*Briefings in Bioinformatics*, 2026, 27, bbag167
https://doi.org/10.1093/bib/bbag167
Published: 17 April 2026
**Problem Solving Protocol**

OXFORD

# CRESCENT: a deep learning framework with multi-scale attention for detecting recurrent copy number alterations

Xikang Feng [1,2,*,‡], Zheng Xu [2,‡], Sisi Peng [2], Jieyi Zheng [2], Chuan Ma [3], Qiangguo Jin [2,*], Lingxi Chen [4,*]

[1]Research & Development Institute, Northwestern Polytechnical University, Sanhang Science & Technology Buliding, No. 45th, Gaoxin South 9th Road, Nanshan District, Shenzhen City, 518063, China
[2]School of Software, Northwestern Polytechnical University, 127 West Youyi Road, Beilin District, Xi'an Shaanxi, 710072, China
[3]School of Civil Aviation, Northwestern Polytechnical University, 127 West Youyi Road, Beilin District, Xi'an Shaanxi, 710072, China
[4]Department of Biomedical Sciences, College of Biomedicine, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Kowloon, Hong Kong SAR, China

*Corresponding authors. Xikang Feng, E-mail: fxk@nwpu.edu.cn; Qiangguo Jin, E-mail: qgking@nwpu.edu.cn; Lingxi Chen, E-mail: lingxi.chen@cityu.edu.hk.
‡The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

## Abstract

**Recurrent copy number alterations (CNAs) are fundamental drivers of tumorigenesis, yet identifying them reliably remains a challenge due to the extreme variability in their genomic scale and context. Current methods often struggle to balance sensitivity across focal, segmental, and arm-level events. Here, we present CRESCENT, a deep learning framework designed to detect recurrent CNAs by integrating multi-scale sampling with convolutional neural networks and self-attention mechanisms. By processing copy number profiles from 7689 cases across 20 The Cancer Genome Atlas (TCGA) cancer projects, CRESCENT learns to distinguish recurrent drivers from background noise through parallel feature fusion. In rigorous leave-one-project-out cross-validation, the model demonstrated robust generalization, achieving area under the curves of 0.894–0.967 for amplifications and 0.804–0.929 for deletions in representative cohorts (Bladder Urothelial Carcinoma, Sarcoma, Glioblastoma Multiforme, Uterine Corpus Endometrial Carcinoma). Finally, extending beyond the TCGA-specific cross-validation, we trained a unified pan-cancer model to assess CRESCENT's generalizability on simulated datasets and independent, non-TCGA cancer cohorts (CGCI and TARGET). Benchmarking against standard tools, including GISTIC2 and RUBIC, reveals that CRESCENT offers superior detection balance, identifying the highest total number of significant events across focal and broad scales. Moreover, extensive focal gene expression validation and pathway annotation, coupled with survival analysis, highlight that CRESCENT identifies critical oncogenic drivers and prognostic markers that conventional statistical methods often overlook. In all, CRESCENT provides a highly sensitive, generalized approach for decoding tumor evolution.**

***Keywords*** copy number alterations, recurrent CNA, focal CNA, cancer analysis, deep learning

## Introduction

Copy number alterations (CNAs) are a hallmark of the cancer genome, playing a pivotal role in tumorigenesis by amplifying oncogenes and deleting tumor suppressor genes [1, 2]. Unlike single nucleotide variants, CNAs affect large segments of the genome, ranging from focal events spanning a few kilobases to broad alterations affecting entire chromosomal arms [3–6]. Identifying recurrent CNAs—alterations that appear across a significant proportion of patients—is critical for distinguishing driver events from random "passenger" mutations [7]. Recent studies have underscored the complexity of these alterations, revealing their role not only in cancer [1, 8, 9] but also in neurodevelopmental disorders and mosaicism in healthy tissues [10, 11]. Consequently, the accurate detection of recurrent CNAs is fundamental to understanding disease etiology and identifying therapeutic targets.

Historically, the identification of recurrent CNAs has relied on statistical frameworks. Early approaches focused on identifying recurrent breakpoints or common regions of overlap based on frequency and amplitude thresholds [12, 13]. These methods laid the groundwork for industry-standard tools such as GISTIC2.0 [7], which identifies significant peaks of alteration by evaluating both the frequency and amplitude of events. Moreover, RUBIC [14] focuses on identifying recurrent breakpoints to define focal events. However, these conventional methods face significant limitations regarding the scale of detection. CNAs exhibit extreme heterogeneity in size [3]; GISTIC2.0 tends to bias towards broad, large-scale events, often obscuring focal drivers in noise, whereas RUBIC effectively captures focal events but frequently fails to detect broad chromosomal alterations [14].

This dichotomy forces researchers to rely on ensemble approaches or manual curation to obtain a complete genomic landscape.

**Received:** December 17, 2025. **Revised:** February 26, 2026. **Accepted:** March 15, 2026

Furthermore, traditional statistical models often struggle with the "noisy" nature of cancer genomic data [7, 12–14], where complex, non-linear patterns and boundary effects can lead to false positives or fragmented calls. The emergence of deep learning in genomics offers a promising alternative. By treating genomic data as visual or sequential patterns, neural networks can learn to recognize the "peak-like" signatures of recurrence without rigid, pre-defined statistical assumptions [15].

To address the limitations of scale-specific detection, we introduce CRESCENT, a deep learning framework equipped with a multi-scale attention mechanism. Unlike previous methods that favor detecting genomic regions at a specific resolution, CRESCENT employs a sliding-window strategy with parallel convolutional neural network (CNN) branches to process genomic contexts at varying resolutions simultaneously. A multi-head self-attention (MHA) mechanism then integrates these features, allowing the model to distinguish focal drivers from broad background alterations dynamically. By training on a comprehensive dataset of 7689 cases across 20 The Cancer Genome Atlas (TCGA) projects, CRESCENT demonstrates superior sensitivity and robustness compared to existing benchmarks, effectively bridging the gap between focal and broad CNA detection. We further validate the robustness of our approach by moving beyond cohort-specific models; we developed a unified pan-cancer model and benchmarked its performance against ground-truth simulated datasets and four independent external cancer cohorts (CGCI-HTMCP-CC [Cancer Genome Characterization Initiative – HIV+ Tumor Molecular Characterization Project, Cervical Cancer], CGCI-HTMCP-DLBCL [Diffuse Large B-Cell Lymphoma], TARGET-ALL-P2 [Therapeutically Applicable Research to Generate Effective Treatments – Acute Lymphoblastic Leukemia, Phase II], and TARGET-AML [Acute Myeloid Leukemia]). Moreover, extensive focal gene expression validation and pathway annotation, coupled with survival analysis, highlight that CRESCENT identifies critical oncogenic drivers and prognostic markers that conventional statistical methods often overlook.

## Materials and methods

### Data collection

We analyzed integer-valued total copy number profiles derived from allele-specific copy number segment data generated via whole-genome sequencing (WGS). The primary cohort included 7689 cases from 20 TCGA projects, downloaded via the Genomic Data Commons (GDC) portal [16] (https://portal.gdc.cancer.gov/) on 12 June 2024 (see Fig. 1a for dataset distribution). To validate the model's generalizability beyond TCGA, we subsequently incorporated data from four non-TCGA projects, including CGCI-HTMCP-CC, CGCI-HTMCP-DLBCL, TARGET-ALL-P2, and TARGET-AML, retrieved from the GDC portal on 6 February 2026. Furthermore, to benchmark performance against ground truth, we utilized simulated BRCA copy number profiles obtained from the RUBIC repository hosted on GitHub (https://github.com/ewaldvandyk/RUBIC-datasets/tree/master/TCGA_SNP6/BRCA_sim). The real datasets were provided in the human reference genome hg38 version. All data handling adhered to GDC usage policies. The complete set of CNAs is now freely available for download from CNAScope [8].

### Data preprocessing

Each TCGA project encompasses multiple cases, where each case may include several samples identified by unique GDC Aliquot identifiers. To align with the model's per-chromosome analysis, data for each sample were segmented by chromosome. The raw data, depicted in Fig. 1b, adhere to the BED format and specify the chromosome, start and end positions, and copy number for each segment. For every unique GDC Aliquot, BED files from potentially multiple workflow types (e.g. ASCAT2, ASCAT3 [17, 18]) were merged. This process entailed sorting segments by position and consolidating adjacent ones. Specifically for the deletion detection task, as shown in the bottom panel of Fig. 1b, segments with copy numbers greater than 2 were ignored, and the remaining values were converted to their absolute deviation from the diploid state (i.e. $|C.N. - 2|$). These processed segments were then mapped to fixed genomic bins to construct an initial sample-by-bin matrix (Fig. 1c). Figure 1d and e illustrate the normalization of this CNA matrix to satisfy the model's input requirement of 40 channels. If the number of samples ($m$) was greater than 40 (Fig. 1e), a folding strategy was applied where samples were averaged in a modulo manner (e.g. averaging samples at indices $i, i+40, i+80\ldots$) to compress the data into 40 channels. Conversely, if $m \leq 40$ (Fig. 1d), the matrix was padded with pseudo-samples containing a constant diploid value of 2 to reach the required dimension. As shown in Fig. 1f, the resulting $40 \times n$ matrix was transposed to serve as the final model input, with rows representing the 40 standardized channels and columns denoting genomic bins. For a detailed justification regarding the selection of 40 channels and the use of padding with a value of 2, please refer to Supplementary Notes 1.1, Supplementary Fig. 1, and Supplementary Tables 1 and 2. Finally, the non-TCGA datasets and simulated data were processed using an identical pipeline to ensure consistency across all analyses.

### Dataset preparation and empirical basis

To develop a robust dataset for training and evaluating our model in detecting recurrent CNAs, we started with an empirical analysis of the preprocessed data. The input comprises chromosome-specific CNA matrices derived from TCGA projects following our preprocessing pipeline. In these matrices, rows represent genomic bins, and columns correspond to the 40 channels as described in the previous section. Each entry encodes the copy number state, enabling visualization of CNAs as patterns across samples. The model's task is to classify whether a target genomic bin exhibits recurrence, defined by coordinated alterations across multiple samples.

