

Discriminating single-molecule binding events from diffraction-limited fluorescence

Corresponding Author: Professor Lipo Wang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Summary

This paper proposes a novel approach to discriminate single-molecule binding events in DNA-PAINT experiments using a temporal-to-channel convolutional neural network (T2C CNN). The method leverages the temporal information encoded in diffraction-limited fluorescent spots, transforming it into channel dimensions for improved classification accuracy. By doing so, the T2C CNN can distinguish between different types of DNA binding domains with high precision, achieving a classification accuracy of approximately 95%, a significant improvement over traditional methods. The technique also dramatically reduces the measurement time required for accurate classification, from 10 minutes to just 5 seconds, thereby enhancing the throughput and practicality of single-molecule localization microscopy (SMLM) for studying biological systems at the nanoscale.

Pros:

- The proposed method demonstrates a substantial increase in classification accuracy, from 75% to around 95%, which is a significant leap forward in the field of SMLM.
- The technique's ability to classify binding events in a fraction of the time previously required represents a major advancement in efficiency, making high-throughput studies more feasible.
- The successful application of deep learning to SMLM data analysis showcases the power of machine learning in advancing microscopy techniques and could inspire further innovations in this intersection.
- The authors have made the data and code used in this study publicly available, which promotes transparency, reproducibility, and further research in the field.

Cons:

- One of the most concerns is the novelty of the proposed T2C CNN, as treating the time dimension as the channel dimension is not a quite new idea in deep learning [Ref1][Ref2][Ref3].
[Ref1] Temporal convolutional networks: A unified approach to action segmentation, ECCVW 2016
[Ref2] Temporal Convolutional Networks for Action Segmentation and Detection, CVPR 2017
[Ref3] No Time to Waste: Squeeze Time into Channel for Mobile Video Understanding, arXiv 2024
- Section 4.3 include too many details about the common components of neural network, e.g., BN, ReLU, cross entropy loss, which make it harder to understand what is newly proposed in this manuscript.
- As shown in Fig. 2, 3D ResNet-18 performs much poorer than the proposed method, with accuracy of around 80%. In Fig. 4, ResNet-18 obtains quite similar performance with the proposed T2C CNN with accuracy of 92%+. It is not quite clear why 3D convolutional networks obtain much worse performance than the 2D one. More discussions and insights are encouraged to be included.
- Moreover, the ResNet-18 seems to have thousands of times FLOPs than the proposed method. It is not quite reasonable since both the proposed method and ResNet-18 are built with 2D convolution.
- No ablation studies are included for the design choices of the proposed method, e.g, the depth/width of the proposed architecture, the effect of Multi-scale feature fusion block.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

In this manuscript, Yin et al. describe their development of a deep neural network for an analysis of single molecule binding events to distinguish molecular species with different binding kinetics. Instead of analyzing binding and dissociation kinetics directly from the time dependent changes of the fluorescence signal (i.e., residence times in the on and off states), which requires trajectories longer than the timescale of the kinetics, the authors developed a method analyzing intensity patterns of diffraction limited spots of a much shorter length of trajectory. The performance of the method is demonstrated using the DNA-PAINT method, in which a mixture of two different DNA binding systems with different binding kinetics and stability was tested. The performance of the method is superior to various other methods compared. The method section of the manuscript is very well written, including details of the neural network and associated analyses. However, I had a hard time understanding the beginning of the result part because some concepts and ideas that are necessary to understand this part are written in the later part of the manuscript. In addition, it is not entirely clear why this method works with short trajectories. I suspect the method analyzes simply diffraction limited spot images of different intensity levels. This has nothing to do with binding kinetics. If this is true, this method will be limited to cases with large intensity variations, which may also be sensitive to various environmental noise in real biological systems. (see comment #5) These points and other minor issues are listed below. Proper address of these points and general applicability will be required for publication of this work.

1. The correlation analysis in Fig. 1b is difficult to understand. How was this calculated? Any equation? I suppose each point was averaged over many molecules (diffraction limited spots). Then, why does high correlation keep recurring at later times? And why are there high correlation blocks? What happened near frame number 320 of the Domain 1 correlation, where the correlation with all other frames are low?

2. I believe it will be greatly helpful for understanding the paper to have a section before starting Results that briefly overviews the method, states what is analyzed by the neural network, and defines variables. For example, it took some time to find out what X, Y, Z mean in Fig. 1c and d. This became clear after reading Methods section 4.3.

(Line 324-326 in section 2.2)

"The PDF method classifies molecule types based on all binding events occurring on each molecule, while image-based methods distinguish binding types for individual binding events."

(Line 346-348 in section 2.3)

"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."

The above will be good to be placed in the overview section.

3. Fig. 2c: How is it possible to get the ground truth in the mixture of the two domains? In addition, this might be an error, but the two sites in the ground truth are classified as Domain 1 (8 nt) at 5 s, but Domain 2 (10 nt) afterwards. And, molecules never dissociate (or photobleach). The mean on times of Domain 1 and 2 are several seconds and ~ 20 s, respectively. In this case, there must be sites that are present in an early time point and disappear at a later time. In addition, while mean on time of Domain 2 is several times longer than that of domain 1, mean off times are comparable. In this case, Domain 2 should be observed more than Domain 1, but the image in Fig. 2c shows the opposite.

4. In Fig. Domain 1 data is behind Domain 2 and is invisible. Also, what causes occasional low accuracy points (< 80% or even much lower) in both Domain 1 and 2? Does this mean there are not many events and one (or a few) misclassification cases make the average accuracy that low?

5. Fig. 3e presents the most important point in the manuscript. This shows the difference of the diffraction limited image of Domain 1 and 2. The authors listed several factors that may affect the diffraction limited image (such as intensity, conformational difference, spectral shift, etc) in Discussion, but to this reviewer, all those factors (except for spectral shift, which is believed to be too small to affect the image) will result in the difference in brightness. The authors also mentioned that (also in the correlation in Fig. 1c and d), domain 1 is less bright than domain 2. So, it looks like the different characteristics of the diffraction limited spots may result from the different intensity and associated statistical fluctuations of the two species.

What is the brightness (i.e., integrated intensity of the spots) distribution of the two domains? Are they well separated?

If the above is true, the method in this manuscript is discriminating between two species with different brightness, possibly better than discriminating simply by an intensity threshold, rather than discriminating different binding events. Probably, this is why it takes only 5 seconds, the timescale much shorter than the binding kinetics.

6. Related with #5, from the practical cellular imaging perspective, there are other issues such as more than two binding kinetics, distribution (intensity) in each species, overlapping spots of different species. It will be great to have some

discussion on these points in terms of the strength and limitation of the method.

7. Figure fonts are too small. Although Nat Commun is an online journal, it is really difficult to read. Letters are small even with 200% magnification.

(Remarks on code availability)

I have checked the codes exist, but have not run for test. The readme file contains information on the installation and creating conda environment, Jupyter configuration and data processing.

Reviewer #3

(Remarks to the Author)

In this manuscript, Yin et al. developed a temporal-to-channel (T2T) as a convolutional neural network (CNN) approach for classifying single-molecule binding events using DNA-PAINT microscopy. Unlike natural videos, which rely on motion patterns of spatial features and spatiotemporal correspondence, fluorescence microscopy movies require analyzing spatial frequency features over extended observation times. To address this, the authors replaced traditional temporal domain convolutions with transformations across a wide time span in this T2T CNN model. This approach uses the same dye for all detecting molecules. Using DNA-PAINT methods, the authors argued that spatial frequency sampling significantly outperforms traditional spatiotemporal models in fluorescence microscopy. The content of this manuscript is written mostly from the machine-learning perspective with potential applications in fluorescence microscopy. This reviewer evaluates the work from the perspective of a single-molecule microscopy user.

The key implementation in the microscopy setting in this report is that it uses two oligo probes with identical dye, but differing in complementary sequence lengths (8-nt and 10-nt). In DNA-PAINT, these imager oligo probes bind and dissociate alternatively, with binding stability determined by the sequence length, resulting in distinct dissociation kinetics. By leveraging these differences in binding frequency and dwell time, the authors distinguished the binding of different oligos using a single color. Combining the T2C CNN model with binding location enabled super-resolution imaging of two targets with one dye, effectively doubling the readout capacity and reducing imaging time. Future optimizations using T2C CNN could further minimize experimental and processing times. From a microscopy perspective, using two probes expands the color limit in fluorescence imaging, though current methods like sequential antibody labeling achieve similar results. However, the imaging data presentation must be included and improved for better clarity and appreciation.

Major Comment

1. Raw images and time courses of imager strand binding and dissociation should be included in Figure 1. Processed images in Figure 3 do not provide sufficient clarity about raw data interpretation.
2. Clear definitions of binding time and off time are necessary. Confidence in assigning signals to Domain 1 or Domain 2 using T2C CNN requires further explanation.
3. Saliency map descriptions as well as the convolution from $1*10*10$ into $64*5*5$ need more detailed explanations.
4. The validation of the proposed T2C CNN method is required, such as sensitivity analysis of key parameters is missing, choice of hyperparameters needs justification, and robustness testing across different experimental conditions is required.
5. As common single-molecule experiments often involve two or three colors, it would benefit to explore the method's performance with multiple binding types. For example, performance with multiple (>2) binding types, applicability to unknown binding patterns, and validation from biochemical systems could demonstrate applicability.

Minor Comments

1. Define how "missing information" improves classification accuracy and explain the large differences between Domain 1 and Domain 2. Enhance labels for "additional information" and "missing information" in Figure 1d for better clarity.
2. The quantitative accuracy data in Figure 2c could be displayed as a bar plot to better emphasize the superior performance of T2C CNN.
3. The role of edges in diffraction patterns should be clarified, as indicated by the saliency map, as shown in Figure 3a.
4. In Figure 3b, the definition of "groups" in the F-test is not clearly stated. It appears the authors divided the two domains into separate groups, but this should be explicitly clarified. Additionally, there is a typographical error in the figure description where "T-test" is written instead of "F-test."
5. In Figure 3e, the bar plot's meaning is not immediately clear. It seems the authors intend to show accuracy comparisons for Domain 1 and Domain 2, but this should be explicitly stated.
6. Figure 4 should include a simple depiction of the experimental design for better understanding.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

1. On the motivation and novelty of in introduced TCN convolution:

As stated by the authors, their design possesses three key features, i.e., "long-term spatial convolutions, skip concatenation

for multi-scale temporal fusion, and a no-pooling strategy to preserve spatiotemporal resolution". To the reviewer's knowledge, this design has a merit to reflect the nature of single-molecule binding event discrimination task. That is to say, this special combination can better capture the long-range spatial binding features (with somehow steady transformation-invariant feature) from variable-length video input with limited computational cost. Therefore, the reviewer has no longer concerns on this part.

2. Ablation studies on network design.

Supplementary Table 2 has elaborated this part. The reviewer acknowledges the effort of the authors on deciphering the function of each components and confirm their motivation.

3. Other revised part.

The reviewer is generally satisfied with the response and the revised part, including the explanation of Flops comparison with ResNet-18. However, the reviewer still finds it particularly hard to understand Figure 8, where too many configurations aggregate on the same page. Therefore, the reviewer kindly requests the authors to revise this Figure before acceptance.

All in all, the reviewer welcomes the author's effort in their response, and hope they can finish the final touch on presentation.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

In the revised manuscript, the authors performed additional analyses and experiments to address reviewers' comments and concerns. I appreciate their effort. I believe I have a clearer picture now than I did in the first round, but still have several questions as listed below. Briefly, I don't understand what causes the small differences in the diffraction limited images between different DNA constructs (i.e., different domains) (comment 4 and 5 below) and how these small differences can achieve such high classification accuracy at the single frame image level (see 1 - 3 below). Since this is a crucial part of the paper, I hope the authors address the issues.

1. Throughout the paper, the importance of using the temporal convolution is emphasized, but the contribution of this component in the classification is uncertain. Although this was used in training, Fig. 6d shows that the analysis of a single frame already reaches a 95% classification accuracy. There is no temporal information involved here. This may indicate that the temporal information (i.e., binding kinetics) does not contribute much to the classification. In other words, just a large number of diffraction limited images may be sufficient for training.

It may be interesting to check if training performed with the scrambled sequence of image frames (i.e., no kinetic information) or training with background images excluded would give a similar classification accuracy. This would assess how important it is to include the temporal information in the analysis.

2. From the above perspective, the relatively lower accuracy in the cell image data analysis (Figs. 9 and 10) may not result from the overlapping binding duration, but from a relatively small difference in the diffraction limited images of the two domains. Is there a plot similar to Fig. 6d and e for the cell image data?

3. Fig. 6a shows that the peripheral region of the diffraction limited image is the most critical part for classification. This is quite counterintuitive. The images in Fig. 6a clearly show intensity variations by shot noise (fluctuation by a finite number of photons in each pixel). The relative shot noise (i.e., signal to noise ratio) should be the largest in the peripheral region due to the low intensity (small number of photons). Then, how can this intrinsically noisy part discriminate between the two domains? In particular, Fig. 6d shows 95% accuracy in the single frame analysis. There is one example in Fig. 6e for explanation, but it is not clear why this intrinsic noise is not transferred to the convoluted images.

In fact, this was one of the reasons that I suspected the average intensity per frame may be different for the two domains in my first round comments. (The intensity of the two input images in Fig. 6e actually looks different.) However, Fig. 5b shows that they are overall almost the same. It may be one of the magics of deep learning, but the high performance of the T2C network in a single frame image is indeed puzzling.

4. Regarding possible reasons for the formation of the different diffraction limited images, it was stated on page 9 that "This design captures frequency variations across spatial regions, effectively encoding DNA binding dynamics that can affect fluorescence intensity, angle, and density. For instance, hybridization-induced changes in fluorophore orientation can alter emission polarization or angular distribution [33, 34]. Leveraging these insights greatly improves the efficiency of binding type classification."

→ This explanation is not sufficient. A fluorophore was not used in Ref. 33. The experiment in Ref. 34 was single-molecule two-color FRET, which can effectively detect the stacking of dyes at the end of DNA duplex because the relative orientation of the two fluorophores is fixed. In one-color experiment performed in this study, however, even if the dye stacks at the end of DNA, due to the rapid movement of DNA strand, the orientation would not be fixed, but averaged in the measurement time scale of 100 ms (one frame). The dye can stick to a rigid part such as the surface, which may fix the orientation or hinder its

rotation. However, if this happens, there must be large intensity variations, which doesn't seem to be the case.

5. In the explanation of the correlation plot of Fig. 3, it was stated on page 5, "These intermittent low-correlation segments in Domain 1 likely stem from peak shifts, spot shape changes, or transient fluctuations-more prevalent due to its weaker binding affinity. Such frame-to-frame correlation patterns provide useful signals for distinguishing binding types in fluorescence videos."

→ It is unlikely that weaker binding affinity would cause those fluctuations unless binding duration is shorter than the frame time of 100 ms as in the case of the new 6-bp data. However, in this case, the average intensity per frame is lower, which is not the case of 8-bp construct (domain 1). Peak shift clearly indicates the construct is stuck to a rigid part.

6. In Fig. 6d, why were an order of magnitude more binding events (~ 70 vs several) observed at the duration of 2 s (20 x 100 ms) for Domain 2?

(Remarks on code availability)

I went through previously, but not for the revised manuscript.

Reviewer #3

(Remarks to the Author)

The authors have made substantial revisions to the manuscript, including a more detailed description of the experimental data and figures, as well as validation of the T2C CNN model and justification of its parameters. They also added experiments illustrating three distinct binding types. I have only a few remaining comments regarding their DNA-PAINT data presentation (in response to point 3.1.1):

(1) It is unclear whether the raw images shown in Figure 1a represent single frames at a specific exposure time or cumulative images compiled from multiple binding events. Please clarify and provide additional details. If the surface-anchored docker strands are loaded at similar densities, the number and intensity of the observed imager strand spots should be comparable, differing only in their dwell durations. Also, please specify the concentrations of the imager strands in the Materials and Methods section.

(2) Figure 1b: The distinction between "length-based methods" and "image-based methods" is difficult to discern from the current illustration. Secondly, as stated in the text and figure caption, "Domain 2 exhibits longer binding durations than Domain 1," but this is not readily apparent. A dwell time analysis is recommended to better demonstrate this point. Finally, the meaning of the (+17133/+18117) annotation on the y-axis and (+6229/+271) on the x-axis is unclear and should be clarified.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Overall, it is not still clear to me why the classification accuracy of the single-frame analysis is so high, but it may not be possible to understand this completely. I appreciate the author's effort to address my questions and I recommend publishing this work.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

Regarding the response to reviewer 3.1 comment, please include the data (such as survival curves of dwell times) to support the ton estimates (as stated in figure caption 1: 2.0 ± 1.2 s for Domain 1 and 8.0 ± 6.5 s for Domain 2 via NUPACK).

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response Letter

Journal: Nature Communications

Ref No.: NCOMMS-24-77281-T

Title: “Discriminating Single-Molecule Binding Event from Diffraction-Limited Fluorescence”

Authors: Yueming Yin, Nithin Pathoor, Kamal Kant Sharma, Shiwen Zhu, Iong Ying Loh, Yan Shan Ang, Shao Ren Sim, Lanry Lin Yue Yung, Thorsten Wohland, Lipo Wang

Dear Reviewers,

Thank you for reviewing. Following your feedback, we have made revisions to improve both the scientific depth and clarity of the manuscript. All changes are highlighted in blue in the revised manuscript and this document. A summary of the revisions and a point-by-point response to your comments are attached below.

Sincerely,

Yueming Yin, Nithin Pathoor, Kamal Kant Sharma, Shiwen Zhu, Iong Ying Loh, Yan Shan Ang, Shao Ren Sim, Lanry Lin Yue Yung, Thorsten Wohland, Lipo Wang

We have made comprehensive revisions to improve both the scientific depth and clarity of the manuscript. A summary of the major changes is provided below:

Wet Lab Experiments:

1. An additional 6nt-10nt binding type has been included to evaluate the classification performance across three types of binding events (Comment 3.1.5).
2. HER2-targeted cell experiments were conducted to validate models' performance in a biochemical system (Comment 3.1.5).

Dry Lab Experiments:

1. We contextualized the proposed method by comparing it with other state-of-the-art approaches that share certain elements. This comparison, presented in the Discussion, Supplementary Note 5, and Supplementary Note 6, underscores the unique strengths and performance benefits of the proposed design (Comment 1.1).
2. An ablation study of T2C CNN was added in the main text and Supplementary Note 6 to highlight its unique contributions (Comment 1.5).
3. The main text and Supplementary Note 5 now include hyperparameter analysis and robustness testing under varied experimental conditions (Comment 3.1.4). Supplementary Table 1 has also been added to report hyperparameter evaluation for baseline models.
4. A figure (Main Text Fig. 5) was added to demonstrate that total intensity distribution alone is insufficient for accurate differentiation. We further explored average frame intensity and found no clear differences between the two domains, suggesting that CNN predictions were not solely based on average intensity levels (Comment 2.5).

Presentation Improvements: To accommodate the added content and stay within the word and figure limits, we reorganized the manuscript and moved extended analyses, figures, and tables to a supplementary document. Key updates include:

1. We renamed the model from *Temporal-to-Channel* CNN to *Temporal-to-Context* (T2C) CNN to better reflect its core mechanisms: long-term spatial convolutions ("Temporal"), contextual transformation via shallow cross-connected blocks, and a pooling-free structure that preserves fine-grained temporal context ("Context"). Corresponding revisions have been made in the Abstract, Introduction and throughout the manuscript (Comments 1.1).
2. The descriptions of standard components in the Method were reduced and relocated to Supplementary Note 9 to better focus on differences from prior work (Comments 1.1, 1.2). We also clarified the definition of classification confidence (Comment 3.1.2).
3. Clarified distinctions between Supplementary Fig. 4 and Main Text Fig. 4 to avoid misinterpretation of performance inconsistencies (Comment 1.3) and added a schematic of the experimental design in Supplementary Fig. 4 (Comment 3.2.6).
4. Supplementary Note 5 now explains why T2C CNN achieves three orders of magnitude fewer GFLOPs than ResNet-18 (Comment 1.4).
5. Supplementary Notes S1.1 and S1.2 include the frame correlation formula (for Fig. 3), explanations for repeated blocks, and analysis of frame 320 in domain 1. Supplementary Fig. 2 was added to visualize frame correlation patterns (Comment 2.1).
6. A section was added to outline the overall workflow and define key symbols and terms before the Result section, including binding time and off time (Comments 2.2, 3.1.2).

7. Clarifications were added to Fig. 4c's caption regarding the causes of observed phenomena and experimental setup. A bar plot was also added for performance visualization, and the color at 5s was corrected (Comments 2.3, 3.2.2).
8. Transparency was increased in the overlapping regions of Fig. 2e to improve visibility, and Supplementary Fig. 2 was added to provide both views (Comment 2.4).
9. In Fig. 6d, we explained why certain accuracies fall below 80%, noting the scarcity of binding events with specific durations despite a large total sample size (Comment 2.4).
10. The Discussion section now reflects on limitations and potential solutions, including extensions to multi-kinetic systems, variable intensities, and complex cellular conditions (Comment 2.6).
11. Font size was increased across all figures for improved readability (Comment 2.7).
12. Raw images and representative time series were added to the added Fig. 1 (Comment 3.1.1).
13. The "Cross-experiment" section now defines missing information and its relevance to classification accuracy. We also enhanced the visual representation of "additional" and "missing" information in Fig. 2 (Comment 3.2.1).
14. Explanations were added to Fig. 6 regarding Saliency maps and the convolutional transition from $1 \times 10 \times 10$ to $64 \times 5 \times 5$ (Comment 3.1.3).
15. The importance of bright spot edges in Fig. 6a was clarified (Comment 3.2.3). Also, "group" was changed to "domain" in Fig. 6b, and a typo ("T-test") was corrected to "F-test" (Comment 3.2.4).
16. The bar plot in Fig. 6e is now explicitly labeled as prediction confidence to avoid confusion with domain accuracy (Comment 3.2.5).

1 Response to the Reviewer#1’s comments.

We would like to thank Reviewer#1 for the suggestion to improve the quality of the manuscript. The following changes have been made to address Reviewer#1’s comments, where all the changes are highlighted in blue in the manuscript and the supplementary material.

1.1 The Reviewer comments that

One of the most concerns is the novelty of the proposed T2C CNN, as treating the time dimension as the channel dimension is not a quite new idea in deep learning [Ref1][Ref2] [Ref3].

[Ref1] Temporal convolutional networks: A unified approach to action segmentation, ECCVW 2016

[Ref2] Temporal Convolutional Networks for Action Segmentation and Detection, CVPR 2017

[Ref3] No Time to Waste: Squeeze Time into Channel for Mobile Video Understanding, arXiv 2024

Response: We sincerely thank the reviewer for this insightful comment.

We fully acknowledge that treating the time dimension as a channel is a well-established concept in the broader deep learning literature. In our work, this strategy serves as a means to facilitate long-term spatial convolutions, enabling the model to aggregate information across an extended temporal context using spatial kernels. The novelty lies in the specific combination and integration of this approach with two additional architectural components—shallow cross-block connections and a pooling-free design—into the purpose-built Temporal-to-Context (T2C) CNN architecture, designed specifically for analyzing diffraction-limited fluorescence spot videos. This combination, optimized for single-molecule fluorescence data, enables the model to effectively capture subtle spatiotemporal patterns that conventional architectures miss.

Specifically, we renamed our model as **T2C CNN—Temporal-to-Context CNN**, where:

- **Temporal** emphasizes the use of long-term spatial convolutions over extended temporal windows;
- **to** signifies the transformation or integration of temporal signals into enriched contextual representations;
- **Context** refers to the combined contributions of shallow cross-connected blocks (which enhance contextual understanding) and a pooling-free structure (which preserves fine-grained temporal precision).

To support this design, we conducted systematic analyses, including ablation studies, hyperparameter evaluations, and statistical significance testing of model performance. The ablation study establishes the synergy among the three core architectural components—long-term spatial convolutions (enabled by treating time as channels), shallow cross-block connections, and the pooling-free design—by removing each component individually. This reveals their respective roles in transforming global features and enabling effective fusion of temporal features across scales. The hyperparameter study further establishes the optimal network depth (evenly two layers per block) and underscores the importance of using a sufficiently long input slice to facilitate effective long-term spatial convolutions. Statistical testing confirms that the observed performance improvements are consistent and not due to random variation, providing additional support for the proposed architectural design.

These long-term spatial convolutions convert extended temporal segments into channel dimensions, which are then processed using spatial convolutions—allowing the model to capture nuanced spatiotemporal dynamics unique to fluorescence signals. Through this, we demonstrate that the combination of long-term spatial convolutions, skip concatenation between intermediate blocks, and a pooling-free design is critical for robust modeling under noisy imaging conditions.

While similar architectural components have been individually studied in prior work, their integration within a structured, four-block framework is novel and uniquely suited to the demands of fluorescence microscopy. This carefully constructed synergy underpins T2C CNN’s accurate, efficient, and generalizable performance across experimental settings.

We have revised the *Abstract*, *Introduction*, and *Discussion* sections to clarify these distinctions and provide comparisons with the most relevant works—[Ref2] (an updated version of [Ref1]) and [Ref3]—covering both architectural differences and quantitative benchmarks. Revised contents are listed on the following pages for ease of reference.

Added contents:

1) In the Abstract, on Page 1,

Single-molecule localization microscopy (SMLM) enables high-resolution imaging of molecular interactions, but discriminating different 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 spots, enabling synchronous single-event classification 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.](#)

2) In the Introduction, on Page 2,

...

[In this paper, 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 \[1\], 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. Overall, T2C CNN provides an accurate, efficient, and generalizable solution for synchronous single-event classification in SMLM.](#)

[We validate the approach using DNA-PAINT \(Points Accumulation for Imaging in Nanoscale Topography \[2, 3\]\), 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 minutes to just 5 seconds, 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 [4], Video Transformer [5], ED-TCN [6], and SqueezeTime [1]—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.

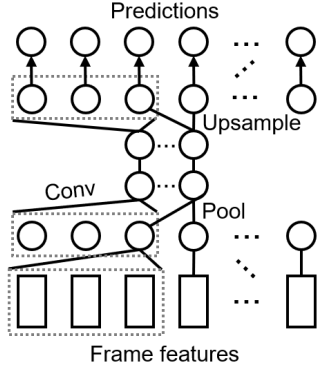
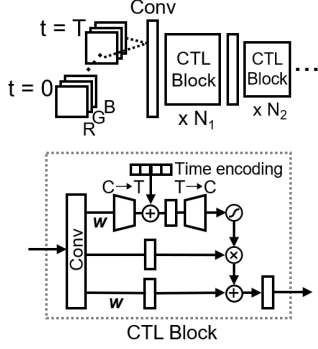
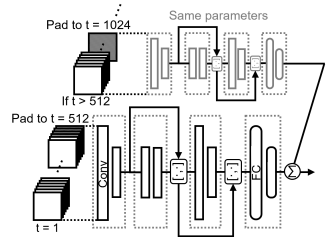
3) In the Discussion, on Page 20,

... 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 [6] and the recently developed SqueezeTime [1], which focuses on reducing computational cost and memory usage for video processing. Supplementary Table 4 highlights their differences and evaluates their suitability for fluorescence video analysis.
...

(continued on next page)

4) In the **Supplementary Table 4**, on **Page 21**,

Table 1: Comparison of methods that are most related to the proposed T2C CNN.

Feature	ED-TCN [6]	SqueezeTime [1]	T2C CNN (Ours)
Input	Pre-prepared image features	Short (16 frames), fixed-length video clips	Variable-length video
Channel composition	Stack image features in chronological order	Stack original image channels in chronological order	Stack single image channel in chronological order
Channel aggregation	Aggregated by a hierarchy of temporal convolutions	Aggregated by applying temporal-adaptive weights	Sliced and aggregated by equal-length temporal kernels
Architecture uniqueness	 Uses hierarchical temporal sequence convolutions ("Conv"), pooling, and upsampling on frame features for fine-grained action segmentation or detection.		
Fitness to fluorescence spot videos	 Short video clips are processed through hierarchical and iterative channel-to-time ("C → T") and time-to-channel ("T → C") spatial convolutions, along with channel weighting (W) and temporal encoding, to capture sequence temporal dynamics.  Processes variable-length videos via spatial convolutions on large channelized temporal slices; utilizes four evenly distributed blocks with cross-connected convolutional and fully connected layers to capture subtle spatial frequency variations over a broad temporal domain.		
	Fragmented associations between frame-level features hinder the capture of long-term variations in fluorescence videos: 1. Feature-based frames: Challenging to perceive the spatiotemporal characteristics of fluorescence videos. 2. Hierarchical temporal convolutions: Low-level detail features are less likely to be retained in high-level representations. 3. Pooling and upsampling: Impede the preservation of subtle fluorescence features and the restoration of detailed fluorescence characteristics.	Repeated conversions between channel-to-time and time-to-channel with weighted convolutions risk losing critical spatiotemporal details: 1. Weighted convolutions: May disrupt the original temporal variation patterns in fluorescence videos. 2. Iterative channel-to-time conversions: In hidden layers, reducing channel dimensions to match time length may not adequately represent temporal sequence information. 3. Multi-level convolutions without cross-links: Pose risks of over-parameterization [7] and gradient vanishing [8] during training.	A small number of cross-connected, long-term spatial convolutions enhance the ability to capture long-term detail variations in fluorescence spot videos: 1. Shallow cross-connections: Help preserve multi-scale spatiotemporal features in fluorescence videos. 2. Long-term spatial convolutions: Facilitate capturing long-term frequency patterns in fluorescence videos. 3. Pooling-free design: Enables all spatiotemporal features to be condensed into channels through large-stride convolutions, maximizing spatiotemporal information integration.

5) In the Hyperparameter analysis of the T2C CNN, on Page 13,

We systematically evaluated the impact of network depth, width, and input slice length on T2C CNN’s performance (Supplementary Note 5, Fig. 1). 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. 1, we also included the baseline models with their optimal depths and widths, which were evaluated on randomly split training and validation datasets (Supplementary Table 1).

6) In the Supplementary Table 1, on Page 18,

Table 2: Hyperparameter evaluation summary for baseline models.

