precision, incorporating genetic variation into the training workflow to assess the impact of sequence variants on APA regulation, and developing subtype-specific models tailored to distinct PAS classes, such as terminal exon versus IPA. Integration with spatial transcriptomics and context-aware learning frameworks represents another promising direction for capturing the tissue-level organization of APA dynamics. Beyond these directions, emerging evidence suggests that APA operates within a broader landscape of post-transcriptional regulation in single cells. For example, RNA modifications such as m [6]. A have been shown to exhibit substantial cell-to-cell heterogeneity and can be effectively modeled using machine learning approaches that integrate sequence features and regulatory signals [57, 58]. These findings indicate that multiple regulatory layers, including APA and RNA modification, may act in a coordinated manner to shape transcript isoform diversity and gene expression programs at single-cell resolution. In addition, RBPs are known to regulate key RNA processing steps, including PAS selection, splicing, RNA stability, and translation, thereby directly influencing APA dynamics [59].

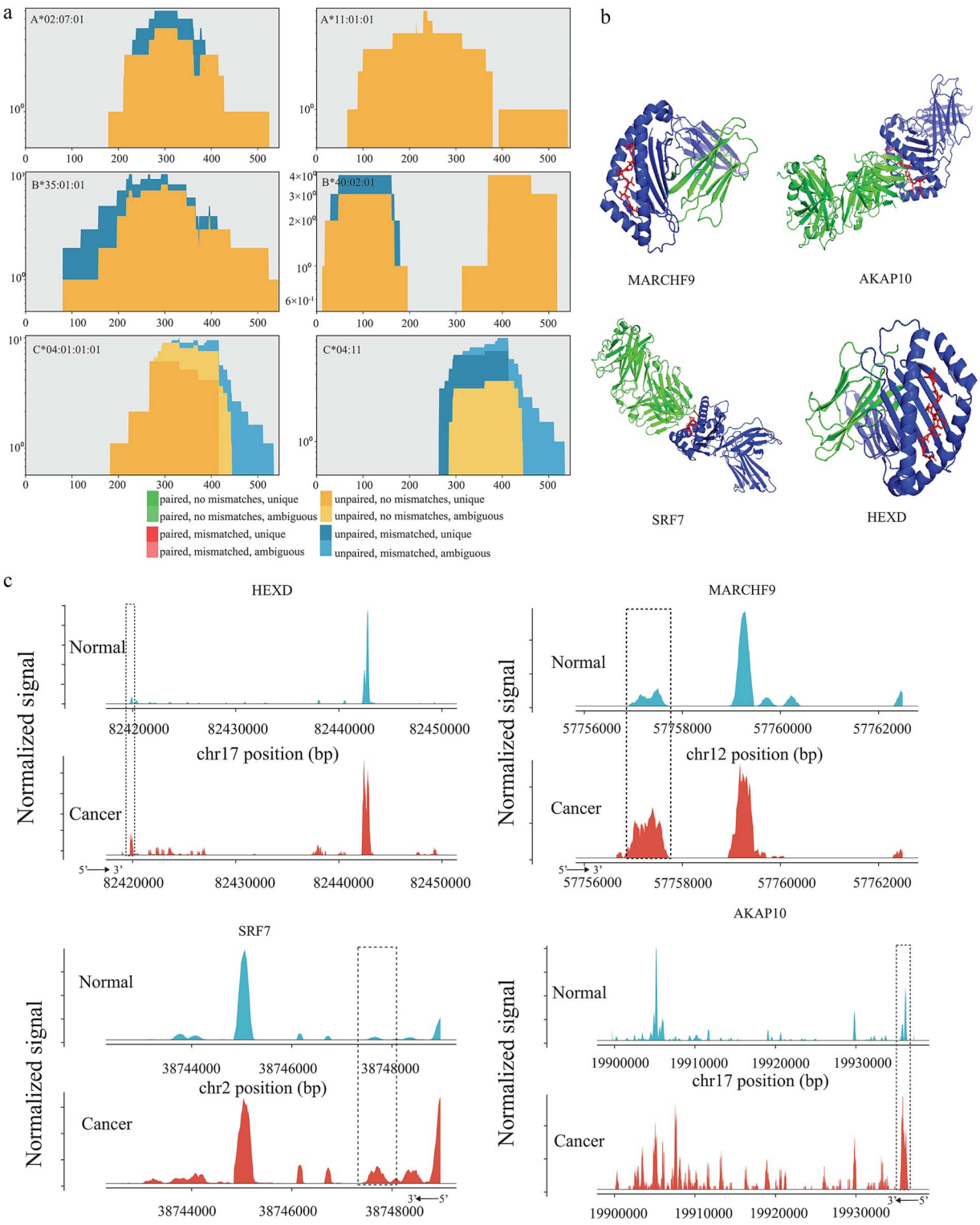

Figure 5 Identification and structural characterization of polyadenylation-derived neoantigens in KRAS-mutant SCLC. (a) Read alignment distributions across representative HLA class I alleles inferred from scRNA-seq data using OptiType. Each panel displays the mapped read distribution for a specific HLA allele, including A*02:07:01, A*11:01:01, B*35:01:01, B*40:02:01, C*04:01:01:01, and C*04:11. Colors denote read pairing status, mismatch rate, and mapping uniqueness. The alleles show high-quality coverage with low mismatch rates and well-distributed read lengths. Based on both binding predictions and alignment quality, A*11:01:01, B*35:01:01, and B*40:02:01 were selected for peptide–MHC structural modeling. (b) Predicted structures of polyadenylation-derived neoantigen peptides bound to patient-specific HLA class I molecules. Representative peptide-HLA complexes were generated, illustrating stable interactions consistent with potential immunogenicity. (c) Coverage plots of polyadenylation regions generating neoantigens. Tumor-specific RE is observed downstream of AKAP10 (− strand), HEXD (+ strand), MARCHF9 (+ strand), and SRSF7 (− strand), consistent with detected polyadenylation events. These tumor-enriched peaks confirm accurat identification of polyadenylation-derived isoforms and neoantigens, supporting their immunogenic potentials.