We developed an interactive visualization tool that renders each chromosome matrix as a heatmap, facilitating systematic exploration of CNA patterns. This tool uncovered key characteristics: CNAs vary dramatically in scale, from tens to thousands of bins, and recurrent events often manifest as distinct "peak-like" signatures with clear boundaries and amplitudes. Optimal detection requires capturing the entire event within a centered window; undersized windows truncate boundaries, while oversized ones introduce noise from surrounding non-recurrent areas. These insights motivated a multi-scale approach for both dataset sampling and model architecture, ensuring adaptability to heterogeneous CNA sizes.

Dataset labeling integrated manual curation with established computational tools to produce high-quality annotations. Positive instances (recurrent CNAs) were identified through: (i) manual selection using the heatmap tool, focusing on regions with evident peak-like patterns indicative of recurrence; and (ii) applying GISTIC2.0 and RUBIC with default parameters to the same matrices, incorporating regions detected by both tools as well as those unique to one but visually confirmed. This hybrid strategy

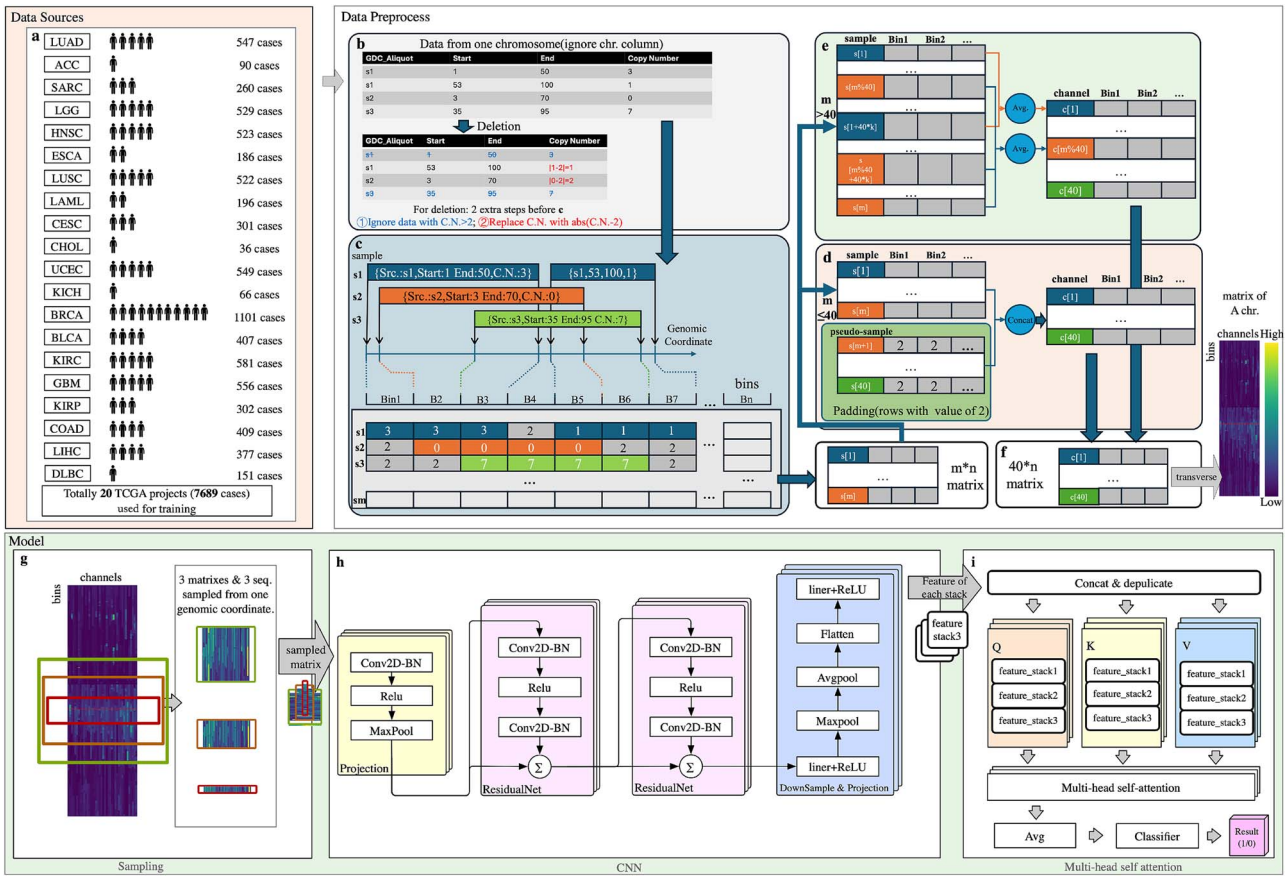


**Figure 1** Workflow of CRESCENT. (a) TCGA data from 20 projects (7689 cases) are utilized, primarily for training, with a LOO validation strategy. (b) Data preprocessing converts raw copy number segments; specifically for deletion detection, segments with copy number $> 2$ are ignored, and others are converted to the absolute deviation from diploidy (abs(C.N.-2)). (c) These segments are mapped to genomic bins to construct an initial sample-by-bin matrix. (d–f) To ensure a fixed input dimension of 40 channels: (d) if the number of samples ($m$) is $\leq 40$, the matrix is padded with pseudo-samples (value of 2); (e) if $m > 40$, samples are folded and averaged into 40 channels. (f) The resulting $40 \times n$ matrix is transposed to serve as the standardized input for the model. (g) Three windows per genomic position are sampled for model input, classifying CNA regions. (h and i) Model structure includes CNN and MHA modules.

leveraged algorithmic efficiency while mitigating tool-specific biases via expert validation. Crucially, this curation pipeline was applied identically to the external non-TCGA datasets (CGCI and TARGET) to ensure consistent evaluation benchmarks across real-world cohorts. For a detailed rationale of this hybrid strategy and a quantitative assessment of label uncertainty, please refer to Supplementary Notes 1.2, Supplementary Fig. 2, and Supplementary Table 3.

Negative instances (non-recurrent regions) were curated to encompass both straightforward and challenging cases: a portion was randomly sampled from areas lacking CNA signals, while others were manually selected as "hard negatives," such as regions adjacent to positives or those with globally elevated copy numbers but lacking peak-like recurrence features. This approach aimed to enhance the model's robustness against false positives.

For the simulated datasets, a distinct labeling strategy was necessitated by the nature of the data. We utilized five independent sets of 1000 simulated breast cancer copy number profiles from the RUBIC study [14]. Since explicit ground-truth labels for recurrent CNAs were not available for these simulations, we established the results identified by the RUBIC algorithm as the reference standard (pseudo-ground truth). This setup provided a controlled environment to benchmark CRESCENT against GISTIC2.0 by measuring the recovery of these RUBIC-defined events.

For each labeled bin, which serves as the center of a target window, we applied multi-scale sampling (detailed in the following section) to generate input instances.

## Multi-scale sampling strategy

To address the empirical scale variability of CNAs, we implemented a multi-scale sliding-window sampling strategy. For each target bin, we center three predefined window sizes on it to extract submatrices from the chromosome matrix: 50×40, 100×40, and 2000×40 (rows × columns) for amplifications; and 20×40, 50×40, and 400×40 for deletions. These window sizes were determined based on a statistical analysis of recurrent CNA segment lengths across major cancer cohorts, which revealed distinct length distributions for amplifications and deletions (see Supplementary Note 1.3 and Supplementary Fig. 3). Each submatrix symmetrically encompasses rows above and below the target, thereby preserving local genomic context at varying resolutions. The resulting triplet of matrices serves as input to parallel branches in our CNN, enabling the model to fuse multi-scale features for effective classification.

Boundary management is essential near chromosome ends, where centering full windows is infeasible. Rather than using constant-value padding (e.g. with 2, which could create artifacts such as artificial contrasts), we duplicate and concatenate the adjacent matrix segment. This approach maintains genuine texture patterns and minimizes false positives.

## Overview of CRESCENT model architecture

We developed a multi-branch CNN, named CRESCENT, to process the multi-scale sampled submatrices. The architecture, as illustrated in Fig. 1, consists of parallel processing branches for each input scale, a cross-branch fusion module employing self-attention, and a final classification head. This design empowers CRESCENT to extract and integrate features across diverse genomic resolutions, enabling effective identification of recurrent CNA patterns.

Each branch handles an input submatrix of shape $(B, 1, H, W)$, where $B$ denotes the batch size, $H$ the number of genomic bins (rows), and $W$ the number of samples (columns). Feature extraction initiates with a projection module: a 2D convolution (Conv2D) followed by batch normalization (BN), rectified linear unit (ReLU) activation, and max pooling to condense spatial dimensions while preserving initial patterns. This is followed by a stack of residual networks (ResidualNets), each featuring two convolutional blocks (Conv2D-BN-ReLU) augmented with skip connections and summation to alleviate gradient vanishing and support deeper feature learning. The branch culminates in a downsampling and projection phase, involving flattening, average pooling, max pooling, and a linear layer with ReLU activation, producing a compact feature vector of uniform dimension $d$ per scale.

Feature vectors from the $N$ branches (typically $N = 3$) are stacked into a tensor of shape $(B, N, d)$, treating scales as a sequence. To capture inter-scale dependencies, we employ a MHA mechanism: the stacked features are concatenated and duplicated to form query (Q), key (K), and value (V) projections through distinct linear layers. The MHA generates attention-weighted representations, modeling high-order interactions across branches, followed by a residual connection and a feed-forward network (FFN) for additional refinement, reminiscent of a Transformer encoder.

For classification, the fused features are globally averaged across the branch dimension to yield a vector of shape $(B, d)$. This feeds into a classifier head, which is a multi-layer perceptron (MLP) incorporating layer normalization, linear mapping, ReLU activation, and dropout. This head outputs a single logit. CRESCENT is trained using binary cross-entropy with logits loss (BCEWithLogitsLoss) for binary classification of recurrent versus non-recurrent CNAs.

## Training and evaluation

We trained the CRESCENT model using a binary classification objective with BCEWithLogitsLoss. Optimization employed the Adam optimizer, initialized with a learning rate of $1 \times 10^{-5}$ and a weight decay of $3 \times 10^{-2}$. To dynamically adjust the learning rate, we incorporated a ReduceLROnPlateau scheduler that monitored validation loss, applying a reduction factor of 0.5, a patience of 3 epochs, and a minimum learning rate of $1 \times 10^{-7}$. Each training run extended over 50 epochs with a batch size of 8, utilizing GPU acceleration where available. To enhance stability, we implemented mixed-precision training via automatic mixed precision and dynamic loss scaling, complemented by gradient clipping with an $\ell_2$ max-norm of 0.9. For reproducibility, we generated and logged random seeds for NumPy [19], and PyTorch [20], while enabling deterministic cuDNN operations.