Video Transformer	Depth (width=128) ¹	1	2	3*	4	5	6	7
	Accuracy ²	66.74	70.16	71.91	69.53	71.61	71.07	69.81
	Width (depth=3)	16	32	48	64	100	128	256
	Accuracy	71.01	67.23	70.21	71.18	71.78	71.91	70.48
SqueezeTime	Depth (width=512 or 2048) ³	10	18	34	50	101	152	200
	Accuracy	64.20	67.07	67.61	71.27	67.87	69.01	66.28
	Width (depth=50)	64	128	204	612	2048	3072	4096
	Accuracy	56.01	47.52	55.65	70.38	71.27	67.15	63.48
ED-TCN	Depth (width=64)	2	4	6	8	10	12	14
	Accuracy	70.78	73.56	73.34	57.37	72.58	67.25	71.28
	Width (depth=4)	16	32	48	64	100	128	256
	Accuracy	68.51	70.12	68.24	73.56	69.24	70.38	69.45
3D ResNet	Depth (width=512 or 2048)	10	18	34	50	101	152	200
	Accuracy	74.80	79.73	78.35	76.47	73.40	79.15	73.25
	Width (depth=18)	32	64	128	256	512	1024	2048
	Accuracy	75.02	78.20	75.02	73.68	79.73	76.58	77.61

¹Model depths and widths were evaluated near their default configurations. ²Class-wise accuracies for single-event classification were recorded. ³For 3D ResNet and the ResNet-based method (SqueezeTime), a width of 512 was used when the depth was below 50; otherwise, a width of 2048 was used. *Bolted values indicate the optimal hyperparameters identified in the trials.

7) In the Supplementary Note 5: Hyperparameter analysis of the T2C CNN, on Page 6,

...

Comparison with baselines. This section compares the proposed T2C CNN with two closely related baseline methods: the classical temporal convolutional network ED-TCN [6] and SqueezeTime [1], which compresses the temporal dimension into the channel dimension. For a fair comparison, we used the same 512-frame input length for SqueezeTime as T2C CNN (its default input is a 16-frame and 48-frame video clip). This adjustment significantly increased SqueezeTime’s model size (28.7 MB → 113.3 MB) and computational cost (5.5 GFLOPs → 19.56 GFLOPs).

SqueezeTime achieved classification accuracies of 71.72% with the default 16-frame input and 72.98% with the 512-frame input on fluorescence videos, indicating that increasing the number of input frames yields only marginal improvement. This is likely because the repeated transformations between channel and temporal dimensions, along with weighted convolutions in SqueezeTime, risk losing discriminative spatiotemporal details, regardless of the input length.

For ED-TCN, we tested both its default 64-frame input and a 512-frame input matching T2C CNN. Its classification accuracies on fluorescence videos were 70.53% and 58.53%, respectively, demonstrating its unsuitability for long fluorescence video inputs. A key limitation is that ED-TCN’s fragmented frame-wise spatial features hinder the capture of long-term spatiotemporal patterns in fluorescence videos. The top-performing ED-TCN and SqueezeTime models are placed in Main Text Fig. 8a.

In contrast, the proposed T2C CNN is tailored for fluorescence video analysis. Its shallow cross-layer connections retain multi-scale spatiotemporal features, while **long-term spatial convolutions** are employed to capture extended spatial frequency patterns. Additionally, the integration of **skip concatenation** with a **no-pooling** design facilitates efficient spatiotemporal information fusion. All spatiotemporal features are compressed into the channel dimension and jointly processed within the T2C layer. The individual and combined contributions of these three components are systematically evaluated in the Ablation Study section. For reference, 3D ResNet-18 and Video Transformer are included in Main Text Fig. 8a. These two models are widely adopted in computer vision for video classification and generally outperform lighter architectures such as SqueezeTime and ED-TCN due to their strong representational capacity. However, as discussed in the “Motivation” of the Main Text Method section, they are still suboptimal for fluorescence video classification, which demands a domain-specific architectural design.

(continued on next page)

8) In the Main Text Fig. 8, on Page 14,

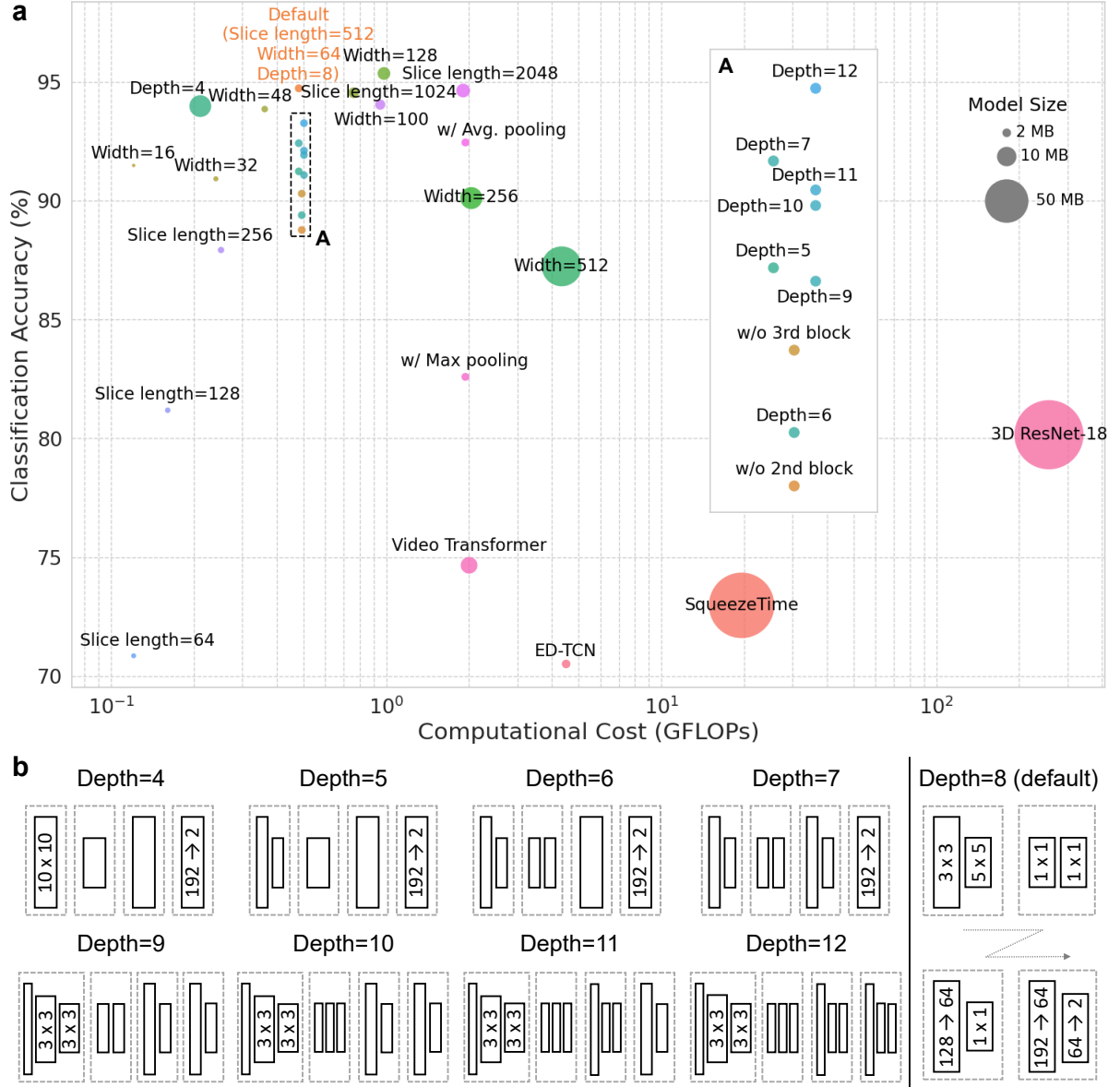


Figure 1: **a) Comparison of classification accuracy, computational cost, and model sizes for variants of the proposed T2C CNN.** Baseline methods, including 3D ResNet-18 [4], Video Transformer [5], ED-TCN [6] and SqueezeTime [1], are also evaluated for comparison (detailed comparison discussed in the 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. **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.

9) In the Ablation study of T2C CNN architecture, on Page 15,

We dissected T2C CNN’s architecture to assess the contributions of its key blocks and design components (Supplementary Note 6, Supplementary Table 1). Removing either the second

(Hidden Transformation) or third (Multi-Scale Feature Fusion) block significantly degraded performance (94.76% $\rightarrow$ 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.

10) In the Statistical significance analysis of model performance, on Page 15,

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

11) In the Supplementary Table 2, on Page 19,

Table 3: Ablation study of T2C CNN components¹.

Settings	Long-term spatial convolutions	Skip concatenations	No pooling	Class-wise accuracy (%) ²		Overall single-event accuracy (%) ³
				Single-event	Single-molecule	
Baseline	–	–	–	72.01 $\pm$ 2.76	79.87 $\pm$ 2.70	72.95 $\pm$ 2.33
Individual component	–	–	✓	70.73 $\pm$ 1.69	77.60 $\pm$ 2.66	74.66 $\pm$ 3.59
	–	✓	–	72.58 $\pm$ 3.36	78.13 $\pm$ 3.96	76.22 $\pm$ 3.09
	✓	–	–	84.14 $\pm$ 3.80	88.86 $\pm$ 4.07	89.07 $\pm$ 2.44
Combined components	–	✓	✓	72.51 $\pm$ 3.82	79.86 $\pm$ 3.19	74.96 $\pm$ 4.56
	✓	✓	–	86.33 $\pm$ 5.16	91.11 $\pm$ 4.21	90.70 $\pm$ 3.65
	✓	–	✓	87.34 $\pm$ 1.89	92.02 $\pm$ 2.38	91.59 $\pm$ 1.75
All components	✓	✓	✓	94.76 $\pm$ 0.47	96.99 $\pm$ 0.46	96.78 $\pm$ 0.50

¹Each component is removed or included to assess its individual and joint contribution to classification accuracy. ²Class-wise accuracy at the single-event and single-molecule levels. Single-molecule predictions derived by confidence-based voting over associated single-event predictions. ³Overall single-event accuracy is calculated across all events without class-wise averaging. Reported values are mean $\pm$ standard deviation over five cross-validation trials.

(continued on next page)

12) In the Supplementary Table 3, on Page 20,

Table 4: 5-fold single-event class-wise accuracies of all tested methods and their average ranks.

Methods	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average rank ($\uparrow$) ¹
T2C CNN (ABC ²)	94.76	94.28	94.36	95.60	94.78	1
T2C CNN (AC)	90.91	86.68	86.85	87.01	85.26	2.8
PDF (All event)	83.88	84.42	84.26	84.42	84.42	4
T2C CNN (AB)	79.35	89.59	86.33	93.97	82.40	4.4
PDF (2 event)	83.80	83.90	83.92	83.90	83.90	5
T2C CNN (A)	81.84	80.70	88.11	89.38	80.69	5.2
PDF (3 event)	82.09	82.18	82.00	82.18	82.18	6.6
3D ResNet-18	80.17	72.58	82.79	80.45	83.72	7.4
PDF (1 event)	75.27	75.27	75.16	75.27	75.27	9.6
T2C CNN (B)	70.22	72.64	69.69	78.98	71.38	12
Video Transformer	74.67	70.85	69.76	72.44	70.85	12.4
T2C CNN (BC)	68.66	68.01	73.26	78.40	74.22	12.4
T2C CNN None	71.61	70.00	69.55	71.61	77.28	12.6
T2C CNN (C)	70.49	69.75	70.35	73.97	69.09	13.2
SqueezeTime	71.72	69.18	71.27	69.55	66.34	13.6
ED-TCN	62.19	63.63	76.01	67.43	71.84	13.8

¹Ranks were computed within each fold using Friedman ranking and then averaged across folds. These ranks were used in the Friedman test and post-hoc analyses reported in the main text. All 16 methods and their variants were sorted in ascending order based on their average ranks. ²The letters “A”, “B”, and “C” following “T2C CNN” indicate the inclusion of long-term spatial convolutions (A), skip concatenations (B), and no-pooling (C) components, respectively. Combinations of these letters represent variants that include the corresponding components. The suffix “None” denotes the variant in which none of these three components are included.

13) In the Supplementary Note 6: Ablation study of the T2C CNN, on Page 7,

To investigate the impact of T2C CNN’s architectural components, we conducted an ablation study focusing on its core building blocks and components. Three key components (**long-term spatial convolutions**, **skip concatenation** and **no-pooling**) were selected for a detailed evaluation of their individual and combined contributions, as presented in Main Text Fig. 8 and Table 3.

Experimental settings. (a) Blocks: T2C CNN consists of four blocks, with two being fundamental: the Temporal-to-Channel (T2C) block for condensing spatial-temporal features and the Feature Classification block. To assess the importance of the remaining blocks—the second (Hidden Transformation) and third (Multi-Scale Feature Fusion)—we individually removed each and evaluated the impact.

(b) Components: The proposed T2C CNN is specifically designed for fluorescence video analysis. It employs **long-term spatial convolutions** to capture extended spatial frequency patterns. The combination of **skip concatenation** and a **no-pooling** design enables effective spatiotemporal information fusion across the network. For evaluation, we compare the optimal configuration (slice length of 512, all skip concatenations enabled, no pooling) with a baseline configuration (slice length of 64, no skip concatenation, average pooling), corresponding respectively to the presence and absence of the three proposed components.

Experimental results (Main Text Fig. 8a). (a) Blocks: Removing the second and third blocks from the T2C CNN architecture reduced classification accuracy from 94.76% to 88.80% and 90.33%, respectively. This significant performance drop indicates that both blocks are integral

components of T2C CNN. Among the two, the second block (Hidden Transformation) plays a more critical role. It consists of two 1×1 convolutional layers that transform the condensed global spatiotemporal features into more discriminative representations, facilitating the differentiation of binding types. Notably, when the third block (Multi-Scale Feature Fusion) operates without the transformed features from the second block, its fusion-based features result in lower accuracy (88.80%) compared to using only the transformed features from the second block (90.33%). This demonstrates the importance of the second block in preparing features for effective multi-scale fusion.

(b) Components: Table 3 presents a systematic evaluation of the key components of T2C CNN—long-term spatial convolutions, skip concatenations, and the no-pooling strategy—on both single-event and single-molecule classification accuracy, with all results averaged over five cross-validation trials.

Among individual components, long-term spatial convolutions contribute the most significant performance gain, boosting class-wise single-event and single-molecule accuracies to 84.14% and 88.86%, respectively, compared to the baseline’s 72.01% and 79.87%. This highlights their pivotal role in capturing extended spatial-frequency patterns essential for distinguishing binding types in fluorescence videos. In contrast, skip concatenations (72.58% / 78.13%) and no-pooling (70.73% / 77.60%) result in slightly lower class-wise accuracies than the baseline, suggesting that their benefit emerges primarily when combined with other architectural elements to support multi-scale feature fusion and preserve spatiotemporal information. The overall single-event accuracy follows a similar trend, with long-term convolutions providing the most substantial standalone gain (89.07% vs. 72.95% baseline).

When components are combined, their effects are complementary. Integrating long-term convolutions with skip concatenations increases class-wise single-event accuracy to 86.33%, while pairing long-term convolutions with no-pooling pushes it to 87.34%. In contrast, combining skip concatenations with no-pooling without long-term convolutions yields only a modest 72.51%, reinforcing the central importance of long-term spatial convolutions.

The full T2C CNN, incorporating all three components, achieves the best performance: 94.76% class-wise single-event accuracy, 96.99% single-molecule accuracy, and 96.78% overall accuracy. Notably, it also exhibits the lowest standard deviations across all metrics ($\pm 0.47\%$, $\pm 0.46\%$, and $\pm 0.50\%$, respectively), indicating stable and consistent predictions across cross-validation folds. This demonstrates that the proposed architectural combination of existing elements enables not only optimal spatiotemporal information fusion and feature preservation, but also robust generalization, tailored specifically for high-precision fluorescence video analysis.

14) In the Robustness test of image-based binding-type classification models, on Page 12,

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 4 and Fig. 2 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 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%.

15) In the Main Text Fig. 7, on Page 13,

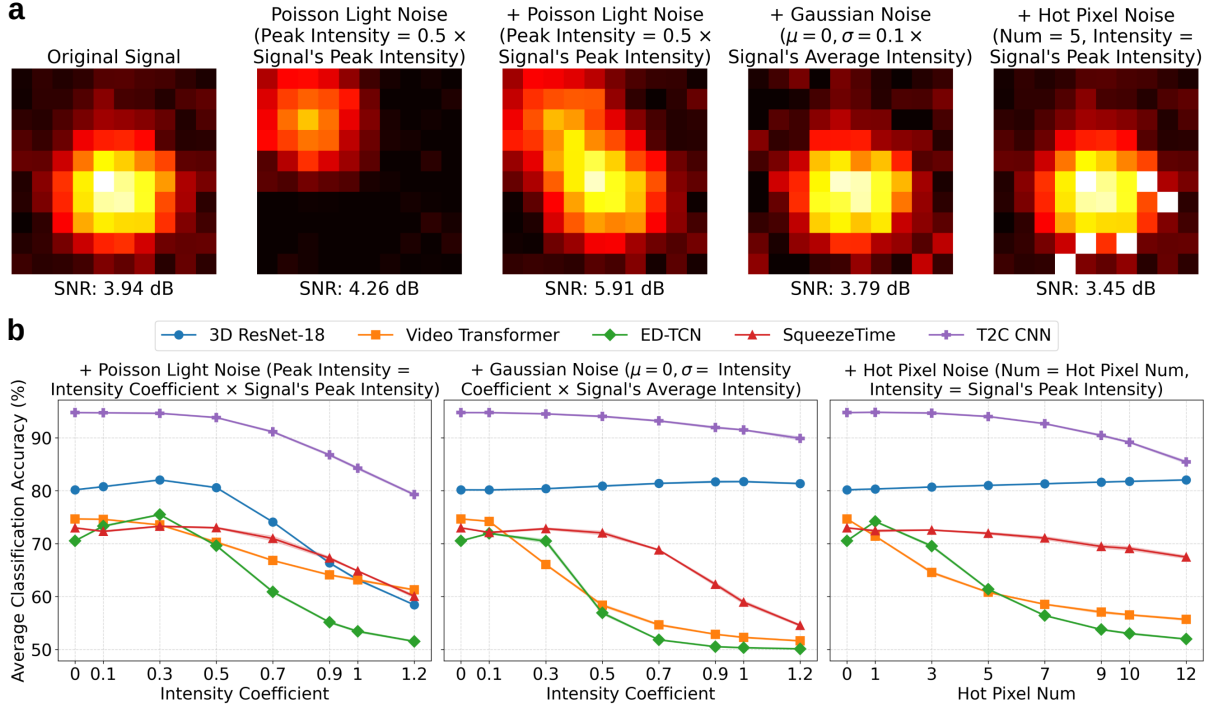


Figure 2: **Robustness test results.** (a) Examples of generated noise added to the original signal, including Poisson light noise with Gaussian decay, Gaussian noise, and hot pixel noise. The symbols μ and σ denote the mean and standard deviation (std.), respectively. (b) Average classification accuracies 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.

16) In the Supplementary Note 4: Robustness test of image-based binding-type classification models, on Page 4,

...

Robustness test results. Main Text Fig. 7(b) records the average classification accuracy of the tested models (3D ResNet-18 [4], Video Transformer [5], ED-TCN [6], SqueezeTime [1], and the proposed T2C CNN). For SqueezeTime and 3D ResNet-18, we use the default 48-frame input, which achieves comparable accuracy to the 512-frame input but with significantly higher efficiency.

(a) Light interference with Poisson distributions. Poisson noise has the most significant impact on classification performance, as it can result from specific or nonspecific binding events occurring in close proximity, causing signal overlap. When the interference spot's peak intensity exceeds 0.5 times that of the signal, all models show a marked performance decline. Increasing the interference intensity from 0.5 to 0.7 times results in accuracy drops of 2.78% to 12.54%. SqueezeTime (2.78%) and T2C CNN (2.88%) exhibit the greatest resilience. Both models employ temporal-to-channel transformations, highlighting the effectiveness of this operation in enhancing robustness against interference spots. One possible reason is that the temporal-to-channel transformation captures long-term fluorescence variations within channels, helping the model focus on the primary signal.

(b) Gaussian noises. The models display polarized performance under Gaussian noise. Video Transformer and ED-TCN, which extract image features before temporal operations, are the most

vulnerable. In contrast, 3D ResNet-18 and T2C CNN, which feature residual connections and cross-block concatenation, show the strongest resistance. Residual connections and cross-block concatenation pass low-level features from earlier layers directly to deeper layers, helping retain structural and spatiotemporal information and making it easier to denoise the signal. This is because Gaussian noise typically affects fine-grained details rather than structural patterns. Conversely, models that first extract image features propagate Gaussian noise into later temporal operations, exacerbating its effect.

(c) Hot pixel noises. 3D ResNet-18 and SqueezeTime, both with large model parameters (>100 MB), show the highest resilience to hot pixel noise. Smaller models (T2C CNN, ED-TCN, and Video Transformer, with parameters from 1.65 to 7.80 MB) are more susceptible, with T2C CNN being the least affected among them. Large models can memorize fine-grained patterns, making them less sensitive to small-scale disruptions such as hot pixels. In contrast, smaller models, with limited parameter capacity, are more susceptible to pixel losses. Despite having the fewest parameters, T2C CNN stands out among small models due to its high parameter efficiency (long temporal kernel and equal-length stride), which enhances its resistance to hot pixel noise.

...

17) In the Multi-class experiments, on Page 9,

We extended the proposed approach to three-class classification by introducing an additional 6nt–6nt R1 strand binding type (Fig. 3a). As shown in Fig. 3c, 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. 2.

(continued on next page)

18) In the **Main Text Fig. 5**, on Page 10,

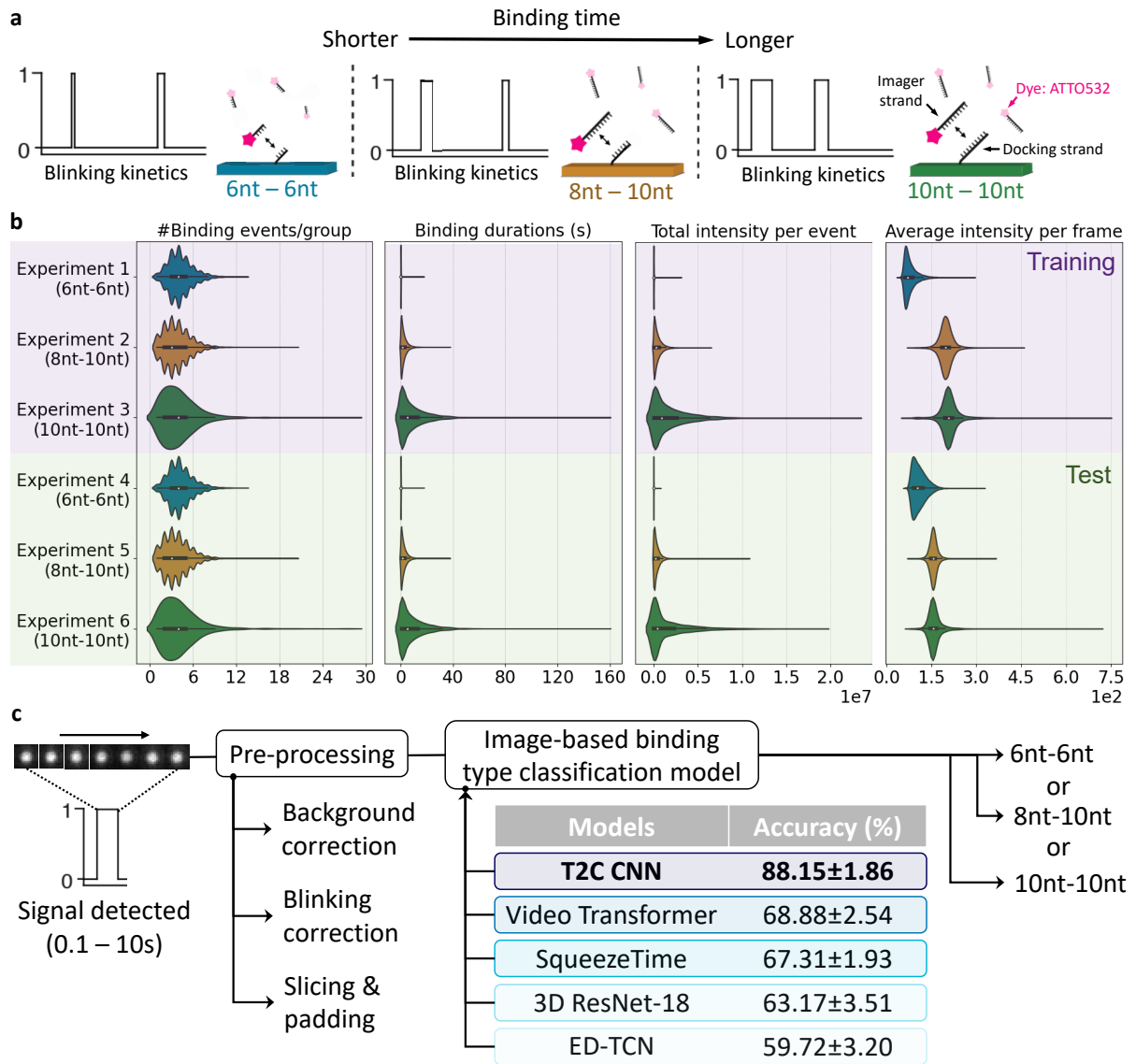


Figure 3: **Multi-Class Data, Workflow, and Results.** **a**) Illustration of the three binding types (6nt-6nt, 8nt-10nt, and 10nt-10nt) used in the multi-class experiments. **b**) Data distributions from the experiments. Notably, the 6nt R1 imager exhibited lower average frame intensity (approximately 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.

19) In the **Supplementary Note 3: Multi-class experiments**, on Page 3,

...

Workflow. Main Text Fig. 5(b) illustrates the workflow for image-based multi-class classification of binding types. When a binding event is detected (typically lasting 0.1–10 seconds), its fluorescence signals undergo background and blinking correction to eliminate background intensity variations and transient disappearances. The processed signals are then sliced, padded to a uniform length, and input into image-based binding classification models, including 3D ResNet-18 [4], Video Transformer [5], ED-TCN [6], SqueezeTime [1], and the proposed T2C CNN.

Results. After training, T2C CNN achieved the highest test accuracy (88.15±1.86%), followed by Video Transformer (68.88±2.54%). These models classify each detected signal into one of three binding types (6nt-6nt, 8nt-10nt, or 10nt-10nt). Using these predictions, we differentiate targets

labeled with distinct docking strands in the final super-resolution (SR) image, generating a three-color SR image, as shown in Fig. 4. T2C CNN consistently demonstrated superior classification accuracy across accumulated time points, with its advantage becoming more pronounced over longer durations. This makes it a powerful tool for accurate single-fluorophore, three-color SR imaging.

...

20) In the [Supplementary Fig. 3](#), on [Page 14](#),

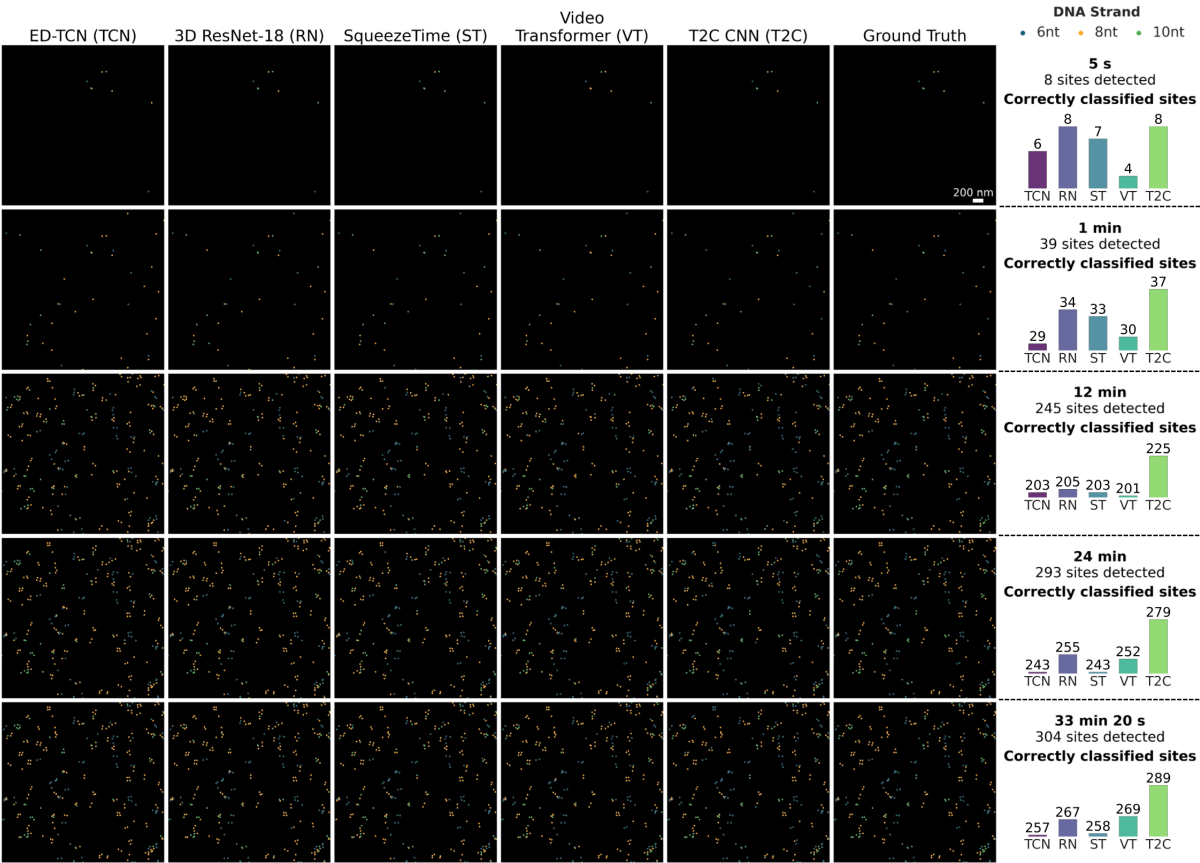


Figure 4: **Multi-Class Predictions.** Pseudo-colored super-resolution images reconstructed at 5 s, 1 min, 12 min, 24 min, and 33 min 20 s using image-based binding type classification models. The number of detected binding sites and correctly classified sites are shown on the right. Note that due to the higher concentration of 6nt and 8nt, they exhibited an increasing number of observed binding sites over time.

