nature communications

Article

https://doi.org/10.1038/s41467-025-64812-1

Discriminating single-molecule binding events from diffraction-limited fluorescence

Received: 22 November 2024

Accepted: 26 September 2025

Published online: 07 November 2025

Check for updates

Yueming Yin[1], Nithin Pathoor[2], Kamal Kant Sharma[2], Shiwen Zhu[2], Iong Ying Loh[3], Yan Shan Ang[3], Shao Ren Sim[2], Lin Yue Lanry Yung[3], Thorsten Wohland[1,2,4] & Lipo Wang[1,5]

Single-molecule localization microscopy enables high-resolution imaging of molecular interactions, but discriminating molecular binding types has traditionally relied on complex strategies, such as multiple dyes, time-division techniques, or kinetic analysis, that are asynchronous, invasive, or time-consuming. Here, we uncover previously overlooked spatiotemporal information embedded within diffraction-limited fluorescence, enabling synchronous classification of individual binding event videos using only a single fluorescent dye. Building on this insight, we propose a Temporal-to-Context Convolutional Neural Network (T2C CNN), which integrates long-term spatial convolutions, shallow cross-connected blocks, and a pooling-free structure to enhance contextual representation while preserving fine-grained temporal features. Applied to DNA-PAINT experiments, T2C CNN achieves up to 94.76% classification accuracy and outperforms state-of-the-art video classification models by 15-25 percentage points. Our approach enables rapid and precise binding-type recognition from fluorescence video data, reducing observation time from minutes to seconds and facilitating high-throughput single-molecule imaging without requiring multiple dye channels or extended kinetic measurements.

Single-molecule localization microscopy (SMLM) has revolutionized the understanding of biological systems by enabling visualization at the nanoscale[1]. Techniques like photoactivated localization microscopy (PALM)[2] and stochastic optical reconstruction microscopy (STORM)[3] have paved the way for this revolution by overcoming the diffraction limit of light and providing insights into molecular structures and dynamics with high resolution.

Traditionally, discriminating single-molecule binding types in SMLM relies on several methods. These include the use of different fluorescent dyes[1,4,5], temporal separation of fluorophores[6–8], or the analysis of blinking kinetics such as the binding on/off times of multiple binding events[9–13]. While effective, these approaches come with significant drawbacks. The use of different fluorescent dyes adds complexity to experimental design and analysis and may introduce potential issues, such as phototoxicity[14] and dye crosstalk[15]. Temporal separation of fluorophores can lead to asynchronization issues, and analyzing blinking kinetics necessitates prolonged observation periods, limiting the practicality of these methods for certain applications, such as high-throughput studies[16], live-cell imaging[17], and real-time imaging[18].

Recent advances in deep learning, particularly convolutional neural networks (CNNs), have made a significant impact across diverse fields, including fluorescence microscopy[19], molecular imaging[20], and single-molecule analysis[21]. While CNNs have been widely adopted in

[1]Institute for Digital Molecular Analytics and Science (IDMxS), Nanyang Technological University, Singapore, Singapore. [2]Centre for Bioimaging Sciences, Department of Biological Sciences, National University of Singapore, Singapore, Singapore. [3]Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore, Singapore. [4]Department of Chemistry, National University of Singapore, Singapore, Singapore. [5]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore. e-mail: twohland@nus.edu.sg; elpwang@ntu.edu.sg

image processing and video classification, their application to single-molecule detection and fluorescence microscopy is still in its early stages. Previous studies[21–24] demonstrated that deep learning models can enhance detection accuracy and reduce analysis time in single-molecule localization microscopy (SMLM) experiments. However, these approaches have yet to fully exploit CNNs' ability to process both temporal and spatial information within a single model.

In this work, we reveal previously overlooked discriminative spatiotemporal information within diffraction-limited fluorescent spots, enabling synchronous classification of binding types at the single-event level using the same fluorescent dye. To leverage this insight, we propose a convolutional neural network architecture, Temporal-to-Context (T2C) CNN, which transforms long temporal fluorescence signals into enriched contextual representations. While temporal-to-channel integration has been explored in generic video analysis[25], its application to fluorescence microscopy presents unique challenges due to low signal-to-noise ratios and subtle spatiotemporal dynamics. In our design, the temporal dimension is reshaped into the channel axis, allowing spatial convolutions to capture long-range temporal dependencies–a strategy we term "long-term spatial convolution". We demonstrate that the combination of three architectural elements–long-term spatial convolutions, shallow cross-connected blocks, and a pooling-free design–enables the effective capture of fine-grained temporal context and multi-scale features essential for binding-type classification. This combination enables robust and generalizable classification performance under noisy conditions, and its effectiveness is supported by ablation experiments and comparisons to state-of-the-art deep learning baselines. We validate the approach using DNA-PAINT (Points Accumulation for Imaging in Nanoscale Topography[26,27]), where the T2C CNN achieves a substantial increase in classification accuracy–from 75% using probability density function (PDF) estimation based on binding time to approximately 95%. This performance gain is accompanied by a significant reduction in measurement time, from 10 min to just 5 s, enabled by the model's efficient spatiotemporal pattern recognition. Moreover, T2C CNN significantly outperforms representative state-of-the-art deep learning methods–including 3D ResNet-18[28], Video Transformer[29], ED-TCN[30], and SqueezeTime[25]–which achieve accuracies ranging from approximately 70% to 80% on the same dataset. Beyond superior accuracy, T2C CNN offers practical advantages: its model size is 1.19 to 77 times smaller, and it requires 4 to 534 times less computation than these alternatives. These features make T2C CNN also efficient, facilitating broader deployment in real-time, resource-constrained, or high-throughput single-molecule imaging and sensing applications. These results underscore the potential of T2C CNN to greatly enhance the analytical capabilities of SMLM and other fluorescence-based techniques for rapid and precise molecular investigations.

Results

Overview of raw data and methods

DNA-PAINT experiments generate raw data as videos, with each frame containing multiple diffraction-limited fluorescent spots. To minimize spot overlap, experimental parameters such as concentration and pH are carefully adjusted to control spot density. Figure 1a displays a representative single frame (100 ms exposure) from such a video. Each frame corresponds to a snapshot in time and captures transient binding events individually. The dynamic appearance and disappearance of spots over consecutive frames reflect the stochastic binding and unbinding of imager strands at docking sites. Traditional analysis methods, such as Picasso[26], detect and localize fluorescent spots, perform drift correction, group spots, and extract multiple binding events at each site to generate a sequence of on-off signals (Fig. 1b).

Binding time, or on-time, is defined as the duration for which an imager strand remains bound to the docking strand before dissociating. It is detected by local intensity maxima with a gradient decrease toward surrounding pixels[26] and depends on the binding affinity and local imager strand concentration. Off-time, also known as dark time, refers to the interval between the dissociation of one imager strand and the subsequent binding of another to the same docking site, influenced by imager strand concentration and diffusion kinetics.

Length-based methods classify binding affinities by comparing average binding durations. While effective for distinguishing interactions with large affinity differences, they struggle to differentiate binding types with similar affinities. To overcome this limitation, we propose an alternative approach that directly extracts discriminative features from raw video signals, referred to as image-based methods. As illustrated in Fig. 1b, length-based methods classify molecule types based on all binding events at a site, whereas image-based methods classify individual binding events. Extracting meaningful features from video signals improves binding type resolution, reduces the need for repeated observations, and enables real-time, high-throughput molecular detection. The proposed T2C CNN, an image-based method, processes videos of diffraction-limited fluorescent spots captured during single DNA binding events and classifies each event into predefined DNA binding types. The following sections present the experimental results, discussion, and methodology on these data and methods.

Binding-type information in diffraction-limited fluorescent spots

A standard DNA binding design in DNA-PAINT[11] is used in this study (as shown in Fig. 2a), with different docking strands for the same 10nt imager strand 5′-GTAATGAAGA-3′: partially complementary 8nt 5′-TT-TCTTCATT-3′ (domain 1) and fully complementary 10nt 5′-TT-TCTTCATTAC-3′ (domain 2), where "-TT-" is a spacer between the docking strands and the DNA origami. Appropriate concentrations were chosen: 1 nM imager strands, 1 nM domain-1 docking strands, and 200 pM domain-2 docking strands. Note that the concentration of domain-2 docking strands was reduced to decrease the probability of spatial and temporal overlap of binding events at different sites, due to its longer binding time. We measured fluorescence microscopy of 20,000 frames (33 min and 20 s duration, 10 fps frame rate) for both domains. After drift correction, we identified 4977 and 1183 binding sites with 19,457 and 6073 binding events for domains 1 and 2, respectively. We analyzed 25,530 binding events from the two domains by applying background and blinking corrections to the diffraction-limited fluorescence spots, computing inter-frame correlations and summary statistics, and evaluating their association with domain labels.

Frame correlation quantifies the similarity between frames based on pixel intensity patterns, revealing temporal and structural consistency (definition in Supplementary Note 1). Figure 3 shows the average inter-frame Pearson correlations for both domains. Initially, both Domain 1 (8nt-10nt) and Domain 2 (10nt-10nt) show stable correlations. However, Domain 2 maintains this stability longer, while Domain 1 rapidly develops alternating blocks of high and low correlation–mainly due to fewer long-duration events. These intermittent low-correlation segments in Domain 1 may arise from increased torsional and lateral fluctuations enabled by its asymmetric partial duplex geometry[31] and flexible overhangs[32], which subtly shift the spot pattern across frames without reducing mean intensity. This is likely due to a shorter binding duration, which reduces the inter-frame persistence of fluorophore localization. Such correlation patterns provide useful cues for identifying heterogeneity in binding site behavior in fluorescence videos.

Common statistics of diffraction-limited fluorescent spots (Fig. 2c) reveal up to 0.5 Pearson correlation with binding type, beyond binding time alone. These include the following. Sum features (e.g., "Sum", "YZ sum", "XZ sum", "XY sum"): Partially matched 8nt-10nt

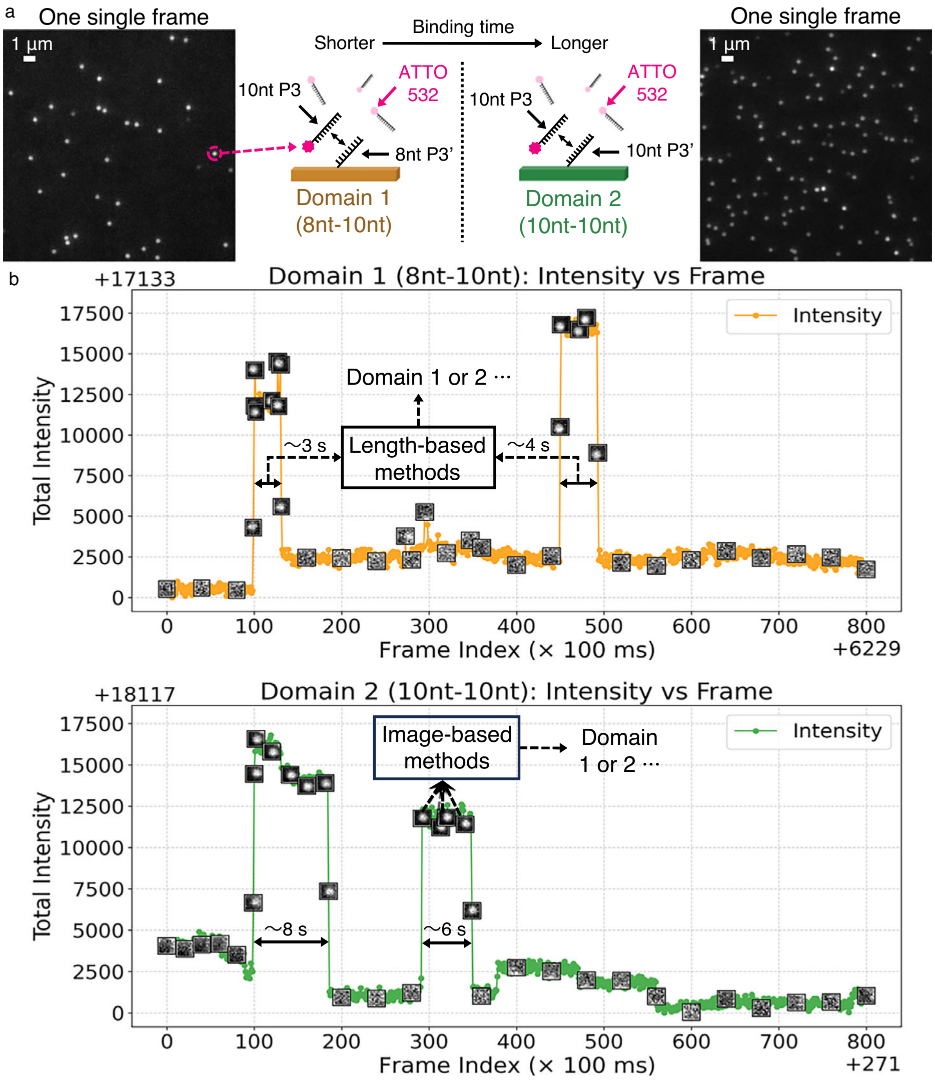


Fig. 1 | An example of raw images and time courses from the DNA-PAINT experiments used in this study. a Schematic of the two DNA-binding domains used to collect diffraction-limited fluorescent spots (adapted from ref. 11). Both domains share the same dye ("ATO532") and strand ("10nt P3"), but differ in their complementary sequences: Domain 1 uses "8nt P3'" (a partial complement of P3), while Domain 2 uses the fully complementary "10nt P3'". Example frames from the raw video data are shown next to the corresponding domains. **b** Example time courses annotated with diffraction-limited fluorescent spots from binding events in Domain 1 and Domain 2. Domain 2 exhibits longer binding durations than Domain 1 (dwell time analysis shows t_{on} as 2.30 ± 2.38 s for Domain 1 and 11.56 ± 16.27 s for Domain 2, with survival curves provided in Supplementary Fig. 1), while both show similar intensity jumps. The length-based method classifies molecules by total binding durations, whereas the image-based method classifies individual events. Axis offsets (+17133/+18117, +6229/+271) indicate base intensity and event start frame, respectively. Source data are provided as a Source Data file.

bindings exhibit lower overall fluorescence intensity than fully matched 10nt-10nt bindings, as observed in projections onto width-time (YZ), height-time (XZ), and height-width (XY) planes. Non-zero features (e.g., "non-zero", "YZ non-zero", "XZ non-zero"): Fully matched 10nt-10nt bindings show fewer blinking or localization losses than partially matched 8nt-10nt bindings. Signal loss leads to zero values in the data since background intensities have been removed. "XY non-zero" is less informative because the time axis is collapsed, masking temporal fluctuations. Non-masked features (e.g., "non-masked", "YZ non-masked", "XZ non-masked"): Partially matched 8nt-10nt bindings exhibit weaker edge intensities, which are masked if their values fall below the average background intensity within the same frame. As with the previous metric, "XY non-masked" provides limited insight due to time axis overlap in the XY projection. XY statistics (e.g., "mean", "std", "min", "max", "median", "range"): 8nt-10nt bindings show lower average intensity, less variation, and more random spatial changes.