For evaluation, we utilized a leave-one-project-out cross-validation strategy across TCGA projects. In each fold, one project served as the test set, with the remaining projects forming the training set. We selected the optimal model checkpoint based on the validation area under the curve (AUC) and reported key performance metrics for the held-out project, including AUC, F1 score, accuracy, precision, and recall. Additionally, we generated receiver operating characteristic (ROC) curves from validation predictions to evaluate model discrimination.

Positive and negative instances for training and validation were chosen according to the labeling guidelines outlined in the preceding section, with detailed instance counts provided in Supplementary Tables 4 and 5.

## Benchmarking settings

For benchmarking, we configured three distinct approaches: GISTIC2 [7], RUBIC [14], and our CRESCENT model. Each underwent tailored preprocessing to ensure compatibility with TCGA data and subsequent gene annotation.

GISTIC2 requires non-overlapping input CNA segments. Given that TCGA provides ASCAT2- and ASCAT3-derived files where segments from the same sample may overlap, we used a publicly available R script to merge these overlapping segments by averaging their copy-number values [21]. We omitted marker information, leaving the marker file empty, and set all other parameters to the defaults of the run_gistic_example script. Results were directly retrieved from the GISTIC output files amp_genes.conf_90.txt and del_genes.conf_90.txt.

For RUBIC, we applied the same deduplication procedure to manage overlapping segments. Since RUBIC mandates a marker input, we supplied a placeholder file that assigned a value of 1 to each segment, ensuring copy number was the only varying factor. All other parameters remained at their default settings, and results were extracted from the specified output files.

In contrast, CRESCENT conducts chromosome-wide segmentation internally. For each window, it calculates copy-number values as averages of overlapping segments, thereby handling deduplication automatically. This enabled direct utilization of raw TCGA files without extra preprocessing.

## Prioritization of focal CNAs with TCGA gene expression support

We employed the hg38 genome build "gencode.v46.annotation.gtf" (https://www.gencodegenes.org/human/release_46.html) for gene annotation, utilizing PyRanges [22] to identify protein-coding genes overlapping with focal CNAs.

To further refine the detected focal CNAs, we evaluated their concordance with matched TCGA gene expression data. For each gene located within a focal region, we performed a one-sample t-test comparing its mean expression against a cohort-specific baseline, defined as the median of all gene means within that cohort. We computed one-sided $P$-values to identify significantly over- or under-expressed genes relative to this background, applying the Benjamini–Hochberg procedure to control for multiple testing.

We subsequently applied a filtering strategy to retain only those focal CNAs with transcriptional support. Focal amplifications were retained only if they overlapped with at least one significantly upregulated gene ($P_{adj} < .05$) and contained zero significantly downregulated genes. Focal deletions were retained only if they overlapped with at

least one significantly low expressed gene ($P_{adj} < .05$) and contained zero significantly high-expressed genes.

This approach ensures that the prioritized focal events are not only recurrent at the genomic level but also demonstrate consistent transcriptional consequences characteristic of driver alterations.

### Pathway annotation for focal genes

We performed pathway enrichment analysis on focal amplified genes with significant overexpression and focal deleted genes with significant underexpression in TCGA. Using GSEApy [23], we interrogated the built-in "KEGG 2021 Human" and "WikiPathway 2023 Human" libraries. Significant pathways were identified using filtration criteria of an adjusted *P*-value <.05 and a minimum gene overlap of 3.

### Clinical validation of focal amplified genes via survival analysis

To assess the prognostic significance of focal amplified genes, we integrated TCGA gene expression profiles with matched clinical overall survival (OS) data. Patients were stratified into two groups based on gene expression levels: a "High Expression" group (expression > median) and a "Low Expression" group (expression ≤ median).

Survival analysis was performed using the lifelines Python package [24]. For each candidate gene, we estimated survival probabilities using the Kaplan-Meier estimator (KaplanMeierFitter). Differences in survival curves between the high and low expression groups were evaluated using the log-rank test. Additionally, we quantified the magnitude of the risk associated with high gene expression by calculating the Hazard Rratio (HR) and its 95% confidence interval using a univariate Cox Proportional Hazards model. Genes were considered statistically significant prognostic markers if the log-rank *P*-value was < .05.

## Result

### Overview of CRESCENT

We developed CRESCENT, a novel deep learning framework designed to detect recurrent copy number alterations (CNAs) by seamlessly integrating multi-scale genomic contexts, thereby addressing the inherent variability in CNA scales observed across diverse cancer types (Fig. 1). Leveraging copy number profiles from TCGA, the workflow commences with the aggregation of high-resolution CNA data from 7689 cases spanning 20 distinct TCGA projects (Fig. 1a; Methods: Data collection). This comprehensive dataset encompasses a wide spectrum of solid tumors, providing a robust foundation for model training and evaluation.

To facilitate chromosome-specific analysis, we first group the raw CNA segments (which are characterized by genomic intervals and integer-valued copy numbers) by sample (defined by unique GDC Aliquot identifiers) and chromosome (Fig. 1b). Specifically for the deletion detection task, segments with copy numbers greater than 2 are ignored, and the remaining values are converted to their absolute deviation from the diploid state. These processed segments are then mapped to fixed genomic bins to construct an initial sample-by-bin matrix (Fig. 1c). To standardize input dimensions for the neural network while preserving signal integrity, the matrix is normalized to a fixed depth of 40 channels: for cohorts exceeding 40 samples, a folding strategy is applied where samples are averaged in a modulo manner; for those with fewer samples, padding with diploid values (value of 2) is employed (Fig. 1d–e). Finally, the matrix is transposed so that rows represent the 40 standardized channels and columns denote genomic bins (Fig. 1f, Methods: Data preprocessing).

Empirical analysis of the input matrices, visualized as heatmaps, revealed that recurrent CNAs manifest as "peak-like" signatures with varying scales, from focal events spanning tens of bins to broad alterations encompassing thousands (Methods: Dataset preparation and empirical basis). This necessitates a multi-scale approach for accurate detection. Accordingly, CRESCENT employs a sliding-window sampling strategy at each genomic bin, extracting submatrices centered on the target position using three predefined window sizes (e.g. 50×40, 100×40, and 2000×40 for amplifications; adjusted for deletions) to capture local contexts (Fig. 1g; Methods: Multi-scale sampling strategy). Boundary effects near chromosome ends are mitigated through duplication and concatenation of adjacent regions, avoiding artifactual padding.

These multi-scale inputs are processed through a parallel neural network architecture comprising dedicated CNN branches for feature extraction (Fig. 1h). Each branch incorporates an initial projection module (2D convolution, BN, ReLU activation, and max pooling) followed by a stack of residual blocks to enable deep, stable learning of hierarchical patterns (Methods: Overview of CRESCENT model architecture). Extracted features from all branches are then fused via a MHA mechanism, which models inter-scale dependencies by computing attention-weighted representations, followed by a FFN and residual connections for refined synthesis (Fig. 1i). The fused features are globally averaged and passed through a compact MLP classifier to output a probability score for recurrence (binary classification: recurrent versus non-recurrent).

Dataset labeling for training integrates manual curation of peak-like patterns with outputs from established tools (GISTIC2 [7] and RUBIC [14]) to generate high-quality positive (recurrent) and negative (non-recurrent) instances (Methods:Dataset Preparation and Empirical Basis). Model training utilizes binary cross-entropy loss with Adam optimization, mixed-precision techniques, and adaptive learning rate scheduling, evaluated via leave-one-project-out cross-validation across TCGA cohorts (Methods: Training and evaluation). For an exhaustive account of preprocessing, architecture, and evaluation protocols, refer to the Methods section. This multi-scale paradigm enables CRESCENT to achieve high sensitivity and generalizability in detecting recurrent CNAs, as demonstrated in subsequent evaluations.

### Evaluation of CRESCENT using leave-one-out cross-validation across 20 TCGA cancer types

We use a leave-one-out (LOO) cross-validation strategy to evaluate the performance of CRESCENT in detecting recurrent amplification (Fig. 2a–h) and deletion (Fig. 2i–p), respectively. Specifically, for each experiment, one TCGA cancer type was held out as the test set, while the remaining 20 cancer types were used as the training dataset (Supplementary Tables 4 and 5). The recurrent amplification and deletion events inferred from the held-out cancer type were used as the ground truth labels for performance evaluation (Methods: Dataset preparation and empirical basis).

Figure 2 focuses on four representative cancer types: BLCA (Bladder Urothelial Carcinoma), SARC (Sarcoma), GBM (Glioblastoma Multiforme), and UCEC (Uterine Corpus Endometrial Carcinoma). As