1.2 The Reviewer comments that

Section 4.3 include too many details about the common components of neural network, e.g., BN, ReLU, cross entropy loss, which make it harder to understand what is newly proposed in this manuscript.

Response: Thank you for the helpful comment.

We have revised the Method section (Section 4.3) to streamline the discussion by minimizing descriptions of standard neural network components (e.g., BatchNorm, ReLU, loss functions), relocating detailed explanations to Supplementary Note 9. This study reveals previously underutilized discriminative information in diffraction-limited fluorescence spot videos for single-molecule binding event identification. To effectively capture this, we designed a domain-specific CNN architecture—T2C CNN—featuring three key architectural innovations: long-term spatial convolutions, skip concatenation for multi-scale temporal fusion, and a no-pooling strategy to preserve spatiotemporal resolution. These elements are systematically integrated into four structured blocks that form the backbone of the network. Jointly optimized for the fluorescence video classification task, these design choices enable T2C CNN to achieve superior performance over existing methods. We have refined the description in the manuscript to emphasize these architectural contributions and their relevance to our specific application. The revised sections are attached below.

Revised parts:

1) In the Motivation, on Page 22,

Unlike conventional video classification tasks [9, 5, 10, 11], 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 [9, 11] or combine small-scale convolutions with sequential models [5, 10]. In general, 3D convolutions use strides smaller than the kernel size [12, 13, 14] to capture fine-grained features. However, such settings can lead to excessive convolutions along the temporal dimension, resulting in a proliferation of transitional features [15]. 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 [5, 10] 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 [1], 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. 6a, 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 2). 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 (Fig. 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.

2) In the T2C layer, on Page 23,

Different fluorescences are typically split into separate channels in fluorescence microscopy [16, 17]. 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 CNN's accuracy in predicting binding types. The output of a T2C layer, denoted by $\text{Conv}^{(\text{T2C})}$, for each fluorescence channel can be represented as:

$$\text{Conv}^{(\text{T2C})}(f, \omega) = \{g_k^{(\text{T2C})}\}_{k=0}^{N_{\text{slice}}}, \quad (1)$$

where

$$g_k^{(\text{T2C})}(i, j, d) = \sum_{u=0}^{W_{\text{kernel}}-1} \sum_{v=0}^{H_{\text{kernel}}-1} \sum_{w=0}^{T_{\text{slice}}-1} \omega_d(u, v, w) \cdot f(s_w \cdot i + u - p_w, s_h \cdot j + v - p_h, T_{\text{slice}} \cdot k + w), \quad (2)$$

Here, $g_k^{(\text{T2C})}(i, j, d)$ 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 $\lceil \frac{D_{\text{in}}}{T_{\text{slice}}} \rceil$. 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 $g_k^{(\text{T2C})}$, 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 [18]. After BN, ReLU activation is applied to introduce non-linearity [19] and accelerate learning convergence [20]. Following its transformation through the T2C layer, the original 3D input video segment is converted into a 2D feature map with fixed channels.

3) In the T2C CNN architecture, on Page 25,

After the T2C layer, we apply 2D convolutions and fully connected (FC) layers [21] to further extract features and make the final prediction. Supplementary Fig. 6b 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 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 [22] is adopted to preserve multi-scale spatiotemporal features. Across all blocks of the T2C CNN, a pooling-free strategy [23, 24] is employed to promote efficient spatiotemporal information fusion. While each of these components has been explored individually in prior work, their integration within 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.

4) In the Training on T2C CNN, on Page 25,

After the construction of a unique CNN architecture consisting of four blocks, the T2C CNN can be trained by a classic cross-entropy loss [25] to achieve an accurate classification of binding types. ...

5) In the Supplementary Note 9: T2C CNN blocks, on Page 10,

- (a) **Temporal2Channel block:** This block integrates the spatiotemporal information of a video by transforming a temporal sequence of grayscale fluorescence images into a channel-rich image representation and applying convolutions with large strides. This design effectively captures long-term spatiotemporal variations in fluorescence videos. The block is defined as follows:

$$l_1 = \text{T2C}(f, \omega_1, \gamma_1, \beta_1), \quad (3)$$

$$l_2 = \text{ReLU} \circ \text{BN} \circ \text{Conv}^{(2D)}(l_1, \omega_2, \gamma_2, \beta_2), \quad (4)$$

where

$$\text{Conv}^{(2D)}(l, \omega)(i, j, d) = \sum_{u=0}^{W_{\text{kernel}}-1} \sum_{v=0}^{H_{\text{kernel}}-1} \sum_{c=0}^{C_{\text{in}}-1} \omega_d(u, v, c) \cdot f(s_w \cdot i + u - p_w, s_h \cdot j + v - p_h, c). \quad (5)$$

Here, l_i , ω_i , γ_i , and β represent the output, weight, and two learnable parameters of batch normalization (BN) for the i -th layer, respectively. The convolutional kernels in these two layers have strides of 2 and 5, respectively, to capture the spatial features of diffraction-limited fluorescent spots over a long time span. They abstract the 512-frame (51.2s) temporal features of the same binding event into 64-dimensional channel features.

- (b) **Hidden transformation block:** This block transforms the hidden layer features element-wise by pointwise convolution (1×1 kernel) with an equal-dimensional output. These transformations improve feature representation for binding type classification. They are defined as:

$$l_3 = \text{ReLU} \circ \text{BN} \circ \text{Conv}^{(2D)}(l_2, \omega_3, \gamma_3, \beta_3), \quad (6)$$

$$l_4 = \text{Cat} \left[\text{ReLU} \circ \text{BN} \circ \text{Conv}^{(2D)}(l_3, \omega_4, \gamma_4, \beta_4), l_2 \right]. \quad (7)$$

The output of l_4 is the concatenation (denoted by $\text{Cat}[\cdot]$) of current convolution result with the output from l_2 , which is the previous Temporal2Channel block's output. This operation provides multi-scale convolutional features for subsequent feature fusion.

- (c) **Multi-scale feature fusion block:** This block integrates features before and after transformations and prepares them for the final step of binding type classification. The definitions are as follows:

$$l_5 = \text{ReLU} \circ \text{BN} \circ \text{Conv}^{(2D)}(l_4, \omega_5, \gamma_5, \beta_5), \quad (8)$$

$$l_6 = \text{Flatten} \left(\text{Cat} \left[\text{ReLU} \circ \text{BN} \circ \text{Conv}^{(2D)}(l_5, \omega_6, \gamma_6, \beta_6), l_4 \right] \right). \quad (9)$$

The function “Flatten($\cdot$)” reshapes the input tensor into a one-dimensional vector for each sample, making it compatible with the subsequent fully connected (FC) layer.

- (d) **Feature classification block:** This block consists of two fully connected (FC) layers: the first refines the features processed by the preceding blocks through fusion, transformation, and abstraction, while the second performs the final classification. They are defined as follows:

$$l_7 = \text{ReLU} \circ \text{BN} \circ \text{FC}(l_6, \omega_7, b_7, \gamma_7, \beta_7). \quad (10)$$

Then, the activated hidden features are passed to the final fully connected layer for classification:

$$\hat{y}_{\text{slice}} = \text{Softmax} \circ \text{FC}(l_7, \omega_8, b_8), \quad (11)$$

$$\hat{y} = \frac{1}{N_{\text{slice}}} \sum_{\text{slice}} \hat{y}_{\text{slice}}. \quad (12)$$

The output $\hat{y}_{\text{slice}}$ from the final layer represents the predicted probabilities for each temporal slice, normalized using the Softmax function [26]. For multiple temporal slices from the same binding event, we take the average ($\hat{y}$) of their predictions.

1.3 The Reviewer comments that

As shown in Fig. 2, 3D ResNet-18 performs much poorer than the proposed method, with accuracy of around 80%. In Fig. 4, ResNet-18 obtains quite similar performance with the proposed T2C CNN with accuracy of 92%+. It is not quite clear why 3D convolutional networks obtain much worse performance than the 2D one. More discussions and insights are encouraged to be included.

Response: Thank you for your thoughtful comment.

In the revised manuscript, we have clarified that the two experiments involve different tasks. Specifically, the 2D model in Supplementary Figure 4 (previously Fig. 4) is used for classifying fluorophores based on single-frame emissions—a relatively easier task—while the Main text Fig. 3 (where the 3D model was used) addresses the more challenging classification of binding types from single-fluorophore videos. The difference in difficulty arises because the PSF characteristics differ more distinctly when different fluorophores are used, due to their distinct emission wavelengths (Fig. 5a). In contrast, when the same fluorophore is used, the variations are much subtler, as they were designed to be influenced by differences in the attached DNA strand binding (Fig. 6). This makes the multicolor classification task more suitable for a broad range of image classification models, such as 2D ResNet-18, which explains its higher performance in Fig. 5. To enhance clarity, we have added a subfigure Fig. 5a to better illustrate the specific task. We have also expanded the discussion to provide more insights into the single-fluorophore experiments.

Added parts:

1) In the Supplementary Note 8: Single-frame discrimination of different dye-labeled binding events, on Page 9,

... Notably, single-frame fluorophore classification is a relatively easier task compared to classifying binding types from single-fluorophore videos. This difference in difficulty stems from the fact that PSF characteristics differ more prominently across fluorophores with distinct emission wavelengths (Fig. 5a), whereas in the binding-type classification task, the same fluorophore is used and the emission patterns are only subtly affected by different DNA strand bindings (Main Text Fig. 1). As a result, models like ResNet-18 achieve higher performance in multicolor classification (Fig. 5) than in the more challenging single-fluorophore task (Main text Fig. 3). ...

(continued on next page)

2) In the **Supplementary Figure 4**, on Page 15,

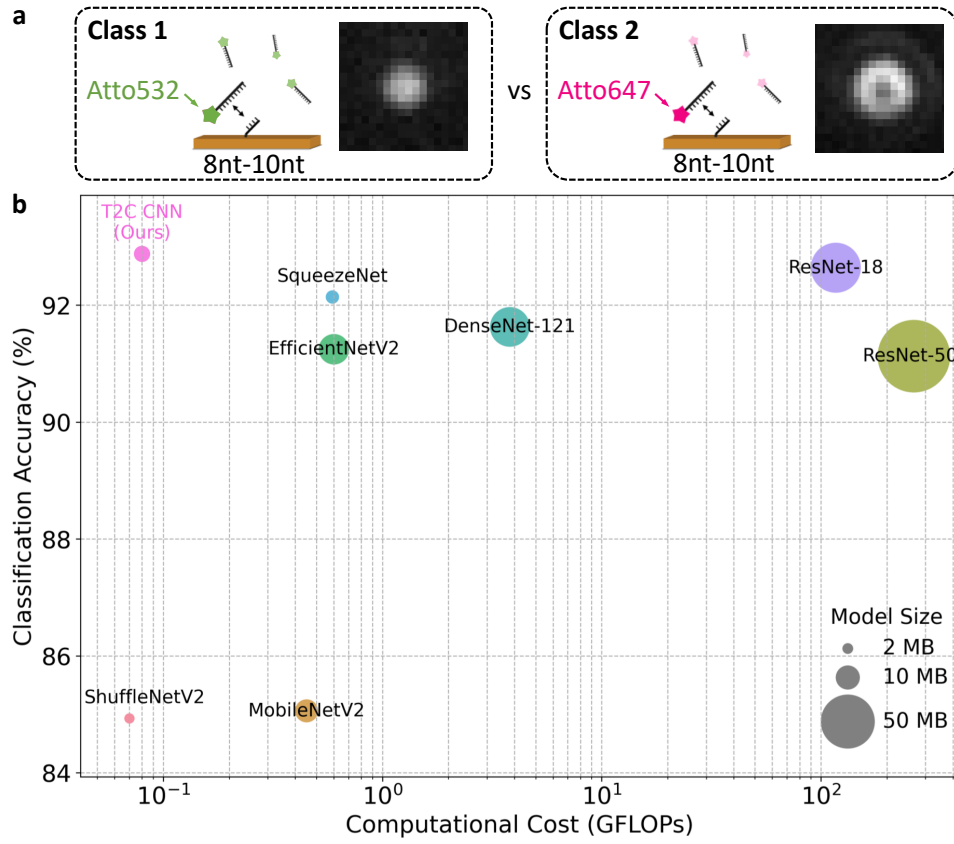


Figure 5: **Single-frame classification of different dye-labeled binding events.** **a)** Experimental setup and example images of two dye types: green (Atto532) and red (Atto647), both using the same 8nt-10nt binding. **b)** Comparison of classification accuracy, computational cost, and model sizes for single-frame discrimination of dye-labeled binding events. GFLOPs (Giga Floating Point Operations per Second) quantify the computational power required, indicating the hardware demands for processing.

(continued on next page)

3) In the **Main Text Figure 1**, on Page 4,

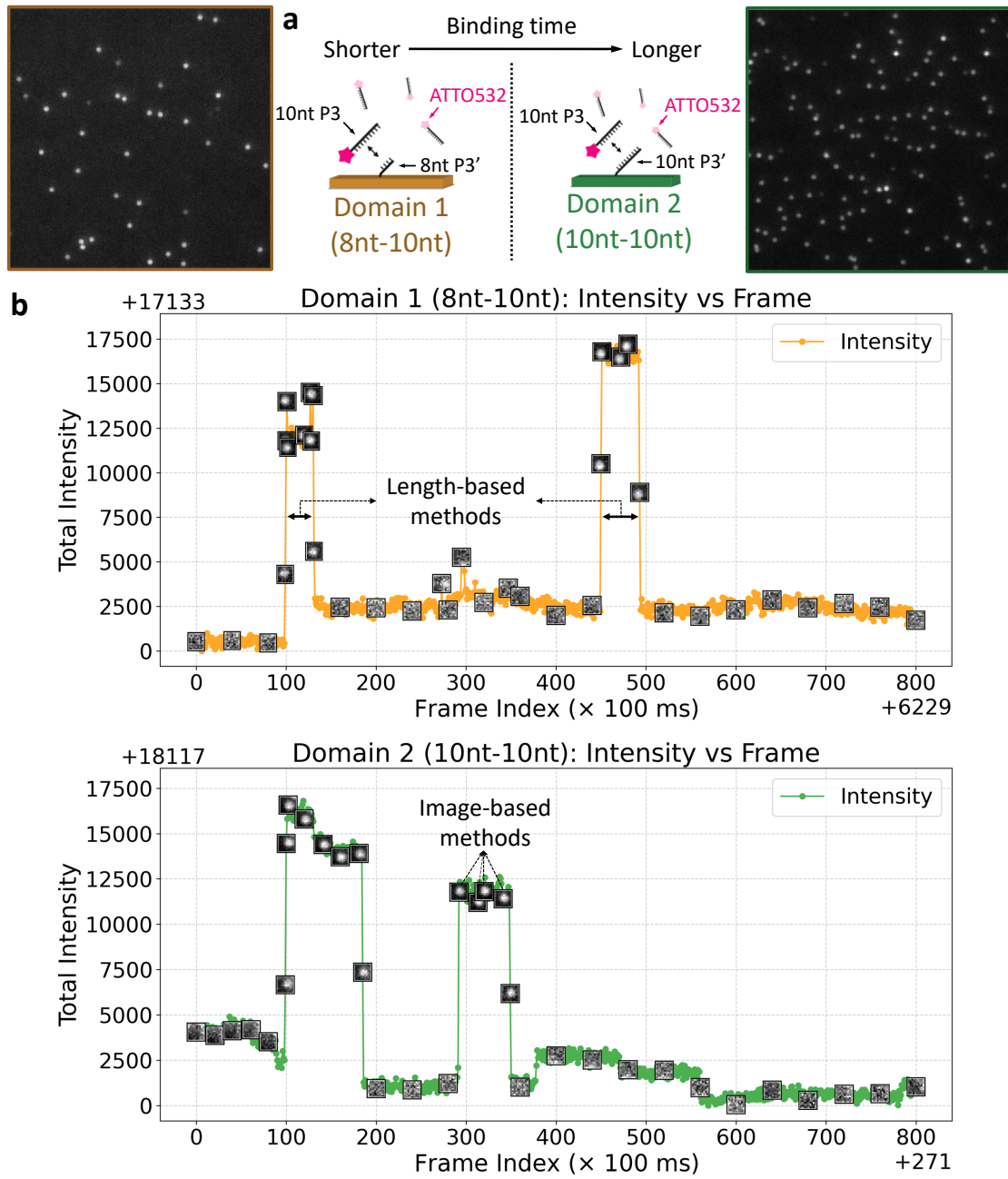


Figure 6: An example of raw images and time courses from the DNA-PAINT experiments used in this study. **a)** Schematic of the two DNA binding domains used to collect diffraction-limited fluorescent spots (adapted from [27]). Both domains share the same dye (“ATTO532”) and strand (“10nt P3”), but differ 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, while both show similar intensity jumps. The length-based method (classic) classifies molecules based on the total binding event durations per molecule, whereas the proposed image-based method classifies individual binding events based on their diffraction-limited fluorescent spots.

4) In the Cross-experiment evaluation of existing methods for discriminating single-dye binding types, on Page 9,

... The proposed T2C CNN attains $\sim 95\%$ accuracy with only one 5-second 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. For instance, hybridization-induced changes in fluorophore orientation can alter emission polarization or angular distribution [28, 29]. Leveraging these insights greatly improves the efficiency of binding type classification. ...

1.4 The Reviewer comments that

Moreover, the ResNet-18 seems to have thousands of times FLOPs than the proposed method. It is not quite reasonable since both the proposed method and ResNet-18 are built with 2D convolution.

Response: Thank you for raising this concern.

In the supplementary material, we have clarified why T2C CNN incurs significantly fewer FLOPs compared to ResNet-18, despite both using 2D convolutions. This difference mainly results from T2C CNN's lightweight architecture: it contains six convolutional layers—one-third the number in ResNet-18—with the last four using 1×1 kernels that are less computationally demanding than the 3×3 kernels used in ResNet-18. Additionally, the intermediate feature maps are kept compact (e.g., 64×64), and the final fully connected layers are smaller (192×64 versus ResNet-18's 512×1000). These design choices reduce the overall computational load, resulting in a FLOP count that is less than one-thousandth of that of ResNet-18.

Added parts of the supplementary material:

1) In the Supplementary Note 8: Single-frame discrimination of different dye-labeled binding events, on Page 10,

... Importantly, the T2C CNN achieved this accuracy with a very small computational cost—only 0.02 Giga Floating Point Operations per Second (GFLOPs)—which is even lower than ShuffleNetV2 (0.07 GFLOPs). This efficiency stems from T2C CNN's compact architecture: it has only six convolutional layers, one-third the number in ResNet-18, with the last four being 1×1 convolutions—requiring one-ninth the operations of ResNet-18's 3×3 convolutions. Additionally, its hidden transformations operate on a 64×64 width— $1/64$ of ResNet-18's 512×512 transformations—and its fully connected layers are only 192×64 , approximately $1/40$ of ResNet-18's 512×1000 configuration. Consequently, the total computational cost of T2C CNN is less than one-thousandth of ResNet-18, making it deployable on lightweight computing devices.

1.5 The Reviewer comments that

No ablation studies are included for the design choices of the proposed method, e.g, the depth/width of the proposed architecture, the effect of Multi-scale feature fusion block.

Response: Thank you for your constructive suggestion.

In response, we have added a section in the supplementary material presenting an ablation study and hyperparameter analysis to assess the design choices of the proposed method. Specifically, we examined the effects of varying the architecture’s depth, width, input slice length, and the inclusion of the multi-scale feature fusion block. A summary of the ablation study and hyperparameter analysis results is included in the main text for clarity and completeness, while detailed experimental designs are provided in the supplementary material to maintain the conciseness of the main manuscript.

Added parts:

1) In the [Hyperparameter analysis of the T2C CNN](#), on Page 13,

We systematically evaluated the impact of network depth, width, and input slice length on T2C CNN’s performance (Fig. 7 and Supplementary Note 5). 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% $\rightarrow$ 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. 7, we also included the baseline models with their optimal depths and widths, which were evaluated on randomly split training and validation datasets (Supplementary Table 1).

(continued on next page)

2) In the **Supplementary Table 1**, on **Page 18**,

Table 5: Hyperparameter evaluation summary for baseline models.

	Depth (width=128) ¹	1	2	3*	4	5	6	7
Video Transformer	Accuracy ²	66.74	70.16	71.91	69.53	71.61	71.07	69.81
	Width (depth=3)	16	32	48	64	100	128	256
	Accuracy	71.01	67.23	70.21	71.18	71.78	71.91	70.48
SqueezeTime	Depth (width=512 or 2048) ³	10	18	34	50	101	152	200
	Accuracy	64.20	67.07	67.61	71.27	67.87	69.01	66.28
	Width (depth=50)	64	128	204	612	2048	3072	4096
	Accuracy	56.01	47.52	55.65	70.38	71.27	67.15	63.48
ED-TCN	Depth (width=64)	2	4	6	8	10	12	14
	Accuracy	70.78	73.56	73.34	57.37	72.58	67.25	71.28
	Width (depth=4)	16	32	48	64	100	128	256
	Accuracy	68.51	70.12	68.24	73.56	69.24	70.38	69.45
3D ResNet	Depth (width=512 or 2048)	10	18	34	50	101	152	200
	Accuracy	74.80	79.73	78.35	76.47	73.40	79.15	73.25
	Width (depth=18)	32	64	128	256	512	1024	2048
	Accuracy	75.02	78.20	75.02	73.68	79.73	76.58	77.61

¹Model depths and widths were evaluated near their default configurations. ²Class-wise accuracies for single-event classification were recorded. ³For 3D ResNet and the ResNet-based method (SqueezeTime), a width of 512 was used when the depth was below 50; otherwise, a width of 2048 was used. * Bolded values indicate the optimal hyperparameters identified in the trials.

(continued on next page)

3) In the **Main Text Fig. 8**, on Page 14,

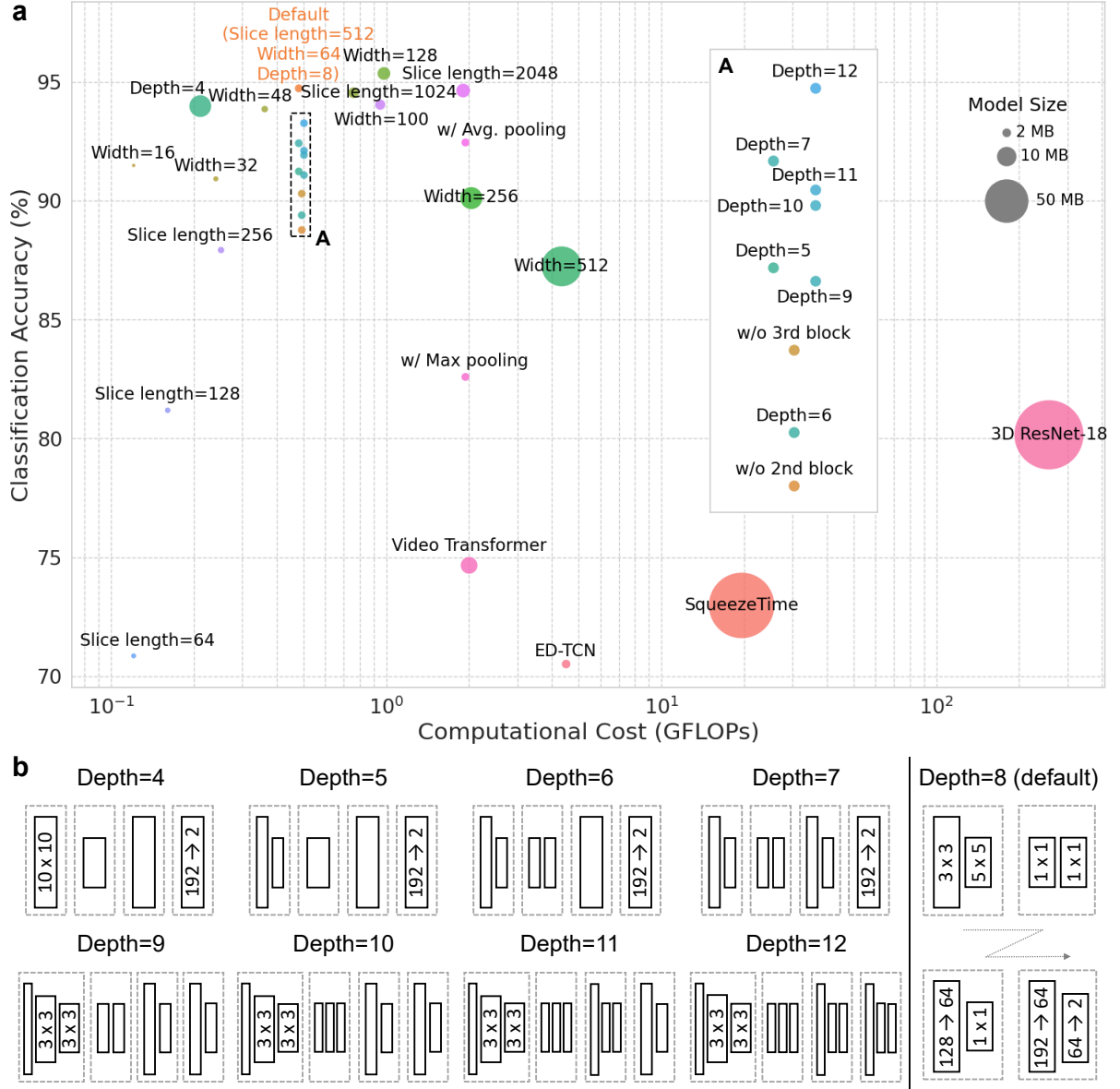


Figure 7: **a) Comparison of classification accuracy, computational cost, and model sizes for variants of the proposed T2C CNN.** Baseline methods, including 3D ResNet-18 [4], Video Transformer [5], ED-TCN [6] and SqueezeTime [1], are also evaluated for comparison (detailed comparison discussed in the 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. **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 $\rightarrow$ M.” Rectangles without labeled kernel sizes or dimension transformations follow the default hidden layer configuration for an 8-layer depth.

4) In the **Supplementary Note 5: Hyperparameter analysis of the T2C CNN**, on Page 5,

To evaluate the impact of hyperparameter choices in T2C CNN, we analyzed its key hyperparam-

eters, including network depth, width, and slice length, as shown in Main Text Fig. 8.

Experimental settings. (a) Depth: Depth is a critical attribute of deep neural networks, significantly affecting performance [30]. The proposed T2C CNN, as shown in Main Text Fig. 8, comprises eight hidden layers (convolutional or fully connected), evenly distributed across four blocks. To investigate the minimum required depth, each block was sequentially assigned 1 to 3 layers, increasing the total depth from 4 to 12 while maintaining the four-block structure. For the first block, which condenses spatial-temporal features, the convolutional kernel size was adjusted for configurations with one or three layers. Main Text Fig. 8b illustrates the corresponding architectures.

(b) Width: Like depth, network width (the number of convolutional kernels or hidden units) plays a critical role in CNN performance [31]. The default T2C CNN configuration uses 64 kernels or hidden units per layer. To explore the effect of width, we tested variations ranging from 16 to 512, including commonly used widths such as 16, 32, 48, 100, 128, 256, and 512.

(c) Slice length: The slice length, which determines the interval at which fluorescence frames are input into T2C CNN, is critical for capturing long-term temporal patterns in fluorescence videos. As shown in Main Text Fig. 2e, most binding events occur within 50 seconds (500 frames), with the longest binding duration not exceeding 2000 frames before averaging. Therefore, the default slice length of 512 frames was varied by halving or doubling to explore alternatives of 64, 128, 256, 1024, and 2048 frames.

Experimental results (Main Text Fig. 8a). (a) Depth: Among the tested depths ranging from 4 to 12 layers, the top-performing configurations were depths of 8 (94.76%), 4 (94.00%), and 12 (93.27%), all multiples of 4. These results indicate that performance is optimal when each block in T2C CNN contains an equal number of hidden layers, with two hidden layers per block yielding the best results, followed by one layer per block. Notably, at a depth of 4, the first block of T2C CNN required a larger convolution kernel (increased from “ 3×3 ” to “ 10×10 ”) to condense all spatial-temporal features within a single hidden layer. This adjustment increased the parameter count approximately 11-fold, significantly contributing to the model size.

(b) Width: Among the tested widths of 16, 32, 48, 100, 128, 256, and 512, a width of 128 slightly outperformed the default 64 (94.76% $\rightarrow$ 95.38%). However, this improvement came with a 2.62-fold increase in model size (1.65 MB $\rightarrow$ 4.32 MB) and a 2.04-fold increase in computational cost (0.48 GFLOPs $\rightarrow$ 0.98 GFLOPs). From an efficiency perspective, the default width of 64 strikes a good balance. If computational resources and latency are not constraints, a width of 128 is also recommended. Widths smaller than 64 or greater than 128 consistently underperformed compared to the default, with performance generally declining as width became too wide.

Interestingly, even with an unusually small width of 16, T2C CNN maintained performance above 90%. This resilience can be attributed to the unique block design of T2C CNN, as breaking the balance of hidden layers within blocks or removing blocks led to performance degradation, even with an optimal width of 64. Remarkably, this suggests that when a fluorescence video spanning 512 frames is condensed into 16 channel dimensions, most discriminative information is retained. This highlights the substantial redundancy in fluorescence videos, much of which stems from identical optical properties of the same fluorescence type. The key discriminative signals arise from subtle changes in the fluorescence microenvironment caused by binding to different targets. These subtle variations are condensed into a small set of discriminative features by T2C CNN, which is fundamental to its functionality.

This also explains why performance significantly declined at widths of 256 and 512. The increased number of hidden nodes likely led to overfitting on redundant information, impairing generalization to test data from different experiments.

(c) Slice length: As the slice length input to T2C CNN increases up to 512 frames, classification accuracy improves significantly. This is because the longer slice length enables T2C CNN to

capture a more complete binding event, aiding in determining the binding type. However, beyond 512 frames, the performance plateaus, as only a small fraction of binding events extend beyond this duration. In practical applications, we recommend setting the slice length to encompass the majority of binding durations (e.g., the maximum binding time excluding outliers) to achieve optimal efficiency.

...

5) In the Ablation study of T2C CNN architecture, on Page 15,

We dissected T2C CNN’s architecture to assess the contributions of its key blocks and design components (Supplementary Note 6, Supplementary Table 1). Removing either the second (Hidden Transformation) or third (Multi-Scale Feature Fusion) block significantly degraded performance (94.76% $\rightarrow$ 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.