In this paper, "missing information" is defined as the gap between the current classification accuracy and an ideal scenario where all discriminative features are captured (i.e., 100% accuracy). This gap quantifies the amount of additional information needed to achieve perfect classification. By voting on binding types using all the statistics in Fig. 2c weighted by their correlation with binding type, we found that the classification accuracy could potentially be improved by 4.05% in Domain 1 and 37.75% in Domain 2. These differences indicate that the available statistics capture most of the necessary discriminative features in Domain 1, but much less so in Domain 2, where a substantial portion of key information is missing. The proposed T2C CNN compensates for the missing information by extracting additional spatio-temporal features from fluorescence videos, recovering 0.25% in Domain 1 and 34.72% in Domain 2, ultimately improving classification accuracy to 96% in both domains. This highlights the effectiveness of T2C CNN in enhancing image-based binding type classification, where traditional image statistical features are insufficient.

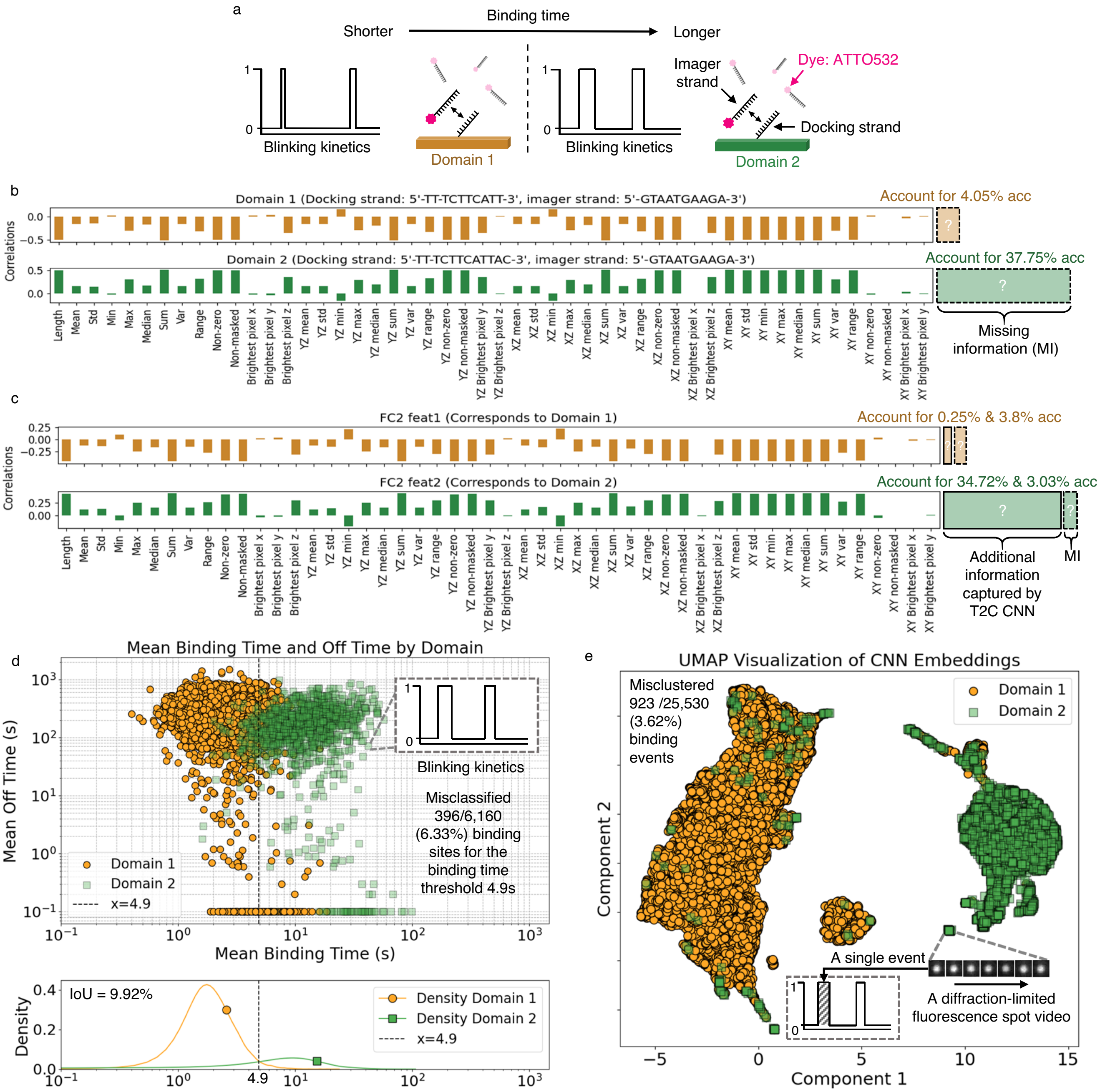


Fig. 2 | Diffraction-limited fluorescent spots hold binding-type information that can be utilized to more accurately differentiate between single binding events. a Two types of DNA-binding domains used to collect diffraction-limited fluorescent spots. Blinking kinetics refers to the binary signals of binding on and off. (Revised from ref. 11). **b** The correlation between domain labels and some common statistics (n = 25,530 independent binding events), as well as the classification accuracy accounted for by the missing information. The "YZ," "XZ," and "XY" statistics refer to those calculated from projections of the video onto the Y-Z, X-Z, and X-Y planes, with X as the width, Y as the height, and Z as the time axis. **c** The correlation between the final classification output of the proposed T2C CNN (FC2 feat1 and FC2 feat2) and some common statistics, as well as the classification accuracy accounted for by the additional information captured by the model and the remaining missing information. **d** Scatter plot of mean on-time versus mean off-time. Below, the Gaussian kernel density estimate of the probability distribution function (PDF) for the mean on-time is displayed. Based on multiple binding events, 6.33% of binding sites are misclassified. **e** A 2D uniform manifold approximation and projection (UMAP) plot of the output embeddings from the convolutional layers of the proposed T2C CNN. This model achieves a lower error rate (3.62%) in more fine-grained recognition of single binding events, significantly reducing the time required to determine the binding type for a binding site. Clearer visualizations with each domain plotted on top are provided in Supplementary Fig. 2. Source data are provided as a Source Data file.

To evaluate the temporal information in fluorescence image sequences, we applied three scrambling strategies: within events (removing order), across events (preserving length but mixing events), and into random-length segments (removing both order and structure). As shown in Supplementary Table 1, the performance of the image-based method (T2C CNN) degrades progressively with increasing disruption. Scrambling within events moderately reduces accuracy (DNA Origami: 94.76% → 89.47%; Cell: 74.09% → 71.59%), highlighting the value of temporal order. Cross-event scrambling further lowers accuracy (DNA Origami: 87.66%; Cell: 70.25%), emphasizing the role of event-level coherence. Random-length scrambling yields the lowest accuracy (DNA Origami: 78.98%; Cell: 62.64%), falling

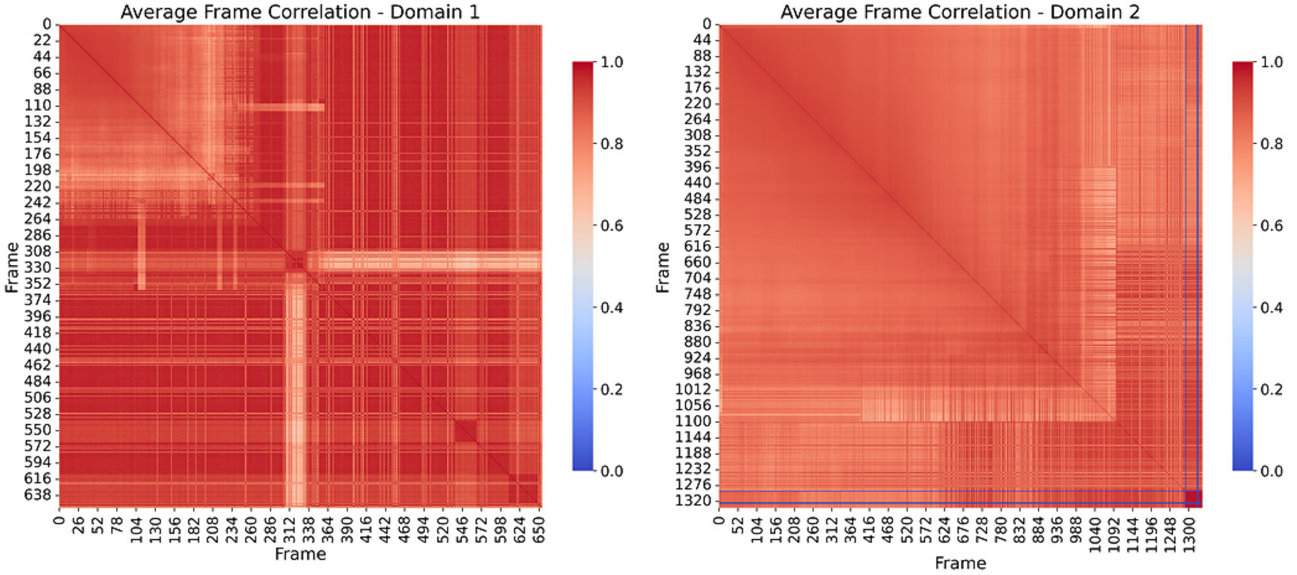


Fig. 3 | Inter-frame correlations. Correlations between diffraction-limited fluorescent spot frames during binding events for Domain 1 and Domain 2. Detailed definitions, analyses, and visualizations are provided in Supplementary Note 1, 2, and Supplementary Fig. 3, respectively. Source data are provided as a Source Data file.

below even length-based methods (DNA Origami: 83.88%; Cell: 66.34%), suggesting that disrupting both order and structure severely degrades discriminative power. These results confirm that both temporal order and event structure of fluorescence image sequences encode meaningful binding information. Additional results are provided in subsequent sections.

Cross-experiment evaluation of existing methods for discriminating single-dye binding types

Using the same DNA binding type design as in previous sections, we increased the domain-2 docking strand concentration from 200 pM to 400 pM and recollected fluorescence microscopy videos of equal duration for both domains. Binding event data from this setup were used to train and evaluate cross-experiment classification models, including: (1) PDF[33]–Gaussian kernel density estimation of binding durations; (2) MLP[34]–trained on sliced binding durations; (3) 3D ResNet-18[28]–a standard video classification CNN; (4) Video Transformer[29]–a self-attention-based video classification model; and (5) T2C CNN–the proposed CNN designed for classifying binding types from stacks of diffraction-limited fluorescent spot images.

Classification accuracy and measurement time (with standard deviations) for all methods are shown in Fig. 4a. The PDF and MLP models, relying on binding durations, reach only ~75% accuracy per event and require at least 3 min of observation to exceed 85%. In contrast, image-based methods (dashed edges in the plot), such as 3D ResNet-18, achieve ~80% accuracy per 5-s binding event. The proposed T2C CNN attains ~95% accuracy with only one 5-s event by using cross-connected long-term spatial convolutions. Unlike traditional 3D CNNs (e.g., 3D ResNet-18) with shallow temporal strides, T2C CNN treats long temporal stacks as channels, enabling 2D convolutions equivalent to long-step 3D convolutions. This design captures frequency variations across spatial regions, effectively encoding DNA binding dynamics that can affect fluorescence intensity, angle, and density. Although orientation-dependent effects are likely averaged out over the 100 ms frame duration, hybridization or partial immobilization may still induce subtle changes in fluorophore behavior. For instance, the linker length and binding strength can influence the mobility[35] of the fluorophore bound to the docking strands. Even though the exact reason is undetermined, the slight changes in the PSF[36] time averaged to 100 ms can be one of the origins of enhanced classification accuracy with diffraction-limited images. Figure 4b shows confusion matrices from the best-performing models across five cross-validation experiments. Duration-based (PDF) and standard image-based models (Video Transformer, 3D ResNet-18) often misclassify Domain 2 events as Domain 1. In contrast, T2C CNN significantly reduces such errors by better distinguishing subtle binding type differences. Figure 4c visualizes predicted binding types with pseudo-colored images across increasing measurement times. T2C CNN consistently reconstructs ground-truth binding site distributions with 95–100% accuracy, enabling reliable interpretation of multi-target fluorescence microscopy data.