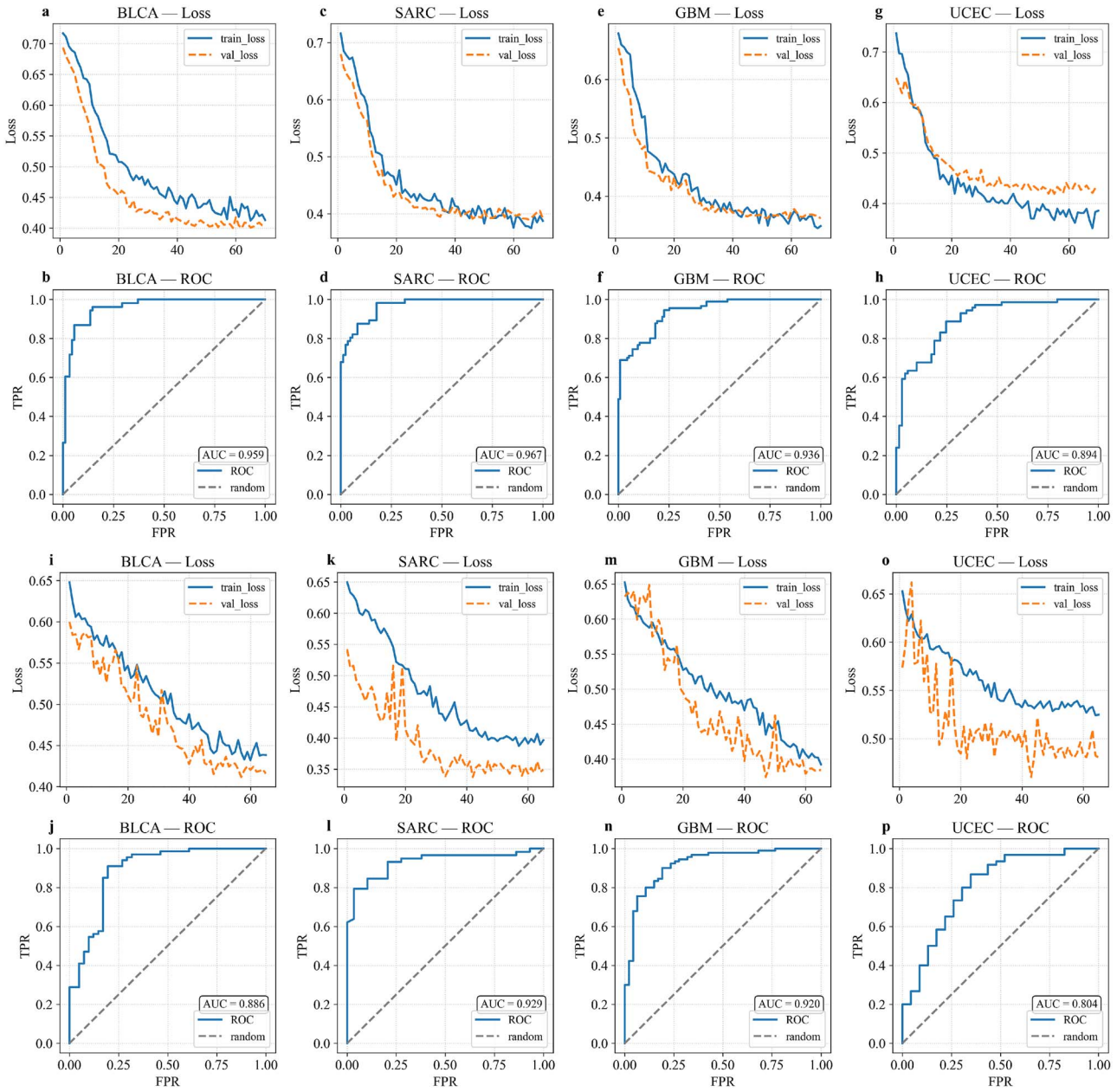


**Figure 2** Performance of CRESCENT Using LOO Cross-Validation. (a, c, e, g) Training and validation loss curves for recurrent amplification prediction in four TCGA cancer types (BLCA, SARC, GBM, and UCEC). (b, d, f, h) ROC curves for recurrent amplification prediction in four cancer types, with high AUC values (BLCA: 0.959, SARC: 0.967, GBM: 0.936, UCEC: 0.894). (i, k, m, o) Training and validation loss curves for recurrent deletion prediction in BLCA, SARC, GBM, and UCEC. (j, l, n, p) ROC curves for recurrent deletion prediction, with corresponding AUCs of 0.886 (BLCA), 0.929 (SARC), 0.920 (GBM), and 0.804 (UCEC). BLCA: Bladder Urothelial Carcinoma. SARC: Sarcoma. GBM: Glioblastoma Multiforme. UCEC: Uterine Corpus Endometrial Carcinoma.

illustrated in Fig. 2a–g, both the training and validation losses decreased and converged over epochs, indicating effective learning and the absence of overfitting. The ROC curves in Fig. 2b–h demonstrate high prediction accuracy for recurrent amplifications in test sets, with the AUC values of 0.959 for BLCA, 0.967 for SARC, 0.936 for GBM, and 0.894 for UCEC.

Similarly, for recurrent deletion prediction, the training and validation losses shown in Fig. 2i–o also converge over epochs, further supporting the robustness and generalizability of the model. The ROC curves in Fig. 2j–p show that the model achieved AUCs of 0.886 for BLCA, 0.929 for SARC, 0.920 for GBM, and 0.804 for UCEC, respectively.

Supplementary Tables 4 and 5 present CRESCENT's performance for all 20 TCGA cancer types. These results collectively demonstrate that CRESCENT provides accurate and generalizable predictions for both recurrent amplification and deletion events across diverse cancer types.

## Detection of focal, medium, and broad CNAs by CRESCENT, GISTIC2, and RUBIC

To assess CRESCENT's performance in detecting recurrent CNAs, we benchmarked it against two baseline tools, GISTIC2 [7] and RUBIC

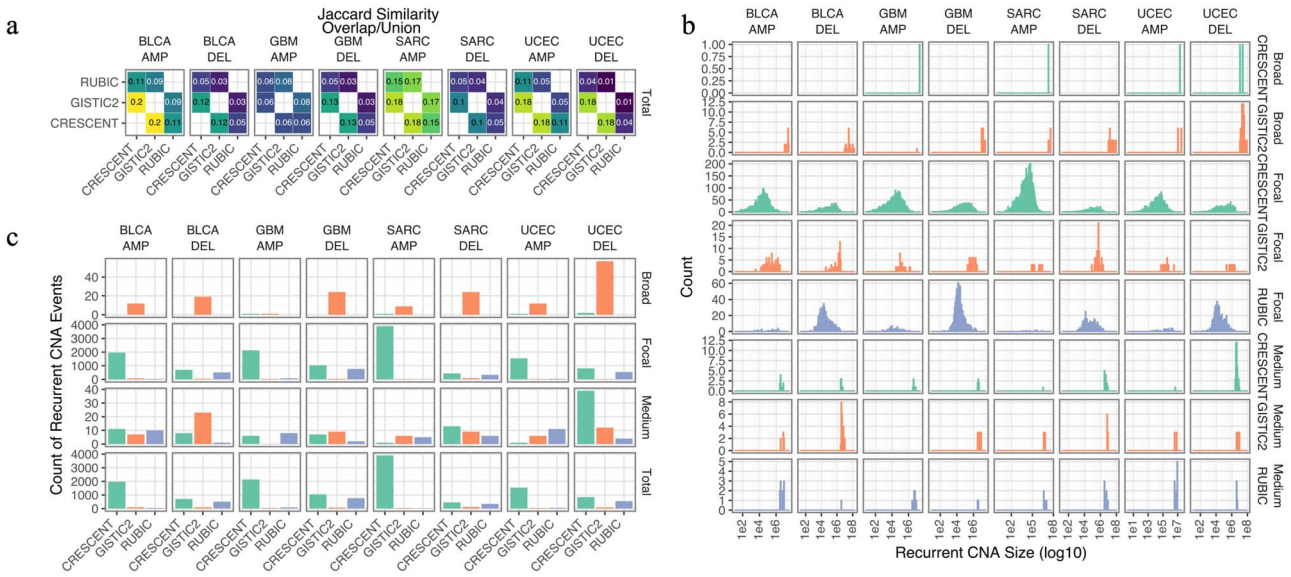


**Figure 3** Comparative summary of recurrent CNA calls across tools. (a) Bar plots showing the count of recurrent CNA events detected by each tool (CRESCENT, GISTIC2, and RUBIC) within every cancer type × CNA type combination. Bars are stratified by CNA size range (Broad, Focal, Medium, Total), allowing direct comparison of event yields across tools. (b) Histogram plots depicting the $\log_1 0$-transformed recurrent CNA segment lengths for each tool. These panels highlight how length profiles differ across tools and ranges. (c) Heatmap of pairwise Jaccard similarities (overlap/union of recurrent CNA base-pairs) between tools for each cancer-type x CNA-type category. Values inside tiles indicate the Jaccard index; yellow shading corresponds to stronger agreement.

[14], using copy number profiles from four representative TCGA cohorts: BLCA (Bladder Urothelial Carcinoma), SARC (Sarcoma), GBM (Glioblastoma Multiforme), and UCEC (Uterine Corpus Endometrial Carcinoma). Baseline methods were run with default parameters (see Methods: Benchmarking settings), and detected events were stratified into focal CNAs (<3 Mb), medium CNAs (3–10 Mb), and broad CNAs (>10 Mb) for scale-specific analysis [3].

Quantitative comparisons across the cohorts (Fig. 3) underscore CRESCENT's robust detection capabilities. In (Fig. 3a), CRESCENT consistently identifies the highest total number of recurrent CNAs in all four datasets. Both CRESCENT and RUBIC predominantly detect focal and medium CNAs, with RUBIC reporting no broad events at all (resulting in their absence from the broad panel in Fig. 3b) and CRESCENT yielding only a minimal number of broad CNAs. In contrast, GISTIC2 focuses primarily on medium and broad CNAs, accounting for nearly all broad events across cohorts, while detecting very few focal ones. This pattern reveals that CRESCENT and RUBIC favor smaller-scale alterations, whereas GISTIC2 biases toward larger ones. Notably, for focal and medium CNAs, CRESCENT detects more events than the other tools in almost all cases, with exceptions only in medium-scale BLCA deletions and GBM deletions. These trends demonstrate CRESCENT's enhanced sensitivity, particularly for finer-scale recurrent CNAs that may be underrepresented by baselines.

This scale-specific performance is further illustrated in the segment length distributions (Fig. 3b), where CRESCENT exhibits a broader spread of log10-transformed lengths, effectively bridging the short-fragment skew of RUBIC (e.g. median <1 Mb in GBM) and the long-segment bias of GISTIC2 (e.g. median >10 Mb in SARC and UCEC).

Pairwise Jaccard similarity analysis (Fig. 3c) reinforces CRESCENT's integrative advantages, as its overlap with RUBIC and GISTIC2 generally exceeds the overlap between RUBIC and GISTIC2 across most datasets and categories. For example, in UCEC amplifications, CRESCENT ∩ RUBIC reaches 0.11 and CRESCENT ∩ GISTIC2 reaches 0.18, both surpassing RUBIC ∩ GISTIC2 at 0.05; similar superiority holds in UCEC deletions (CRESCENT ∩ RUBIC: 0.04; CRESCENT ∩ GISTIC2: 0.18; versus RUBIC ∩ GISTIC2: 0.01). Exceptions occur in GBM amplifications (CRESCENT ∩ RUBIC: 0.06; CRESCENT ∩ GISTIC2: 0.06; slightly below RUBIC ∩ GISTIC2: 0.08) and SARC amplifications (CRESCENT ∩ RUBIC: 0.15; CRESCENT ∩ GISTIC2: 0.18; on par with RUBIC ∩ GISTIC2: 0.17). Overall, these indices indicate that CRESCENT captures consensus events while expanding coverage to scales overlooked by the other tools.