(continued on next page)

6) In the Supplementary Table 2, on Page 19,

Table 6: Ablation study of T2C CNN components¹.

Settings	Long-term spatial convolutions	Skip concatenations	No pooling	Class-wise accuracy (%) ²		Overall single-event accuracy (%) ³
				Single-event	Single-molecule	
Baseline	–	–	–	72.01±2.76	79.87±2.70	72.95±2.33
Individual component	–	–	✓	70.73±1.69	77.60±2.66	74.66±3.59
	–	✓	–	72.58±3.36	78.13±3.96	76.22±3.09
	✓	–	–	84.14±3.80	88.86±4.07	89.07±2.44
Combined components	–	✓	✓	72.51±3.82	79.86±3.19	74.96±4.56
	✓	✓	–	86.33±5.16	91.11±4.21	90.70±3.65
	✓	–	✓	87.34±1.89	92.02±2.38	91.59±1.75
All components	✓	✓	✓	94.76±0.47	96.99±0.46	96.78±0.50

¹Each component is removed or included to assess its individual and joint contribution to classification accuracy. ²Class-wise accuracy at the single-event and single-molecule levels. Single-molecule predictions derived by confidence-based voting over associated single-event predictions. ³Overall single-event accuracy is calculated across all events without class-wise averaging. Reported values are mean ± standard deviation over five cross-validation trials.

7) In the Supplementary Table 3, on Page 20,

Table 7: 5-fold single-event class-wise accuracies of all tested methods and their average ranks.

Methods	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average rank (↑) ¹
T2C CNN (ABC ²)	94.76	94.28	94.36	95.60	94.78	1
T2C CNN (AC)	90.91	86.68	86.85	87.01	85.26	2.8
PDF (All event)	83.88	84.42	84.26	84.42	84.42	4
T2C CNN (AB)	79.35	89.59	86.33	93.97	82.40	4.4
PDF (2 event)	83.80	83.90	83.92	83.90	83.90	5
T2C CNN (A)	81.84	80.70	88.11	89.38	80.69	5.2
PDF (3 event)	82.09	82.18	82.00	82.18	82.18	6.6
3D ResNet-18	80.17	72.58	82.79	80.45	83.72	7.4
PDF (1 event)	75.27	75.27	75.16	75.27	75.27	9.6
T2C CNN (B)	70.22	72.64	69.69	78.98	71.38	12
Video Transformer	74.67	70.85	69.76	72.44	70.85	12.4
T2C CNN (BC)	68.66	68.01	73.26	78.40	74.22	12.4
T2C CNN None	71.61	70.00	69.55	71.61	77.28	12.6
T2C CNN (C)	70.49	69.75	70.35	73.97	69.09	13.2
SqueezeTime	71.72	69.18	71.27	69.55	66.34	13.6
ED-TCN	62.19	63.63	76.01	67.43	71.84	13.8

¹Ranks were computed within each fold using Friedman ranking and then averaged across folds. These ranks were used in the Friedman test and post-hoc analyses reported in the main text. All 16 methods and their variants were sorted in ascending order based on their average ranks. ²The letters “A”, “B”, and “C” following “T2C CNN” indicate the inclusion of long-term spatial convolutions (A), skip concatenations (B), and no-pooling (C) components, respectively. Combinations of these letters represent variants that include the corresponding components. The suffix “None” denotes the variant in which none of these three components are included.

8) In the Supplementary Note 6: Ablation study of the T2C CNN, on Page 7,

To investigate the impact of T2C CNN’s architectural components, we conducted an ablation study focusing on its core building blocks and components. Three key components (**long-term spatial convolutions**, **skip concatenations** and **pooling-free**) were selected for a detailed evaluation of their individual and combined contributions, as presented in Main Text Fig. 8 and Table 6.

Experimental settings. (a) Blocks: T2C CNN consists of four blocks, with two being fundamental: the Temporal-to-Channel (T2C) block for condensing spatial-temporal features and the Feature Classification block. To assess the importance of the remaining blocks—the second (Hidden Transformation) and third (Multi-Scale Feature Fusion)—we individually removed each and evaluated the impact.

(b) Components: The proposed T2C CNN is specifically designed for fluorescence video analysis. It employs **long-term spatial convolutions** to capture extended spatial frequency patterns. The combination of **skip concatenations** and a **pooling-free** design enables effective spatiotemporal information fusion across the network. For evaluation, we compare the optimal configuration (slice length of 512, all skip concatenations enabled, no pooling) with a baseline configuration (slice length of 64, no skip concatenation, average pooling), corresponding respectively to the presence and absence of the three proposed components.

Experimental results (Main Text Fig. 8a). (a) Blocks: Removing the second and third blocks from the T2C CNN architecture reduced classification accuracy from 94.76% to 88.80% and 90.33%, respectively. This significant performance drop indicates that both blocks are integral components of T2C CNN. Among the two, the second block (Hidden Transformation) plays a more critical role. It consists of two 1×1 convolutional layers that transform the condensed global spatiotemporal features into more discriminative representations, facilitating the differentiation of binding types. Notably, when the third block (Multi-Scale Feature Fusion) operates without the transformed features from the second block, its fusion-based features result in lower accuracy (88.80%) compared to using only the transformed features from the second block (90.33%). This demonstrates the importance of the second block in preparing features for effective multi-scale fusion.

(b) Components: Table 6 presents a systematic evaluation of the key components of T2C CNN—long-term spatial convolutions, skip concatenations, and the pooling-free strategy—on both single-event and single-molecule classification accuracy, with all results averaged over five cross-validation trials.

Among individual components, long-term spatial convolutions contribute the most significant performance gain, boosting class-wise single-event and single-molecule accuracies to 84.14% and 88.86%, respectively, compared to the baseline’s 72.01% and 79.87%. This highlights their pivotal role in capturing extended spatial-frequency patterns essential for distinguishing binding types in fluorescence videos. In contrast, skip concatenations (72.58% / 78.13%) and no-pooling (70.73% / 77.60%) result in slightly lower class-wise accuracies than the baseline, suggesting that their benefit emerges primarily when combined with other architectural elements to support multi-scale feature fusion and preserve spatiotemporal information. The overall single-event accuracy follows a similar trend, with long-term convolutions providing the most substantial standalone gain (89.07% vs. 72.95% baseline).

When components are combined, their effects are complementary. Integrating long-term convolutions with skip concatenations increases class-wise single-event accuracy to 86.33%, while pairing long-term convolutions with no-pooling pushes it to 87.34%. In contrast, combining skip concatenations with no-pooling without long-term convolutions yields only a modest 72.51%, reinforcing the central importance of long-term spatial convolutions.

The full T2C CNN, incorporating all three components, achieves the best performance: 94.76% class-wise single-event accuracy, 96.99% single-molecule accuracy, and 96.78% overall accuracy.

Notably, it also exhibits the lowest standard deviations across all metrics ($\pm 0.47\%$, $\pm 0.46\%$, and $\pm 0.50\%$, respectively), indicating stable and consistent predictions across cross-validation folds. This demonstrates that the proposed architectural combination of existing elements enables not only optimal spatiotemporal information fusion and feature preservation, but also robust generalization, tailored specifically for high-precision fluorescence video analysis.

2 Response to the Reviewer#2’s comments.

We would like to thank Reviewer#2 for the suggestion to improve the quality of the manuscript. The following changes have been made to address Reviewer#2’s comments, where all the changes are highlighted in blue in the manuscript and the supplementary material.

2.1 The Reviewer comments that

The correlation analysis in Fig. 1b is difficult to understand. How was this calculated? Any equation? I suppose each point was averaged over many molecules (diffraction limited spots). Then, why does high correlation keep recurring at later times? And why are there high correlation blocks? What happened near frame number 320 of the Domain 1 correlation, where the correlation with all other frames are low?

Response: Thank you for the insightful comment. In the supplementary material, we have clarified the calculation of the frame correlation matrix shown in Fig. 1b. Specifically, we compute the Pearson correlation coefficient between pairs of non-zero frames by first flattening their pixel intensities into one-dimensional vectors (in consistent row-major order), and then treating the corresponding pixel values across frames as paired observations from two random variables. These pairwise correlations are then aggregated across samples within each class to obtain a class-wise average correlation matrix that reflects temporal binding stability.

To explain the recurring high-correlation blocks and the anomaly around frame 320 in Domain 1, we performed additional analysis and included the corresponding figure in the supplementary material. This figure illustrates representative fluorescence patterns from long binding events, highlighting how shifts in spot position, blinking, photobleaching, or interference from nearby emitters influence the inter-frame correlation structure.

To aid reproducibility and interpretation, we have included the exact equations and extended explanations in the supplementary material, as shown below.

Added parts of the supplementary material:

1) In the **Supplementary Note 1: Frame correlation definition**, on Page 2,

Given a video represented as a sequence of frames, we compute the Pearson correlation coefficient (PCC) between all pairs of non-zero frames. The correlation matrix is then accumulated across different video samples belonging to the same class. Let $\mathbf{x}_t$ and $\mathbf{x}_s$ be the vectorized pixel intensities of frames t and s . The Pearson correlation coefficient between two frames is defined as:

$$\rho_{t,s} = \frac{\sum_i (\mathbf{x}_{t,i} - \mu_t)(\mathbf{x}_{s,i} - \mu_s)}{\sqrt{\sum_i (\mathbf{x}_{t,i} - \mu_t)^2} \sqrt{\sum_i (\mathbf{x}_{s,i} - \mu_s)^2}} \quad (13)$$

where:

- $\mathbf{x}_{t,i}$ is the intensity of pixel i in frame t .
- μ_t and μ_s are the mean pixel intensities of frames t and s , respectively.
- The denominator normalizes the covariance, ensuring the correlation values are within the range $[-1, 1]$.

For each class label y , the accumulated frame correlation matrix is computed as:

$$C_y(t, s) = \sum_{n \in \text{class } y} \rho^n(t, s) \quad (14)$$

where $\rho^n(t, s)$ is the correlation coefficient between frames t and s in the n -th video sample. Since different frame pairs appear with varying frequencies in the dataset, the final averaged correlation matrix is computed as:

$$\bar{C}_y(t, s) = \frac{C_y(t, s)}{N_y(t, s) + \epsilon} \quad (15)$$

where:

- $N_y(t, s)$ is the number of times the correlation between frames t and s was computed.
- ϵ is a small positive constant (e.g., 10^{-6}) to avoid division by zero.

This ensures that correlation values are normalized by their respective occurrence counts, providing an unbiased estimate of frame-wise dependencies.

2) In the **Supplementary Note 2: Analysis of frame correlation patterns**, on **Page 2**,

A particularly notable deviation occurs around frame 320 in Domain 1, where frames exhibit high mutual correlation but lower correlation with subsequent frames. Upon examining the raw data (Fig. 8), we found that among the few binding events lasting over 300 frames, the longest event (655 frames) displayed two closely positioned spots between frames 308 and 334. This interference caused a fluorescence pattern distinct from the typical single-fluorescence pattern observed in other frames, leading to a drop in correlation with later frames, especially when fewer binding events followed.

To illustrate these fluorescence features, we sampled representative inter-frame correlation images (Fig. 8). Beyond the correlation variations, subtle differences in fluorescence images can also be observed. For example, Domain 1 (8nt-10nt) exhibits slightly greater fluctuations in peak positions and shapes compared to Domain 2 (10nt-10nt). Frames such as 198 and 620 in Domain 1 show a fluorescence shift toward the right, reducing their correlation with other frames. The shorter binding duration in Domain 1 also minimizes the impact of photobleaching. In contrast, the first long binding event in Domain 2 undergoes noticeable photobleaching in its later stages, reducing the contrast between the fluorescence spot and the background. However, peak positions and shapes remain relatively stable compared to Domain 1, providing a basis for differentiation. The final two blue lines in the Domain 2 inter-frame correlations result from transient blinking (signal loss) near the end of the longest binding event.

(continued on next page)

3) In the **Supplementary Figure 1**, on Page 12,

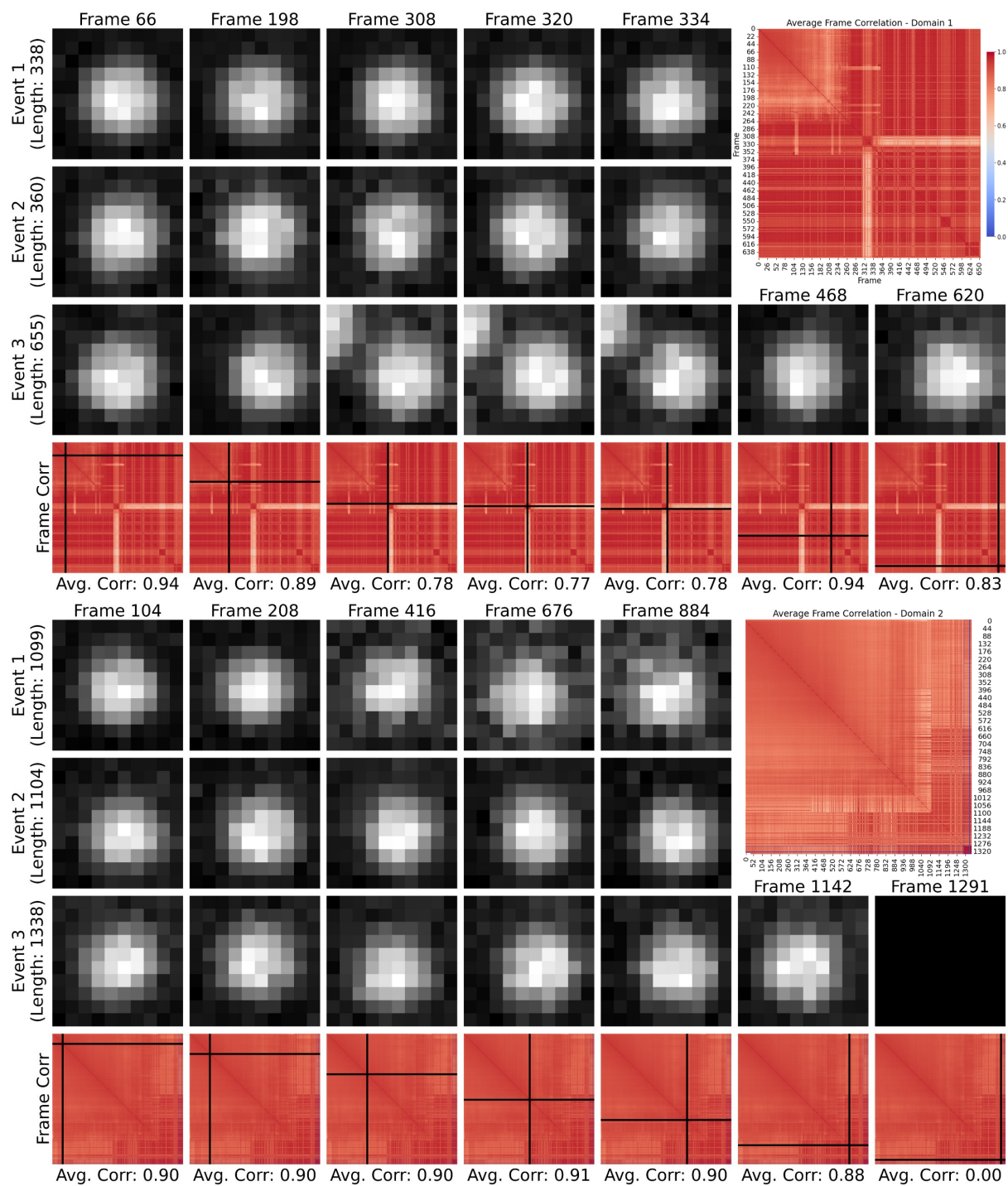


Figure 8: **Sampled fluorescence spots from the Top-3 longest binding events in each domain, selected based on typical inter-frame correlation patterns.** Domain 1 is shown above and Domain 2 below, with inter-frame correlation maps displayed in the upper right corner of each domain, highlighting the temporal fluorescence variations associated with binding stability.

2.2 The Reviewer comments that

I believe it will be greatly helpful for understanding the paper to have a section before starting Results that briefly overviews the method, states what is analyzed by the neural network, and defines variables. For example, it took some time to find out what X, Y, Z mean in Fig. 1c and d. This became clear after reading Methods section 4.3.

(Line 324-326 in section 2.2) “The PDF method classifies molecule types based on all binding events occurring on each molecule, while image-based methods distinguish binding types for individual binding events.”

(Line 346-348 in section 2.3) “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.”

The above will be good to be placed in the overview section.

Response: Thank you for the insightful suggestion.

To improve clarity and ease of understanding, we have added an “Overview of raw data and methods” section in front of the Results section. This section briefly introduces the raw data, the task setup, the proposed method, and the interpretation of key variables. This addition helps readers quickly grasp the core methodology before diving into detailed analyses. Additionally, we have updated the caption of Fig. 2 (previously Fig. 1) to explicitly define the meanings of the X, Y, and Z axes. The added section is included below for reference.

Added parts of the manuscript:

1) In the **Overview of raw data and methods**, on Page 3,

DNA-PAINT experiments generate raw data as videos, with each frame containing multiple diffraction-limited fluorescent spots (Fig. 9a). To minimize spot overlap, experimental parameters such as concentration and pH are carefully adjusted to control spot density. Traditional analysis methods, such as Picasso [2], 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. 9b).

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 [2] 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. 9b, 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.

2) In the **Main Text Fig. 1**, on Page 4,

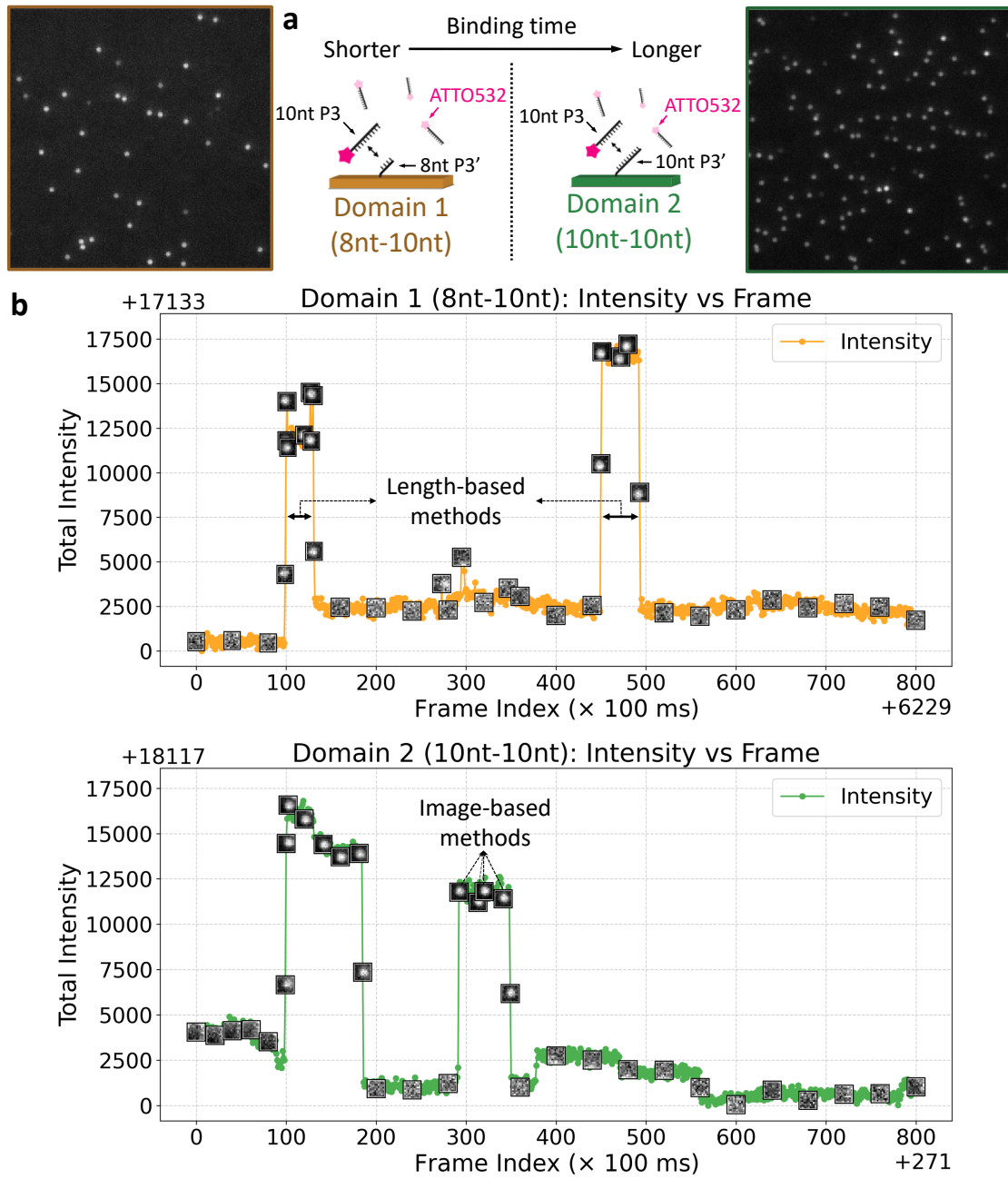


Figure 9: An example of raw images and time courses from the DNA-PAINT experiments used in this study. **a)** Schematic of the two DNA binding domains used to collect diffraction-limited fluorescent spots (adapted from [27]). Both domains share the same dye (“ATTO532”) and strand (“10nt P3”), but differ 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, while both show similar intensity jumps. The length-based method (classic) classifies molecules based on the total binding event durations per molecule, whereas the proposed image-based method classifies individual binding events based on their diffraction-limited fluorescent spots.

3) In the Main Text Fig. 2 (previously Fig. 1), on Page 6,

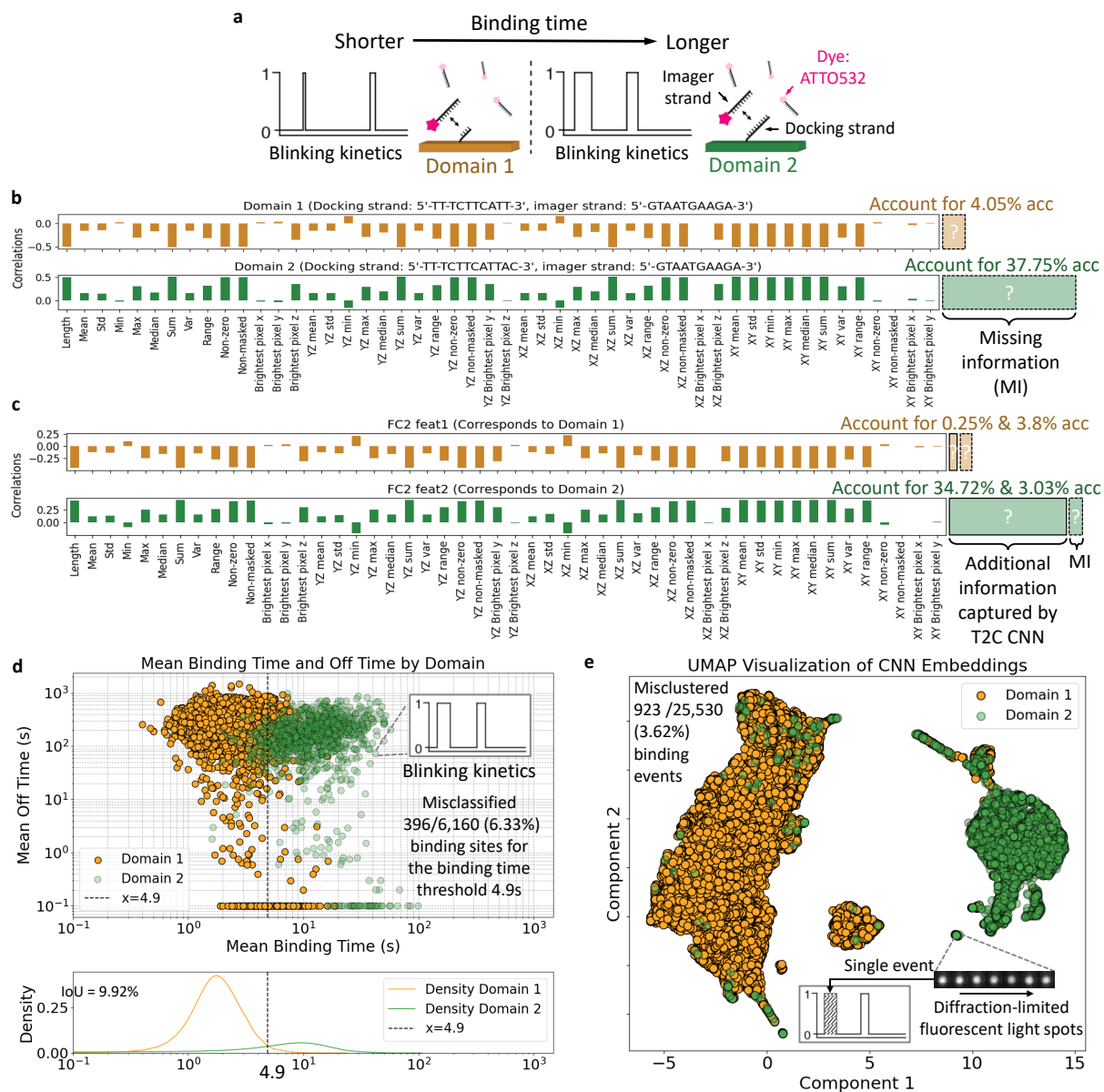


Figure 10: Diffraction-limited fluorescent spots hold binding-type information that can be utilized to more accurately differentiate between single binding events. **a**, Two types of DNA binding domains used to collect diffraction-limited fluorescent spots. Blinking kinetics refers to the binary signals of binding on and off. (Revised from [27]) **b**, The correlation between domain labels and some common statistics, 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 mean on-time is displayed. Based on multiple binding events, 6.33% of binding sites are misclassified. **e**, UMAP dimensionality reduction scatter 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.

2.3 The Reviewer comments that

Fig. 2c: How is it possible to get the ground truth in the mixture of the two domains? In addition, this might be an error, but the two sites in the ground truth are classified as Domain 1 (8nt) at 5 s, but Domain 2 (10nt) afterwards. And, molecules never dissociate (or photobleach). The mean on times of Domain 1 and 2 are several seconds and ~ 20 s, respectively. In this case, there must be sites that are present in an early time point and disappear at a later time. In addition, while mean on time of Domain 2 is several times longer than that of domain 1, mean off times are comparable. In this case, Domain 2 should be observed more than Domain 1, but the image in Fig. 2c shows the opposite.

Response: Thank you for your careful and insightful observations. Below, we offer clarifications and explanations in response to your observations.

- (a) **How we obtain the ground truth:** We collected separate videos for the two binding types and generated the ground truth for Fig. 2c using fluorescence additivity [32]. The pseudo-colored composite fluorescence images were created by frame-wise summation of the two videos. By localizing fluorescence spots before summation, we could accurately assign each spot in the composite video to its corresponding binding type. Image-based methods were trained on the mixed data from the composite video without knowledge of the original intensities in the separate videos. In the caption of Fig. 2c, we clarify the meaning of “pseudo-colored composite fluorescence images” and have attached the revised figure below.
- (b) **Inconsistency in color labeling over time:** We investigated why two sites were incorrectly marked as Domain 1 (8nt) at 5 s. This was due to a color selection bug in the plotting code, which assigned colors based on unique label indices. Since only Domain 2 (10nt) was present at 5 s, it was mistakenly assigned the first available color (which should belong to Domain 1). We have corrected this issue in the revised Fig. 2, as shown below.
- (c) **Dissociation, disappearance, and photobleaching:** In the experiments, molecules typically dissociate within seconds, causing fluorescence signals to disappear. However, because the imaging buffer contains a high concentration of fluorophore-labeled imagers, and the laser illuminates a thin focal plane near the binding sites, newly bound imagers are likely to carry non-photobleached fluorophores. Once a signal is detected at a binding site, we retain its position and binding type in the reconstructed super-resolution image, accumulating observations over time. As a result, the number of detected binding sites increases with time.
- (d) **Higher observed binding site count for Domain 1:** To control the density of fluorescence spots per frame and minimize overlap, we used a lower concentration (400 pM) for Domain 2 (10nt) due to its longer binding duration, while using 1 nM for Domain 1 (8nt) with shorter binding times. Because longer binding durations lead to events that persist across many frames, using the same or higher concentration would increase the risk of overlapping spots, complicating detection and analysis. This difference resulted in more detected binding sites for Domain 1. We have added this clarification in the revised figure captions and data descriptions, as shown below.