In summary, scDeepAPA enables accurate PAS identification and quantitative characterization of APA dynamics in single-cell RNA-seq data. By providing single-cell–level resolution, supporting intronic APA analysis, and enabling integration with regulatory and immunogenic features, scDeepAPA facilitates in-depth downstream investigations ranging from mechanistic studies of post-transcriptional regulation to the discovery and prioritization of neoantigen candidates. As single-cell and spatial profiling technologies continue to evolve, scDeepAPA will further support comprehensive isoform-level analyses in complex tissues and diverse disease contexts.

Key Points

- Alternative polyadenylation (APA) is an important post-transcriptional regulatory mechanism, but existing computational tools are largely designed for bulk RNA-seq and have limited ability to analyze APA at single-cell resolution.
- We developed scDeepAPA, a deep learning method optimized for single-cell RNA sequencing data that accurately identifies polyadenylation sites (PAS) and quantifies APA dynamics across individual cells.
- scDeepAPA integrates convolutional layers, Mamba-based state-space modeling, and bidirectional long short-term memory (LSTM) networks, achieving superior performance over five state-of-the-art PAS prediction tools across multiple evaluation metrics.
- Applications to Alzheimer's disease mouse brain and Kirsten rat sarcoma viral oncogene homolog (KRAS)-mutant lung cancer datasets reveal widespread cell-type-specific APA remodeling and identify tumor-specific intronic APA events that may generate neoantigenic peptides.

Supplementary material

Supplementary material is available at Briefings in Bioinformatics online.

Conflicts of interest

None declared.

Funding

Q.S. is supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM151089). M.X. is supported by National Institute of General Medical Sciences of the National Institutes of Health (R35GM128753). This work partially used Jetstream2 [60] through allocation CIS230237 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) [61] program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Data availability

The scRNA-seq datasets analyzed in this study are publicly available from the NCBI Sequence Read Archive. The accession numbers are as follows: SCLC mutant samples (SRR1119782932), SCLC control samples (SRR2140777033), mouse WT control brain samples (SRR1471295034), and mouse AD model brain samples (SRR1471295134). Public PAS annotations were obtained from the PolyASite v3.0 database, accessible at https://polyasite.unibas.ch/3.0/. All analysis code used in this study is publicly available on GitHub at https://github.com/QSong-github/scDeepAPA and has been archived on Zenodo under the DOI 10.5281/zenodo.15066940.