Chromosome-level heatmaps (Fig. 4) provide visual evidence of CRESCENT's advantages in resolving complex recurrent patterns. These maps overlay CRESCENT's probability scores (upper panels) with observed copy number values (lower panels) for selected chromosomes, juxtaposed against GISTIC2 and RUBIC calls (Fig. 4a). For straightforward focal CNAs with prominent recurrent signals and minimal surrounding noise, such as the amplification in GBM chr7 (Fig. 4b), all three tools successfully identify the event. In more complex regions with substantial upstream or downstream noise (Fig. 4c–i), CRESCENT consistently outperforms the baselines, showcasing greater robustness. For amplifications, consider SARC chr11 (Fig. 4d): at the central position, a clear recurrent signal is detected by CRESCENT and GISTIC2 but missed by RUBIC; conversely, on the right side, a continuous prominent signal is captured by CRESCENT and RUBIC, yet overlooked by GISTIC2. For deletions, in GBM chr9 (Fig. 4h), the evident signal in the middle region is identified by CRESCENT and GISTIC2, but not by RUBIC. These examples, along with others like BLCA chr13 deletions (Fig. 4f) and UCEC chr10 deletions (Fig. 4g), illustrate CRESCENT's ability to resolve intricate patterns across a spectrum of scales while mitigating noise-induced misses or fragments seen in GISTIC2 and RUBIC.

Further validation of CRESCENT's multi-scale design comes from Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations on positive validation cases (Supplementary Figs 4–5).

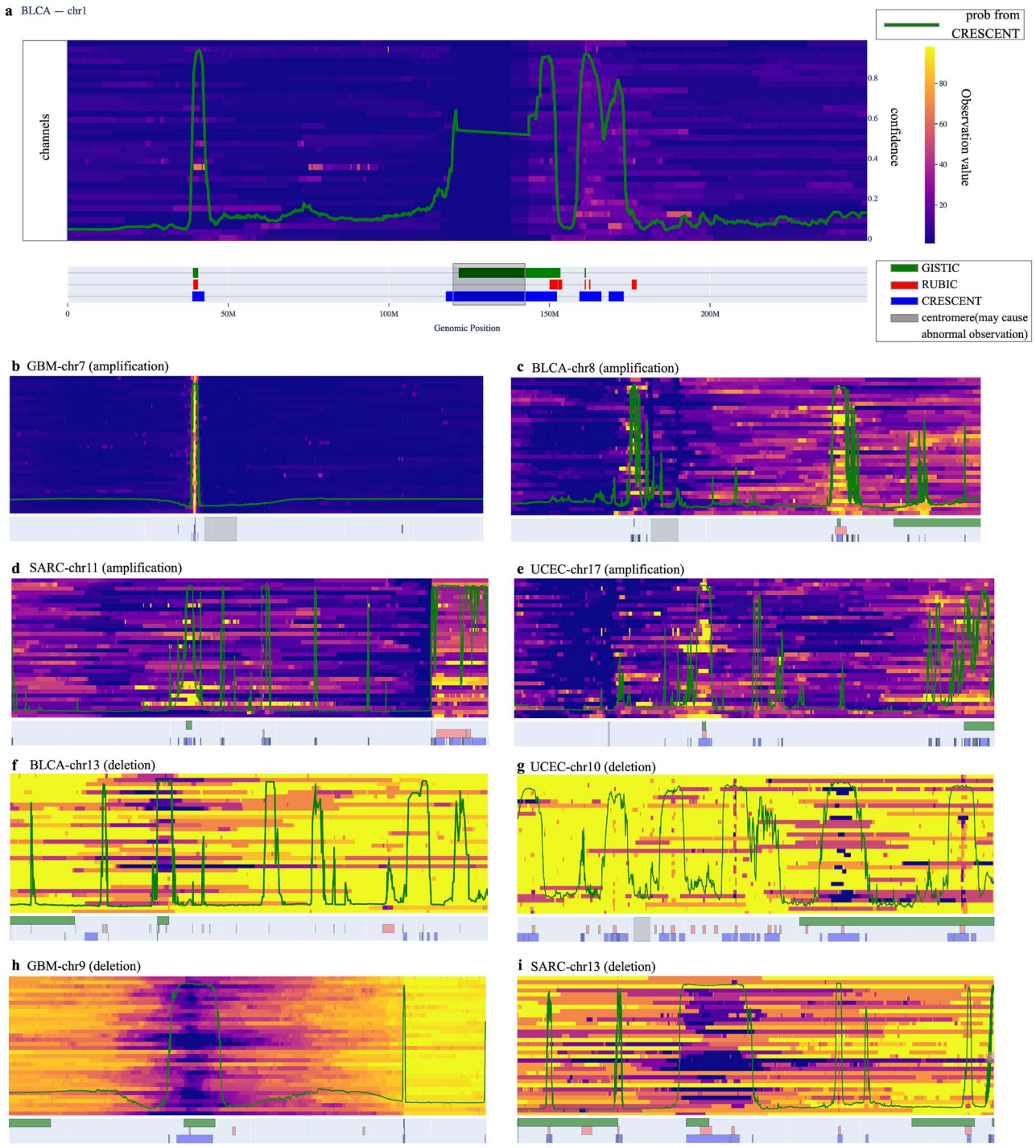


**Figure 4** Chromosome-level heatmaps illustrating predicted recurrent copy number amplifications and deletions across selected TCGA cancer types. (a) BLCA amplifications on chr1; (b) GBM amplifications on chr7; (c) BLCA amplifications on chr8; (d) SARC amplifications on chr11; (e) UCEC amplifications on chr17; (f) BLCA deletions on chr13; (g) UCEC deletions on chr10; (h) GBM deletions on chr9; (i) SARC deletions on chr13. Each heatmap integrates CRESCENT's probability outputs (upper panel) with observed copy number values (lower panel), highlighting focal recurrent events. Overlaid annotations from GISTIC2 and RUBIC demonstrate regions where CRESCENT identifies additional focal CNAs undetected by these methods. Centromeric regions, which may introduce anomalous observations, are indicated.

These heatmaps, overlaid on multi-branch inputs, illustrate how narrower windows (e.g. branch0:50×40) emphasize focal peaks in BLCA amplifications (Supplementary Fig. 4), while broader windows (e.g. branch2:2000×40) capture expansive signals in GBM (Supplementary Fig. 5). To substantiate these qualitative observations with systematic evidence, we further conducted a dataset-level quantitative analysis comparing positive instances ($n = 984$) against a negative noise baseline ($n = 922$). We evaluated the attribution maps using localization consistency (quantifying focus on the expected central region) and attribution concentration

(top-k energy ratio). Positive instances exhibited significantly stronger localization consistency (Central-mass ratio: AUC $\approx$ 0.789, Cliff's $\delta \approx 0.578$) and higher attribution concentration (AUC $\approx$ 0.683, Cliff's $\delta \approx 0.367$) compared to the baseline. These results confirm that the model consistently prioritizes the intended central genomic signal rather than background noise. Collectively, these visualizations and quantitative metrics confirm that CRESCENT's parallel branches and attention mechanism enable robust integration of scale-varying CNAs, outperforming the scale-specific limitations of GISTIC2 and RUBIC.

## Comparative analysis of focal CNA detection by CRESCENT, GISTIC2, and RUBIC

We performed a comprehensive comparative analysis and pathway annotation of the focal CNA regions identified by CRESCENT, GISTIC2, and RUBIC. Figure 5a illustrates the total count of focal amplified and deleted genes detected by each method prior to expression filtering. In general, CRESCENT consistently identified the highest number of focal genes among the three tools. While RUBIC generally detected more focal deletions than GISTIC2 across the cohorts, GISTIC2 showed lower sensitivity in identifying focal events in this comparison. Importantly, when filtering for genes with expression support (Fig. 5b, Methods: Prioritization of focal CNAs with TCGA gene expression support), the total number of candidates decreased substantially for all methods; however, CRESCENT retained the highest volume of validated genes, suggesting it captures a broader landscape of transcriptionally active driver alterations.

We subsequently analyzed the overlap of expression-supported genes across multiple cancer cohorts to evaluate consensus on canonical cancer drivers. The concordance among the three tools is visualized in Supplementary Fig. 6. In the amplification analysis, the intersection of all three methods (CRESCENT $\cap$ RUBIC $\cap$ GISTIC2) consistently captured well-established oncogenes and tumor suppressors across cohorts. Specifically, the consensus set included critical cell-cycle and signaling regulators such as *CCND1* in BLCA, *EGFR* and *CDK4* in GBM, *CCNE* in UCEC, and *MDM2* in SARC. This high degree of overlap on known drivers suggests that CRESCENT is highly effective at capturing biologically significant focal events validated by established benchmarks. In the analysis of focal deletions, concordance was primarily observed between CRESCENT and RUBIC, as the majority of genes detected by GISTIC2 lacked transcriptional support. We observed shared candidates such as *OR4P4* (BLCA and SARC) and *RBFOX1* (UCEC). Notably, CRESCENT identified a significantly larger proportion of unique focal events compared to the other tools (e.g. 87.9% and 93.9% unique amplifications in BLCA and GBM, respectively). These unique fractions contained potential drivers like *SDK1* (BLCA) and *HSDL2* (GBM). Conversely, GISTIC2 and RUBIC detected distinct but smaller sets of genes (e.g. *CD24* and *SOX4* in BLCA, respectively), highlighting that while consensus methods capture core drivers, CRESCENT provides a more expansive view of the focal copy number landscape.

To further validate the biological relevance of the detected focal CNAs, we performed pathway enrichment analysis on the consensus genes identified by all three tools (Fig. 5c). Analysis using the WikiPathways and the KEGG libraries (Methods: Pathway annotation for focal genes) revealed that these consensus genes are significantly enriched in classical oncogenic drivers and core survival pathways, most notably the PI3K-Akt signaling pathway, glioma signaling, and general pathways in cancer. This confirms that all three tools successfully capture the fundamental, high-frequency driver alterations common in glioma. Furthermore, we examined the pathway enrichment for genes uniquely identified by CRESCENT (Fig. 5d) to assess whether these additional candidates possess biological functional relevance. The analysis demonstrated that CRESCENT-unique genes map to distinct and critical biological processes that extend beyond core oncogenic signaling. For instance, we observed significant enrichment in metabolic pathways such as the TCA cycle and oxidative phosphorylation, as well as broader signaling networks like Wnt, MAPK, and Rap1 signaling. Notably, in the GBM cohort, CRESCENT uniquely identified genes associated with tissue-specific neural functions, including synaptic signaling pathways, GABA receptor signaling, and neuroactive ligand-receptor interactions. This suggests that CRESCENT provides a higher-resolution view of the genomic landscape, capturing subtle but biologically significant alterations—such as metabolic reprogramming and tissue-specific dependencies—that are missed by other methods.