Revised parts:

1) In the Fig. 4 (previously Fig. 2), on Page 8,

(continued on next page)

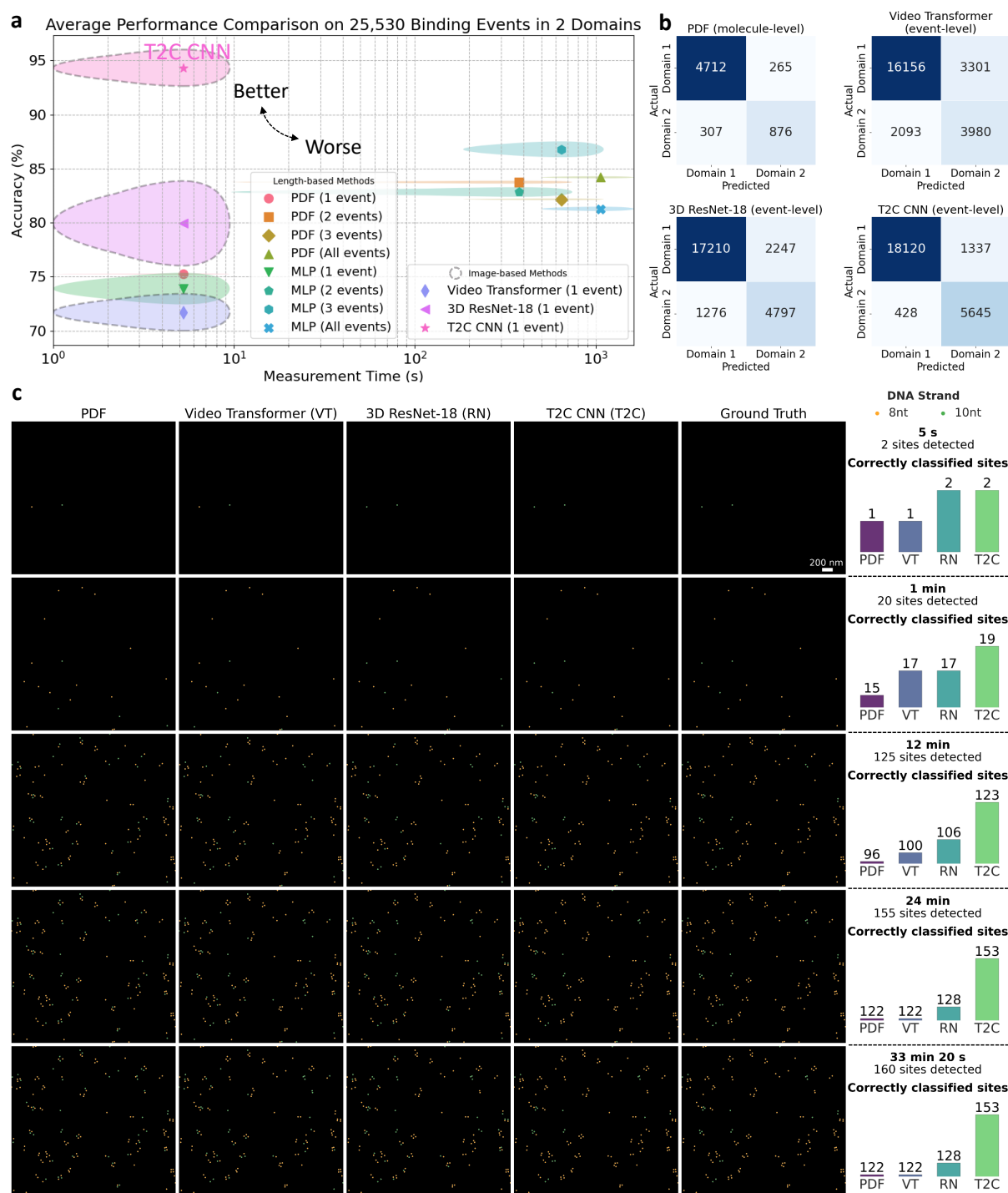


Figure 11: Cross-experiment results for discriminating binding types. **a**, Comparison of average performance on 25,530 binding events across the two domains. The width of the colored regions in the horizontal and vertical directions represents the standard deviation of measurement time and accuracy, 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.

2) In the Sample preparation, on Page 21,

... To limit fluorescence spots per frame and reduce overlap, we used 400 pM for the 10nt docking strand (with longer binding time) and 1 nM for the 8nt and 6nt strands (with shorter binding times). All experiments were conducted on separate samples.

2.4 The Reviewer comments that

In Fig. Domain 1 data is behind Domain 2 and is invisible. Also, what causes occasional low accuracy points (< 80% or even much lower) in both Domain 1 and 2? Does this mean there are not many events and one (or a few) misclassification cases make the average accuracy that low?

Response: Thank you for your helpful comment.

To improve the visibility of overlapping points, we increased the transparency of Domain 2 data in Figures 1e and 1f so that Domain 1 data points are more clearly visible. Additionally, to further improve clarity, we have included reordered versions with each domain pinned on top in Supplementary Fig. 2.

As shown in Fig. 2e and 2f, a total of 25,530 binding events were collected from 6,160 binding sites across both domains, indicating that the dataset is not limited by a low number of events. However, for a small number of long-duration binding events, the amount of available samples was limited. This led to occasional dips in classification accuracy below 80%, as observed in Fig. 6d for events longer than 10 seconds. To clarify this, the updated Fig. 6d now includes the number of binding events per duration bin, and we have revised the discussion accordingly. The revised figures and descriptions are attached below.

Revised parts:

1) In the Main Text Fig. 2 (previously Fig. 1), on Page 6,

(continued on next page)

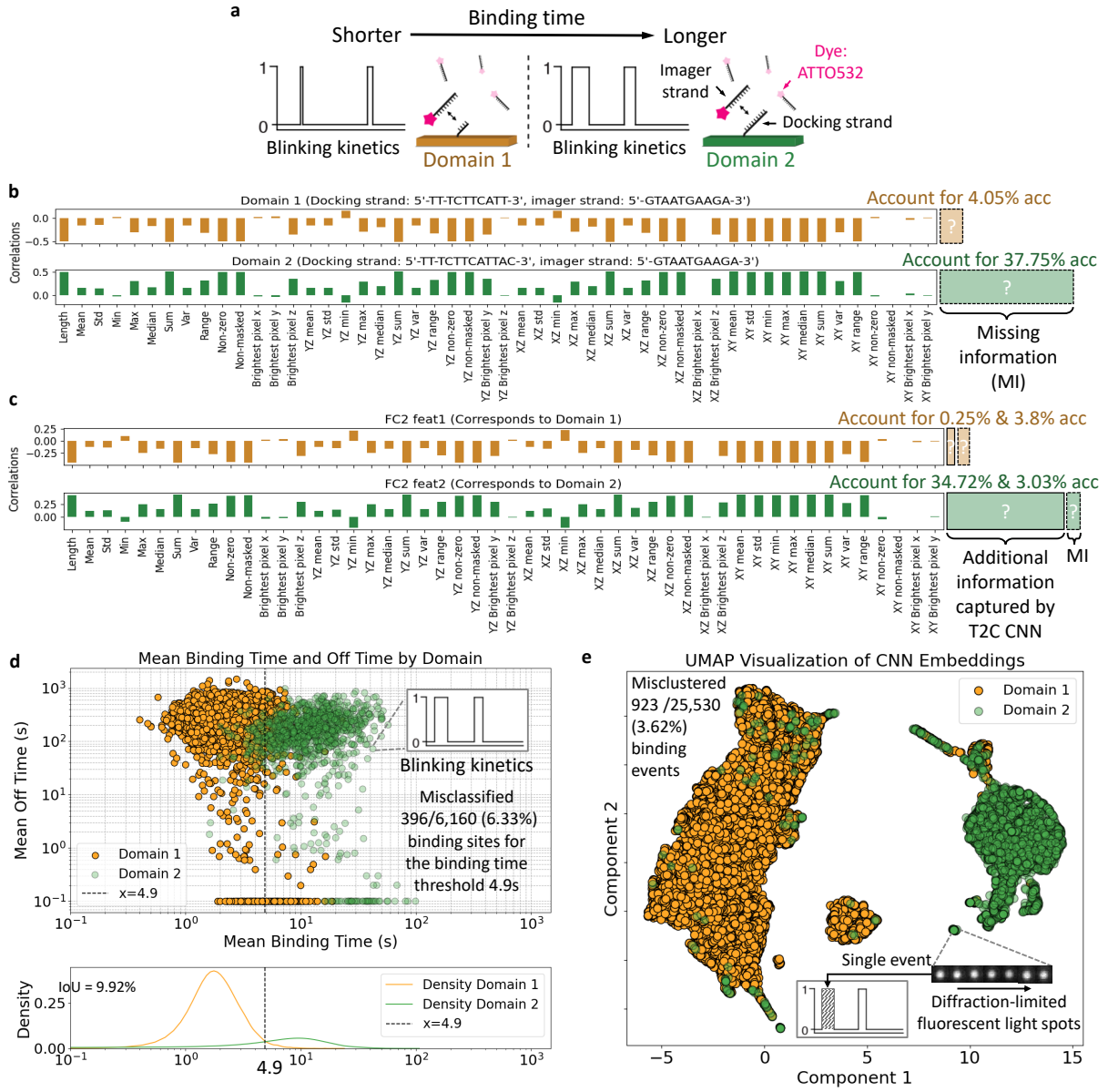


Figure 12: Diffraction-limited fluorescent spots hold binding-type information that can be utilized to more accurately differentiate between single binding events. **a**, Two types of DNA binding domains used to collect diffraction-limited fluorescent spots. Blinking kinetics refers to the binary signals of binding on and off. (Revised from [27]) **b**, The correlation between domain labels and some common statistics, 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 mean on-time is displayed. Based on multiple binding events, 6.33% of binding sites are misclassified. **e**, UMAP dimensionality reduction scatter 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.

2) In the [Supplementary Fig. 2](#), on Page 13,

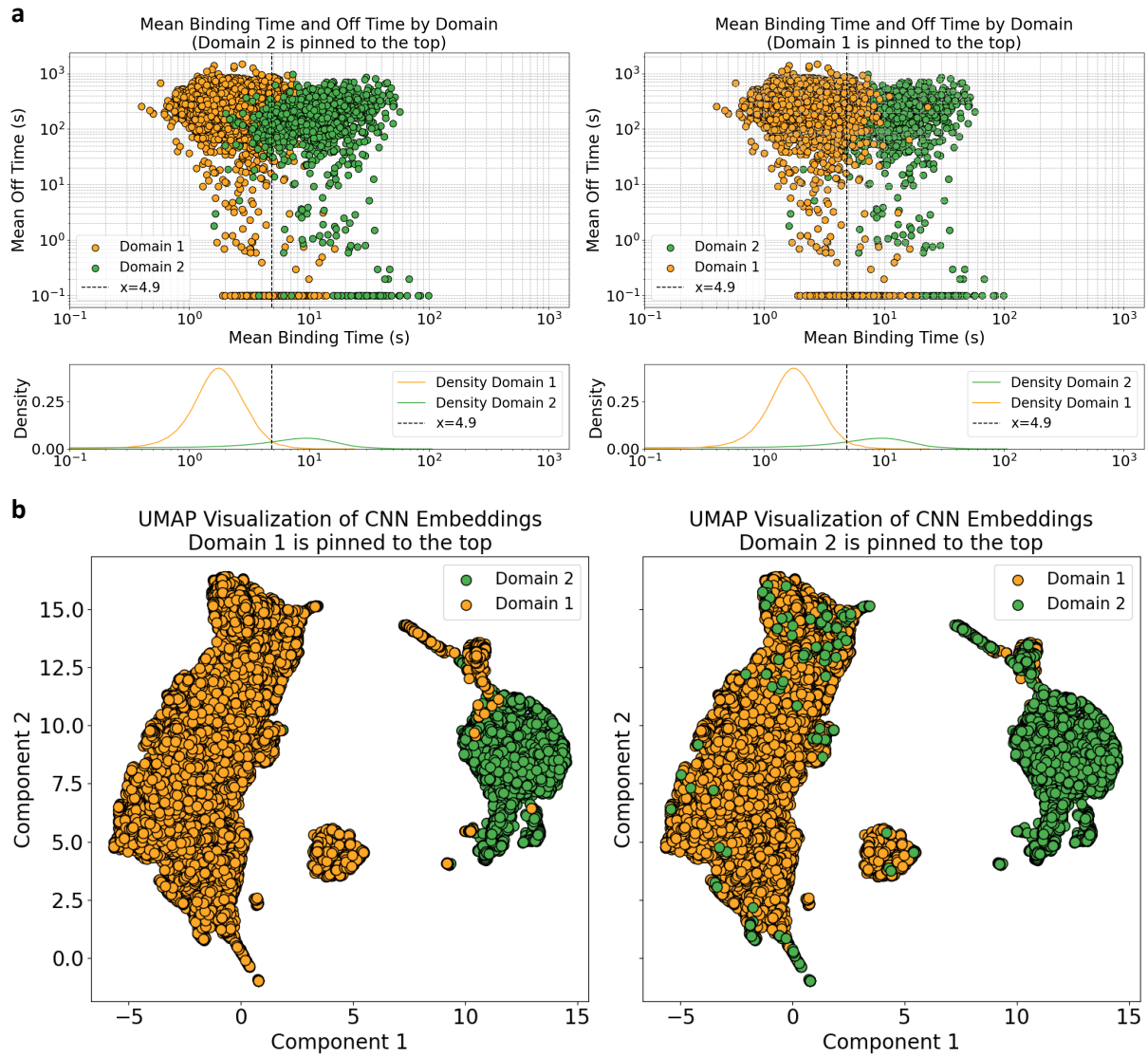


Figure 13: **Binding kinetics and CNN embedding visualization.** Scatter plots show the **a**, mean binding time and off time and **b**, UMAP-reduced embeddings with different domain types plotted last to ensure visibility. Each subplot highlights one domain by rendering it above the other, enabling clearer inspection of the distribution overlap.

3) In the Main Text Fig. 6 (previously Fig. 3), on Page 11,

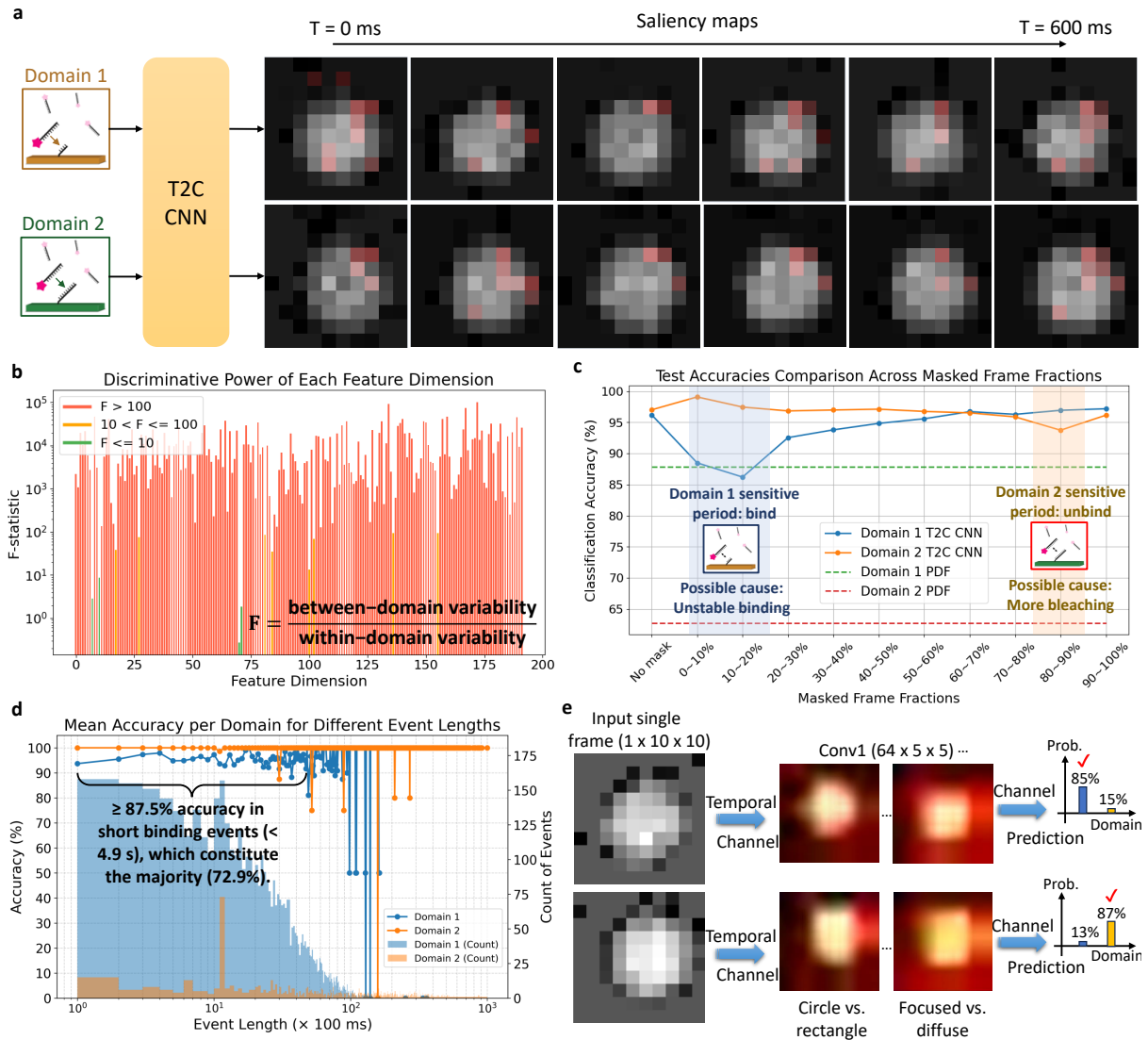


Figure 14: Interpretations of the temporal-to-context convolutional neural network (T2C CNN). **a**, T2C CNN saliency maps for diffraction-limited fluorescent spots in different DNA binding domains. The saliency map visually highlights the regions within each frame that T2C CNN considers most critical 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. **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).

4) In the Interpretations of the T2C CNN, on Page 11,

...

Finally, we evaluated T2C CNN on DNA binding events of varying durations. As shown in Fig. 14d, T2C CNN achieves $\geq 87.5\%$ accuracy for events shorter than 4.9 seconds (49 frames), which constitute 72.9% of all 25,530 events. The probability density function in Fig. 12e 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.

...

2.5 The Reviewer comments that

Fig. 3e presents the most important point in the manuscript. This shows the difference of the diffraction limited image of Domain 1 and 2. The authors listed several factors that may affect the diffraction limited image (such as intensity, conformational difference, spectral shift, etc) in Discussion, but to this reviewer, all those factors (except for spectral shift, which is believed to be too small to affect the image) will results in the difference in brightness. The authors also mentioned that (also in the correlation in Fig. 1c and d), domain 1 is less bright than domain 2. So, it looks like the different characteristics of the diffraction limited spots may result from the different intensity and associated statistical fluctuations of the two species.

What is the brightness (i.e., integrated intensity of the spots) distribution of the two domains? Are they well separated?

If the above is true, the method in this manuscript is discriminating between two species with different brightness, possibly better than discriminating simply by an intensity threshold, rather than discriminating different binding events. Probably, this is why it takes only 5 seconds, the timescale much shorter than the binding kinetics.

Response: Thank you for this thoughtful comment.

To address your question, we have plotted the distributions of both integrated intensity and frame-average intensity in the newly added Fig. 5. The distribution of integrated intensity closely mirrors that of binding time, which is expected since the per-frame intensity remains similar between the two domains (8 nt and 10 nt). As such, differentiating between the two domains based solely on integrated intensity would be as challenging as relying only on binding time. Furthermore, the frame-average intensity distributions are largely overlapping between the two test domains, further limiting their discriminability.

Importantly, the discrimination achieved by the proposed method is based on the full fluorescence spot videos, which capture richer temporal dynamics beyond static features. These include, but are not limited to, inter-frame intensity correlations shown in Fig. 3 and other correlated statistics presented in Fig. 2c. The revised manuscript and supplementary material now include these additional analyses and the corresponding figure.

(continued on next page)

Revised parts:

1) In the **Main Text Fig. 5**, on Page 10,

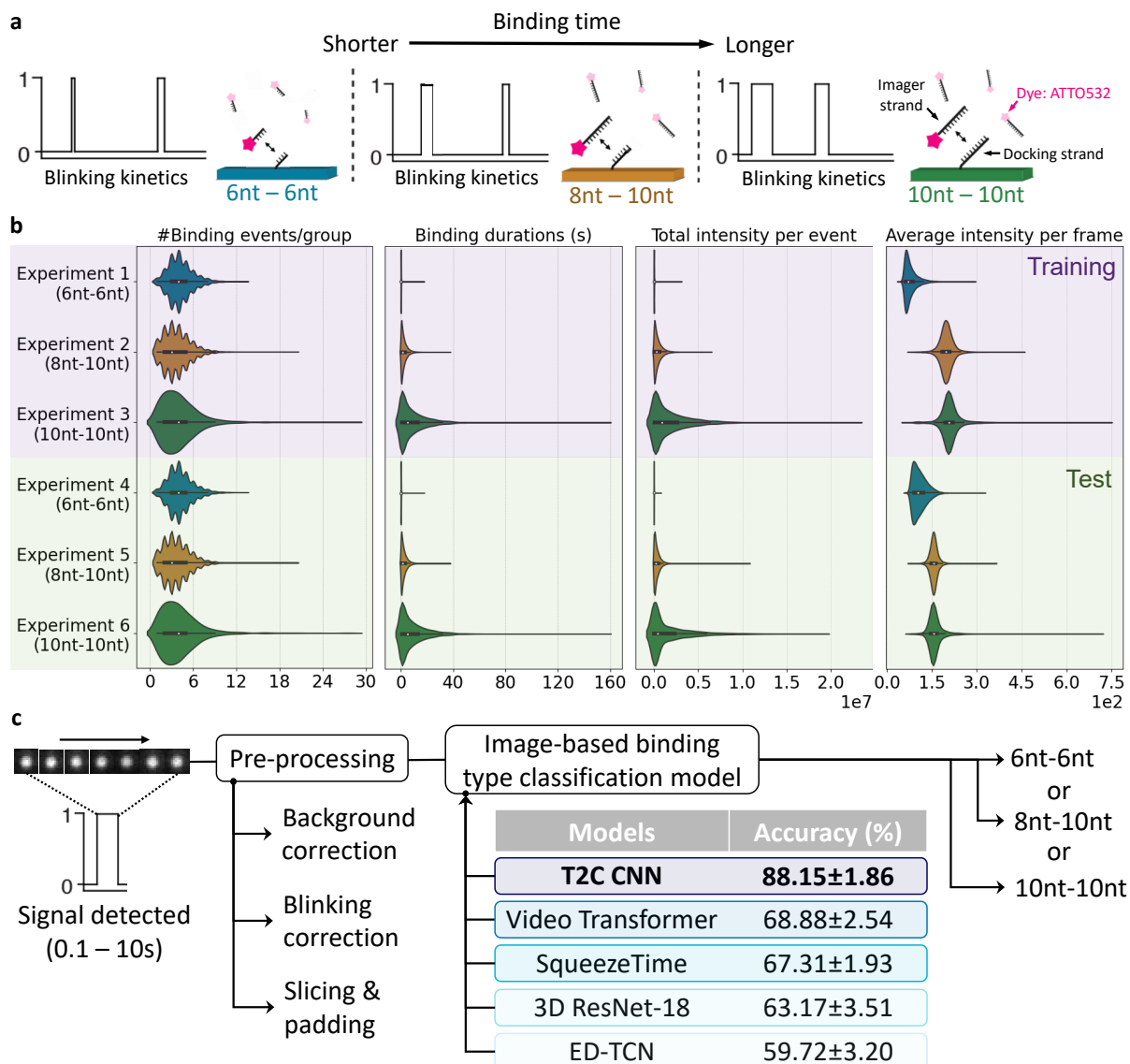


Figure 15: **Multi-Class Data, Workflow, and Results.** **a)** Illustration of the three binding types (6nt-6nt, 8nt-10nt, and 10nt-10nt) used in the multi-class experiments. **b)** Data distributions from the experiments. Notably, the 6nt R1 imager exhibited lower average frame intensity (approximately 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.

2) In the Supplementary Note 3: Multi-class experiments, on Page 3,

Data analysis. Main Text Fig. 5(b) presents the distribution of experimental data, corrected for background and blinking, for three binding types across six experiments. We controlled the concentration so that each binding site exhibited a similar number of binding events within the same 20k frames (10 fps), with most sites showing 1–9 binding events. The detected binding sites for the three binding types were 19,457, 6,073, and 10,888, respectively. Their binding durations increased sequentially, primarily ranging from 0.1 to 10 seconds.

To further investigate the discriminability of binding events based on brightness, we plotted both the integrated intensity and the mean frame intensity distributions in Main Text Fig. S4b. The integrated intensity distribution primarily reflects the underlying distribution of binding durations, as expected given the relatively stable per-frame intensity within each binding event. Consequently, CNN models need to distinguish binding types by learning from the full fluorescence video dynamics rather than relying on static intensity metrics. This involves leveraging temporal features such as inter-frame correlations (Main Text Fig. 2b) and other temporally correlated statistics (Main Text Fig. 2c). For the 6nt R1 imager, the average frame intensity was lower (approximately 50 and 100 for training and testing sets, respectively), likely due to short binding durations that sometimes result in incomplete fluorescence accumulation within a single frame.

(continued on next page)

3) In the Main Text Fig. 2, on Page 6,

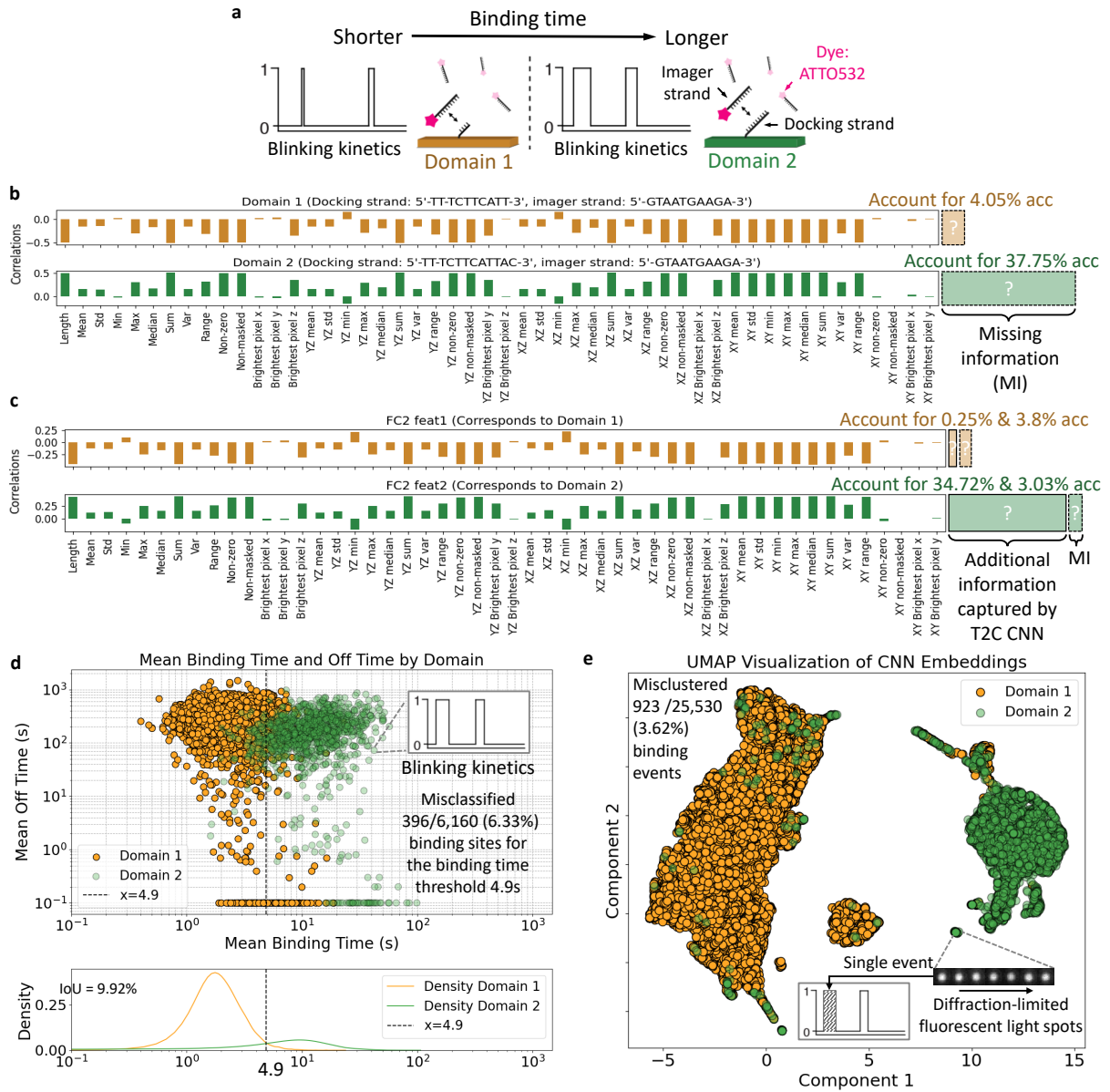


Figure 16: Diffraction-limited fluorescent spots hold binding-type information that can be utilized to more accurately differentiate between single binding events. **a**, Two types of DNA binding domains used to collect diffraction-limited fluorescent spots. Blinking kinetics refers to the binary signals of binding on and off. (Revised from [27]) **b**, The correlation between domain labels and some common statistics, 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 mean on-time is displayed. Based on multiple binding events, 6.33% of binding sites are misclassified. **e**, UMAP dimensionality reduction scatter 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.

4) In the Main Text Fig. 3, on Page 7,

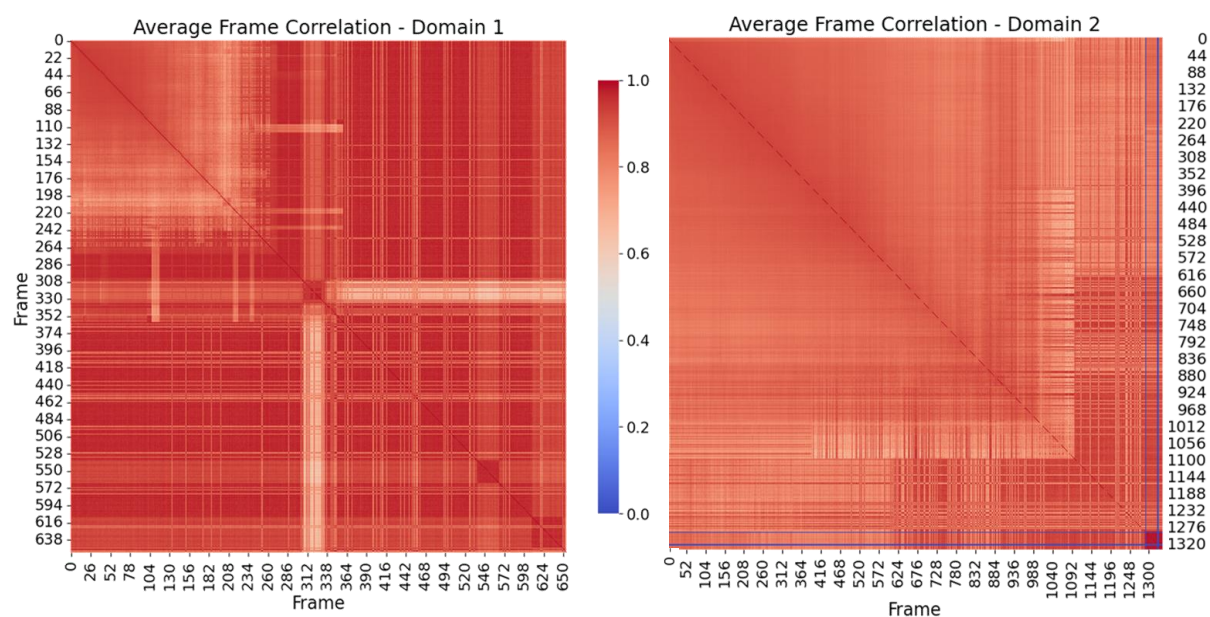


Figure 17: Correlations between diffraction-limited fluorescent spot frames during binding events for Domain 1 and Domain 2. Detailed definitions, analyses, and visualizations are provided in Supplementary Note 1, S1.2, and Supplementary Figure S1, respectively.