Multi-class experiments

We extended the proposed approach to three-class classification by introducing an additional 6nt–6nt R1 strand binding type (Fig. 5a). As shown in Fig. 5c, T2C CNN outperforms state-of-the-art video models by a large margin, offering a robust tool for high-accuracy, single-fluorophore, three-target SR imaging. Additional results are provided in Supplementary Note 3 and Supplementary Fig. 4.

Interpretations of the T2C CNN

The proposed T2C CNN takes as input the video of diffraction-limited fluorescent spots, captured by a microscope, that result from the excitation of conjugated dyes in single DNA binding events. The output is the classification of the DNA binding event into predefined DNA binding types. This process does not require manually defined video features, which are difficult to discern with the naked eye. In this section, we attempt to explain the additional information captured by the T2C CNN to differentiate between different DNA-binding types, building on the cross-experiment studies.

First, we sought to identify which part of the diffraction-limited fluorescent spot most accurately reflects the DNA binding type. By overlaying the model input with its saliency map, we found that the edge region of the diffraction-limited fluorescence spot contributes most to the model's prediction (an example is shown in Fig. 6a). This is likely because the edge plays a crucial role in forming the overall diffraction pattern. In diffraction-limited fluorescence imaging, the edges of the spots often exhibit unique interference patterns and intensity gradients that encode critical information about the underlying molecular interactions[37]. The saliency map generated by T2C CNN shows that these edge regions are particularly influential in the model's decision-making process. This suggests that T2C CNN leverages the

https://doi.org/10.1038/s41467-025-64812-1

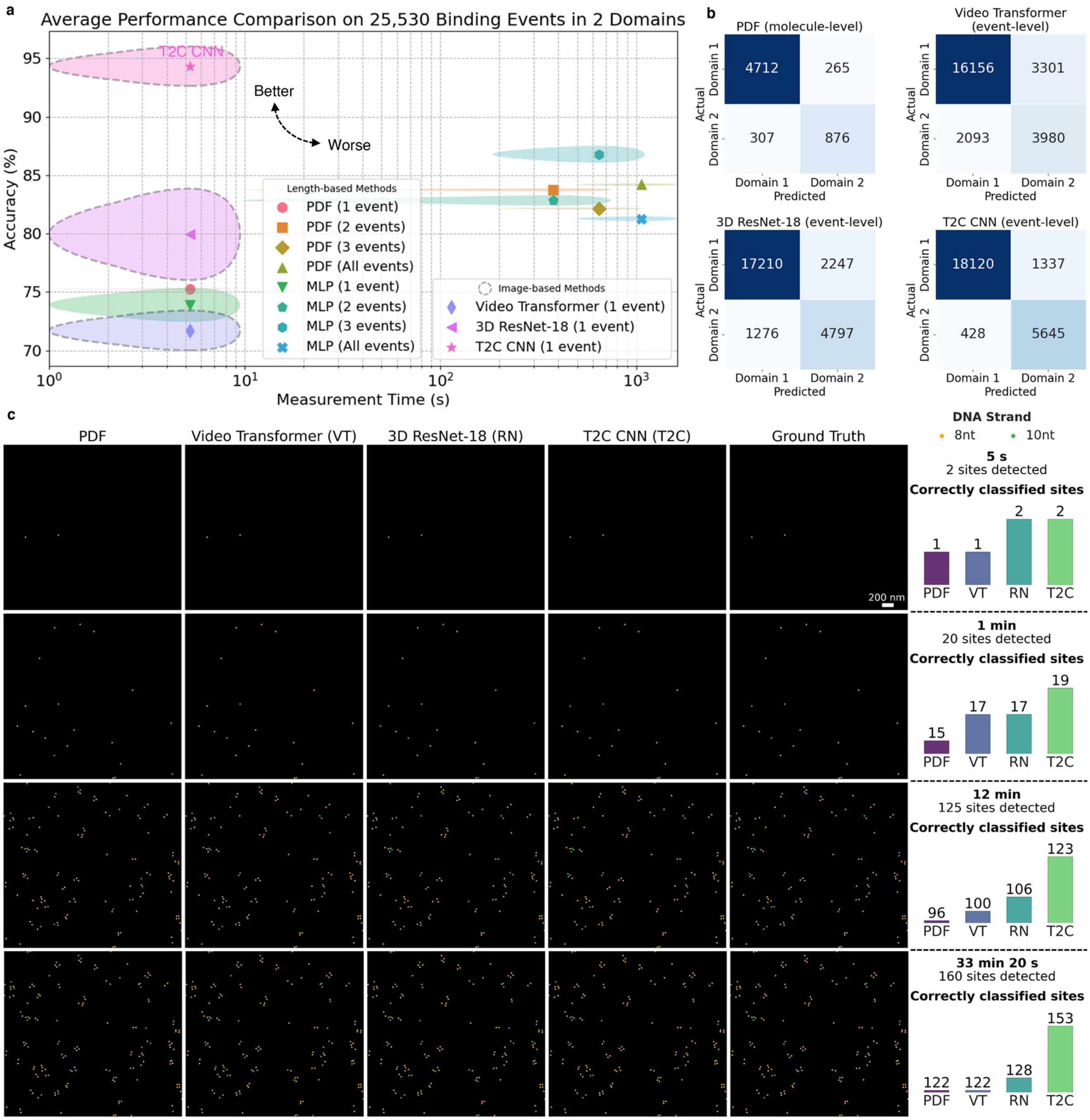
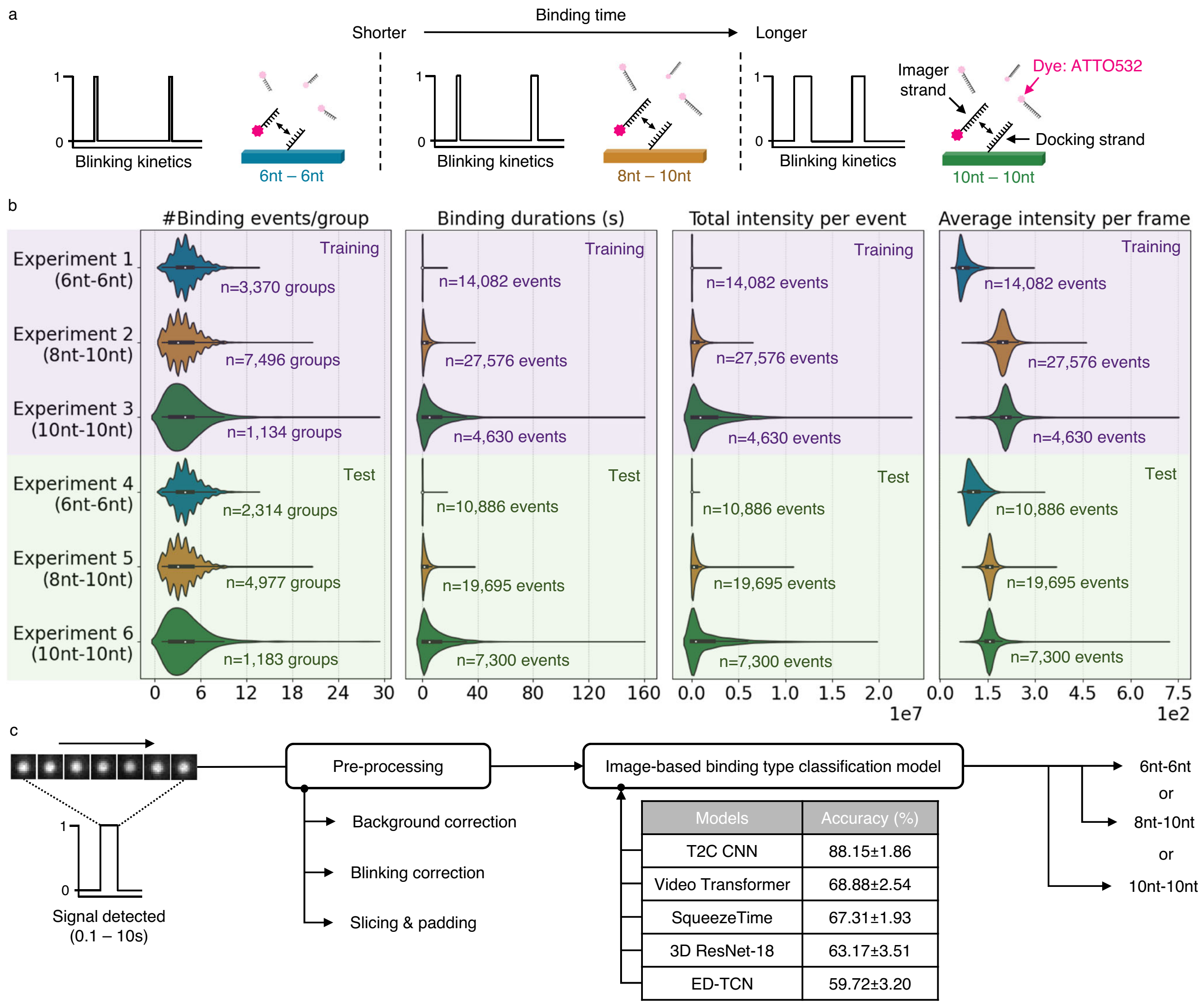


Fig. 4 | Cross-experiment results for discriminating binding types. **a** Comparison of average performance on 25,530 binding events across the two domains. The width of the colored regions in the horizontal and vertical directions represents the standard deviation of measurement time and accuracy, respectively, across five cross-validation trials. **b** Comparison of confusion matrices. The PDF method provides molecule-level predictions based on multiple binding events. In contrast, other methods are image-based and offer predictions at the single-event level. **c** Comparison of pseudo-colored composite fluorescence images, created by frame-wise summation of the two separated videos representing different binding types. By localizing fluorescence spots before summation, we accurately assigned each spot in the composite video to its corresponding binding type, establishing the ground truth. Image-based methods were trained on the mixed data from the composite video, without knowledge of the original intensities in the separate videos. The rightmost column displays the observation time and the classification accuracy for binding sites, representing the cumulative accuracy of binding type classification for all events occurring at the site. To reduce spot overlap from longer binding events, a lower concentration was used for Domain 2 (10nt) than for Domain 1 (8nt), leading to a higher number of observed 8nt binding sites over time. Source data are provided as a Source Data file.

subtle variations along the edges–such as differences in brightness, shape, and gradient–that arise from diffraction effects to distinguish between binding types. By focusing on these edge features, the network can capture nuanced differences that may not be apparent in the central region of the spots, leading to more accurate classification outcomes.

We also analyzed the correlation between the model output and common statistics of the diffraction-limited fluorescent spot video (Fig. 2d), which generally reflects the true domain correlations (Fig. 2c). Additionally, we performed an F-test on the features output by T2C CNN (Fig. 6b), revealing that the majority of features have a high F-statistic (F >100). This indicates that the model has learned



Models	Accuracy (%)
T2C CNN	88.15±1.86
Video Transformer	68.88±2.54
SqueezeTime	67.31±1.93
3D ResNet-18	63.17±3.51
ED-TCN	59.72±3.20

Fig. 5 | Multi-class data, workflow, and results. a Illustration of the three binding types (6nt–6nt, 8nt–10nt, and 10nt–10nt) used in the multi-class experiments. **b** Data distributions from the experiments. Each of the three binding types includes two biologically independent replicates (Experiments 1 & 4, 2 & 5, and 3 & 6), with one replicate used for training and the other for testing. Violin plots show the full distribution of the data. The embedded boxplots indicate the 25th-75th percentiles (bounds of the box), the median (center line), and the minima and maxima (whiskers). Notably, the 6nt R1 imager exhibited lower average frame intensity (approximately 50 and 100 for training and testing) due to its shorter binding time, which sometimes resulted in incomplete fluorescence accumulation within a single frame. **c** Workflow and results of image-based binding type classification. Source data are provided as a Source Data file.

discriminative features from the diffraction-limited fluorescent spot videos to distinguish DNA binding types.

Next, we explored the importance of different periods of the DNA-binding events in determining the DNA-binding type. For each DNA-binding event's diffraction-limited fluorescent spot video, we sequentially masked out 10% of the frames (not masking fewer than 10 frames). As shown in Fig. 6c, when we masked the first 10% or between 10% and 20% of the frames, the classification accuracy of domain 1 DNA binding types dropped from around 96% to below 90%. We hypothesize that this is because the early binding stages of a partially matched 8nt docking strand with a 10nt imager strand in domain 1 are still unstable, compared to the more stable early binding stages of a fully matched 10nt docking strand in domain 2. This hypothesis is consistent with the conclusions in Fig. 3.

Finally, we evaluated T2C CNN on DNA binding events of varying durations. As shown in Fig. 6d, T2C CNN achieves ≥87.5% accuracy for events shorter than 4.9 s (49 frames), which constitute 72.9% of all 25,530 events. The probability density function in Fig. 2e shows substantial overlap between the two domains in this range, underscoring the challenge addressed by the proposed method. Occasionally, lower accuracies (<80%) arise in a small subset (1.6%) of longer events. Notably, only 1–6 events per binding length category (across 12 of 463 categories, or 2.6%) yielded 0% or 50% accuracy, remaining within acceptable limits for typical applications.