## Prognostic significance of CRESCENT-unique and consensus focal amplified genes in TCGA cohorts

We further validated the clinical relevance of the detected focal amplifications by correlating gene expression with OS using matched TCGA clinical data (Methods: Clinical validation of focal amplified genes via survival analysis). In Fig. 6a, the consensus gene set—limited by the conservative intersection of three tools—yielded only a single significant prognostic marker: *CCNE1* in the UCEC cohort. High expression of *CCNE1* was strongly associated with poor prognosis [Hazard Ratio (HR) = 2.60, log-rank $p = 9.05 \times 10^{-6}$], as confirmed by the Kaplan-Meier analysis (Fig. 6b), where patients with high *CCNE1* levels exhibited significantly reduced survival times. Biologically, *CCNE1* (Cyclin E1) is a canonical cell-cycle regulator whose amplification drives G1/S phase transition and genomic instability, a well-documented mechanism of aggressiveness in endometrial cancers and associated with poor prognosis [25].

In contrast to the limited yield of the consensus approach, the CRESCENT-unique gene set revealed a rich landscape of prognostic markers across all four cohorts. CRESCENT identified a substantial number of significant (log-rank $P < 0.05$) high-risk (HR > 1) and low-risk (HR < 1) genes: 84 low-risk/117 high-risk in BLCA, 46 low-risk/44 high-risk in GBM, 128 low-risk/92 high-risk in SARC, and 114 low-risk/87 high-risk in UCEC. This demonstrates that CRESCENT captures a broader spectrum of clinically relevant alterations missed by other tools.

Several top candidates from the CRESCENT-unique set were validated via Kaplan-Meier analysis (Fig. 6b). In UCEC, *TRIM46* emerged as a top high-risk gene (HR = 2.60, $P = 8.78 \times 10^{-6}$). In Fig. 6b, patients with elevated TRIM46 expression showed a marked decrease in survival probability. *TRIM46* is an E3 ubiquitin ligase known to regulate microtubule organization and has been implicated in promoting cell proliferation and chemotherapy resistance in various carcinomas [26]. In BLCA, *PATZ1* was identified as a significant low-risk gene (HR = 0.55, $P = 4.51 \times 10^{-5}$), with higher expression correlating with better survival outcomes (Fig. 6b). *PATZ1* acts as a transcriptional repressor and tumor suppressor in several contexts, often inhibiting the p53 pathway or epithelial-mesenchymal transition [27], consistent with a protective role in bladder cancer. In GBM, *RPL39L* was a top high-risk factor (HR = 2.43, $P = 5.99 \times 10^{-7}$), while *GDI2* served as a

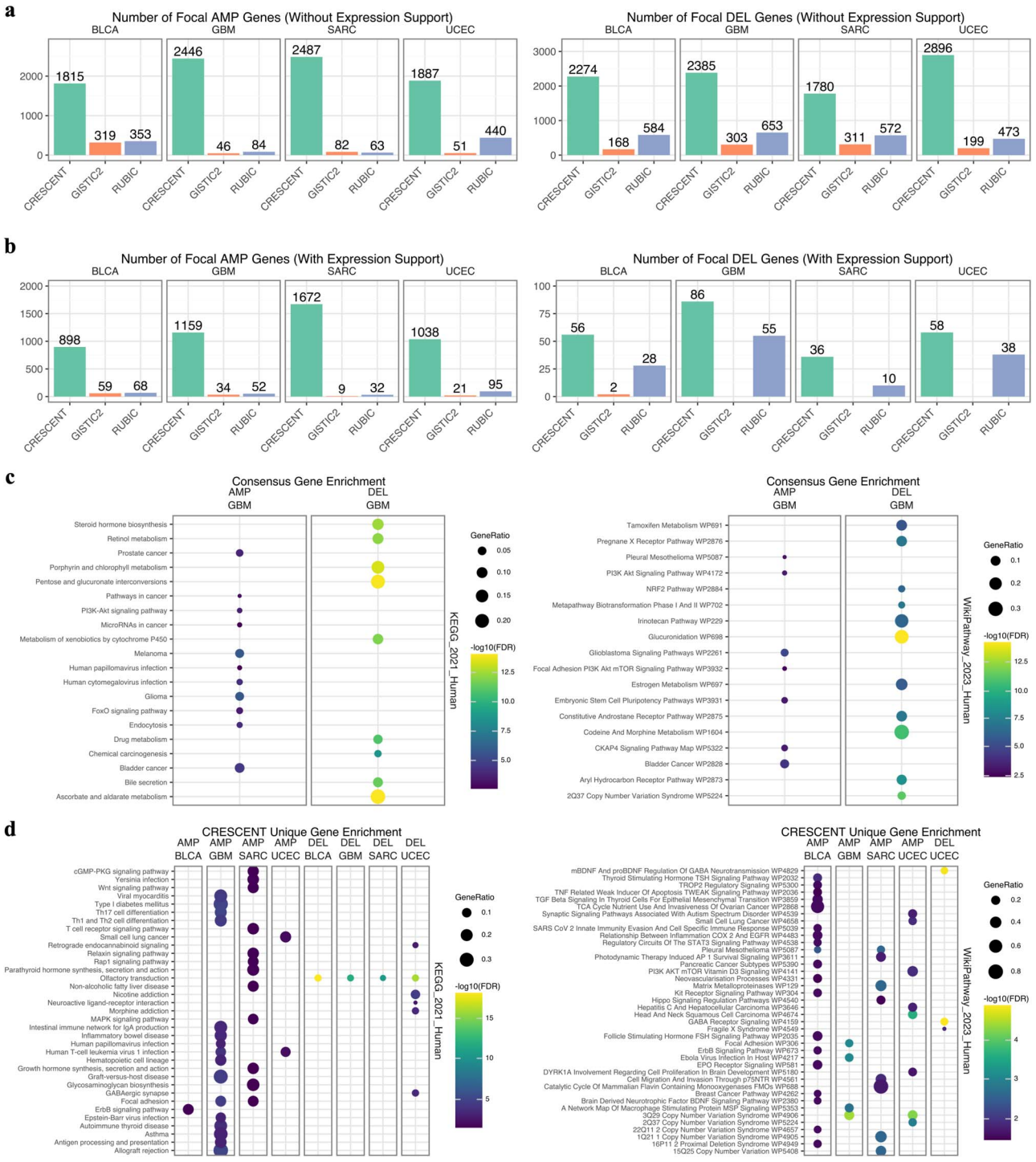


**Figure 5** Gene and pathway annotation of detected focal CNAs. (a–b) The number of focal amplified and deleted genes detected by the three tools without gene expression validation (a) and with expression support (b). (c–d) Pathway enrichment analysis of consensus (c) and CRESCENT unique (d) focal CNA genes using the WikiPathways and KEGG databases.

protective, low-risk marker (HR = 0.57, $P = .0011$) (Fig. 6b). *RPL39L* is a ribosomal protein paralog often upregulated in highly proliferative tumor cells to support increased protein synthesis demands [28], fitting the aggressive nature of glioblastoma. Conversely, *GDI2* (Rab GDP dissociation inhibitor beta) regulates vesicle transport; its association with better prognosis suggests it may modulate membrane trafficking pathways that limit invasive potential in specific glioma subtypes. In SARC, *LIMS2* was identified as a protective, low-risk gene (HR = 0.40, $P = 6.04\times10^{-6}$), whereas *MEX3A* was a significant high-risk factor (HR = 2.44, $P = 1.32 \times 10^{-5}$) (Fig. 6b). LIMS2 is involved in focal adhesion and cell spreading, often acting as a tumor suppressor by stabilizing cell-matrix interactions and preventing metastasis [29]. In contrast,