References

1. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. *Front Genet* 2019;**10**:317. https://doi.org/10.3389/fgene.2019.00317
2. Kharchenko PV. The triumphs and limitations of computational methods for scRNA-seq. *Nat Methods* 2021;**18**:723–32. https://doi.org/10.1038/s41592-021-01171-x
3. Zheng GX, Terry JM, Belgrader P *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**:14049. https://doi.org/10.1038/ncomms14049
4. Macosko EZ, Basu A, Satija R *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;**161**:1202–14. https://doi.org/10.1016/j.cell.2015.05.002
5. Klein AM, Mazutis L, Akartuna I *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;**161**:1187–201. https://doi.org/10.1016/j.cell.2015.04.044
6. Han X, Wang R, Zhou Y *et al.* Mapping the mouse cell atlas by microwell-seq. *Cell* 2018;**173**:1307. https://doi.org/10.1016/j.cell.2018.05.012
7. Picelli S, Faridani OR, Björklund ÅK *et al.* Full-length RNA-seq from single cells using smart-seq2. *Nat Protoc* 2014;**9**:171–81. https://doi.org/10.1038/nprot.2014.006
8. Cao J, Packer JS, Ramani V *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 2017;**357**:661–7. https://doi.org/10.1126/science.aam8940
9. Xu Z, Wang Y, Sheng K *et al.* Droplet-based high-throughput single microbe RNA sequencing by smRandom-seq. *Nat Commun* 2023;**14**:5130. https://doi.org/10.1038/s41467-023-40137-9
10. Gierahn TM, Wadsworth MH II, Hughes TK *et al.* Seq-well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods* 2017;**14**:395–8. https://doi.org/10.1038/nmeth.4179
11. Zhang Y, Huang Z, Lu W *et al.* Alternative polyadenylation in cancer: molecular mechanisms and clinical application. *Crit Rev Oncol Hematol* 2025;**206**:104599. https://doi.org/10.1016/j.critrevonc.2024.104599
12. Fu Y, Chen L, Chen C *et al.* Crosstalk between alternative polyadenylation and miRNAs in the regulation of protein translational efficiency. *Genome Res* 2018;**28**:1656–63. https://doi.org/10.1101/gr.231506.117
13. Erson-Bensan AE. Alternative polyadenylation and RNA-binding proteins. *J Mol Endocrinol* 2016;**57**:F29–34. https://doi.org/10.1530/JME-16-0070
14. Norman KL, Chen TC, Zeiner G *et al.* Precursor microRNA-122 inhibits synthesis of Insig1 isoform mRNA by modulating polyadenylation site usage. *RNA* 2017;**23**:1886–93. https://doi.org/10.1261/rna.063099.117
15. Sena RM, Twiss JL, Gardiner AS *et al.* The RNA-binding protein HuD regulates alternative splicing and alternative polyadenylation in the mouse neocortex. *Molecules* 2021;**26**:2836. https://doi.org/10.3390/molecules26102836
16. Liu X, Jin W, Bao D *et al.* DIPAN: detecting personalized intronic polyadenylation derived neoantigens from RNA sequencing data.