Interestingly, T2C CNN achieves approximately 95% accuracy in distinguishing binding types on those single-frame events (see Fig.6d at $x = 10^0$). Accordingly, Fig.6e illustrates a representative example of single-frame events from both domains, including the input image, a representative feature map, and the corresponding classification probability. Feature maps reveal differences in spot shape (e.g., round vs. square) and pixel intensity distribution (e.g., concentrated vs. dispersed). The model consistently outputs high-confidence predictions

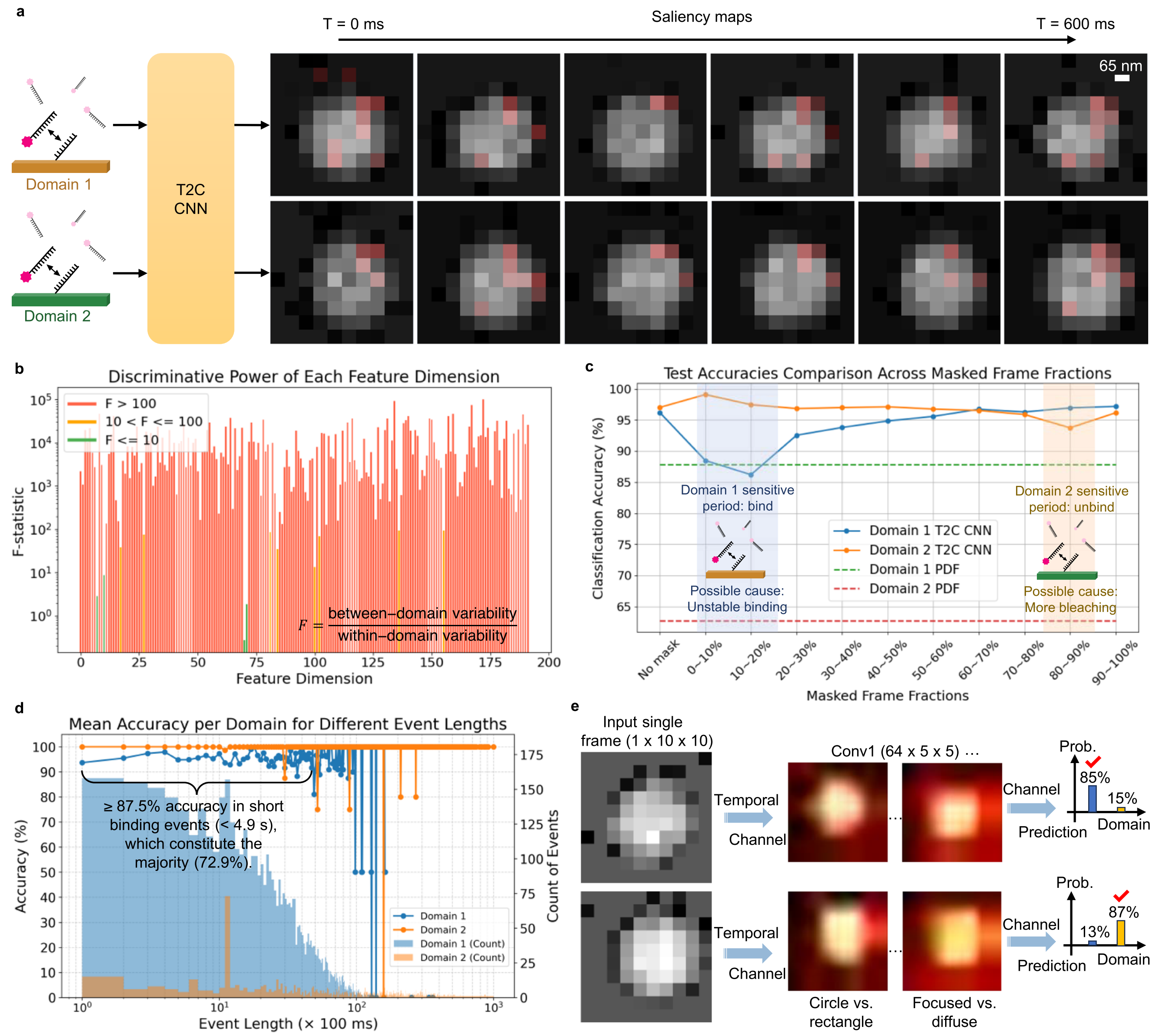


Fig. 6 | Interpretations of the temporal-to-context convolutional neural network (T2C CNN). a T2C CNN saliency maps for diffraction-limited fluorescent spots in different DNA binding domains. The saliency map visually highlights the regions within each frame that T2C CNN considers most critical for its classification decisions. **b** F-test on T2C CNN convolutional layer outputs, with each domain treated as a separate group. **c** Comparison of T2C CNN binding type classification accuracies with different masked frame sections. **d** Average classification accuracies of T2C CNN for binding events of varying lengths. The peak at 1.1 s arises from merged short events due to blinking correction thresholds (1 s). **e** Example T2C CNN feature heatmaps and classification probabilities (Prob.) for single-frame diffraction-limited fluorescent spots, demonstrating that T2C CNN assigns a classification probability of over 85% to similar single-frame fluorescence spots in the correct domain, attributable to its ability to capture discriminative features through learnable convolutions. As the first learnable convolution layer, "Conv1" applies 64 convolutional kernels of size (3 × 3) with a (2 × 2) stride and (1,1,1,1) padding, converting the (1 × 10 × 10) input images into 64 feature maps with spatial dimensions of (5 × 5). More examples and analysis are provided in Supplementary Figs. 5, 6, 7, and Supplementary Note 4. Source data are provided as a Source Data file.

(with the confidence ≥85%). These results explain how the T2C CNN works at the single-frame level. For multi-frame spots, the model integrates temporal variations of these features. Given its ability to differentiate single-frame spots from the same dye, it should, in principle, also distinguish different dyes–explored further in the last section of Results.

Robustness test of image-based binding-type classification models

In single-molecule fluorescence binding experiments, nonspecific binding, background intensity noise, and camera defects are common issues, making it crucial to evaluate the robustness of various methods under noisy conditions. Supplementary Note 5 and Fig. 7 assess the robustness of image-based binding-type classification models (3D ResNet, Video Transformer, ED-TCN, SqueezeTime, and the proposed T2C CNN) by simulating these three types of noise interference. Among all tested models, the proposed T2C CNN consistently demonstrated strong robustness, particularly against Poisson and Gaussian noise, due to its temporal-to-channel transformation and efficient parameter design. These results suggest that T2C CNN can reliably handle signal disturbances commonly encountered in fluorescence imaging, provided that interference spot intensity is kept

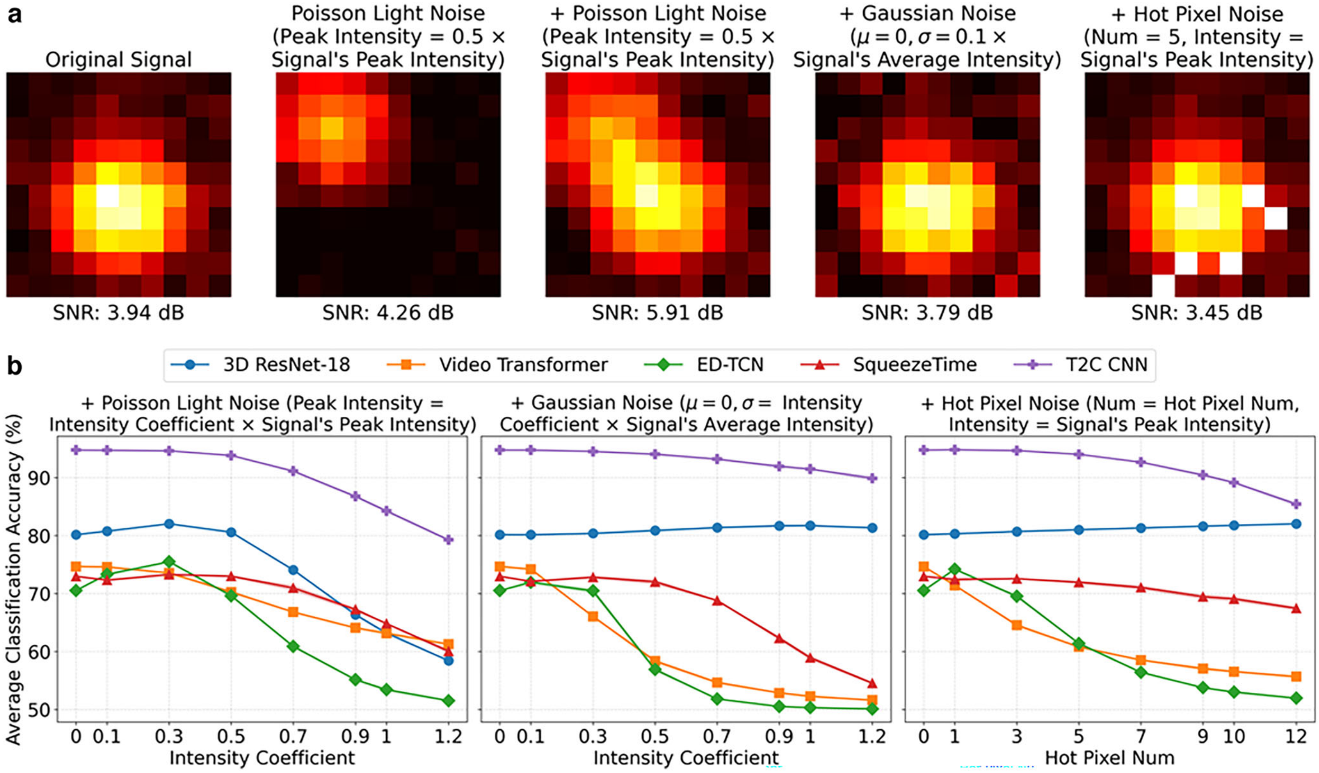


Fig. 7 | Robustness test results. a Examples of generated noise added to the original signal, including Poisson light noise with Gaussian decay, Gaussian noise, and hot pixel noise. The symbols μ and σ denote the mean and standard deviation (std.), respectively. **b** Average classification accuracies of image-based binding type classification models under different noise conditions. The shaded regions represent the std. of accuracies from noises generated with 5 random seeds. Models exhibit stable performance across different random seeds (with minimal std., resulting in nearly invisible shading) but show performance degradation with increasing noise intensity. Source data are provided as a Source Data file.

below half that of the signal and the effective pixel ratio—the proportion of the signal area that remains unobstructed by camera defects—exceeds 93%.

Hyperparameter analysis of the T2C CNN

We systematically evaluated the impact of network depth, width, and input slice length on T2C CNN's performance (Fig. 8 and Supplementary Note 6). An optimal depth of 8 layers (two per block) achieved the highest accuracy (94.76%). Configurations with evenly distributed layers (4, 8, and 12) consistently outperformed uneven ones (e.g., 5, 6, 7, 9, 10, 11), underscoring the importance of architectural symmetry. The default width of 64 provided the best trade-off between accuracy and efficiency, with 128 yielding only marginal gains (94.76% → 95.38%) at the cost of doubling model size and computation. Remarkably, even with a small width of 16, accuracy remained above 90%, indicating T2C CNN's robust feature extraction and the redundancy in fluorescence video data. Regarding input length, increasing the slice length improved performance up to 512 frames by capturing complete binding events, with performance plateauing beyond that.

In Fig. 8, we also included the baseline models with their optimal depths and widths, which were evaluated on randomly split training and validation datasets (Supplementary Table 2).

Ablation study of T2C CNN architecture

We dissected T2C CNN's architecture to assess the contributions of its key blocks and design components (Supplementary Note 7, Supplementary Table 2). Removing either the second (Hidden Transformation) or third (Multi-Scale Feature Fusion) block significantly degraded performance (94.76% → 90.33% and 88.80%, respectively), confirming their critical roles. The second block proved particularly essential for transforming global features into discriminative representations that facilitate effective multi-scale fusion. Further analysis of individual components—long-term spatial convolutions, skip concatenations, and no-pooling—revealed that each contributed incrementally to performance, with their combined use achieving the highest accuracy. Notably, the default configuration also yielded the lowest standard deviation, reflecting stable generalization. The optimal configuration consistently outperformed the baseline across both single-event and single-molecule tasks. These results demonstrate that the full T2C CNN architecture forms a synergistic and domain-adapted design optimized for robust classification under noisy single-dye fluorescence conditions.

Statistical significance analysis of model performance

A Wilcoxon signed-rank test conducted on single-event class-wise accuracies across five folds confirmed that T2C CNN significantly outperformed all of its ablated variants ($p = 0.031$). To assess performance across a broader set of baseline models—including all seven ablated variants, four PDF variants (1-event, 2-event, 3-event, and all-event) and four image-based methods—we performed a Friedman test on the same single-event accuracies (Supplementary Table 4), which revealed a statistically significant difference among the 16 methods ($\chi^2 = 64.11$, $p = 4.89 \times 10^{-8}$). All models were evaluated using five-fold cross-validation and tested on an independent sample. Although additional random seeds were not explored, the cross-validation results were consistent across folds, indicating stable performance and low sensitivity to model initialization. Post-hoc analysis using the Nemenyi test further showed that T2C CNN significantly outperformed the state-of-the-art baselines, including SqueezeTime ($p = 0.0030$), ED-TCN ($p = 0.0023$), and Video Transformer ($p = 0.0145$). These results