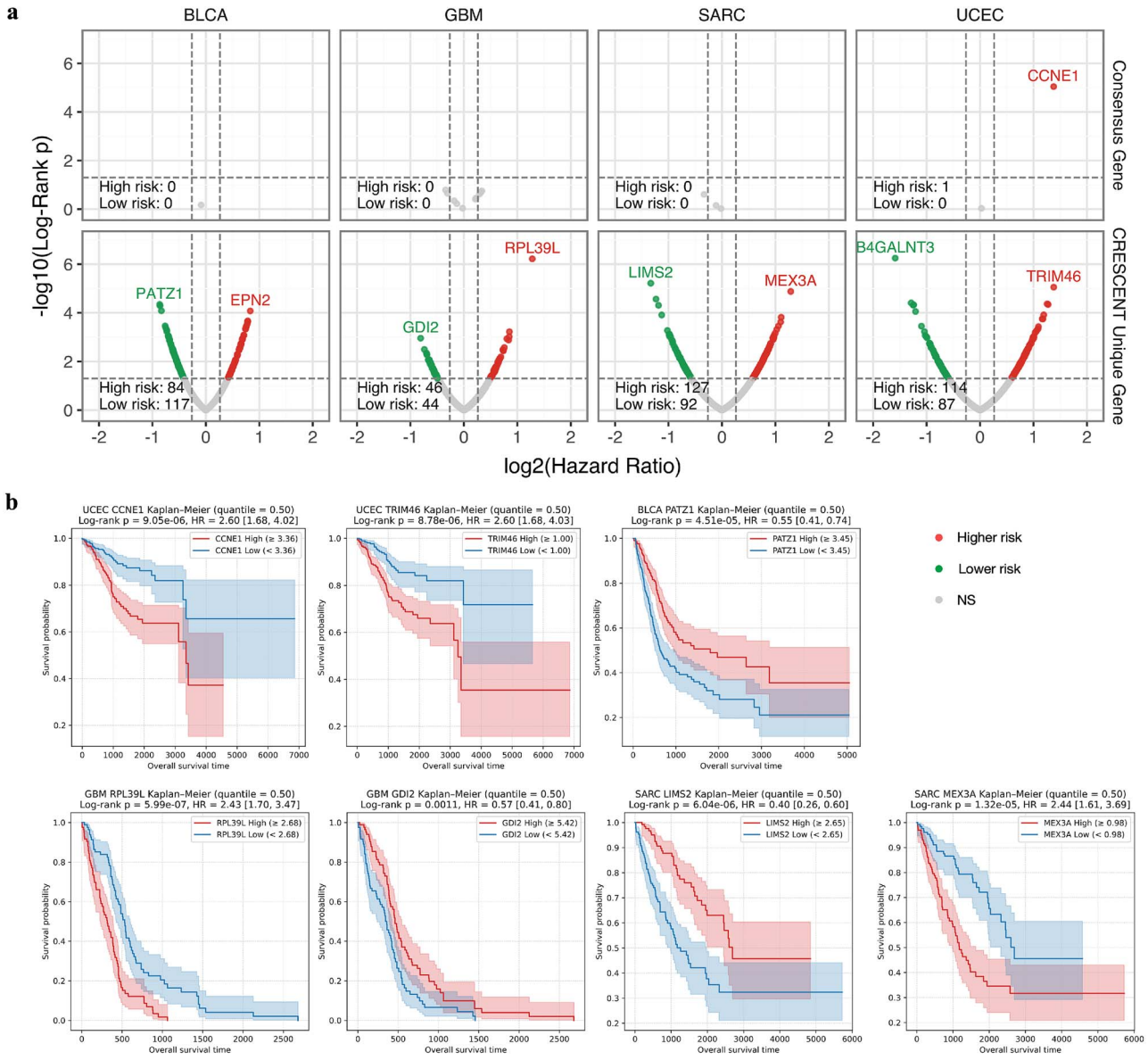


**Figure 6** Prognostic significance of CRESCENT-unique and consensus focal amplified genes in TCGA cohorts. (a) Volcano plots illustrating the relationship between gene expression and patient survival across four cancer types. The x-axis represents the log2 HR, and the y-axis represents the -log10 Log-rank $P$-value. The top row displays consensus genes, while the bottom row displays CRESCENT-unique genes. Genes associated with significantly lower risk (HR $<$ 1) are colored green, while those associated with higher risk (HR $>$ 1) are colored red. (b) Kaplan-Meier survival curves stratifying patients into high- and low-expression groups based on median gene expression levels.

MEX3A is an RNA-binding protein that regulates mRNA stability and has been linked to stemness and resistance to apoptosis [30], driving poor outcomes in sarcomas.

Collectively, these findings highlight that CRESCENT not only recovers canonical drivers but also uncovers a vast array of novel, clinically significant focal amplifications that possess distinct prognostic value.

## Evaluation of CRESCENT's pan-cancer generalizability on simulated and independent cohorts

To further evaluate robustness, we moved beyond cohort-specific cross-validation and trained a unified pan-cancer model. This allowed us to test CRESCENT's generalizability using both simulated data and independent external cancer cohorts.

The pan-cancer model was trained on the aggregated dataset of all 20 TCGA cancer types using a five-fold cross-validation strategy. As shown in Supplementary Figs 7 and 8 and Supplementary Table 6, the pan-cancer model achieved robust performance with a mean AUC of 0.9468 (±0.0125) for amplifications and 0.9435 (±0.0165) for deletions, demonstrating that CRESCENT can effectively learn shared recurrent CNA patterns across diverse cancer types.

First, we validated the pan-cancer model using simulated datasets from the RUBIC study [14]. We utilized the five independent sets of 1000 simulated breast cancer copy number profiles provided by the authors. Since explicit ground-truth labels for recurrent CNAs were not available for these simulations, we established the results identified by RUBIC as the reference standard (pseudo-ground truth). We then benchmarked CRESCENT against GISTIC2.0 by measuring how

well each method recovered these RUBIC-defined events. As presented in Supplementary Table 7, CRESCENT demonstrated superior concordance with the reference standard compared to GISTIC2.0. For amplifications, CRESCENT achieved a mean F1 score of 0.9822, significantly outperforming GISTIC2.0 (0.8810). Similarly, for deletions, CRESCENT showed higher consistency with the reference, yielding a mean F1 of 0.8461 compared to 0.8103 for GISTIC2.0. These results demonstrate that CRESCENT achieves robust performance on the simulated datasets, exhibiting superior sensitivity and precision compared to GISTIC2.0 in recovering the reference events.

Next, we demonstrate the pan-cancer model's generalizability to four external WGS datasets from CGCI-HTMCP-CC, CGCI-HTMCP-DLBCL, TARGET-ALL-P2, and TARGET-AML projects. As shown in Supplementary Table 8, CRESCENT maintained high performance on these unseen datasets, with a mean AUC of 0.9757 for amplifications and 0.9649 for deletions. This indicates that the features learned by CRESCENT are not specific to TCGA batch effects but represent generalized biological signals of recurrence.

### Computational efficiency of CRESCENT

In addition to predictive performance, we assessed the computational efficiency of CRESCENT on large-scale datasets. We benchmarked the runtime and resource usage against RUBIC and GISTIC2.0 using three TCGA cohorts (BLCA, SARC, and UCEC) on a standard workstation. CRESCENT demonstrated high efficiency, processing each cohort in ∼1.1–1.3 h, which is significantly faster than RUBIC (∼7.5 h) and comparable to the heuristic-based GISTIC2.0. Notably, CRESCENT achieved this with minimal memory overhead (∼1.4 GB peak RAM versus RUBIC's 28–67 GB), confirming its scalability for genome-wide analysis (detailed hardware specifications and benchmarks are provided in Supplementary Tables 9–11).

## Discussion

In this study, we presented CRESCENT, a unified deep learning framework for the detection of recurrent CNAs. By leveraging a multi-scale attention architecture, CRESCENT addresses a persistent challenge in cancer genomics: the simultaneous and accurate detection of alterations that vary by orders of magnitude in length. Our extensive evaluation across 20 TCGA cancer types demonstrates that CRESCENT not only matches the performance of established tools on consensus drivers but also uncovers significant recurrent events that conventional methods miss. Crucially, to ensure our findings were not artifacts of the TCGA source distribution, we validated a unified pan-cancer version of CRESCENT on simulated data and independent external cohorts (CGCI and TARGET). The model maintained high performance, confirming that CRESCENT learns generalized signatures of recurrence rather than dataset-specific batch effects.

To properly interpret the comparative performance of CRESCENT against GISTIC2 and RUBIC, it is essential to acknowledge the fundamental methodological differences in how these tools detect recurrent CNAs. GISTIC2.0 relies on amplitude and frequency thresholds, utilizing a "peel-off" algorithm to separate overlapping events [7]. While effective for identifying high-amplitude peak summits, this iterative subtraction can struggle in noisy regions or with lower-amplitude signals. Consequently, if a sharp focal peak is not distinct enough to be "peeled off," GISTIC2.0 often defaults to aggregating the signal into broad, arm-level events, thereby obscuring the underlying focal drivers. Conversely, RUBIC focuses on "recurrent breaks" (transitions between copy number states) rather than peaks [14]. This allows RUBIC to excel at identifying focal events with sharp boundaries but inherently limits its sensitivity to broad, arm-level alterations where boundaries are often diffuse or span centromeres. A primary advantage of CRESCENT is its architectural capacity to emulate expert visual inspection. Much like a human annotator who zooms in to identify focal peaks and zooms out to contextualize broad arm-level changes, CRESCENT utilizes a multi-branch CNN to process genomic data across resolutions simultaneously. By utilizing scale-specific inputs (50, 100, and 2000 bins for amplifications; 20, 50, and 400 for deletions), the model effectively synthesizes local-to-global contexts, allowing for the precise mapping of focal (<3 Mb), medium (3–10 Mb), and broad (>10 Mb) events.

Our TCGA LOOCV benchmarking of CRESCENT against GISTIC2.0 and RUBIC revealed distinct biases in the baseline tools. GISTIC2.0 predominantly identified broad segments (>10 Mb), likely due to difficulties in resolving focal boundaries within noisy signals, while RUBIC remained restricted to focal alterations—findings that are consistent with their respective methodological constraints. In contrast, CRESCENT exhibited a balanced detection profile, identifying a higher total number of events while maintaining a length distribution that spans both extremes. This capability is particularly evident in complex genomic regions (e.g. SARC chr11 and GBM chr9), where CRESCENT successfully resolved boundaries that appeared fragmented or were missed entirely by statistical baselines due to signal noise.

We acknowledge that our ground truth labeling strategy may influence benchmarking metrics, as specific labeling criteria can inherently favor certain segmentation logics. However, the ultimate measure of utility is the recovery of biologically validated and clinically relevant drivers. In this context, CRESCENT demonstrated superior performance: it identified a significantly larger number of focal events supported by gene expression data compared to the baseline tools. Furthermore, the unique focal amplifications detected by CRESCENT exhibited stronger prognostic associations, validating their clinical significance.

Beyond computational metrics, the ultimate utility of a recurrent CNA caller is its ability to identify biologically relevant targets. When we filtered focal CNA candidates based on transcriptional support (significant over/under-expression relative to copy number status), CRESCENT retained the highest volume of validated focal CNAs. While GISTIC2 and RUBIC provide a conservative "core" set of drivers (e.g. *EGFR*, *CCND1*, *CDK4*), further survival analysis revealed that CRESCENT offers a more comprehensive catalog of potential therapeutic targets, bridging the gap between statistical recurrence and biological impact. For example, we identified *TRIM46* in UCEC and *RPL39L* in GBM as high-risk factors associated with poor survival, and *PATZ1* in BLCA as a protective factor—associations that were missed by the other two tools.

Despite its robust performance, CRESCENT relies on supervised deep learning, necessitating high-quality training data. Although we leveraged massive WGS cohorts from TCGA and demonstrated generalizability across simulated data and independent cohorts (CGCI and TARGET), the model's efficacy on rare cancer types with limited training samples warrants further exploration. Furthermore, while our preprocessing pipeline accommodates diverse input formats, downstream recurrence detection is inevitably influenced by the quality of upstream segmentation tools like ASCAT [17, 18]. The current

CRESCENT architecture is optimized for WGS data. Direct application to single-cell DNA sequencing [31–33] or Whole-Exome Sequencing (WES) [8] may result in resolution artifacts due to inherent sparsity, noise, and discontinuity. Future iterations will incorporate imputation or transfer learning techniques to address these challenges. Furthermore, we aim to extend the framework to detect recurrent complex structural variants [34] beyond simple copy number changes. Finally, recognizing that automated analysis and interactive visualization are critical for benchmarking and biological discovery [35], we aim to integrate CRESCENT and its visual outputs into our existing CNA database, CNAScope [8].