2.6 The Reviewer comments that

Related with #5, from the practical cellular imaging perspective, there are other issues such as more than two binding kinetics, distribution (intensity) in each species, overlapping spots of different species. It will be great to have some discussion on these points in terms of the strength and limitation of the method.

Response: Thank you for your insightful comment.

We have expanded the Discussion section in the revised manuscript to reflect on the strengths and limitations of the proposed method in the context of more complex cellular imaging scenarios. In the revision, we included a three-class experiment to examine the model's ability to handle multiple binding kinetics. In the revised discussion, we now address potential challenges that may arise in even more complex settings, such as the presence of additional binding kinetics, variations in fluorescence intensity within each species, and overlapping spots from different species. These points are discussed in terms of both methodological limitations and potential directions for improvement. The added discussion is provided below.

Added parts of the manuscript:

1) In the Discussion, on Page 20,

...

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.

2.7 The Reviewer comments that

Fig. fonts are too small. Although Nat Commun is an online journal, it is really difficult to read. Letters are small even with 200% magnification.

Response: Thank you for pointing this out.

In response to your comment, we have carefully revised the figures to increase the font sizes throughout, aiming to enhance readability on screen. We hope that the updated figures now meet the expected clarity standards and would appreciate your feedback on their legibility.

3 Response to the Reviewer#3's comments.

We would like to thank Reviewer#3 for the suggestion to improve the quality of the manuscript. The following changes have been made to address Reviewer#3's comments, where all the changes are highlighted in blue in the manuscript and the supplementary material.

3.1 Response to the Reviewer#3's major comments.

3.1.1 The Reviewer comments that

Raw images and time courses of imager strand binding and dissociation should be included in Fig. 1. Processed images in Fig. 3 do not provide sufficient clarity about raw data interpretation.

Response: Thank you for this helpful suggestion.

In response, we have added a figure (Fig. 1) and an accompanying section in the revised manuscript to present representative raw images and time courses of imager strand binding and dissociation. This addition aims to provide greater clarity on the nature of the raw data and help the reader better understand the basis of the analysis before introducing the processed results. The added content is provided below.

Added parts of the manuscript:

1) In the **Overview of raw data and methods**, on Page 3,

DNA-PAINT experiments generate raw data as videos, with each frame containing multiple diffraction-limited fluorescent spots (Fig. 18a). To minimize spot overlap, experimental parameters such as concentration and pH are carefully adjusted to control spot density. Traditional analysis methods, such as Picasso [2], 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. 18b).

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 [2] 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. 18b, 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 pre-defined DNA binding types. The following sections present the experimental results, discussion, and methodology on these data and methods.

2) In the [Main Text Fig. 1](#), on Page 4,

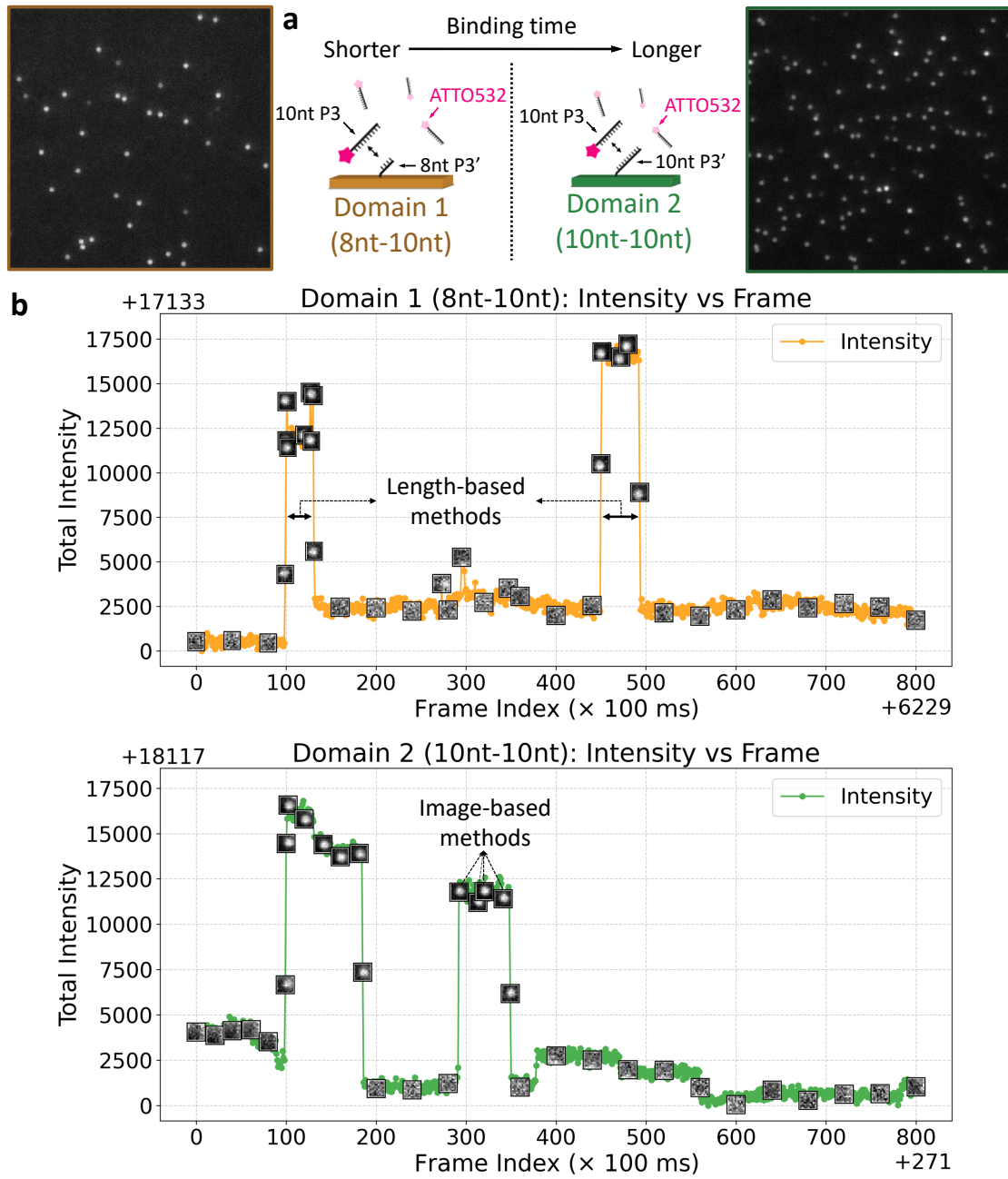


Figure 18: **An example of raw images and time courses from the DNA-PAINT experiments used in this study.** **a)** Schematic of the two DNA binding domains used to collect diffraction-limited fluorescent spots (adapted from [27]). Both domains share the same dye (“ATTO532”) and strand (“10nt P3”), but 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, while both show similar intensity jumps. The length-based method (classic) classifies molecules based on the total binding event durations per molecule, whereas the proposed image-based method classifies individual binding events based on their diffraction-limited fluorescent spots.

3.1.2 The Reviewer comments that

Clear definitions of binding time and off time are necessary. Confidence in assigning signals to Domain 1 or Domain 2 using T2C CNN requires further explanation.

Response: Thank you for your valuable comment.

In the revised manuscript, we have provided clearer definitions for binding time and off time to improve clarity. Additionally, we have elaborated on how confidence scores are derived from the T2C CNN model to support the assignment of signals to Domain 1 or Domain 2. The added content is listed below.

Added parts:

1) In the **Overview of raw data and methods**, on Page 4,

...

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 [2] 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.

...

2) In the **Test on T2C CNN**, on Page 25,

...

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) [33, 34], 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.

...

(For reference) In the **Supplementary Note 9: T2C CNN blocks**, on Page 10,

...

Feature classification block: This block consists of two fully connected (FC) layers: the first refines the features processed by the preceding blocks through fusion, transformation, and abstraction, while the second performs the final classification. They are defined as follows:

$$l_7 = \text{ReLU} \circ \text{BN} \circ \text{FC}(l_6, \omega_7, b_7, \gamma_7, \beta_7). \quad (16)$$

Then, the activated hidden features are passed to the final fully connected layer for classification:

$$\hat{y}_{\text{slice}} = \text{Softmax} \circ \text{FC}(l_7, \omega_8, b_8), \quad (17)$$

$$\hat{y} = \frac{1}{N_{\text{slice}}} \sum_{\text{slice}} \hat{y}_{\text{slice}}. \quad (18)$$

The output $\hat{y}_{\text{slice}}$ from the final layer represents the predicted probabilities for each temporal slice, normalized using the Softmax function [26]. For multiple temporal slices from the same binding event, we take the average ($\hat{y}$) of their predictions.

3.1.3 The Reviewer comments that

Saliency map descriptions as well as the convolution from $1 \times 10 \times 10$ into $64 \times 5 \times 5$ need more detailed explanations.

Response: Thank you for your helpful comment.

To improve clarity, we have revised the caption of Fig. 6 to include more detailed explanations of both the saliency map and the convolution operation, including the transformation from a $1 \times 10 \times 10$ input to a $64 \times 5 \times 5$ output. The updated Fig. 6 (on Page 11) is shown below.

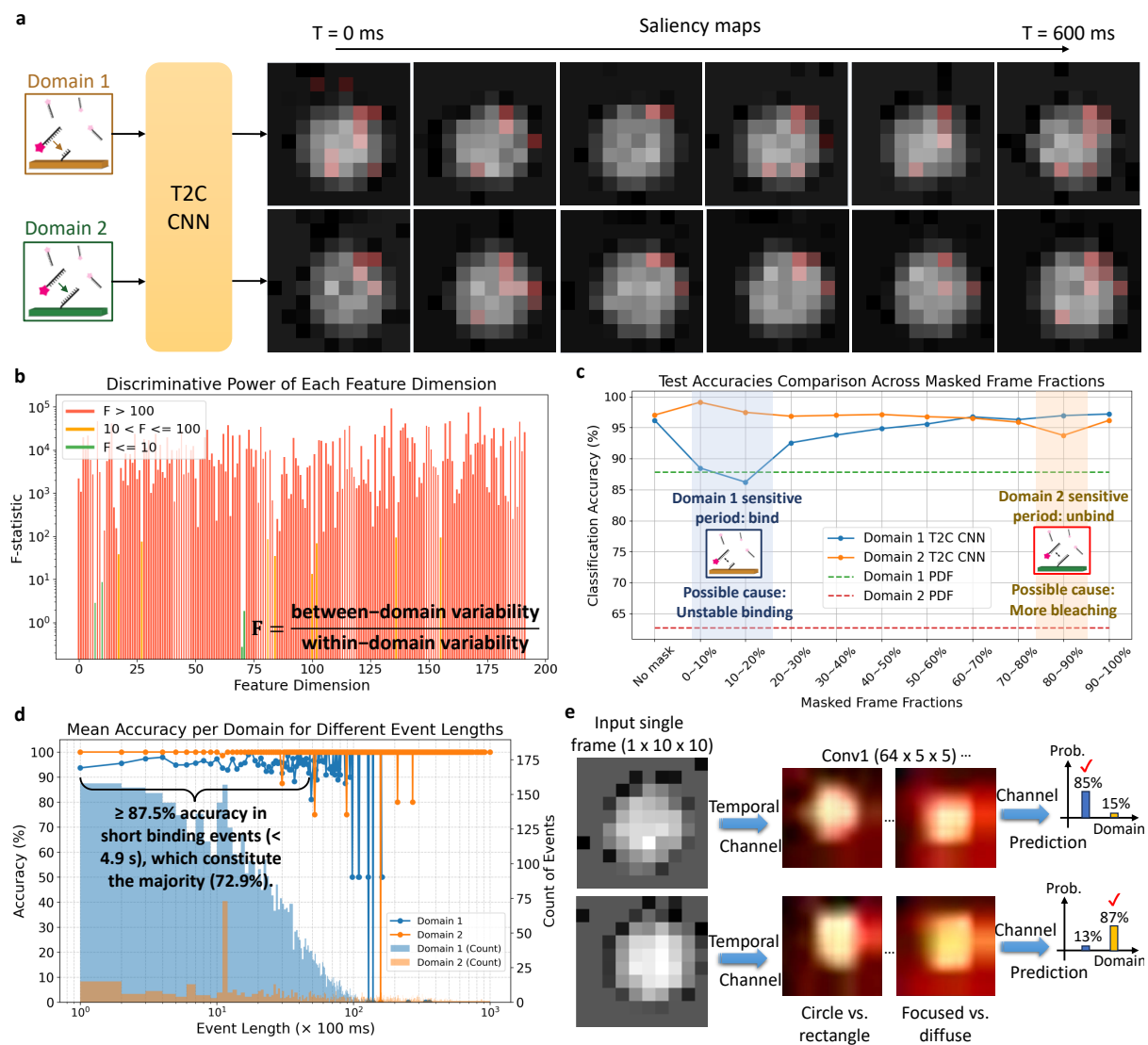


Figure 19: Interpretations of the temporal-to-channel convolutional neural network (T2C CNN). **a**, T2C CNN saliency maps for diffraction-limited fluorescent spots in different DNA binding domains. The saliency map visually highlights the regions within each frame that T2C CNN considers most critical 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. **e**, Example T2C CNN feature heatmaps and classification probabilities (Prob.) for single-frame diffraction-limited fluorescence 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 $(1 \times 10 \times 10)$ input images into 64 feature maps with spatial dimensions of (5×5) .

3.1.4 The Reviewer comments that

The validation of the proposed T2C CNN method is required, such as sensitivity analysis of key parameters is missing, choice of hyperparameters needs justification, and robustness testing across different experimental conditions is required.

Response: Thank you for raising this important point.

In the revised manuscript, we have included a comprehensive ablation study to assess the contribution of each component of the proposed T2C CNN model. We further analyzed the sensitivity of key hyperparameters, including network depth, width, architectural configurations, and input slice length. In addition, we evaluated the robustness of the model across varying experimental conditions to assess its generalizability. To complement these analyses, we also performed statistical significance testing on model performance to ensure that observed differences were not due to chance. Supplementary Table 1 has been added to report hyperparameter evaluations for the baseline models. The corresponding results and discussions have been added to the revised manuscript, as detailed below.

Added sections:

1) In the Robustness test of image-based binding-type classification models, on Page 12,

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 4 and Fig. 20 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 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%.

(continued on next page)

2) In the **Main Text Fig. 7**, on **Page 13**,

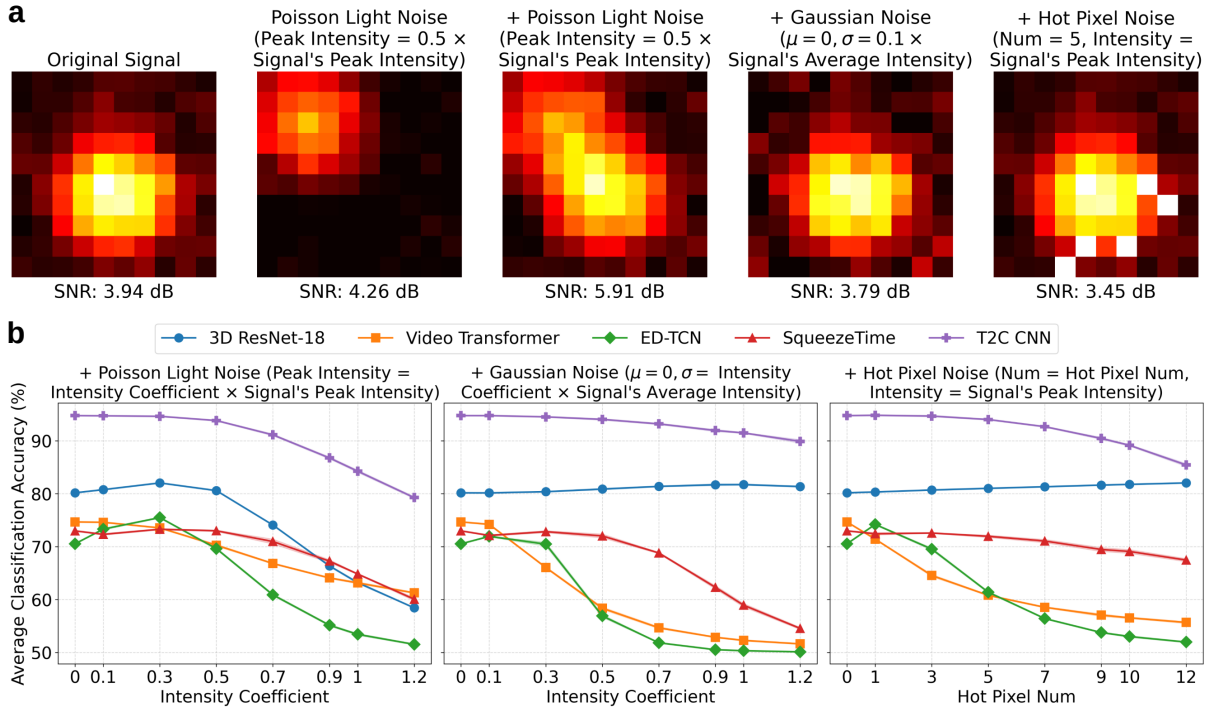


Figure 20: **Robustness test results.** (a) Examples of generated noise added to the original signal, including Poisson light noise with Gaussian decay, Gaussian noise, and hot pixel noise. The symbols μ and σ denote the mean and standard deviation (std.), respectively. (b) Average classification accuracies 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.

3) In the **Supplementary Note 4: Robustness test of image-based binding-type classification models**, on **Page 4**,

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. This section assesses 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.

Robustness test settings. (a) **Light interference with Poisson distributions:** Nonspecific molecular binding events generate short, weak fluorescence signals, which can overlap with specific signals [35] and disrupt accurate classification. Considering the discrete nature of photon detection [36] and a radial intensity drop-off based on physical optics [37], we simulate interference signals using Poisson distributions with Gaussian decay. As shown in the first three images of Main Text Fig. 7(a), an artificial light spot is overlaid on the original signal, producing the third image with interference. Despite increasing the signal-to-noise ratio (SNR) (defined as the ratio of the image mean to its standard deviation [38]), the added spot is not fluorescence from molecular binding. To mimic experimental fluorescence images, we set the interference spot radius and Gaussian decay standard deviation to 5 and 1.8, respectively. The interference spot's peak intensity is varied from 0.1 to 1.2 times the original signal's peak intensity in increments of 0.1 or 0.2, with positions randomized near the original signal to test the robustness of the classification models.

(b) Gaussian noises: Fluorescence images often suffer from Gaussian noise caused by detector electronic noise, such as thermal and readout noise [39]. We simulate Gaussian noise with a mean of 0 and variances ranging from 0.1 to 1.2 times the original signal’s mean intensity. The fourth image in Main Text Fig. 7(a) displays an example with Gaussian noise variance set to 0.1 times the original signal’s mean intensity, reducing the SNR from 3.94 to 3.79.

(c) Hot pixel noises: Prolonged camera usage can cause sensor defects, producing pixels with abnormally high intensity (hot pixels) [40]. To evaluate the robustness of classification models against hot pixel noise, we randomly introduce 1, 3, 5, 7, 9, 10, and 12 hot pixels with intensities equal to the original signal’s peak intensity. The last image in Main Text Fig. 7(a) shows an example with 5 hot pixels, which mask parts of the original signal.

Robustness test results. Main Text Fig. 7(b) records the average classification accuracy of the tested models (3D ResNet-18 [4], Video Transformer [5], ED-TCN [6], SqueezeTime [1], and the proposed T2C CNN). For SqueezeTime and 3D ResNet-18, we use the default 48-frame input, which achieves comparable accuracy to the 512-frame input but with significantly higher efficiency.

(a) Light interference with Poisson distributions. Poisson noise has the most significant impact on classification performance, as it can result from specific or nonspecific binding events occurring in close proximity, causing signal overlap. When the interference spot’s peak intensity exceeds 0.5 times that of the signal, all models show a marked performance decline. Increasing the interference intensity from 0.5 to 0.7 times results in accuracy drops of 2.78% to 12.54%. SqueezeTime (2.78%) and T2C CNN (2.88%) exhibit the greatest resilience. Both models employ temporal-to-channel transformations, highlighting the effectiveness of this operation in enhancing robustness against interference spots. One possible reason is that the temporal-to-channel transformation captures long-term fluorescence variations within channels, helping the model focus on the primary signal.

(b) Gaussian noises. The models display polarized performance under Gaussian noise. Video Transformer and ED-TCN, which extract image features before temporal operations, are the most vulnerable. In contrast, 3D ResNet-18 and T2C CNN, which feature residual connections and cross-block concatenation, show the strongest resistance. Residual connections and cross-block concatenation pass low-level features from earlier layers directly to deeper layers, helping retain structural and spatiotemporal information and making it easier to denoise the signal. This is because Gaussian noise typically affects fine-grained details rather than structural patterns. Conversely, models that first extract image features propagate Gaussian noise into later temporal operations, exacerbating its effect.

(c) Hot pixel noises. 3D ResNet-18 and SqueezeTime, both with large model parameters (>100 MB), show the highest resilience to hot pixel noise. Smaller models (T2C CNN, ED-TCN, and Video Transformer, with parameters from 1.65 to 7.80 MB) are more susceptible, with T2C CNN being the least affected among them. Large models can memorize fine-grained patterns, making them less sensitive to small-scale disruptions such as hot pixels. In contrast, smaller models, with limited parameter capacity, are more susceptible to pixel losses. Despite having the fewest parameters, T2C CNN stands out among small models due to its high parameter efficiency (long temporal kernel and equal-length stride), which enhances its resistance to hot pixel noise.

Based on the above robustness test results, we recommend maintaining the interference spot intensity below half of the signal spot intensity and using cameras with an effective pixel ratio above 93% (tolerating up to 7 hot pixels per 100 pixels) when applying the T2C CNN.

4) In the Hyperparameter analysis of the T2C CNN, on Page 13,

We systematically evaluated the impact of network depth, width, and input slice length on T2C CNN’s performance (Fig. 21 and Supplementary Note 5). 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% $\rightarrow$ 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. 21, we also included the baseline models with their optimal depths and widths, which were evaluated on randomly split training and validation datasets (Supplementary Table 1).

5) In the Supplementary Table 1, on Page 18,

Table 8: Hyperparameter evaluation summary for baseline models.

	Depth (width=128) ¹	1	2	3*	4	5	6	7
Video Transformer	Accuracy ²	66.74	70.16	71.91	69.53	71.61	71.07	69.81
	Width (depth=3)	16	32	48	64	100	128	256
	Accuracy	71.01	67.23	70.21	71.18	71.78	71.91	70.48
SqueezeTime	Depth (width=512 or 2048) ³	10	18	34	50	101	152	200
	Accuracy	64.20	67.07	67.61	71.27	67.87	69.01	66.28
	Width (depth=50)	64	128	204	612	2048	3072	4096
	Accuracy	56.01	47.52	55.65	70.38	71.27	67.15	63.48
ED-TCN	Depth (width=64)	2	4	6	8	10	12	14
	Accuracy	70.78	73.56	73.34	57.37	72.58	67.25	71.28
	Width (depth=4)	16	32	48	64	100	128	256
	Accuracy	68.51	70.12	68.24	73.56	69.24	70.38	69.45
3D ResNet	Depth (width=512 or 2048)	10	18	34	50	101	152	200
	Accuracy	74.80	79.73	78.35	76.47	73.40	79.15	73.25
	Width (depth=18)	32	64	128	256	512	1024	2048
	Accuracy	75.02	78.20	75.02	73.68	79.73	76.58	77.61

¹Model depths and widths were evaluated near their default configurations. ²Class-wise accuracies for single-event classification were recorded. ³For 3D ResNet and the ResNet-based method (SqueezeTime), a width of 512 was used when the depth was below 50; otherwise, a width of 2048 was used. * Bolded values indicate the optimal hyperparameters identified in the trials.

(continued on next page)

6) In the **Main Text Fig. 8**, on **Page 14**,

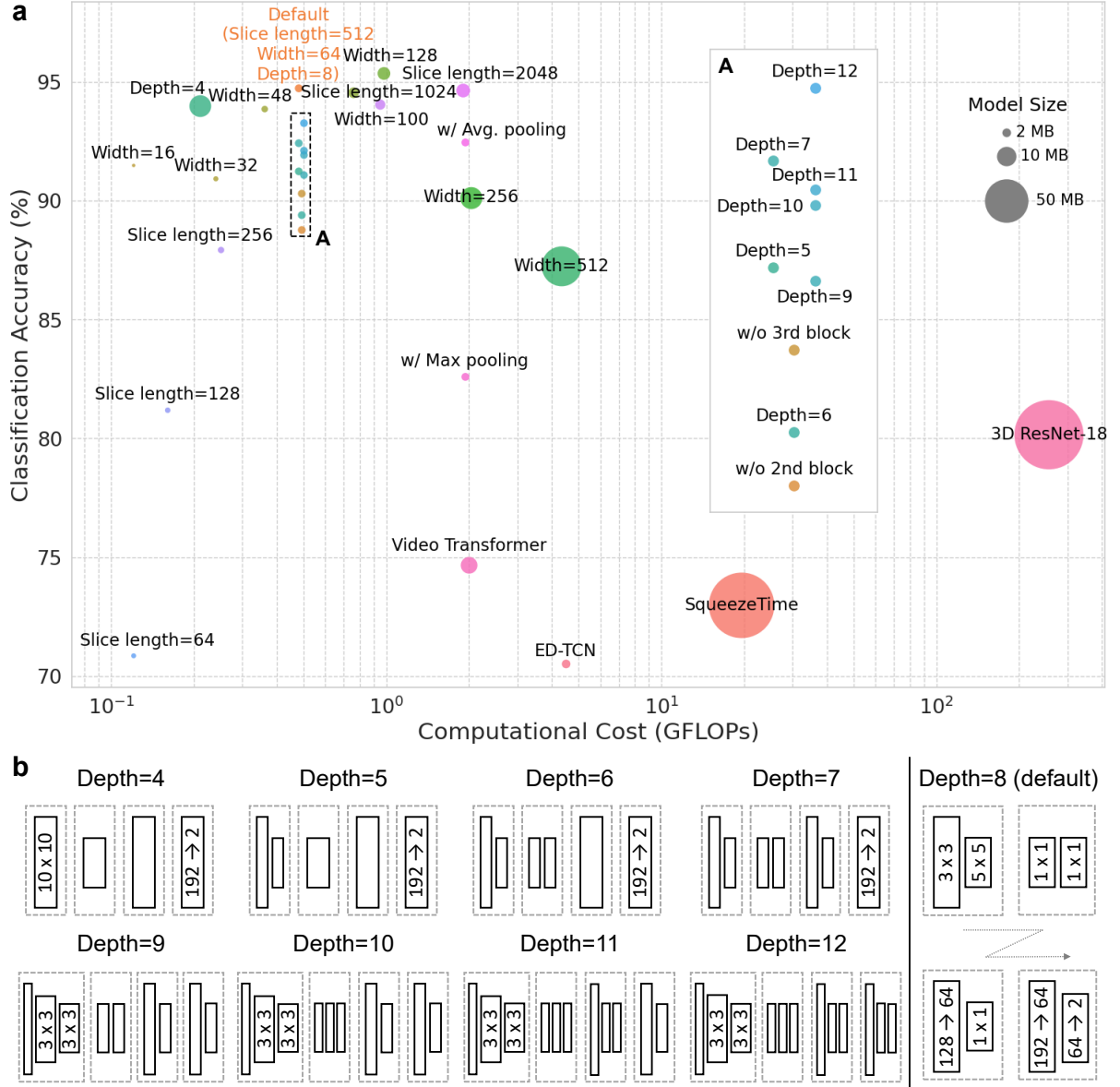


Figure 21: **a) Comparison of classification accuracy, computational cost, and model sizes for variants of the proposed T2C CNN.** Baseline methods, including 3D ResNet-18 [4], Video Transformer [5], ED-TCN [6] and SqueezeTime [1], are also evaluated for comparison (detailed comparison discussed in the 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. **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.

7) In the **Supplementary Note 5: Hyperparameter analysis of the T2C CNN**, on **Page 5**,

To evaluate the impact of hyperparameter choices in T2C CNN, we analyzed its key hyperparameters, including network depth, width, and slice length, as shown in Main Text Fig. 8.

Experimental settings. (a) Depth: Depth is a critical attribute of deep neural networks, significantly affecting performance [30]. The proposed T2C CNN, as shown in Main Text Fig. 8, comprises eight hidden layers (convolutional or fully connected), evenly distributed across four blocks. To investigate the minimum required depth, each block was sequentially assigned 1 to 3 layers, increasing the total depth from 4 to 12 while maintaining the four-block structure. For the first block, which condenses spatial-temporal features, the convolutional kernel size was adjusted for configurations with one or three layers. Main Text Fig. 8b illustrates the corresponding architectures.

(b) Width: Like depth, network width (the number of convolutional kernels or hidden units) plays a critical role in CNN performance [31]. The default T2C CNN configuration uses 64 kernels or hidden units per layer. To explore the effect of width, we tested variations ranging from 16 to 512, including commonly used widths such as 16, 32, 48, 100, 128, 256, and 512.

(c) Slice length: The slice length, which determines the interval at which fluorescence frames are input into T2C CNN, is critical for capturing long-term temporal patterns in fluorescence videos. As shown in Main Text Fig. 2e, most binding events occur within 50 seconds (500 frames), with the longest binding duration not exceeding 2000 frames before averaging. Therefore, the default slice length of 512 frames was varied by halving or doubling to explore alternatives of 64, 128, 256, 1024, and 2048 frames.

Experimental results (Main Text Fig. 8a). (a) Depth: Among the tested depths ranging from 4 to 12 layers, the top-performing configurations were depths of 8 (94.76%), 4 (94.00%), and 12 (93.27%), all multiples of 4. These results indicate that performance is optimal when each block in T2C CNN contains an equal number of hidden layers, with two hidden layers per block yielding the best results, followed by one layer per block. Notably, at a depth of 4, the first block of T2C CNN required a larger convolution kernel (increased from “ 3×3 ” to “ 10×10 ”) to condense all spatial-temporal features within a single hidden layer. This adjustment increased the parameter count approximately 11-fold, significantly contributing to the model size.