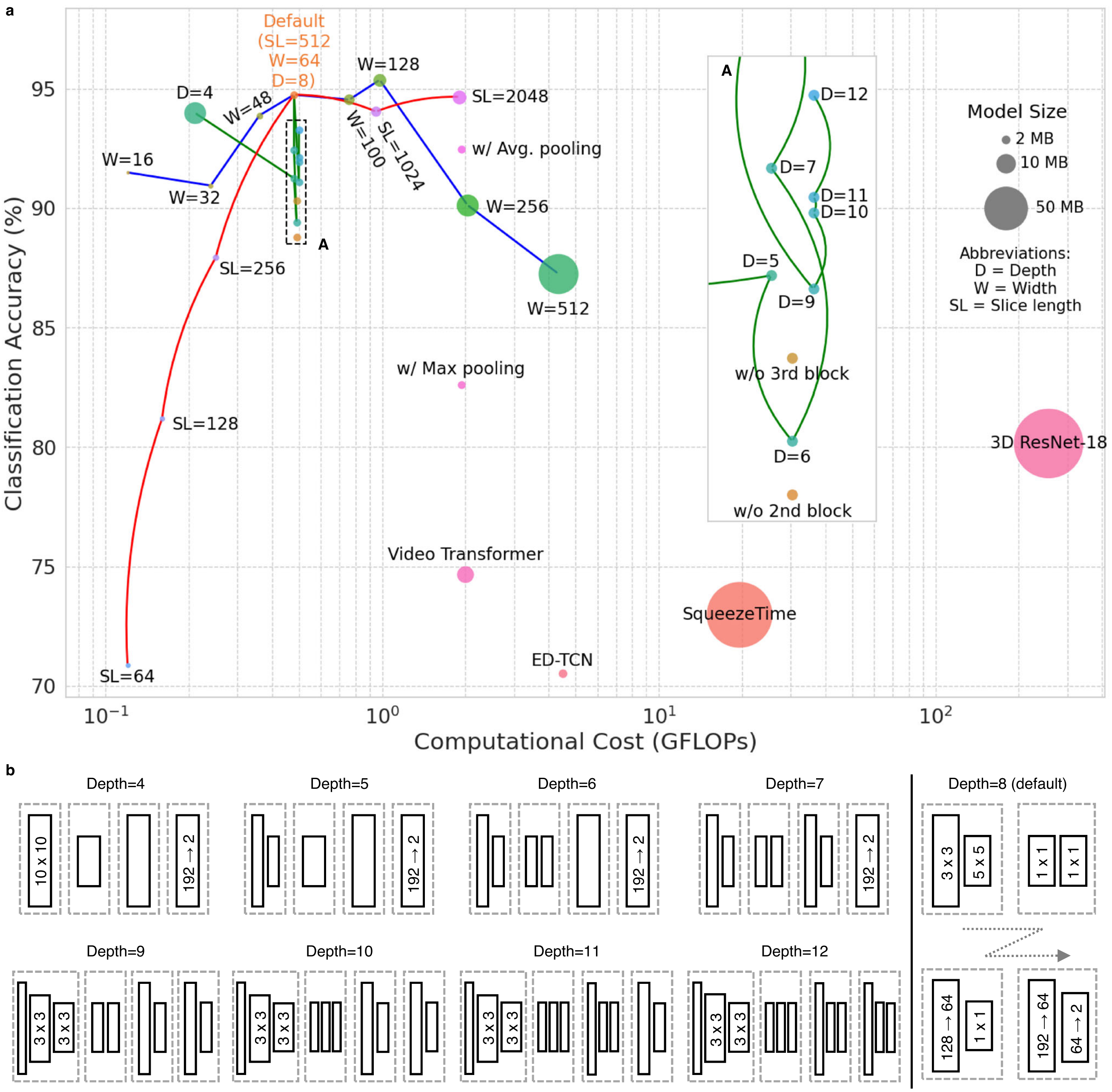


Fig. 8 | Ablation study and hyperparameter analysis of T2C CNN. a Comparison of classification accuracy, computational cost, and model sizes for T2C CNN variants. Baseline methods, including 3D ResNet-18[28], Video Transformer[29], ED-TCN[30], and SqueezeTime[25], are also evaluated for comparison (detailed comparison discussed in the Discussion and Motivation sections). The abbreviation "w/o" represents "without." Giga Floating Point Operations per Second (GFLOPs) quantify the computational power required for processing, indicating hardware demands. Curved arrows visualize the performance trend across variants with the same hyperparameter type. **b** T2C CNN Architecture Variants Investigated in the Ablation Study. Dashed boxes highlight the four blocks of T2C CNN, while solid rectangles represent a convolutional or fully connected layer followed by BN and ReLU layers (except for the final output layer, which omits BN and ReLU). The convolution kernel size is labeled as "N x N," and the rectangle length indicates the kernel size or the number of hidden units (i.e., width). Hidden dimension transformations are denoted as "N → M." Rectangles without labeled kernel sizes or dimension transformations follow the default hidden layer configuration for an 8-layer depth. Source data are provided as a Source Data file.

underscore the necessity of domain-specific architectural design in fluorescence video classification.

HER2-targeted cell experiments

As illustrated in the central panel of Fig. 9a, two groups of HER2-positive *AU565* cell samples (Domain 1 and Domain 2) were immobilized on separate glass slides. Herceptin antibodies conjugated with a 10nt P3' docking strand (complementary to the 10nt P3 strand) were used to specifically bind to the HER2 proteins on the cell membranes. Identical fluorophores (ATTO532)-labeled imagers were introduced to both samples. The same truncated 8nt P3 strand and full-length 10nt P3 strand were used as in the previous experiments. TIRF microscopy was used to image a thin layer at the cell-glass interface, where imagers can access the contact region through gaps between the cell and the slide. A higher number of binding events is expected near the edges of the contact area, which are more exposed to the imaging buffer. Details of the experimental preparation procedures can be found in the "Sample preparation" section of Methods.

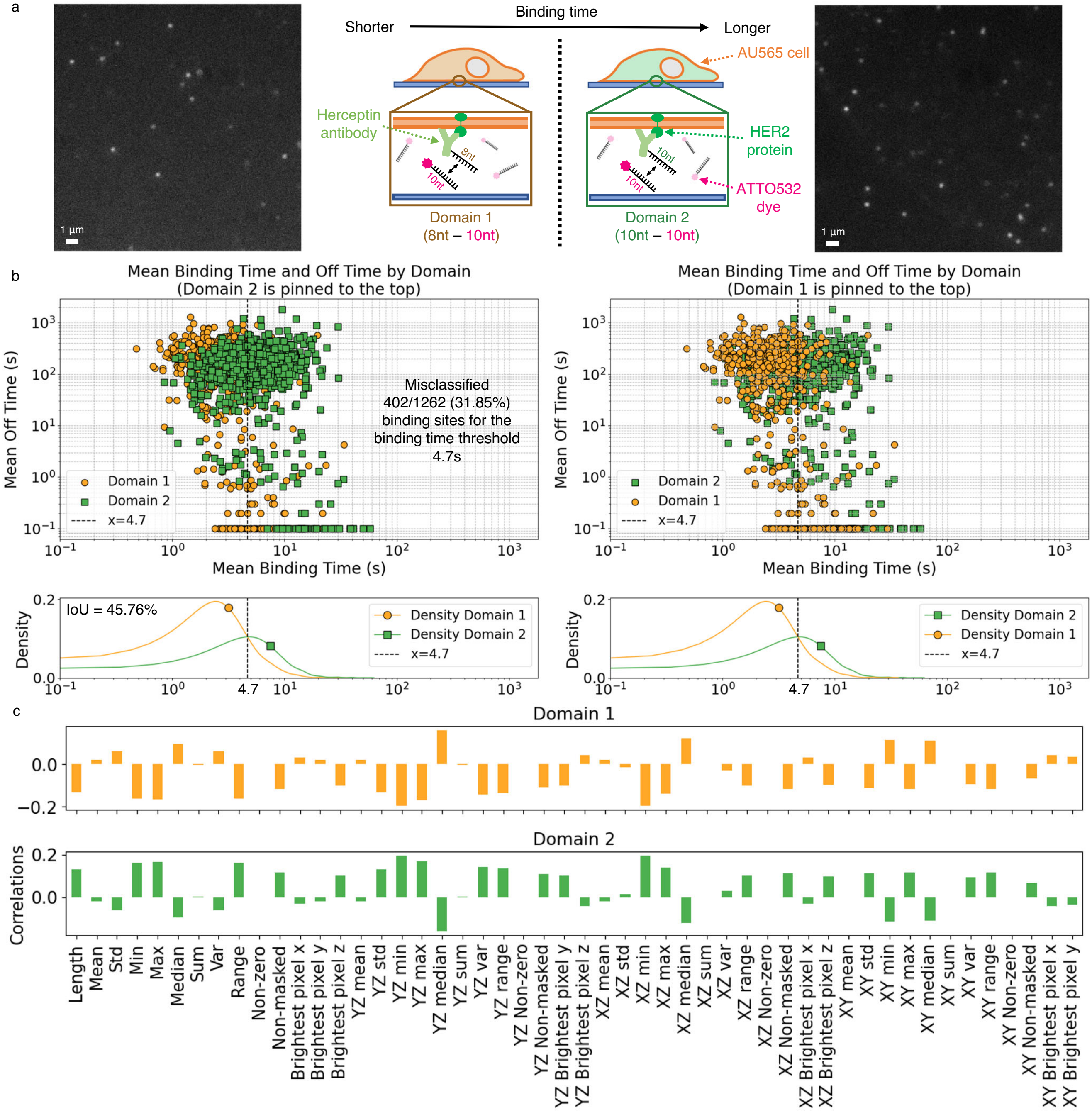


Fig. 9 | Cell experimental settings and data analysis. **a** Settings: HER2-positive *AU565* cells were cultured, fixed, and immunolabeled with custom antibody-oligo conjugates (8nt or 10nt P3' strands), followed by incubation with imager strands (10nt P3 strands) in salt-buffered PBS for fluorescence imaging. Example fluorescence images are shown on the corresponding sides of the two domains. Experiments were repeated twice biologically and five times technically (5-fold validation) with similar results. **b** Binding time analysis: The average binding durations for Domain 1 (8nt–10nt) and Domain 2 (10nt-10nt) overlap substantially, with an intersection-over-union (IoU) of 45.76% between the probability density functions (PDFs) estimated using Gaussian kernels. This implies that nearly half of the binding times cannot be clearly assigned to either Domain 1 or Domain 2. **c** Fluorescence spot video analysis: Correlations were calculated between domain labels and several common statistical features ($n = 19{,}695$ independent binding events). The “YZ,” “XZ,” and “XY” statistics refer to those computed from projections of the video onto the Y--Z, X--Z, and X--Y planes, where X represents the width, Y height, and Z time. Compared to the Origami data in Fig. 2c, the cell data is markedly more difficult to distinguish--both in terms of the overlapping binding times and in the reduced correlation of fluorescence spot video features (dropping from 0.5 in Origami data to 0.2 in cell data), likely due to the increased molecular complexity of the cellular surface. Source data are provided as a Source Data file.

Figure 9a shows raw fluorescence images from two hybridization types: 8nt-10nt and 10nt-10nt, where 10nt imager strands bind to HER2-targeted docking strands on *AU565* cells. By carefully adjusting the concentration ratio between the docking and imager strands, we minimize spot overlap to accurately reveal the locations of individual HER2 proteins. After drift correction over 33 min 20 s (20,000 frames at 10 fps), super-resolution reconstruction (Fig. 10a, middle) reveals HER2 positions (orange/green dots).

Binding events at these sites were tracked to compute average on/off times. As shown in Fig. 9b, Domain 1 (8nt-10nt) shows shorter

https://doi.org/10.1038/s41467-025-64812-1

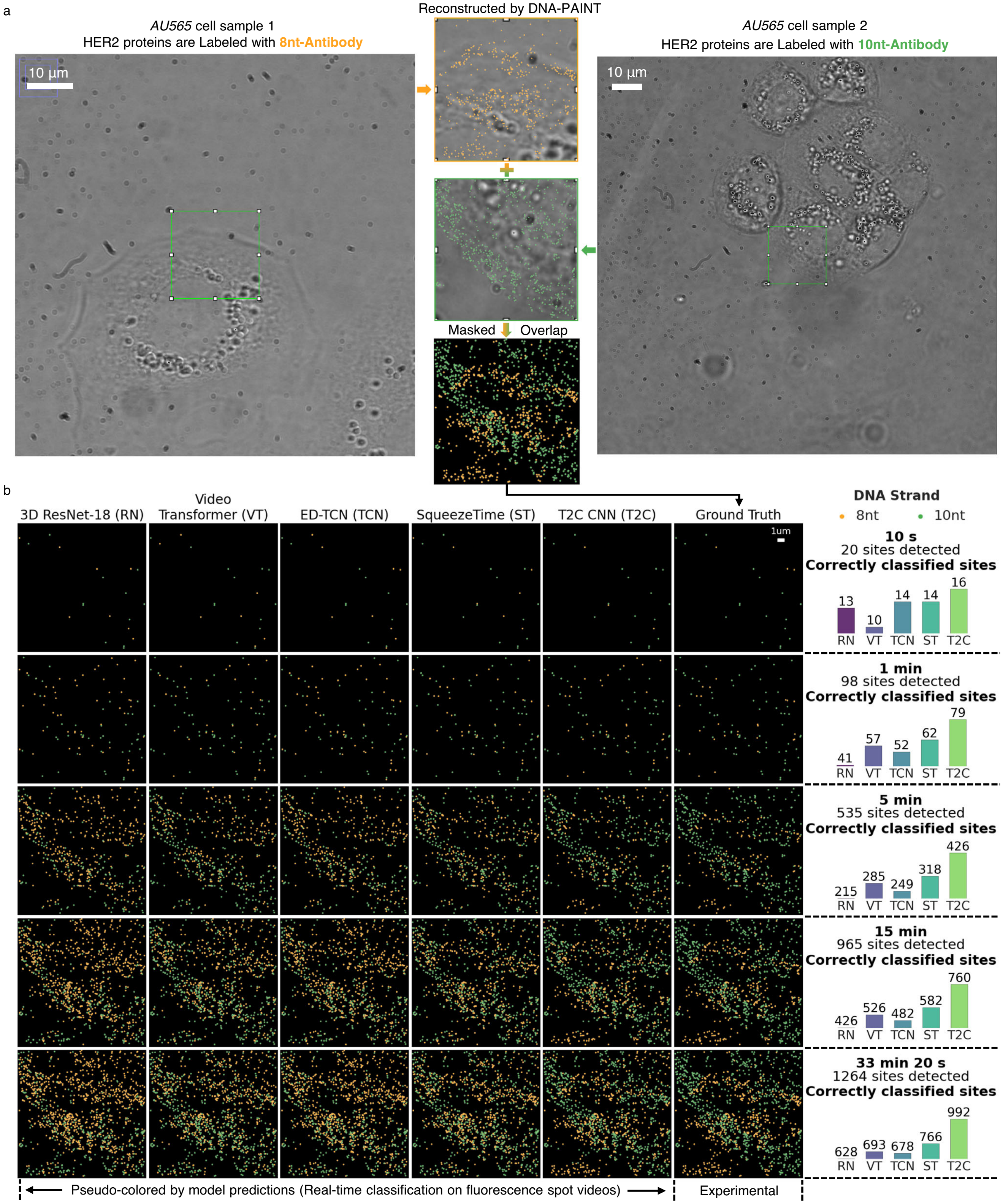


Fig. 10 | Cell experimental results. a Bright-field and super-resolution microscopy images of two test HER2-positive *AU565* cell samples, with HER2 proteins immunolabeled using 8nt and 10nt DNA strands. The DNA-PAINT technique[26] is used to reconstruct HER2 protein locations. Fluorescence videos from the two samples are aligned frame-by-frame to synthesize the multiplexed scenario, serving as ground truth. Experiments were repeated twice biologically and five times technically (5-fold validation) with similar results. **b** Real-time classification of binding events from fluorescence spot videos on the two test samples. The models were trained on another two separate HER2-positive samples prepared under the same experimental conditions. Results show that the proposed T2C CNN outperforms the state-of-the-art video classification models by a significant margin. Source data are provided as a Source Data file.

binding durations than Domain 2 (10nt-10nt), with similar off-times. Compared to the Origami setup (Fig. 2e), domain durations in cells overlap more, likely due to environmental factors such as membrane interactions or local field instabilities. Kernel density estimation (Fig. 2c, bottom) reveals a 45.76% intersection-over-union (IoU) between domains. Using the intersection point of the distributions (4.7s) as a threshold, 31.85% (402/1262) of events are misclassified, highlighting the limits of duration-based classification in cells.