Comput Struct Biotechnol J 2024;**23**:2057–66. https://doi.org/10.1016/j.csbj.2024.05.008
17. Ha KCH, Blencowe BJ, Morris Q. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol* 2018;**19**:45. https://doi.org/10.1186/s13059-018-1414-4
18. Li GW, Nan F, Yuan GH *et al.* SCAPTURE: a deep learning-embedded pipeline that captures polyadenylation information from 3′ tag-based RNA-seq of single cells. *Genome Biol* 2021;**22**:221. https://doi.org/10.1186/s13059-021-02437-5
19. Yu H, Dai Z. SANPolyA: a deep learning method for identifying poly (A) signals. *Bioinformatics* 2020;**36**:2393–400. https://doi.org/10.1093/bioinformatics/btz970
20. Kalkatawi M, Magana-Mora A, Jankovic B *et al.* DeepGSR: an optimized deep-learning structure for the recognition of genomic signals and regions. *Bioinformatics* 2019;**35**:1125–32. https://doi.org/10.1093/bioinformatics/bty752
21. Gao X, Zhang J, Wei Z *et al.* DeepPolyA: a convolutional neural network approach for polyadenylation site prediction. *IEEE Access* 2018;**6**:24340–9. https://doi.org/10.1109/ACCESS.2018.2825996
22. Xia Z, Li Y, Zhang B *et al.* DeeReCT-PolyA: a robust and generic deep learning method for PAS identification. *Bioinformatics* 2019;**35**:2371–9. https://doi.org/10.1093/bioinformatics/bty991
23. Gao Y, Li L, Amos CI *et al.* Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. *Genome Res* 2021;**31**:1856–66. https://doi.org/10.1101/gr.271346.120
24. Bai Y, Qin Y, Fan Z *et al.* scMAPA: identification of cell-type-specific alternative polyadenylation in complex tissues. *Gigascience* 2022;**11**:giac033. https://doi.org/10.1093/gigascience/giac033
25. Patrick R, Humphreys DT, Janbandhu V *et al.* Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome Biol* 2020;**21**:167. https://doi.org/10.1186/s13059-020-02071-7
26. Wu X, Liu T, Ye C *et al.* scAPAtrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. *Brief Bioinform* 2021;**22**:bbaa273. https://doi.org/10.1093/bib/bbaa273
27. Shulman ED, Elkon R. Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Res* 2019;**47**:10027–39. https://doi.org/10.1093/nar/gkz781
28. Moon Y, Herrmann CJ, Mironov A *et al.* PolyASite v3.0: a multi-species atlas of polyadenylation sites inferred from single-cell RNA-sequencing data. *Nucleic Acids Res* 2025;**53**:D197–204. https://doi.org/10.1093/nar/gkae1043
29. Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023.
30. Greff K, Srivastava RK, Koutník J *et al.* LSTM: a search space odyssey. *IEEE Trans Neural Netw Learn Syst* 2016;**28**:2222–32. https://doi.org/10.1109/TNNLS.2016.2582924
31. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* 2022;**13**:1246. https://doi.org/10.1038/s41467-022-28803-w
32. Miura P, Shenker S, Andreu-Agullo C *et al.* Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res* 2013;**23**:812–25. https://doi.org/10.1101/gr.146886.112
33. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res* 2017;**27**:491–9. https://doi.org/10.1101/gr.209601.116
34. Heinz S, Benner C, Spann N *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;**38**:576–89. https://doi.org/10.1016/j.molcel.2010.05.004
35. Wang R, Tian B. APAlyzer: a bioinformatics package for analysis of alternative polyadenylation isoforms. *Bioinformatics* 2020;**36**:3907–9. https://doi.org/10.1093/bioinformatics/btaa266
36. Stuart T, Butler A, Hoffman P *et al.* Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031
37. Aran D, Looney AP, Liu L *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;**20**:163–72. https://doi.org/10.1038/s41590-018-0276-y
38. Li JH, Liu S, Zhou H *et al.* starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 2014;**42**:D92–7. https://doi.org/10.1093/nar/gkt1248
39. Liao J-Y, Yang B, Zhang YC *et al.* EuRBPDB: a comprehensive resource for annotation, functional and oncological investigation of eukaryotic RNA binding proteins (RBPs). *Nucleic Acids Res* 2020;**48**:D307–13. https://doi.org/10.1093/nar/gkz823
40. Agarwal V, Bell GW, Nam JW *et al.* Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 2015;**4**:e05005. https://doi.org/10.7554/eLife.05005
41. Casper J, Speir ML, Raney BJ *et al.* The UCSC Genome Browser database: 2026 update. *Nucleic Acids Res* 2025;**54**:D1331–5. https://doi.org/10.1093/nar/gkaf1250
42. Lex A, Gehlenborg N, Strobelt H *et al.* UpSet: visualization of intersecting sets. *IEEE Trans Vis Comput Graph* 2014;**20**:1983–92. https://doi.org/10.1109/TVCG.2014.2346248
43. Jurtz V, Paul S, Andreatta M *et al.* NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol* 2017;**199**:3360–8. https://doi.org/10.4049/jimmunol.1700893
44. Evans R *et al.* Protein complex prediction with AlphaFold-Multimer. *biorxiv*, 2021.2010. 2004.463034. 2021.
45. Chen Y, Zhang H, Wang W *et al.* Rapid generation of high-quality structure figures for publication with PyMOL-PUB. *Bioinformatics* 2024;**40**:btae139. https://doi.org/10.1093/bioinformatics/btae139
46. Lodder C, Scheyltjens I, Stancu IC *et al.* CSF1R inhibition rescues tau pathology and neurodegeneration in an A/T/N model with combined AD pathologies, while preserving plaque associated microglia. *Acta Neuropathol Commun* 2021;**9**:108. https://doi.org/10.1186/s40478-021-01204-8
47. Clarke JP, Thibault PA, Salapa HE *et al.* A comprehensive analysis of the role of hnRNP A1 function and dysfunction in the pathogenesis of neurodegenerative disease. *Front Mol Biosci* 2021;**8**:659610. https://doi.org/10.3389/fmolb.2021.659610
48. Bobkova NV, Lyabin DN, Medvinskaya NI *et al.* The Y-box binding protein 1 suppresses Alzheimer's disease progression in two animal models. *PLoS One* 2015;**10**:e0138867. https://doi.org/10.1371/journal.pone.0138867
49. Lin P, Yang J, Wu S *et al.* Current trends of high-risk gene Cul3 in neurodevelopmental disorders. *Front Psychiatry* 2023;**14**:1215110. https://doi.org/10.3389/fpsyt.2023.1215110
50. Gao C, Jiang J, Tan Y *et al.* Microglia in neurodegenerative diseases: mechanism and potential therapeutic targets. *Signal Transduct Target Ther* 2023;**8**:359. https://doi.org/10.1038/s41392-023-01588-0
51. Zhang C, Yin K, Liu SY *et al.* Multiomics analysis reveals a distinct response mechanism in multiple primary lung adenocarcinoma after neoadjuvant immunotherapy. *J Immunother Cancer* 2021;**9**:e002312. https://doi.org/10.1136/jitc-2020-002312