(b) Width: Among the tested widths of 16, 32, 48, 100, 128, 256, and 512, a width of 128 slightly outperformed the default 64 (94.76% $\rightarrow$ 95.38%). However, this improvement came with a 2.62-fold increase in model size (1.65 MB $\rightarrow$ 4.32 MB) and a 2.04-fold increase in computational cost (0.48 GFLOPs $\rightarrow$ 0.98 GFLOPs). From an efficiency perspective, the default width of 64 strikes a good balance. If computational resources and latency are not constraints, a width of 128 is also recommended. Widths smaller than 64 or greater than 128 consistently underperformed compared to the default, with performance generally declining as width became too wide.

Interestingly, even with an unusually small width of 16, T2C CNN maintained performance above 90%. This resilience can be attributed to the unique block design of T2C CNN, as breaking the balance of hidden layers within blocks or removing blocks led to performance degradation, even with an optimal width of 64. Remarkably, this suggests that when a fluorescence video spanning 512 frames is condensed into 16 channel dimensions, most discriminative information is retained. This highlights the substantial redundancy in fluorescence videos, much of which stems from identical optical properties of the same fluorescence type. The key discriminative signals arise from subtle changes in the fluorescence microenvironment caused by binding to different targets. These subtle variations are condensed into a small set of discriminative features by T2C CNN, which is fundamental to its functionality.

This also explains why performance significantly declined at widths of 256 and 512. The increased number of hidden nodes likely led to overfitting on redundant information, impairing generalization to test data from different experiments.

(c) Slice length: As the slice length input to T2C CNN increases up to 512 frames, classification accuracy improves significantly. This is because the longer slice length enables T2C CNN to capture a more complete binding event, aiding in determining the binding type. However, beyond

512 frames, the performance plateaus, as only a small fraction of binding events extend beyond this duration. In practical applications, we recommend setting the slice length to encompass the majority of binding durations (e.g., the maximum binding time excluding outliers) to achieve optimal efficiency.

Comparison with baselines. This section compares the proposed T2C CNN with two closely related baseline methods: the classical temporal convolutional network ED-TCN [6] and SqueezeTime [1], which compresses the temporal dimension into the channel dimension. For a fair comparison, we used the same 512-frame input length for SqueezeTime as T2C CNN (its default input is a 16-frame and 48-frame video clip). This adjustment significantly increased SqueezeTime’s model size (28.7 MB $\rightarrow$ 113.3 MB) and computational cost (5.5 GFLOPs $\rightarrow$ 19.56 GFLOPs). SqueezeTime achieved classification accuracies of 71.72% with the default 16-frame input and 72.98% with the 512-frame input on fluorescence videos, indicating that increasing the number of input frames yields only marginal improvement. This is likely because the repeated transformations between channel and temporal dimensions, along with weighted convolutions in SqueezeTime, risk losing discriminative spatiotemporal details, regardless of the input length.

For ED-TCN, we tested both its default 64-frame input and a 512-frame input matching T2C CNN. Its classification accuracies on fluorescence videos were 70.53% and 58.53%, respectively, demonstrating its unsuitability for long fluorescence video inputs. A key limitation is that ED-TCN’s fragmented frame-wise spatial features hinder the capture of long-term spatiotemporal patterns in fluorescence videos. The top-performing ED-TCN and SqueezeTime models are placed in Main Text Fig. 8a.

In contrast, the proposed T2C CNN is tailored for fluorescence video analysis. Its shallow cross-layer connections retain multi-scale spatiotemporal features, while long-term spatial convolutions are employed to capture extended spatial frequency patterns. Additionally, the integration of shallow cross-connected blocks with a pooling-free design facilitates efficient spatiotemporal information fusion. All spatiotemporal features are compressed into the channel dimension and jointly processed within the T2C layer. The individual and combined contributions of these three components are systematically evaluated in the Ablation Study section. For reference, 3D ResNet-18 and Video Transformer are included in Main Text Fig. 8a. These two models are widely adopted in computer vision for video classification and generally outperform lighter architectures such as SqueezeTime and ED-TCN due to their strong representational capacity. However, as discussed in the “Motivation” of the Main Text Method section, they are still suboptimal for fluorescence video classification, which demands a domain-specific architectural design.

8) In the Ablation study of T2C CNN architecture, on Page 15,

We dissected T2C CNN’s architecture to assess the contributions of its key blocks and design components (Supplementary Note 6, Supplementary Table 1). Removing either the second (Hidden Transformation) or third (Multi-Scale Feature Fusion) block significantly degraded performance (94.76% $\rightarrow$ 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.

9) In the Statistical significance analysis of model performance, on Page 15,

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

10) In the Supplementary Table 2, on Page 19,

Table 9: Ablation study of T2C CNN components¹.

Settings	Long-term spatial convolutions	Skip concatenations	No pooling	Class-wise accuracy (%) ²		Overall single-event accuracy (%) ³
				Single-event	Single-molecule	
Baseline	–	–	–	72.01±2.76	79.87±2.70	72.95±2.33
Individual component	–	–	✓	70.73±1.69	77.60±2.66	74.66±3.59
	–	✓	–	72.58±3.36	78.13±3.96	76.22±3.09
	✓	–	–	84.14±3.80	88.86±4.07	89.07±2.44
Combined components	–	✓	✓	72.51±3.82	79.86±3.19	74.96±4.56
	✓	✓	–	86.33±5.16	91.11±4.21	90.70±3.65
	✓	–	✓	87.34±1.89	92.02±2.38	91.59±1.75
All components	✓	✓	✓	94.76±0.47	96.99±0.46	96.78±0.50

¹Each component is removed or included to assess its individual and joint contribution to classification accuracy. ²Class-wise accuracy at the single-event and single-molecule levels. Single-molecule predictions derived by confidence-based voting over associated single-event predictions. ³Overall single-event accuracy is calculated across all events without class-wise averaging. Reported values are mean ± standard deviation over five cross-validation trials.

(continued on next page)

11) In the Supplementary Table 3, on Page 20,

Table 10: 5-fold single-event class-wise accuracies of all tested methods and their average ranks.

Methods	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average rank ($\uparrow$) ¹
T2C CNN (ABC ²)	94.76	94.28	94.36	95.60	94.78	1
T2C CNN (AC)	90.91	86.68	86.85	87.01	85.26	2.8
PDF (All event)	83.88	84.42	84.26	84.42	84.42	4
T2C CNN (AB)	79.35	89.59	86.33	93.97	82.40	4.4
PDF (2 event)	83.80	83.90	83.92	83.90	83.90	5
T2C CNN (A)	81.84	80.70	88.11	89.38	80.69	5.2
PDF (3 event)	82.09	82.18	82.00	82.18	82.18	6.6
3D ResNet-18	80.17	72.58	82.79	80.45	83.72	7.4
PDF (1 event)	75.27	75.27	75.16	75.27	75.27	9.6
T2C CNN (B)	70.22	72.64	69.69	78.98	71.38	12
Video Transformer	74.67	70.85	69.76	72.44	70.85	12.4
T2C CNN (BC)	68.66	68.01	73.26	78.40	74.22	12.4
T2C CNN None	71.61	70.00	69.55	71.61	77.28	12.6
T2C CNN (C)	70.49	69.75	70.35	73.97	69.09	13.2
SqueezeTime	71.72	69.18	71.27	69.55	66.34	13.6
ED-TCN	62.19	63.63	76.01	67.43	71.84	13.8

¹Ranks were computed within each fold using Friedman ranking and then averaged across folds. These ranks were used in the Friedman test and post-hoc analyses reported in the main text. All 16 methods and their variants were sorted in ascending order based on their average ranks. ²The letters “A”, “B”, and “C” following “T2C CNN” indicate the inclusion of long-term spatial convolutions (A), skip concatenations (B), and no-pooling (C) components, respectively. Combinations of these letters represent variants that include the corresponding components. The suffix “None” denotes the variant in which none of these three components are included.

12) In the Supplementary Note 6: Ablation study of the T2C CNN, on Page 7,

To investigate the impact of T2C CNN’s architectural components, we conducted an ablation study focusing on its core building blocks and components. Three key components (**long-term spatial convolutions**, **skip concatenations** and **pooling-free**) were selected for a detailed evaluation of their individual and combined contributions, as presented in Main Text Fig. 8 and Table 9.

Experimental settings. (a) Blocks: T2C CNN consists of four blocks, with two being fundamental: the Temporal-to-Channel (T2C) block for condensing spatial-temporal features and the Feature Classification block. To assess the importance of the remaining blocks—the second (Hidden Transformation) and third (Multi-Scale Feature Fusion)—we individually removed each and evaluated the impact.

(b) Components: The proposed T2C CNN is specifically designed for fluorescence video analysis. It employs **long-term spatial convolutions** to capture extended spatial frequency patterns. The combination of **skip concatenations** and a **pooling-free** design enables effective spatiotemporal information fusion across the network. For evaluation, we compare the optimal configuration (slice length of 512, all skip concatenations enabled, no pooling) with a baseline configuration (slice length of 64, no skip concatenation, average pooling), corresponding respectively to the presence and absence of the three proposed components.

Experimental results (Main Text Fig. 8a). (a) Blocks: Removing the second and third blocks from the T2C CNN architecture reduced classification accuracy from 94.76% to 88.80% and 90.33%, respectively. This significant performance drop indicates that both blocks are integral

components of T2C CNN. Among the two, the second block (Hidden Transformation) plays a more critical role. It consists of two 1×1 convolutional layers that transform the condensed global spatiotemporal features into more discriminative representations, facilitating the differentiation of binding types. Notably, when the third block (Multi-Scale Feature Fusion) operates without the transformed features from the second block, its fusion-based features result in lower accuracy (88.80%) compared to using only the transformed features from the second block (90.33%). This demonstrates the importance of the second block in preparing features for effective multi-scale fusion.

(b) Components: Table 9 presents a systematic evaluation of the key components of T2C CNN—long-term spatial convolutions, skip concatenations, and the pooling-free strategy—on both single-event and single-molecule classification accuracy, with all results averaged over five cross-validation trials.

Among individual components, long-term spatial convolutions contribute the most significant performance gain, boosting class-wise single-event and single-molecule accuracies to 84.14% and 88.86%, respectively, compared to the baseline’s 72.01% and 79.87%. This highlights their pivotal role in capturing extended spatial-frequency patterns essential for distinguishing binding types in fluorescence videos. In contrast, skip concatenations (72.58% / 78.13%) and no-pooling (70.73% / 77.60%) result in slightly lower class-wise accuracies than the baseline, suggesting that their benefit emerges primarily when combined with other architectural elements to support multi-scale feature fusion and preserve spatiotemporal information. The overall single-event accuracy follows a similar trend, with long-term convolutions providing the most substantial standalone gain (89.07% vs. 72.95% baseline).

When components are combined, their effects are complementary. Integrating long-term convolutions with skip concatenations increases class-wise single-event accuracy to 86.33%, while pairing long-term convolutions with no-pooling pushes it to 87.34%. In contrast, combining skip concatenations with no-pooling without long-term convolutions yields only a modest 72.51%, reinforcing the central importance of long-term spatial convolutions.

The full T2C CNN, incorporating all three components, achieves the best performance: 94.76% class-wise single-event accuracy, 96.99% single-molecule accuracy, and 96.78% overall accuracy. Notably, it also exhibits the lowest standard deviations across all metrics ($\pm 0.47\%$, $\pm 0.46\%$, and $\pm 0.50\%$, respectively), indicating stable and consistent predictions across cross-validation folds. This demonstrates that the proposed architectural combination of existing elements enables not only optimal spatiotemporal information fusion and feature preservation, but also robust generalization, tailored specifically for high-precision fluorescence video analysis.

3.1.5 The Reviewer comments that

As common single-molecule experiments often involve two or three colors, it would benefit to explore the method's performance with multiple binding types. For example, performance with multiple (>2) binding types, applicability to unknown binding patterns, and validation from biochemical systems could demonstrate applicability.

Response: Thank you for this insightful suggestion.

In the revised manuscript, we have extended the experiments to include three distinct binding types to demonstrate the applicability of the proposed method beyond binary classification. We have also included preliminary applications to cellular data to illustrate its potential in more complex biological settings. These additions are intended to provide a broader perspective on the method's utility in realistic multi-color single-molecule imaging scenarios. The added contents are listed below.

Added contents:

1) In the [Multi-class experiments](#), on Page 9,

...

We extended the proposed approach to three-class classification by introducing an additional 6nt–6nt R1 strand binding type (Fig. 22a). As shown in Fig. 22c, 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. 2.

(continued on next page)

2) In the Main Text Fig. 5, on Page 10,

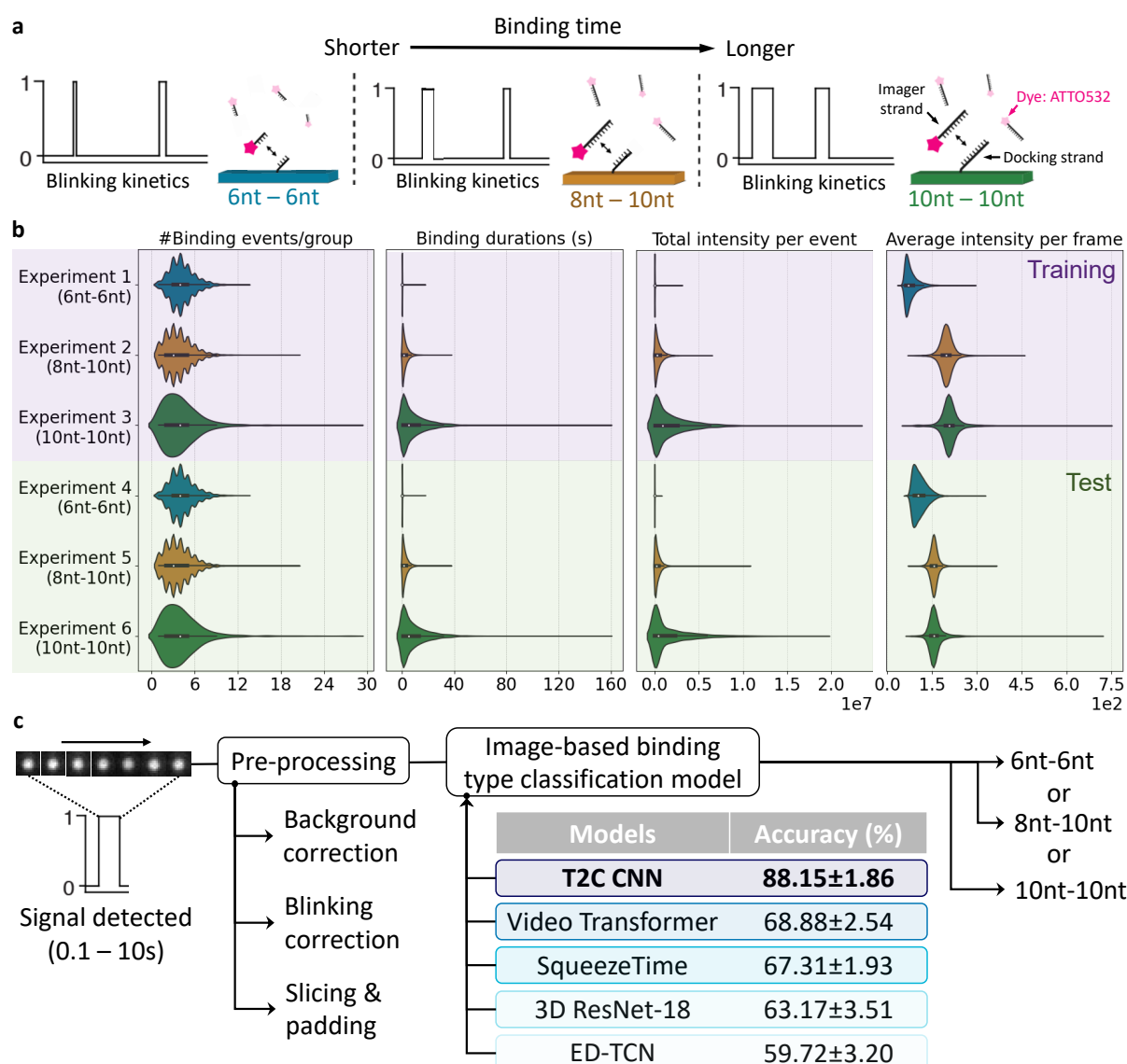


Figure 22: Multi-Class Data, Workflow, and Results. **a)** Illustration of the three binding types (6nt-6nt, 8nt-10nt, and 10nt-10nt) used in the multi-class experiments. **b)** Data distributions from the experiments. Notably, the 6nt R1 imager exhibited lower average frame intensity (approximately 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.

3) In the Supplementary Note 3: Multi-class experiments, on Page 3,

Settings. Single-molecule experiments often involve two or three target types. To evaluate the method's performance in multi-class settings, we introduce an additional 6nt-6nt R1 strand binding type. As shown in Main Text Fig. 5(a), we shorten the target strand from 8nt to 6nt. However, due to the short binding duration of a 6nt strand with the original 10nt P3 imager, we instead use a complementary 6nt R1 strand as both the docking and imager strand. Notably, the 6nt R1 imager carries the same ATTO532 dye to enable multiplexing.

Data analysis. Main Text Fig. 5(b) presents the distribution of experimental data, corrected for background and blinking, for three binding types across six experiments. We controlled the concentration so that each binding site exhibited a similar number of binding events within the same 20k frames (10 fps), with most sites showing 1–9 binding events. The detected binding sites for

the three binding types were 19,457, 6,073, and 10,888, respectively. Their binding durations increased sequentially, primarily ranging from 0.1 to 10 seconds.

To further investigate the discriminability of binding events based on brightness, we plotted both the integrated intensity and the mean frame intensity distributions in Main Text Fig. S4b. The integrated intensity distribution primarily reflects the underlying distribution of binding durations, as expected given the relatively stable per-frame intensity within each binding event. Consequently, CNN models need to distinguish binding types by learning from the full fluorescence video dynamics rather than relying on static intensity metrics. This involves leveraging temporal features such as inter-frame correlations (Main Text Fig. 2b) and other temporally correlated statistics (Main Text Fig. 2c). For the 6nt R1 imager, the average frame intensity was lower (approximately 50 and 100 for training and testing sets, respectively), likely due to short binding durations that sometimes result in incomplete fluorescence accumulation within a single frame.

Workflow. Main Text Fig. 5(b) illustrates the workflow for image-based multi-class classification of binding types. When a binding event is detected (typically lasting 0.1–10 seconds), its fluorescence signals undergo background and blinking correction to eliminate background intensity variations and transient disappearances. The processed signals are then sliced, padded to a uniform length, and input into image-based binding classification models, including 3D ResNet-18 [4], Video Transformer [5], ED-TCN [6], SqueezeTime [1], and the proposed T2C CNN.

Results. After training, T2C CNN achieved the highest test accuracy ($88.15 \pm 1.86\%$), followed by Video Transformer ($68.88 \pm 2.54\%$). These models classify each detected signal into one of three binding types (6nt-6nt, 8nt-10nt, or 10nt-10nt). Using these predictions, we differentiate targets labeled with distinct docking strands in the final super-resolution (SR) image, generating a three-color SR image, as shown in Fig. 23. T2C CNN consistently demonstrated superior classification accuracy across accumulated time points, with its advantage becoming more pronounced over longer durations. This makes it a powerful tool for accurate single-fluorophore, three-color SR imaging.

When multiple fluorophores are available, T2C CNN can first distinguish fluorophores frame by frame (will be discussed in Supplementary Note 8) and then leverage accurate single-fluorophore classification ability to multiply the number of identifiable targets. This significantly enhances both analysis capacity and speed.

4) In the **Supplementary Fig. 3**, on Page 14,

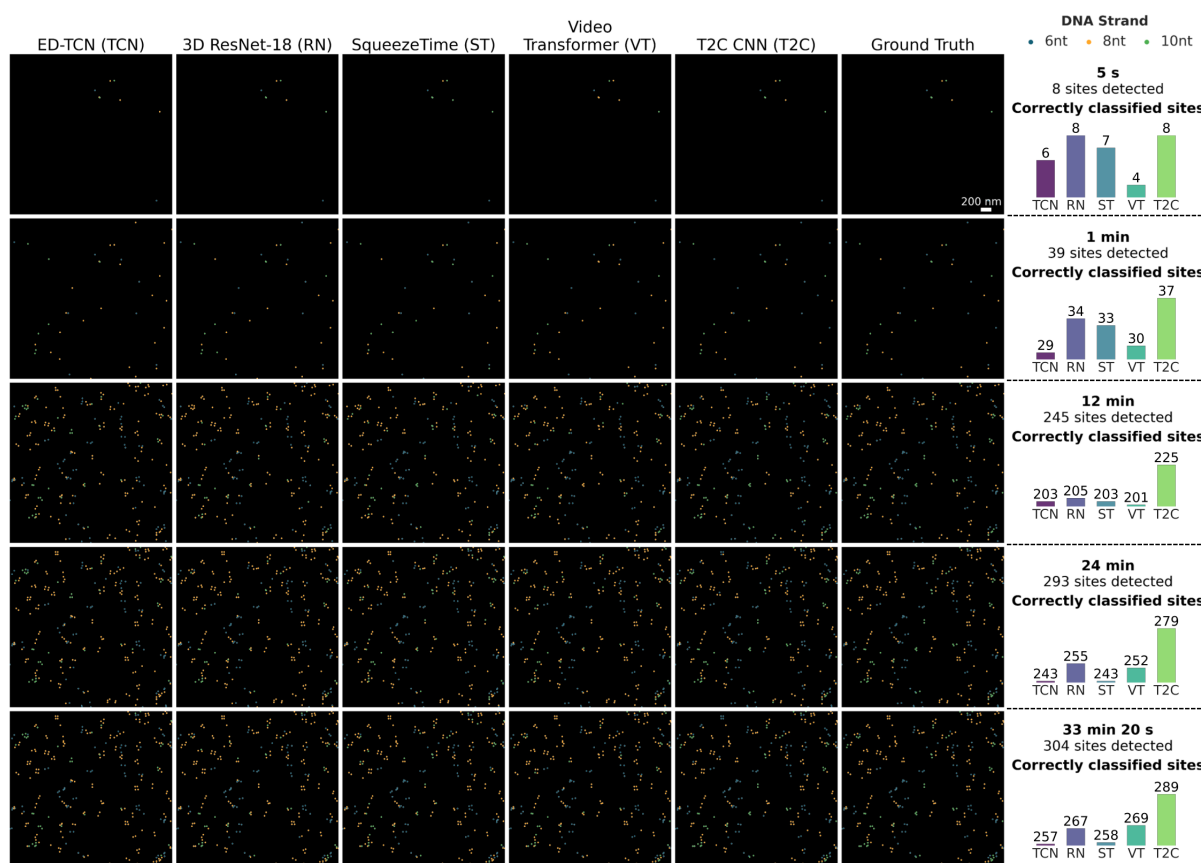


Figure 23: **Multi-Class Predictions.** Pseudo-colored super-resolution images reconstructed at 5 s, 1 min, 12 min, 24 min, and 33 min 20 s using image-based binding type classification models. The number of detected binding sites and correctly classified sites are shown on the right. Note that due to the higher concentration of 6nt and 8nt, they exhibited an increasing number of observed binding sites over time.

5) In the **HER2-targeted cell experiments**, on Page 14,

As illustrated in the central panel of Fig. 24a, 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 Supplementary Note 7.

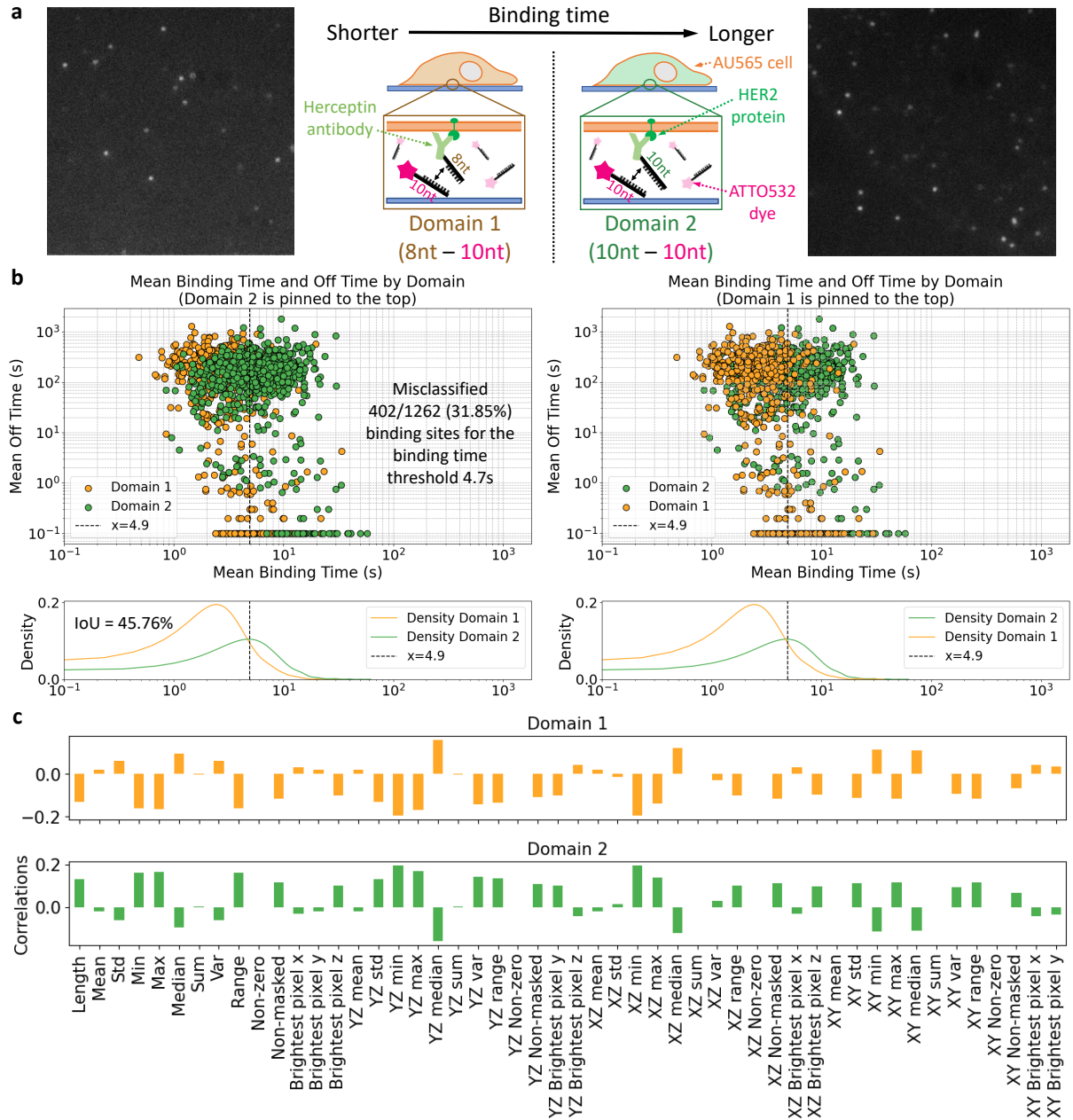


Figure 24: Cell experimental settings and data analysis. **a) Settings:** HER2-positive AU565 cells were cultured, fixed, and immunolabeled with custom antibody-oligo conjugates (8nt or 10nt P3' strands), followed by incubation with imager strands (10nt P3 strands) in salt-buffered PBS for fluorescence imaging. Example fluorescence images are shown on the corresponding sides of the two domains. **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. 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. 26c, 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.

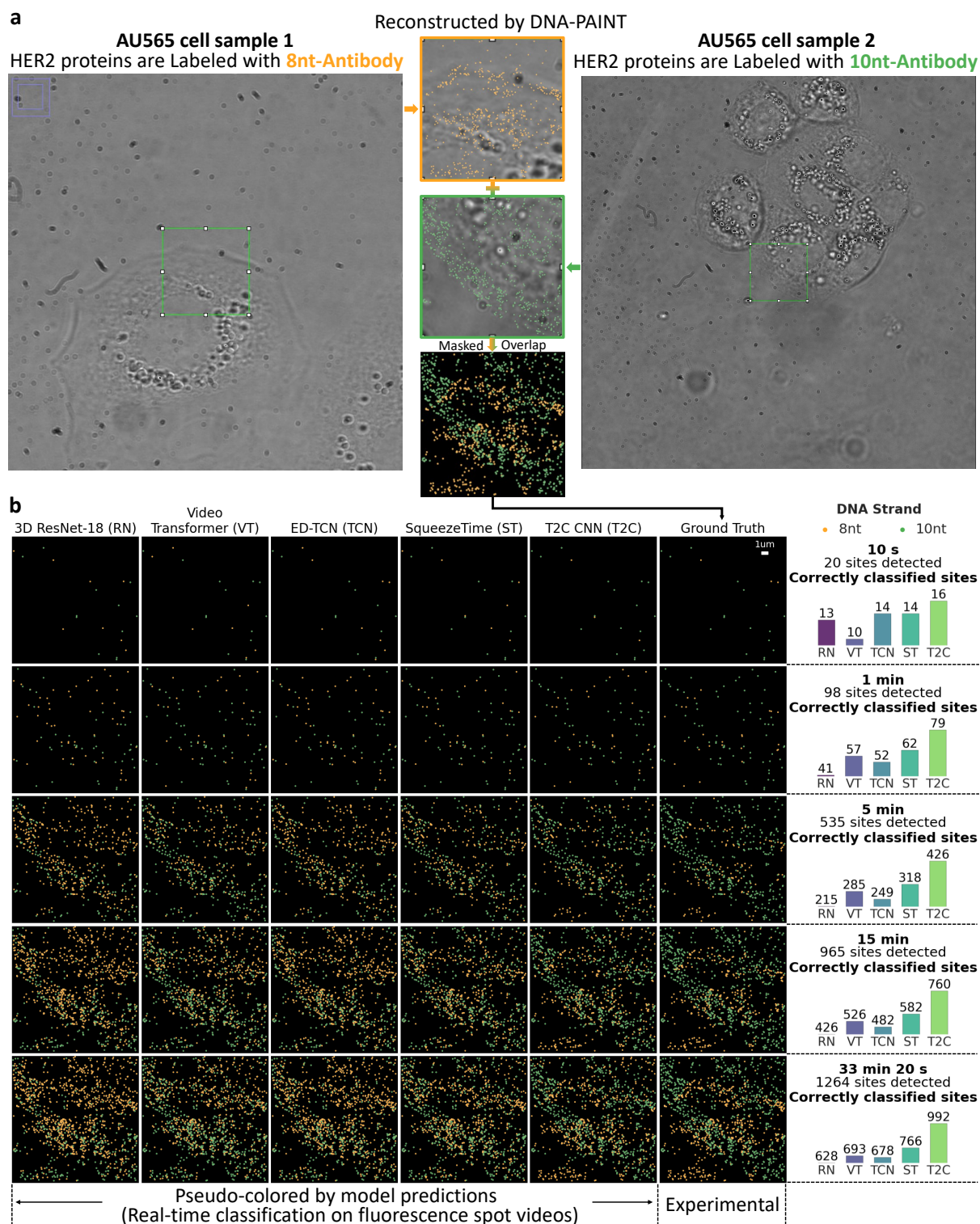


Figure 25: **Cell experimental results.** **a)** Bright-field and super-resolution microscopy images of two test HER2-positive AU565 cell samples, with HER2 proteins immunolabeled using 8nt and 10nt DNA strands. The DNA-PAINT technique [2] is used to reconstruct HER2 protein locations. Fluorescence videos from the two samples are aligned frame-by-frame to synthesize the multiplexed scenario, serving as ground truth. **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.