To enhance feature richness, each event is converted into a video by stacking cropped regions across frames. Figure 9c shows feature-label correlations: duration correlates weakly (± 0.13), while other features–e.g., min/max intensity and Y-Z median intensity–show stronger trends (up to ± 0.17), offering improved discriminative power. Still, these cell-derived features are less informative than those from the Origami dataset (correlations ≈ ± 0.5 in Fig. 2c), underscoring the increased complexity of cellular classification.

HER2 site-localized videos (Fig. 10a) were background-corrected and normalized before being input to video classifiers. Labels (Domain 1 vs. Domain 2) guided model training via parameterized operations (convolution, matrix multiplication, etc.). During testing, models classified unseen videos. To evaluate multiplexing, we overlaid super-resolution images from two samples (Fig. 10b), reconstructed at multiple durations (10s to 33min20s). Accuracy trends across time are summarized in the right-hand bar plot.

Despite cellular complexity lowering overall performance, the proposed T2C CNN consistently outperformed other models, reaching 78.48–80.61% accuracy versus 40.19–70.00% for others. The performance gap widened with longer measurements, with T2C CNN surpassing the second-best by 10.00–20.18%. This enables more accurate reconstruction of binding types in real time, advancing high-accuracy classification for single-molecule fluorescence in cells.

Single-frame discrimination of different dye-labeled binding events

As a fluorescence classification model, T2C CNN can be used for multiplexing using the wavelength dependence of the emission PSF. This is an alternative to multi-fluorophore experiments using spectral separation[38,39]. As shown in Supplementary Note 8 and Supplementary Fig. 8, T2C CNN demonstrated superior classification accuracy (92.88%), comparable to larger models like VGG16. At this level of accuracy, T2C CNN requires less than 0.1% of ResNet-18's computational cost, making it well-suited for deployment on lightweight devices. This outcome confirms the ability to differentiate multicolor data at the single fluorophore level by analyzing the PSF patterns of emission wavelengths. This finding lays the groundwork for future multicolor microscopy techniques that do not require wavelength-specific analysis.

Discussion

In this study, we reveal that beyond binding time, diffraction-limited fluorescence spots also contain information related to different types of binding interactions. These interactions may arise from various molecular processes that affect fluorescence. For instance, in DNA binding, the fluorescence intensity of conjugated dyes may either increase or decrease upon hybridization, depending on the sequence and position of the dye[40]. Beyond DNA binding, other molecular interactions and conformational changes have also been shown to influence fluorescence. For example, acrylodan fluorescence emission is sensitive to its local environment; when bound to a protein, it exhibits changes in both intensity and emission wavelength, reflecting the degree of solvent exclusion and the effective dielectric constant of the fluorophore's environment[41]. The fluorescence spectrum of a conjugated fluorophore can be sensitive to microenvironmental changes, such as solvent variations, and may fluctuate over time due to slow, spontaneous conformational changes in the protein molecule[42]. Certain ligands can induce distinct conformational states in the binding protein, altering the environment around the conjugated fluorophore side chain[41]. Additionally, the conformational changes in conjugated polymers can modify the distance between the polymer (acting as an energy donor) and the reporter dye molecule (acting as an energy acceptor). These detection mechanisms typically result in fluorescence turn-on or turn-off, or changes in either the visible color or fluorescence emission color of the conjugated polymer[43]. This work opens the door to exploring the effects of a broader range of molecular binding events on fluorescence, thereby simplifying molecular detection and enhancing specificity.

Some deep learning models that perform well in natural video classification, such as 3D ResNet[28] and Video Transformer[29], do not perform as well on fluorescence videos. A key reason for this discrepancy lies in the way video semantics are represented. Natural videos emphasize the motion patterns of spatial features over time, which involves spatiotemporal correspondence[44]. In contrast, fluorescence videos require attention to the frequency domain of spatial features over a long time span, providing information that is at least as valuable as what can be obtained from the temporal domain[45]. Based on this distinction, we replaced conventional temporal-domain convolutions with transformations spanning a wide temporal range, leading to the development of long-term spatial convolutions in the Temporal-to-Context (T2C) CNN model. While treating the time dimension as the channel dimension, T2C CNN features a design distinct from the classical temporal convolutional network architecture ED-TCN[30] and the recently developed SqueezeTime[25], which focuses on reducing computational cost and memory usage for video processing. Supplementary Table 5 highlights their differences and evaluates their suitability for fluorescence video analysis. Experimental results demonstrated that the T2C CNN, which captures spatial frequency features, significantly outperforms traditional spatiotemporal models in fluorescence videos with low spatial-temporal ratios. This discovery provides valuable insights into the processing and recognition of fluorescence videos.

While this study demonstrates the feasibility of distinguishing binding types from diffraction-limited fluorescence videos, several practical challenges remain, particularly in complex cellular imaging scenarios. These include the presence of more than three distinct binding kinetics, variations in fluorescence intensity across molecular species, and the occurrence of overlapping fluorescence spots from multiple species. Below, we outline these limitations and suggest possible avenues for addressing them.

Although the current experiments involve three binding types, future applications may involve even more complex kinetic behaviors. In such cases, the T2C CNN could be extended by incorporating training data from additional purified target samples. For inference on mixed or unknown samples, one strategy is to flag videos that deviate significantly from known profiles as ’unknown’ or ’nonspecific’, thereby preventing overconfident misclassification.

Differences in fluorescence emission profiles across species can introduce variability in the input signals. To mitigate this, calibration data from purified targets across different species may help the model learn invariant features, potentially improving its robustness and generalization in heterogeneous biological environments.

In multi-species imaging, partial overlap between spots is often unavoidable. For moderate overlap, image deconvolution using Gaussian fitting or related techniques may resolve individual events. However, for more severe overlap, additional modeling strategies may be required. These may include training the model on simulated composite signals, identifying distinct dynamic patterns within overlapping videos, or using outlier detection methods to identify atypical fluorescence dynamics.

Overall, while the approach has shown promising results under controlled conditions, further refinements and methodological extensions will be essential to enable its broader applicability in more complex and variable cellular contexts.

Methods

This study did not involve human participants, animal experiments, or other procedures requiring ethics oversight. Therefore, no ethics approval was required.

Sample preparation

The DNA-origami configuration used in this paper is obtained from the design module of Picasso[26]. We used 4 docking strands per origami at the 4 corners, with the sequence of docking and imager strands adapted from previous research[11]. The origami is prepared by mixing the M13mp18 single-stranded scaffold DNA (10 nM), core staples (100 nM), biotinylated staples (100 nM), and staples with docking strand (1 μM) and annealing in a thermocycler. First, the temperature is raised to 80 °C, followed by a thermal gradient from 60 °C to 4 °C in 3 h. The origami was then purified by centrifugal filtration using Amicon 0.5ml 50k MWCO, at 5000 × *g* for 6 min followed by collection at 5000 × *g* for 5 min.

The substrate is prepared by cleaning a coverslip and sticking it with a CoverWell perfusion chamber. The origami is immobilized on the coverslip using a BSA-biotin-streptavidin linkage, and the chamber is filled with an imager solution (1 nM imager strands in imaging buffer consisting of 1 × PBS, 500 mM NaCl, and saturated with Trolox). The same 1 nM ATO532-labeled imager strand solution was used for both domains to ensure consistent imager availability and minimize potential differences in binding frequency due to imager concentration. To limit the number of fluorescence spots per frame and reduce overlap, we adjusted the surface density of docking strands: 400 pM for the 10nt docking strand (which has a longer binding time) and 1 nM for the 8nt and 6nt strands (which have shorter binding times).

The antibody-oligo conjugate probes were prepared using the method described in[46]. The 5′-amine-modified docking strands were conjugated with Herceptin monoclonal antibody, sourced from Trastuzumab (Roche, Graz Steiermark, Austria), which targets the HER2 protein on the cell membrane, using a disuccinimidyl suberate (DSS) linker. For cell experiments, the antibody-oligo conjugates were prepared at a 1:1 labeling ratio and used at a final concentration of 1 nM. The oligo (200 μM in nuclease-free water) was first mixed with an equal volume of acetonitrile, DSS (25 mM, dissolved in dimethylformamide), and 1:800 (v/v) of triethylamine for 15 min at room temperature, and then purified via ethanol precipitation. Sodium acetate (0.3 M, pH 5.2) and magnesium chloride (10 mM) were added to the conjugation product and mixed with three times the volume of cold absolute ethanol. After 1 h incubation at −20 °C, the slurry was centrifuged at 24,100 g for 15 min at 4 °C. The pellet was washed once with ice-cold 75% ethanol and reconstituted in nuclease-free water. The activated oligo was then incubated with a 3-fold molar excess of Herceptin antibody in 50 mM phosphate buffer (pH 7.2) for 12 h at room temperature.

The antibody-oligo conjugates were purified by ion exchange chromatography (IEX) using an Agilent Bio SAX NP3 (4.6 × 50 mm) column on an Agilent 1260 Infinity HPLC system. Elution began with 100% buffer A (50 mM phosphate buffer, pH 7.2), followed by a step change to 70% buffer A and 30% buffer B (50 mM phosphate buffer, pH 7.2, supplemented with 1.0 M NaCl). A linear gradient was then applied, increasing the proportion of buffer B to 65% (corresponding to 35% buffer A) over 13 min to achieve progressive salt-mediated elution. Each collected fraction, containing antibodies conjugated with discrete numbers of oligo strand(s), was concentrated using Amicon ultrafiltration columns with a 50 kDa cut-off. The final antibody-oligo conjugates were quantified using the Qubit™ ssDNA Assay Kit. The purified conjugates were stored at 4 °C in buffer containing phosphate saline (pH 7.4) and 1 mM EDTA.

HER2 receptor protein-expressing *AU565* cells (CRL-2351™, ATCC; Manassas, Virginia, USA) were cultivated in Roswell Park Memorial Institute 1640 Medium (RPMI/ATCC Modification #A1049101, Gibco, Thermo Fisher Scientific, Massachusetts, USA). The source of AU565 was from ATCC (CRL-2351), isolated from a pleural effusion of a 43-year-old, White, female, patient with breast adenocarcinoma. The medium was supplemented with 1% penicillin-streptomycin (#15070063, Gibco) and 10% fetal bovine serum (FBS; #10270106, Gibco). The cells were seeded at a density of 50,000 cells per dish onto culture dishes (#P35G-1.5-20-C, MatTek, Massachusetts, USA). The cultures were maintained in a heated CO_2 incubator at 37 °C and 5% (v/v) CO_2 concentration (Forma Steri-Cycle CO_2 incubator, Thermo Fisher Scientific) for 36 h.

Cells were then fixed with 4% paraformaldehyde (PFA, EMS) in PBS for 15 min, followed by quenching with 1 mg/mL sodium borohydride for 7 min. Cells were washed three times (5 min per wash) with PBS, and incubated with 1% BSA and 0.05% Tween-20 in PBS for 2 h to block non-specific binding. The blocked cells were washed four times with 1% BSA and incubated with conjugated antibodies (1 nM in PBS containing 500 mM NaCl) for 1 h at room temperature. Cells were then washed four times with PBS supplemented with 500 mM NaCl. Imager solution (1–5 nM in PBS with 500 mM NaCl and 2 mM Trolox) was added immediately before imaging.

All experiments were conducted on separate samples.