52. Adams TS, Schupp JC, Poli S *et al.* Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv* 2020;**6**:eaba1983. https://doi.org/10.1126/sciadv.aba1983
53. Li J, Wu Z, Pan Y *et al.* GNL3L exhibits pro-tumor activities via NF-kappaB pathway as a poor prognostic factor in acute myeloid leukemia. *J Cancer* 2024;**15**:4072–80. https://doi.org/10.7150/jca.95339
54. Tian Y, Zhou J, Chai X *et al.* TCF12 activates TGFB2 expression to promote the malignant progression of melanoma. *Cancers (Basel)* 2023;**15**. https://doi.org/10.3390/cancers15184505
55. Chen J, Mai H, Chen H *et al.* Pan-cancer analysis identified C1ORF112 as a potential biomarker for multiple tumor types. *Front Mol Biosci* 2021;**8**:693651. https://doi.org/10.3389/fmolb.2021.693651
56. Szolek A, Schubert B, Mohr C *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* 2014;**30**:3310–6. https://doi.org/10.1093/bioinformatics/btu548
57. Li Y, Li J, Li W *et al.* Scm6A: a fast and low-cost method for quantifying m6A modifications at the single-cell level. *Genomics Proteomics Bioinformatics* 2024;**22**:qzae039. https://doi.org/10.1093/gpbjnl/qzae039
58. Lin Y, Li SJ, Xu X *et al.* Intermittent fasting inhibits Tp53-driven glioma through gut microbiota-mediated methionine-m6A regulation. *Nat Commun* 2026;**17**:1804. https://doi.org/10.1038/s41467-026-68512-2
59. Xu P, Mo Y, Tan R *et al.* RNA-binding proteins: biological mechanisms and their impact on osteoporosis development. *ICell* 2024;**1**35–44. https://doi.org/10.71373/LIGH9788
60. Hancock DY *et al.* Jetstream2: Accelerating cloud computing via Jetstream. In *Practice and Experience in Advanced Research Computing 2021: Evolution Across All Dimensions*, pp. 1–8.
61. Boerner TJ, Deems S, Furlani TR *et al.* Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support. *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good* 2023;173–6.

Briefings in Bioinformatics, 2026, **27**, bbag339
https://doi.org/10.1093/bib/bbag339
Problem Solving Protocol