In all, CRESCENT represents a significant advancement in the computational analysis of cancer genomes. By integrating multi-scale features through deep learning, it provides a comprehensive and robust solution for detecting recurrent CNAs. This tool offers researchers a more accurate lens through which to view the genomic landscape of cancer, facilitating the discovery of novel drivers and therapeutic targets.

**Key Points**

- CRESCENT utilizes parallel convolutional neural networks and self-attention to detect recurrent CNAs at multiple resolutions simultaneously.
- CRESCENT outperforms traditional recurrent CNA tools by achieving higher sensitivity and generalization in detecting both focal and broad CNAs.
- CRESCENT identifies therapeutic targets that conventional statistical methods often overlook.

## Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions.

## Author contributions

Xikang Feng (Conceptualization, Supervision, Software, Formal analysis, Validation, Visualization, Investigation, Methodology, Funding acquisition, Writing—original draft), Zheng Xu (Software, Methodology, Formal analysis, Validation, Visualization, Writing—original draft), Sisi Peng (Data curation, Writing—review & editing), Jieyi Zheng (Formal analysis, Writing—review & editing), Chuan Ma (Investigation), Qiangguo Jin (Investigation, Supervision, Funding acquisition), Lingxi Chen (Conceptualization, Supervision, Formal analysis, Validation, Visualization, Investigation, Methodology, Funding acquisition, Writing—original draft)

## Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

## Conflicts of interest

The authors declare no competing interests.

## Funding

This work is supported in part by funds from the National Natural Science Foundation of China (No. 32300527; No. 32400519; No. 62572401), the Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515110784), the Research Grants Council of Hong Kong (No. 21200425), the CityUHK Start-Up Grant (No. 9610687), and the Basic Research Programs of Taicang, 2024 (No. TC2024JC43).

## Data and code availability

The copy number profiles analyzed in this study were obtained from 20 TCGA projects, comprising 7689 cases, via the Genomic Data Commons (GDC) portal (https://portal.gdc.cancer.gov/) on 12 June 2024. Additional datasets, including CGCI-HTMCP-CC, CGCI-HTMCP-DLBCL, TARGET-ALL-P2, and TARGET-AML, were subsequently downloaded from the GDC portal on February 6, 2026. The simulated datasets were retrieved from the RUBIC repository hosted on GitHub (https://github.com/ewaldvandyk/RUBIC-datasets/tree/master/TCGA_SNP6/BRCA_sim). The complete set of CNAs is now freely available for download from CNAScope [8]. CRESCENT software is available open-source from https://github.com/BioThinkLab/CRESCENT.

## References

1. Steele CD, Ammal Abbasi SM, Islam A *et al.* Signatures of copy number alterations in human cancer. *Nature* 2022;**606**:984–91. https://doi.org/10.1038/s41586-022-04738-6
2. Hastings PJ, Lupski JR, Rosenberg SM *et al.* Mechanisms of change in gene copy number. *Nat Rev Genet* 2009;**10**:551–64. https://doi.org/10.1038/nrg2593
3. Krijgsman O, Carvalho B, Meijer GA *et al.* Focal chromosomal copy number aberrations in cancer—needles in a genome haystack. *Biochim Biophys Acta* 2014;**1843**:2698–704. https://doi.org/10.1016/j.bbamcr.2014.08.001
4. Agarwala S, Veerappa AM, Ramachandra NB. Identification of primary copy number variations reveal enrichment of calcium, and mapk pathways sensitizing secondary sites for autism. *Egypt J Med Hum Genet* 2020;**21**:55. https://doi.org/10.1186/s43042-020-00091-3
5. Gao Y, Ni X, Guo H *et al.* Single-cell sequencing deciphers a convergent evolution of copy number alterations from primary to circulating tumor cells. *Genome Res* 2017;**27**:1312–22. https://doi.org/10.1101/gr.216788.116
6. Harbers L, Agostini F, Nicos M *et al.* Somatic copy number alterations in human cancers: An analysis of publicly available data from the cancer genome atlas. *Front Oncol* 2021;**11**:700568. https://doi.org/10.3389/fonc.2021.700568
7. Mermel CH, Schumacher SE, Hill B *et al.* Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011;**12**:1–14. https://doi.org/10.1186/gb-2011-12-4-r41
8. Feng X, Zheng J, Peng S *et al.* CNAScope: pan-cancer copy number aberration database with functional annotation and interactive visualization. *Nucleic Acids Res* 2025;**54**:D1364–75. https://doi.org/10.1093/nar/gkaf1242
9. Zhang L, Chen L, Li SC *et al.* Heterogeneity in lung cancers by single-cell dna sequencing. *Clin Transl Med* 2023;**13**:e1388. https://doi.org/10.1002/ctm2.1388
10. Sun C, Kathuria K, Emery SB *et al.* Mapping recurrent mosaic copy number variation in human neurons. *Nat Commun* 2024;**15**:4220.

11. Smajlagić D, Lavrichenko K, Berland S *et al.* Population prevalence and inheritance pattern of recurrent CNVS associated with neurodevelopmental disorders in 12,252 newborns and their parents. *Eur J Hum Genet* 2021;**29**:205–15. https://doi.org/10.1038/s41431-020-00707-7
12. Mei TS, Salim A, Stefano Calza K *et al.* Identification of recurrent regions of Copy-Number variants across multiple individuals. *BMC Bioinformatics* 2010;**11**:147. https://doi.org/10.1186/1471-2105-11-147
13. Ritz A, Paris PL, Ittmann MM *et al.* Detection of recurrent rearrangement breakpoints from copy number data. *BMC Bioinformatics* 2011;**12**:114. https://doi.org/10.1186/1471-2105-12-114
14. Van Dyk E, Hoogstraat M, Ten Hoeve J *et al.* RUBIC identifies driver genes by detecting recurrent DNA copy number breaks. *Nat Commun* 2016;**7**:12159. https://doi.org/10.1038/ncomms12159
15. Montalbano S, Sánchez XC, Vaez M *et al.* Accurate and effective detection of recurrent copy number variants in large SNP genotype datasets. *Current protocols* 2022;**2**:e621. https://doi.org/10.1002/cpz1.621
16. Zhang Z, Hernandez K, Savage J *et al.* Uniform genomic data analysis in the NCI Genomic Data Commons. *Nat Commun* 2021;**12**:1226. https://doi.org/10.1038/s41467-021-21254-9
17. Van Loo P, Nordgard SH, Lingjærde OC *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci* 2010;**107**:16910–5.
18. Raine KM, Van Loo P, Wedge DC *et al.* ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr Protoc Bioinformatics* 2016;**56**:15–9.
19. Harris CR, Jarrod Millman K, Van Der Walt SJ *et al.* Array programming with numpy. *Nature* 2020;**585**:357–62. https://doi.org/10.1038/s41586-020-2649-2
20. Paszke A, Gross S, Massa F *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv Neural Inf Proces Syst* 2019;**32**:8024–8035.
21. Wang S, He Z, Wang X *et al.* Antigen presentation and tumor immunogenicity in cancer immunotherapy response prediction. *Elife* 2019;**8**:e49020. https://doi.org/10.7554/eLife.49020
22. Stovner EB, Sætrom P. PyRanges: efficient comparison of genomic intervals in Python. *Bioinformatics* 2020;**36**:918–9.
23. Fang Z, Liu X, Peltz G. GSEApy: A comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* 2023;**39**:btac757.
24. Davidson-Pilon C. Lifelines: survival analysis in Python. *J Open Source Softw* 2019;**4**:1317. https://doi.org/10.21105/joss.01317
25. Zheng X, Chen L, Liu W *et al.* CCNE1 is a predictive and immunotherapeutic indicator in various cancers including UCEC: a pan-cancer analysis. *Hereditas* 2023;**160**:13. https://doi.org/10.1186/s41065-023-00273-0
26. Meng J, Zong C, Wang M *et al.* Constructing a prognostic model of uterine corpus endometrial carcinoma and predicting drug-sensitivity responses using programmed cell death-related pathways. *J Cancer* 2024;**15**:2948–59. https://doi.org/10.7150/jca.92201
27. Chiappetta G, Valentino T, Vitiello M *et al.* PATZ1 acts as a tumor suppressor in thyroid cancer via targeting p53-dependent genes involved in EMT and cell migration. *Oncotarget* 2015;**6**:5310–23. https://doi.org/10.18632/oncotarget.2776
28. Dave B, Gonzalez DD, Liu Z-B *et al.* Role of RPL39 in metaplastic breast cancer. *J Natl Cancer Inst* 2017;**109**:djw292. https://doi.org/10.1093/jnci/djw292
29. Chenying S, Cai X, Taotao X *et al.* Lims2 is downregulated in osteosarcoma and inhibits cell growth and migration. *J Oncol* 2022;**2022**:1–13. https://doi.org/10.1155/2022/4811260
30. Yang D, Jiao Y, Li Y *et al.* Clinical characteristics and prognostic value of MEX3A mRNA in liver cancer. *PeerJ* 2020;**8**:e8252. https://doi.org/10.7717/peerj.8252
31. Wang R, Lin D-Y, Jiang Y. SCOPE: a normalization and copy-number estimation method for single-cell DNA sequencing. *Cell Syst* 2020;**10**:445–452.e6. https://doi.org/10.1016/j.cels.2020.03.005
32. Feng X, Chen L, Qing Y *et al.* SCYN: single cell CNV profiling method using dynamic programming. *BMC Genomics* 2021;**22**:651.
33. Feng X, Chen L. SCSilicon: a tool for synthetic single-cell DNA sequencing data generation. *BMC Genomics* 2022;**23**:359. https://doi.org/10.1186/s12864-022-08566-w
34. Li C, Chen L, Pan G *et al.* Deciphering complex breakage-fusion-bridge genome rearrangements with ambigram. *NatCommun* 2023;**14**:5528. https://doi.org/10.1038/s41467-023-41259-w
35. Chen L, Qing Y, Li R *et al.* Somatic variant analysis suite: copy number variation clonal visualization online platform for large-scale single-cell genomics. *Brief Bioinform* 2022;**23**:bbab452.





*Briefings in Bioinformatics*, 2026, **27**, bbag167
https://doi.org/10.1093/bib/bbag167
Problem Solving Protocol