Fluorescence microscopy

The intermittent binding of an imager and corresponding immobilization of the fluorophore to the docking strand on the origami are imaged in total internal reflection fluorescence (TIRM) mode. We used an inverted microscope (IX-71, Olympus, Tokyo, Japan) with an oil immersion objective (PlanApo, 100 × , NA 1.5, Olympus). A 532 nm laser (Samba, Cobolt, AB, Sweden) is coupled with TIRF illuminator model IX2-RFAEVA-2 (Olympus) through an optical fiber. The laser is focused on the back-focal plane of the objective, away from the center to achieve TIRF. The laser power density is measured after the objective and determined to be 125 W/cm2. The emission is collected through the same objective, passed through a dichroic beamsplitter (Semrock, Di03-R405/488/532/635-t1-25 × 36) and emission filter (Semrock, NF03-405/488/532/635E-25), imaged with a scientific complementary metal-oxide semiconductor (sCMOS) camera (ORCA-Flash 4.0, Hamamatsu, Shizuoka, Japan). The image sequences are collected at a frame rate of 10 Hz, and analyzed without further processing.

Data preprocessing

This section describes the data preprocessing steps from raw microscope images to the model input. Initially, due to the continuous use of laser excitation for fluorescence, the number of photons received by the microscope gradually decreases and fluctuates over time. This causes the overall intensity of the raw microscope images to vary over time, and there are noticeable intensity differences between images from different experiments, depending on the laser and dye conditions during the experiments. To estimate the background intensity of each frame of the raw microscope images, we used Picasso v0.7.5[26] to localize diffraction-limited fluorescent spots on the raw microscope images and masked out all detected fluorescent spots to calculate the average intensity of the remaining pixels. As shown in Supplementary Fig. 9a, there are different frame-level variations and significant gaps between experiments in the background intensity across four experiments in the two domains. To reduce the interference of background intensity variations on fluorescence intensity, we subtracted the background intensity from the detected fluorescent spots and assigned a small value (such as 100, to distinguish it from the zeros used in padding) to pixels below the background intensity. An example of background intensity correction is shown in Supplementary Fig. 9b. This example demonstrates a significant change in background intensity within just three frames (0.3 s), making the fluorescent spots appear much brighter. After correction, we found that the fluorescence intensity remained at the same level after 0.3 s. Before inputting the background-corrected fluorescent spots into the model, it is recommended to divide all intensity values by a scaling constant (e.g., 500) or do the event-level normalization to reduce the numerical scale to a suitable range (e.g., not greater than 10).

Fluorescence blinking and detection algorithm errors can cause the loss of a few fluorescent spots in the same binding event, including some at the edges and those with high overlaps. As shown in Supplementary Fig. 9c, we consider two binding-on signals detected very close in space and time to be from the same binding event and pad zeros between them. By reading the drift of the slides using Picasso Render[26] (as shown in Supplementary Fig. 9d), we obtained a series of binding events grouped by binding site location after drift correction. Subsequently, in Supplementary Fig. 9e, we analyzed the number of binding events per location (denoted as “#Binding events/group”) and the duration of each binding event over 33 min and 20 s across four experiments in the two domains. The number of binding events in both domains is concentrated between 2 and 4, and the binding durations are concentrated between 0 and 10 s. Compared to the partially matched 8nt-10nt binding events in domain 1, the fully matched 10nt-10nt binding events in domain 2 have a smoother distribution and longer binding times. In the cross-experiment binding type classification task, we used data from the first two experiments to train models and data from the last two experiments to test models.

Motivation

Unlike conventional video classification tasks[29,47–49], diffraction-limited fluorescence spot videos used for DNA binding type classification are characterized by low spatial resolution and long temporal sequences. We refer to these as low spatial-temporal ratio objects. Traditional video models typically rely on narrow 3D convolutional kernels[47,49] or combine small-scale convolutions with sequential models[29,48]. In general, 3D convolutions use strides smaller than the kernel size[50–52] to capture fine-grained features. However, such settings can lead to excessive convolutions along the temporal dimension, resulting in a proliferation of transitional features[53]. For sparse, pattern-like fluorescence spots, these transitional features can introduce redundancy and increase the risk of overfitting. On the other hand, small-scale convolutions with sequential methods[29,48] emphasize token dependencies but often fail to capture the global temporal patterns essential for identifying distinct binding events.

To address these challenges, we propose a domain-specific architecture optimized for grayscale fluorescence videos with low spatial-temporal ratios: the temporal-to-context convolutional neural network (T2C CNN). Unlike prior work such as[25], which also treats the temporal dimension as channels, the T2C CNN incorporates long temporal strides that match the convolutional kernel’s temporal extent, facilitating long-term spatial convolutions. This design effectively abstracts global temporal features while reducing unnecessary transitional representations. As illustrated in Supplementary Fig. 10a, treating the time dimension as the channel dimension allows T2C CNN’s 2D convolution to achieve a receptive field equivalent to that of a 3D convolution, while enabling more efficient model size and flexible input handling.

Moreover, the T2C CNN includes additional domain-specific innovations: long-term spatial convolutions to capture frequency-like patterns in the spatial domain, shallow cross-connected blocks to retain multi-scale spatiotemporal features, and a pooling-free strategy to preserve spatial detail and temporal continuity. These components collectively enhance the model’s ability to discriminate between binding types (Supplementary Table 3). For long binding events, the T2C CNN performs slice-wise inference and aggregates predictions to yield a robust final output. This architecture achieves superior accuracy and stability (Figs. 4, 5, 6, 7, and 10) compared to general-purpose video models. In the following section, we detail the design and implementation of the T2C CNN.

Notations

Let the function $f(x, y, z, c)$ represent the input video, where $f(x = X, y = Y, z = Z, c = C)$ denotes the pixel value at the three-dimensional coordinates (X, Y, Z) and the C-th channel in the input video. Assume the input video $f(x, y, z, c)$ has a frame width of $\mathrm{W_{in}}$, a frame height of $\mathrm{H_{in}}$, a number of frames $\mathrm{D_{in}}$, and a number of channels $\mathrm{C_{in}}$. Therefore, $x, y, z, c \in \mathbb{N}_0$, $x < \mathrm{W_{in}}$, $y < \mathrm{H_{in}}$, $z < \mathrm{D_{in}}$, and $c < \mathrm{C_{in}}$. For values of x, y, z, c outside these ranges, $f(x, y, z, c)$ is considered to be 0 to facilitate the compatibility of padding in expressions.

Let the convolution kernel ω have a width of $\mathrm{W_{kernel}}$, a height of $\mathrm{H_{kernel}}$, a depth of $\mathrm{D_{kernel}}$, and a channel size of $\mathrm{C_{in}}$, with its value at the three-dimensional coordinates (u, v, w) and the channel index c given by $\omega(u, v, w, c)$. Assume the convolution strides in width, height, and depth are s_w, s_h, and s_d, respectively, and the padding sizes on both sides for width, height, and depth are p_w, p_h, p_d, respectively. Let $\mathrm{W_{out}}$, $\mathrm{H_{out}}$, and $\mathrm{D_{out}}$ represent the output dimensions in width, height, depth, and channel, respectively. Assume the number of output channels is $\mathrm{C_{out}}$, each corresponds to a unique kernel $\omega_d(u, v, w, c)$ where d ranges from 1 to $\mathrm{C_{out}}$. Then, the output of a typical 3D CNN convolution[54] is represented as:

$$\mathrm{Conv}^{(3D)}(f, \omega) = g^{(3D)}, \tag{1}$$

where

$$g^{(3D)}(i, j, k, d) = \sum_{u=0}^{\mathrm{W_{kernel}}-1} \sum_{v=0}^{\mathrm{H_{kernel}}-1} \sum_{w=0}^{\mathrm{D_{kernel}}-1} \sum_{c=0}^{\mathrm{C_{in}}-1} \omega_d(u, v, w, c) \cdot f(s_w \cdot i + u - p_w, s_h \cdot j + v - p_h, s_d \cdot k + w - p_d, c). \tag{2}$$

Here, $g^{(3D)}(i, j, k, d)$ represents the output of the 3D convolution operation at position (i, j, k, d) in the output volume. The indices i, j, k iterate over the output dimensions $\mathrm{W_{out}}$, $\mathrm{H_{out}}$, $\mathrm{D_{out}}$, and $\mathrm{C_{out}}$, respectively. The sums iterate over the kernel dimensions $\mathrm{W_{kernel}}$, $\mathrm{H_{kernel}}$, $\mathrm{D_{kernel}}$, $\mathrm{C_{in}}$, and u, v, w, c index into the kernel. The function $f(x, y, z, c)$ represents the input video, where values outside the video dimensions are assumed to be 0. The convolution kernel is denoted by $\omega_d(u, v, w, c)$.

T2C layer

Different fluorescences are typically split into separate channels in fluorescence microscopy[55,56]. Further distinguishing different binding types within a single channel can significantly increase the labeling capacity of the fluorescence and help eliminate potential non-specific binding events. In this study, the proposed T2C layer reinterprets the extensive temporal dimension inherent in fluorescence microscopy data by treating it as the channel dimension of a grayscale image. This transformation allows the application of spatial convolutions across the broad temporal domain, effectively capturing frequency variations. This approach enhances the efficiency of spatiotemporal information fusion, thereby improving the CNN’s accuracy in predicting binding types. The output of a T2C layer, denoted by $\mathrm{Conv}^{(T2C)}$, for each fluorescence channel can be represented as:

$$\mathrm{Conv}^{(T2C)}(f, \omega) = \{g_k^{(T2C)}\}_{k=0}^{\mathrm{N_{slice}}}, \tag{3}$$

where

$$g_k^{(T2C)}(i, j, d) = \sum_{u=0}^{\mathrm{W_{kernel}}-1} \sum_{v=0}^{\mathrm{H_{kernel}}-1} \sum_{w=0}^{\mathrm{T_{slice}}-1} \omega_d(u, v, w) \cdot f(s_w \cdot i + u - p_w, s_h \cdot j + v - p_h, \mathrm{T_{slice}} \cdot k + w), \tag{4}$$

Here, $g_k^{(T2C)}(i,j,d)$ represents the output of the temporal-to-channel convolution operation at position (i, j, d) in the output volume for the k-th temporal slice with length T_{slice}. The number of slices is denoted by N_{slice}, which is defined as $\lceil \frac{D_{in}}{T_{slice}} \rceil$. The indices i, j iterate over the output dimensions W_{out} and H_{out}, respectively. The sums iterate over the kernel dimensions W_{kernel}, H_{kernel}, T_{slice}, and u, v, w index into the kernel. The function $f(x, y, z)$ represents the input single-channel video, where values outside the video dimensions are assumed to be 0. The convolution kernel is denoted by $\omega_d(u, v, w)$.

At the beginning of the T2C CNN, a batch of images is input to the T2C convolution to produce the output $g_k^{(T2C)}$, which then undergoes batch normalization (BN) and rectified linear unit (ReLU) activation. BN is used to accelerate T2C CNN training by reducing internal covariate shift[57]. After BN, ReLU activation is applied to introduce non-linearity[58] and accelerate learning convergence[59]. Following its transformation through the T2C layer, the original 3D input video segment is converted into a 2D feature map with fixed channels.

T2C CNN architecture

After the T2C layer, we apply 2D convolutions and fully connected (FC) layers[60] to further extract features and make the final prediction. Supplementary Fig. 10b shows the dimensions of each convolutional and FC layer in the T2C CNN used in this study, which can be functionally divided into four blocks: Temporal2Channel block, Hidden transformation block, Multi-scale feature fusion block, and Feature classification block. The definition of the four blocks is detailed in Supplementary Note 9.

The architecture of the T2C CNN is specifically tailored for fluorescence video analysis. In the first block, long-term spatial convolutions are used to capture extended spatial frequency patterns. In the intermediate blocks (i.e., neither input nor output), skip concatenations[61] are adopted to preserve multi-scale spatiotemporal features. Across all blocks of the T2C CNN, a pooling-free strategy[62,63] is employed to promote efficient spatiotemporal information fusion. While each of these components has been explored individually in prior work, their integration within the T2C CNN constitutes an architecture specifically optimized for fluorescence video analysis. The individual and combined contributions of these three components are systematically analyzed in the ablation study.

Training on T2C CNN

After the construction of a unique CNN architecture consisting of four blocks, the T2C CNN can be trained by a classic cross-entropy loss[64] to achieve an accurate classification of binding types. We use stochastic gradient descent[65] on a single Tesla V100 32G GPU to update the parameters of the T2C CNN. The training set is randomly divided into five equal parts. For each experiment, one part is used for validation to assess the model's classification performance, while the remaining four parts are used for training. Training stops when the validation performance no longer improves for 100 epochs, and the model with the highest validation performance is selected for testing. In the training environment (CPU: Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz, GPU: Tesla V100 32G), preparing the data for 3000 binding events takes approximately 5 min, and training each epoch takes about 3 s. The optimal model is typically determined after around 200 epochs, so the entire training process for 3000 binding events, including data preparation, takes about 15 min.

Test on T2C CNN

The aggregated prediction $\hat{y}_{ij}$ is derived from the average softmax output (as shown in the last block in Supplementary Note 9), it can generally be interpreted as the confidence score for the i-th binding event (f_i) belonging to the j-th binding type (y_j)[66,67], i.e., $\text{Confidence}(f_i \in y_j | f_i) : = \hat{y}_{ij}$. During inference, we classify each binding event by selecting the binding type with the highest confidence score. That is, after model training is completed, T2C CNN predicts the binding type $\hat{c}_i$ for each single binding event (where i denotes its index) based on the sequence of its diffraction-limited fluorescent spots: $\hat{c}_i = \arg\max_j \text{Confidence}(f_i \in y_j | f_i)$. This classification approach ensures that the final prediction is based on the most probable binding type, reflecting the model's learned feature representations.

In the testing environment (CPU: Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz, GPU: Tesla V100 32G), extracting data for all 31,479 binding events from a 33-min and 20-s observation video takes approximately 25 min, while predicting their binding types takes around 18 s.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw DNA-PAINT data generated in this study have been deposited in the Figshare database under accession code 28606874 at https://doi.org/10.6084/m9.figshare.28606874.v1[68]. Details on the data structure and processing are available in the GitHub repository: https://github.com/Yueming-Yin/T2C-CNN[69]. Source data are provided with this paper. No clinical datasets or third-party data were used in this study. Correspondence and requests for materials should be addressed to Lipo Wang or Thorsten Wohland. Source data are provided with this paper.

Code availability

The code for data preparation, model inference, and result visualization in this study is publicly available at https://github.com/Yueming-Yin/T2C-CNN[69], along with instructions. The best model parameters are provided. The Picasso software[26] for generating super-resolution images is publicly available at https://github.com/jungmannlab/picasso. Version 0.7.5 was employed in this study. The ImagingFCS v1.613 plugin in ImageJ 1.54p was used to control the camera. ImageJ 1.54p can be downloaded from https://imagej.net/ij/download.html.

References

1. Lelek, M. et al. Single-molecule localization microscopy. *Nat. Rev. Methods Prim.* **1**, 39 (2021).
2. Betzig, E. et al. Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–1645 (2006).
3. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). *Nat. Methods* **3**, 793–796 (2006).
4. Finan, K., Flottmann, B. & Heilemann, M. Photoswitchable fluorophores for single-molecule localization microscopy. *Nanoimag. Methods Protoc.* **950**, 131–151 (2013).
5. Schirripa Spagnolo, C. & Luin, S. Choosing the probe for single-molecule fluorescence microscopy. *Int. J. Mol. Sci.* **23**, 14949 (2022).
6. Yankelevich, D.R. et al. Design and evaluation of a device for fast multispectral time-resolved fluorescence spectroscopy and imaging. *Rev. Sci. Instrum.* **85**, 034303 (2014).
7. Li, H. & Vaughan, J. C. Switchable fluorophores for single-molecule localization microscopy. *Chem. Rev.* **118**, 9412–9454 (2018).
8. Coelho, S. et al. Ultraprecise single-molecule localization microscopy enables in situ distance measurements in intact cells. *Sci. Adv.* **6**, 8271 (2020).
9. Jungmann, R. et al. Single-molecule kinetics and super-resolution microscopy by fluorescence imaging of transient binding on DNA origami. *Nano Lett.* **10**, 4756–4761 (2010).
10. Johnson-Buck, A. et al. Kinetic fingerprinting to identify and count single nucleic acids. *Nat. Biotechnol.* **33**, 730–732 (2015).
11. Wade, O. K. et al. 124-color super-resolution imaging by engineering DNA-paint blinking kinetics. *Nano Lett.* **19**, 2641–2646 (2019).

12. Morozumi, A. et al. Spontaneously blinking fluorophores based on nucleophilic addition/dissociation of intracellular glutathione for live-cell super-resolution imaging. *J. Am. Chem. Soc.* **142**, 9625–9633 (2020).
13. Chatterjee, T. et al. Direct kinetic fingerprinting and digital counting of single protein molecules. *Proc. Natl Acad. Sci. USA* **117**, 22815–22822 (2020).
14. Liu, T. et al. Gentle rhodamines for live-cell fluorescence microscopy. *ACS Cent. Sci.* **10**, 1933–1944 (2024).
15. Arppe, R., Carro-Temboury, M. R., Hempel, C., Vosch, T. & Just Sørensen, T. Investigating dye performance and crosstalk in fluorescence enabled bioimaging using a model system. *PloS One* **12**, 0188359 (2017).
16. Yang, L. et al. High-throughput methods in the discovery and study of biomaterials and materiobiology. *Chem. Rev.* **121**, 4561–4677 (2021).
17. Cuny, A. P., Schlottmann, F. P., Ewald, J. C., Pelet, S. & Schmoller, K. M. Live cell microscopy: from image to insight. *Biophys. Rev.* **3**, 021302 (2022).
18. Lauwerends, L. J. et al. Real-time fluorescence imaging in intraoperative decision making for cancer surgery. *Lancet Oncol.* **22**, 186–195 (2021).
19. Wang, Q., Li, Z., Zhang, S., Chi, N. & Dai, Q. A versatile wavelet-enhanced CNN-transformer for improved fluorescence microscopy image restoration. *Neural Netw.* **170**, 227–241 (2024).
20. Zhong, S., Hu, J., Yu, X. & Zhang, H. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. *Chem. Eng. J.* **408**, 127998 (2021).
21. Liu, X., Jiang, Y., Cui, Y., Yuan, J. & Fang, X. Deep learning in single-molecule imaging and analysis: recent advances and prospects. *Chem. Sci.* **13**, 11964–11980 (2022).
22. Nehme, E. et al. Deepstorm3d: dense 3d localization microscopy and PSF design by deep learning. *Nat. methods* **17**, 734–740 (2020).
23. Ouyang, W., Aristov, A., Lelek, M., Hao, X. & Zimmer, C. Deep learning massively accelerates super-resolution localization microscopy. *Nat. Biotechnol.* **36**, 460–468 (2018).
24. Narayanasamy, K. K., Rahm, J. V., Tourani, S. & Heilemann, M. Fast DNA-paint imaging using a deep neural network. *Nat. Commun.* **13**, 5047 (2022).
25. Zhai, Y., Li, W., Tang, Y., Chen, X., Wang, Y. No time to waste: squeeze time into channel for mobile video understanding. arXiv preprint arXiv:2405.08344 (2024).
26. Schnitzbauer, J., Strauss, M. T., Schlichthaerle, T., Schueder, F. & Jungmann, R. Super-resolution microscopy with dna-paint. *Nat. Protoc.* **12**, 1198–1228 (2017).
27. Nieves, D. J., Gaus, K. & Baker, M. A. Dna-based super-resolution microscopy: Dna-paint. *Genes* **9**, 621 (2018).
28. Tran, D. et al. A closer look at spatiotemporal convolutions for action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459 (2018).
29. Selva, J. et al. Video transformers: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 12922–12943 (2023).
30. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D. Temporal convolutional networks for action segmentation and detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 156–165 (2017).
31. Clowsley, A. H. et al. Repeat DNA-paint suppresses background and non-specific signals in optical nanoscopy. *Nat. Commun.* **12**, 501 (2021).
32. Ma, G., Wan, Z., Yang, Y., Jing, W. & Wang, S. Three-dimensional tracking of tethered particles for probing nanometer-scale single-molecule dynamics using a plasmonic microscope. *ACS Sens.* **6**, 4234–4243 (2021).
33. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **33**, 1065–1076 (1962).
34. Riedmiller, M. & Lernen, A. Multi layer perceptron. *Machine Learning Lab Special Lecture* Vol. **24** (University of Freiburg, 2014).
35. Lew, M. D., Backlund, M. P. & Moerner, W. Rotational mobility of single molecules affects localization accuracy in super-resolution fluorescence microscopy. *Nano Lett.* **13**, 3967–3972 (2013).
36. Hulleman, C. N. et al. Simultaneous orientation and 3d localization microscopy with a vortex point spread function. *Nat. Commun.* **12**, 5934 (2021).
37. Hou, Y. et al. Multi-resolution analysis enables fidelity-ensured deconvolution for fluorescence microscopy. *eLight* **4**, 14 (2024).
38. Bates, M., Huang, B., Dempsey, G. T. & Zhuang, X. Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science* **317**, 1749–1753 (2007).
39. Gómez-García, P. A., Garbacik, E. T., Otterstrom, J. J., Garcia-Parajo, M. F. & Lakadamyali, M. Excitation-multiplexed multicolor super-resolution imaging with fm-storm and fm-dna-paint. *Proc. Natl. Acad. Sci. USA* **115**, 12991–12996 (2018).
40. Nazarenko, I., Pires, R., Lowe, B., Obaidy, M. & Rashtchian, A. Effect of primary and secondary structure of oligodeoxyribonucleotides on the fluorescent properties of conjugated dyes. *Nucleic Acids Res.* **30**, 2089–2195 (2002).
41. Hibbs, R. E., Talley, T. T. & Taylor, P. Acrylodan-conjugated cysteine side chains reveal conformational state and ligand site locations of the acetylcholine-binding protein. *J. Biol. Chem.* **279**, 28483–28491 (2004).
42. Wazawa, T., Ishii, Y., Funatsu, T. & Yanagida, T. Spectral fluctuation of a single fluorophore conjugated to a protein molecule. *Biophys. J.* **78**, 1561–1569 (2000).
43. Lee, K., Povlich, L. K. & Kim, J. Recent advances in fluorescent and colorimetric conjugated polymer-based biosensors. *Analyst* **135**, 2179–2189 (2010).
44. Han, Z. et al. Active trace: a sparse spatiotemporal representation for videos. *IEEE Access* **5**, 22433–22442 (2017).
45. Lakowicz, J. R., Gratton, E., Cherek, H., Maliwal, B. & Laczko, G. Determination of time-resolved fluorescence emission spectra and anisotropies of a fluorophore-protein complex using frequency-domain phase-modulation fluorometry. *J. Biol. Chem.* **259**, 10967–10972 (1984).
46. Ang, Y. S. & Yung, L.-Y. L. Protein–dna conjugates with a discrete number of oligonucleotide strands for highly reproducible protein quantification by the dna proximity assay. *Anal. Chem.* **95**, 12071–12079 (2023).
47. Karpathy, A. et al. Large-scale video classification with convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732 (2014).
48. Rayan, A. et al. Utilizing CNN-LSTM techniques for the enhancement of medical systems. *Alex. Eng. J.* **72**, 323–338 (2023).
49. Ur Rehman, A., Belhaouari, S. B., Kabir, M. A. & Khan, A. On the use of deep learning for video classification. *Appl. Sci.* **13**, 2007 (2023).
50. Zaniolo, L. & Marques, O. On the use of variable stride in convolutional neural networks. *Multimed. Tools Appl.* **79**, 13581–13598 (2020).
51. Lin, B., Zhang, S. & Bao, F. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3054–3062 (2020).
52. Wang, C. A review on 3d convolutional neural network. In: *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)*, pp. 1204–1208 (IEEE, 2023).
53. Wahab, A., Ali, S. D., Tayara, H. & Chong, K. T. iim-cnn: intelligent identifier of 6ma sites on different species by using convolution neural network. *IEEE Access* **7**, 178577–178583 (2019).

54. Singh, R. D., Mittal, A. & Bhatia, R. K. 3d convolutional neural network for object recognition: a review. *Multimed. Tools Appl.* **78**, 15951–15995 (2019).
55. Combs, C. A. & Shroff, H. Fluorescence microscopy: a concise guide to current imaging methods. *Curr. Protoc. Neurosci.* **79**, 2–1 (2017).
56. Rodríguez-Sevilla, P., Thompson, S. A. & Jaque, D. Multichannel fluorescence microscopy: advantages of going beyond a single emission. *Adv. NanoBiomed Res.* **2**, 2100084 (2022).
57. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning*, pp. 448–456 (PMLR, 2015).
58. Karnewar, A., Ritschel, T., Wang, O. & Mitra, N. Relu fields: The little non-linearity that could. In: *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–9 (2022).
59. Mukherjee, A., Basu, A., Arora, R. & Mianjy, P. Understanding deep neural networks with rectified linear units. In: *International Conference on Learning Representations* (2018).
60. Basha, S. S., Dubey, S. R., Pulabaigari, V. & Mukherjee, S. Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing* **378**, 112–119 (2020).
61. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708 (2017).
62. Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015).
63. Yu, F., Koltun, V. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015).
64. Mao, A., Mohri, M. & Zhong, Y. Cross-entropy loss functions: theoretical analysis and applications. In: *International Conference on Machine Learning*, pp. 23803–23828 (PMLR, 2023).
65. Amari, S.-i Backpropagation and stochastic gradient descent method. *Neurocomputing* **5**, 185–196 (1993).
66. Kízíltepe, R. S. & Gan, J. Q. Simple effective methods for decision-level fusion in two-stream convolutional neural networks for video classification. In: *Intelligent Data Engineering and Automated Learning–IDEAL 2020: 21st International Conference, Guimaraes, Portugal, November 4–6, 2020, Proceedings, Part I 21*, pp. 77–87 (Springer, 2020).
67. Pearce, T., Brintrup, A. & Zhu, J. Understanding softmax confidence and uncertainty. arXiv preprint arXiv:2106.04972 (2021).
68. Yin, Y. et al. On_time_varied_data.tar.xz https://doi.org/10.6084/m9.figshare.28606874.v1 (2025).
69. Yin, Y. et al. Yueming-yin/t2c-cnn: Release for the paper “discriminating single-molecule binding event from diffraction-limited fluorescence” https://doi.org/10.5281/zenodo.17070444 (2025).

Acknowledgements

We would like to acknowledge the support received from the Ministry of Education, Singapore, under its Research Center of Excellence award to the Institute for Digital Molecular Analytics (IDMxS, grant: EDUN C-33-18-279-V12) (Y.Y., L.W., T.W., and S.S.). This work was also supported by the Ministry of Education, Singapore, under Grant No. MOE T2EP50120-0037 (N.P., I.L., Y.A., L.Y., and T.W.), and by the National University of Singapore (NUS) Graduate Student Research Scholarship (S.S.).

Author contributions

L.W. and T.W. conceived the project and supervised the method design. I.L. and L.Y. provided the original DNA-PAINT idea and prepared all the materials. S.Z. and Y.A. prepared the biological samples. N.P., K.S., and S.Z. designed and performed the DNA-PAINT experiments. Y.Y. developed the data processing and machine-learning methods. Y.Y. carried out the model training and analyzed the results. S.S. advised on method design and experimental setup. Y.Y. wrote the initial manuscript, N.P. and T.W. edited the multi-dye and DNA-PAINT microscopy sections. S.Z. edited the cell experiment preparation section. All authors discussed the results and contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-64812-1.

Correspondence and requests for materials should be addressed to Thorsten Wohland or Lipo Wang.

Peer review information *Nature Communications* thanks Xinghao Chen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.