Fig. 24a 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 minutes 20 seconds (20,000 frames at 10 fps), super-resolution reconstruction (Fig. 25a, middle) reveals HER2 positions (orange/green dots).

Binding events at these sites were tracked to compute average on/off times. As shown in Fig. 24b, Domain 1 (8nt–10nt) shows shorter binding durations than Domain 2 (10nt–10nt), with similar off-times. Compared to the Origami setup (Fig. 26e), domain durations in cells overlap more—likely due to environmental factors such as membrane interactions or local field instabilities. Kernel density estimation (Fig. 26c, 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. Fig. 24c 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 $\approx \pm 0.5$ in Fig. 26c), underscoring the increased complexity of cellular classification.

HER2 site-localized videos (Fig. 25a) 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. 25b), 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.

6) In the Supplementary Note 7: Cell experiment preparation, on Page 8,

Preparation of Antibody-Oligo conjugate probes. The antibody-oligo conjugate probes were prepared using the method described in [41]. The 5'-amine-modified docking strands were conjugated with Herceptin monoclonal antibody (targeting HER2 protein on the cell membrane) (Trastuzumab, Roche, Graz Steiermark, Austria) using disuccinimidyl suberate (DSS) linker. The oligo (200 μ M 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 minutes 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 hour incubation at -20°C , the slurry was centrifuged at 24,100 g for 15 minutes 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 hours 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 minutes 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.

Cell culture and immunofixation. 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 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 CO₂ incubator at 37 °C and 5% (v/v) CO₂ concentration (Forma Steri-Cycle CO₂ incubator, Thermo Fisher Scientific) for 36 hours.

Cells were then fixed with 4% paraformaldehyde (PFA, EMS) in PBS for 15 minutes, followed by quenching with 1 mg/mL sodium borohydride for 7 minutes. Cells were washed three times (5 minutes per wash) with PBS, and incubated with 1% BSA and 0.05% Tween-20 in PBS for 2 hours 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 hour 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.

3.2 Response to the Reviewer#3's minor comments.

3.2.1 The Reviewer comments that

Define how “missing information” improves classification accuracy and explain the large differences between Domain 1 and Domain 2. Enhance labels for “additional information” and “missing information” in Fig. 1d for better clarity.

Response: We appreciate the Reviewer's insightful suggestions.

In the revised manuscript, we have provided a clearer definition of “missing information” in the context of the classification task and elaborated on how it contributes to improved accuracy by highlighting differences in domain-specific fluorescence dynamics. We also clarified the observed discrepancies between Domain 1 and Domain 2 and refined the labels in the current Fig. 2d (previously Fig. 1d) to enhance clarity and interpretability. The revised content is included in the updated figure and corresponding text below.

Revised parts:

1) In the Binding-type information in diffraction-limited fluorescent spots, on Page 7,

...

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. 26c 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 spatiotemporal 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.

...

2) In the Main Text Fig. 2, on Page 6,

(continued on next page)

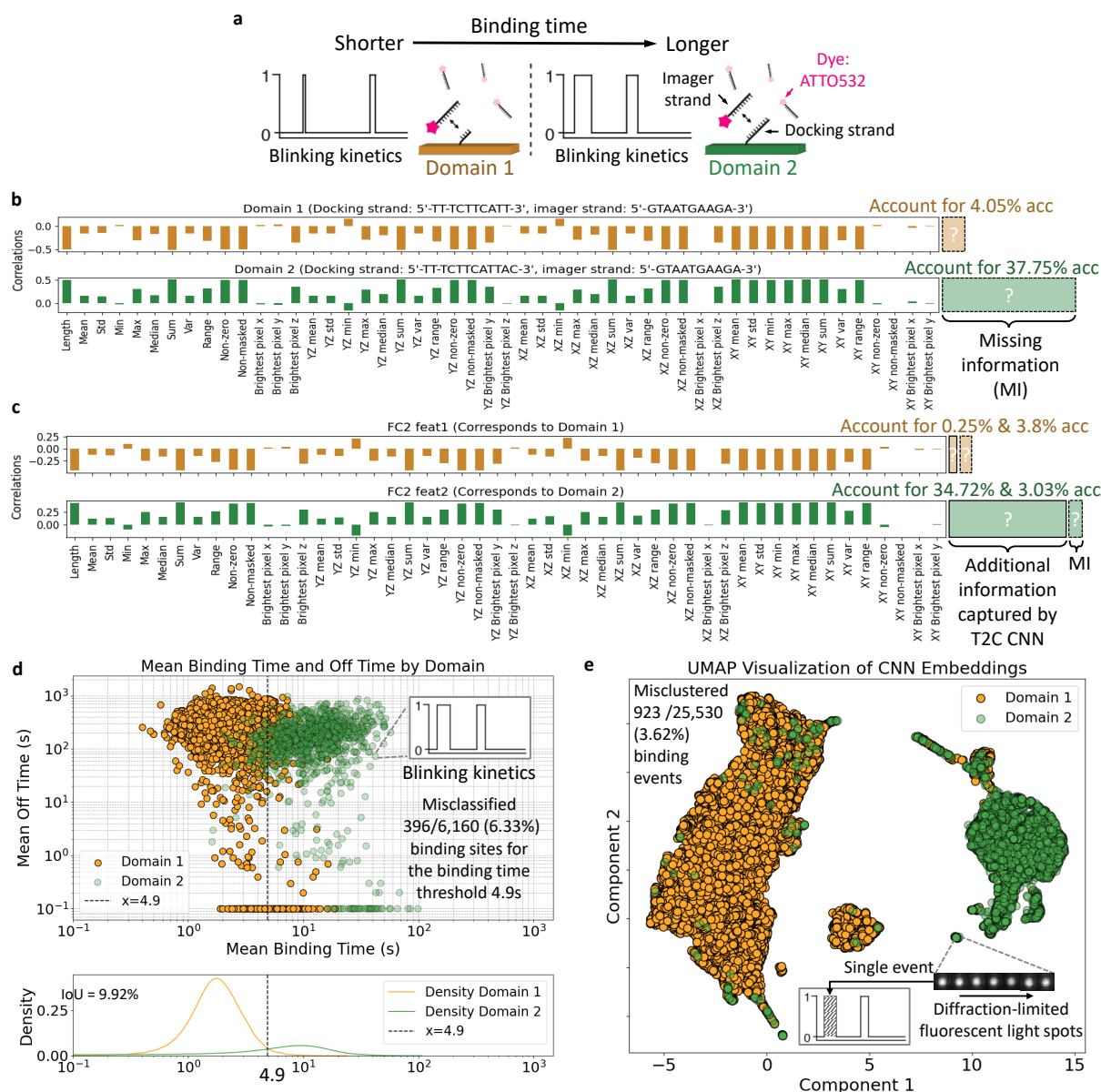


Figure 26: Diffraction-limited fluorescent spots hold binding-type information that can be utilized to more accurately differentiate between single binding events. **a**, Two types of DNA binding domains used to collect diffraction-limited fluorescent spots. Blinking kinetics refers to the binary signals of binding on and off. (Revised from [27]) **b**, The correlation between domain labels and some common statistics, 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 mean on-time is displayed. Based on multiple binding events, 6.33% of binding sites are misclassified. **e**, UMAP dimensionality reduction scatter 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.

3.2.2 The Reviewer comments that

The quantitative accuracy data in Fig. 2c could be displayed as a bar plot to better emphasize the superior performance of T2C CNN.

Response: We thank the Reviewer for this helpful suggestion.

In the revised version of Fig. 2c, we have updated the presentation to use bar plots, which more clearly highlight the comparative performance and better emphasize the strengths of the T2C CNN method.

Revised parts:

1) In the Fig. 4 (previously Fig. 2), on Page 8,

(continued on next page)

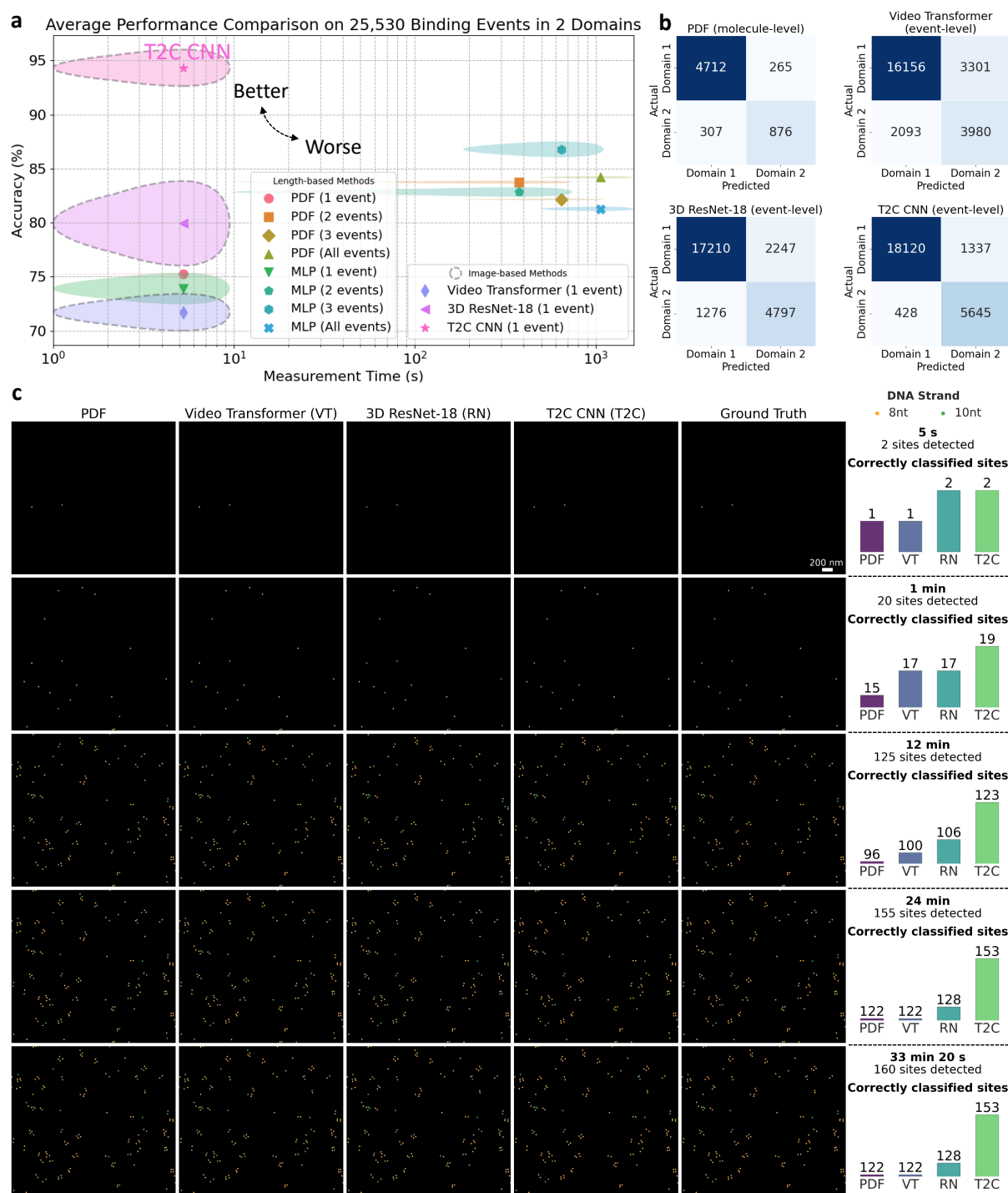


Figure 27: Cross-experiment results for discriminating binding types. **a**, Comparison of average performance on 25,530 binding events across the two domains. The width of the colored regions in the horizontal and vertical directions represents the standard deviation of measurement time and accuracy, 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.

3.2.3 The Reviewer comments that

The role of edges in diffraction patterns should be clarified, as indicated by the saliency map, as shown in Fig. 3a.

Response: We appreciate the Reviewer's observation.

In the revised manuscript, we have elaborated on the role of edges in fluorescence diffraction patterns, as highlighted by the saliency maps in current Fig. 6a (previously Fig. 3a). This clarification provides additional insight into how the model leverages spatiotemporal features at the spot boundaries to enhance classification performance.

Added parts of the manuscript:

1) In the Interpretations of the T2C CNN, on Page 10,

...

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. 28a). 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 [42]. 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 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.

...

3.2.4 The Reviewer comments that

In Fig. 3b, the definition of “groups” in the F-test is not clearly stated. It appears the authors divided the two domains into separate groups, but this should be explicitly clarified. Additionally, there is a typographical error in the figure description where “T-test” is written instead of “F-test.”

Response: Thank you for pointing this out.

In the revised caption for current Fig. 6b (previously Fig. 3b), we have clarified the definition of “groups” used in the F-test analysis and corrected the typographical error where “T-test” was mistakenly written instead of “F-test.”

(continued on next page)

Revised parts:

1) In the Fig. 6 (previously Fig. 3), on Page 11,

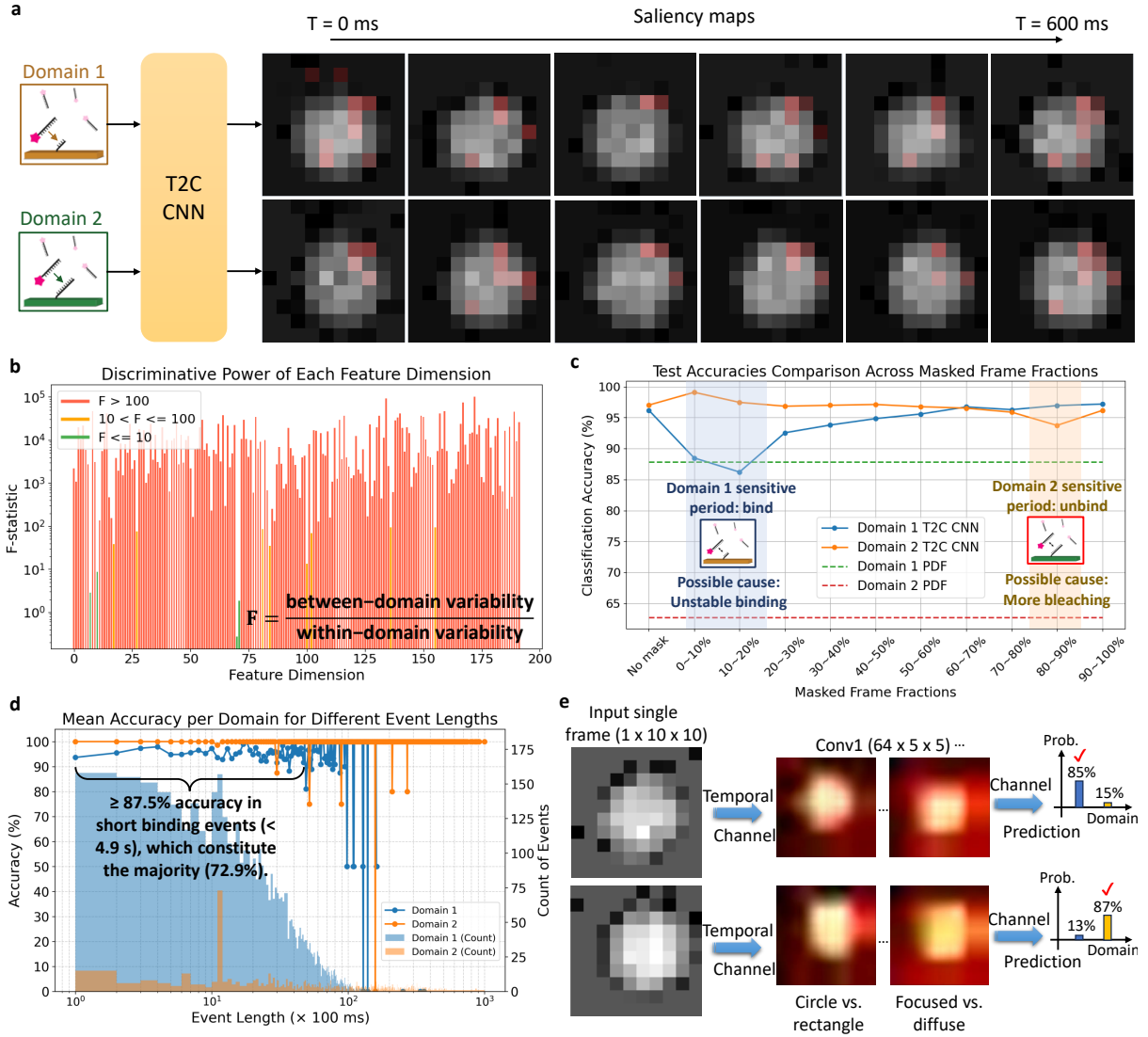


Figure 28: Interpretations of the temporal-to-context convolutional neural network (T2C CNN). **a**, T2C CNN saliency maps for diffraction-limited fluorescent spots in different DNA binding domains. The saliency map visually highlights the regions within each frame that T2C CNN considers most critical 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. **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 (3x3) with a (2x2) stride and (1,1,1,1) padding, converting the (1x10x10) input images into 64 feature maps with spatial dimensions of (5x5).

3.2.5 The Reviewer comments that

In Fig. 3e, the bar plot's meaning is not immediately clear. It seems the authors intend to show accuracy comparisons for Domain 1 and Domain 2, but this should be explicitly stated.

Response: Thank you for your comment.

In the revised caption of the current Fig. 6e (previously Fig. 3e), we have clarified that the bar plot represents the classification probabilities (Prob.) for the example single-frame diffraction-limited fluorescent spots from Domain 1 and Domain 2. To avoid potential misunderstanding, we have also rephrased the relevant sentences in the main text describing this figure. The revised parts are provided below.

(continued on next page)

Revised parts:

1) In the Fig. 6 (previously Fig. 3), on Page 11,

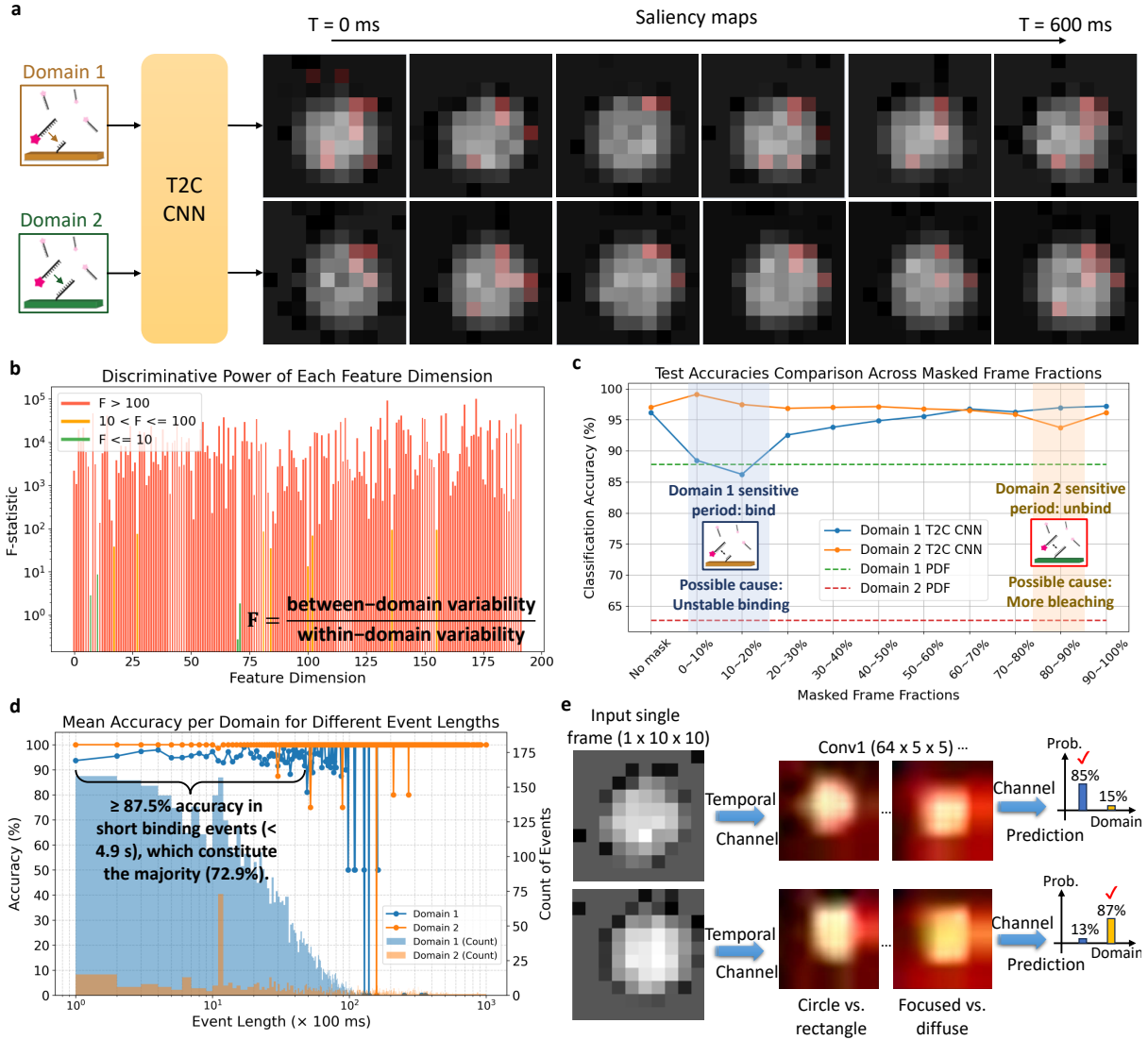


Figure 29: Interpretations of the temporal-to-context convolutional neural network (T2C CNN). **a**, T2C CNN saliency maps for diffraction-limited fluorescent spots in different DNA binding domains. The saliency map visually highlights the regions within each frame that T2C CNN considers most critical 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 of T2C CNN for binding events of varying lengths. **d**, Average classification accuracies of T2C CNN for binding events of varying lengths. **e**, Example T2C CNN feature heatmaps and classification probabilities (Prob.) for single-frame events of 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 (3x3) with a (2x2) stride and (1,1,1,1) padding, converting the (1x10x10) input images into 64 feature maps with spatial dimensions of (5x5).

2) In the Interpretations of the T2C CNN, on Page 12,

...

Interestingly, T2C CNN achieves approximately 95% accuracy in distinguishing binding types on those single-frame events (see Fig.29d at $x = 10^0$). Accordingly, Fig.29e 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 (with

the confidence $\geq 85\%$). These results explain how 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 Supplementary Note 9.

...

3.2.6 The Reviewer comments that

Fig. 4 should include a simple depiction of the experimental design for better understanding.

Response: Thank you for your comment.

In the revised Supplementary Fig. 4 (previously Fig. 4), we have included a subfigure that provides a simplified overview of the experimental design to improve clarity and aid interpretation. The revised figure is attached below.

Revised parts:

1) In the Supplementary Figure 4 (previously Fig. 4), on Page 15,

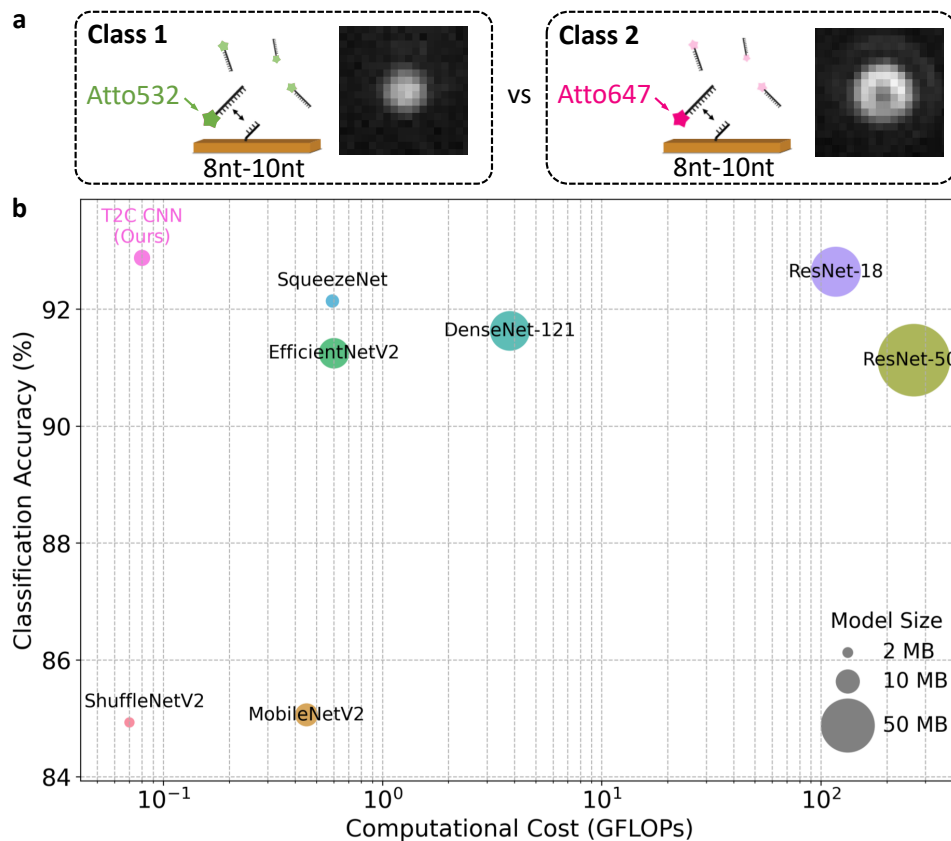


Figure 30: **Single-frame classification of different dye-labeled binding events.** **a)** Experimental setup and example images of two dye types: green (Atto532) and red (Atto647), both using the same 8nt-10nt binding. **b)** Comparison of classification accuracy, computational cost, and model sizes for single-frame discrimination of dye-labeled binding events. GFLOPs (Giga Floating Point Operations per Second) quantify the computational power required, indicating the hardware demands for processing.

References

- [1] Yingjie Zhai, Wenshuo Li, Yehui Tang, Xinghao Chen, and Yunhe Wang. No time to waste: Squeeze time into channel for mobile video understanding. *arXiv preprint arXiv:2405.08344*, 2024.
- [2] Joerg Schnitzbauer, Maximilian T Strauss, Thomas Schlichthaerle, Florian Schueder, and Ralf Jungmann. Super-resolution microscopy with dna-paint. *Nature protocols*, 12(6):1198–1228, 2017.
- [3] Daniel J Nieves, Katharina Gaus, and Matthew AB Baker. Dna-based super-resolution microscopy: Dna-paint. *Genes*, 9(12):621, 2018.
- [4] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.
- [5] Javier Selva, Anders S Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B Moeslund, and Albert Clapés. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12922–12943, 2023.
- [6] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.
- [7] Eshaan Nichani, Adityanarayanan Radhakrishnan, and Caroline Uhler. Do deeper convolutional networks perform better? In *International Conference on Machine Learning*, 2021.
- [8] A Supani, Y Andriani, and H Indarto. Enhancing deeper layers with residual network on cnn architecture: A review. In *Proceedings of the 6th FIRST 2022 International Conference (FIRST-ESCSI 2022)*, volume 14, page 449. Springer Nature, 2023.
- [9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
- [10] Alanazi Rayan, Alaa S Alaerjan, Saad Alanazi, Ahmed I Taloba, Osama R Shahin, Mostafa Salem, et al. Utilizing cnn-lstm techniques for the enhancement of medical systems. *Alexandria Engineering Journal*, 72:323–338, 2023.
- [11] Atiq Ur Rehman, Samir Brahim Belhaouari, Md Alamgir Kabir, and Adnan Khan. On the use of deep learning for video classification. *Applied Sciences*, 13(3):2007, 2023.
- [12] Luiz Zaniolo and Oge Marques. On the use of variable stride in convolutional neural networks. *Multimedia Tools and Applications*, 79(19):13581–13598, 2020.
- [13] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3054–3062, 2020.
- [14] Chuqi Wang. A review on 3d convolutional neural network. In *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)*, pages 1204–1208. IEEE, 2023.
- [15] Abdul Wahab, Syed Danish Ali, Hilal Tayara, and Kil To Chong. iim-cnn: intelligent identifier of 6ma sites on different species by using convolution neural network. *IEEE Access*, 7:178577–178583, 2019.

- [16] Christian A Combs and Hari Shroff. Fluorescence microscopy: a concise guide to current imaging methods. *Current protocols in neuroscience*, 79(1):2–1, 2017.
- [17] Paloma Rodríguez-Sevilla, Sebastian A Thompson, and Daniel Jaque. Multichannel fluorescence microscopy: advantages of going beyond a single emission. *Advanced NanoBiomed Research*, 2(5):2100084, 2022.
- [18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.
- [19] Animesh Karnewar, Tobias Ritschel, Oliver Wang, and Niloy Mitra. Relu fields: The little non-linearity that could. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022.
- [20] Anirbit Mukherjee, Amitabh Basu, Raman Arora, and Poorya Mianjy. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018.
- [21] SH Shabbeer Basha, Shiv Ram Dubey, Viswanath Pulabaigari, and Snehasis Mukherjee. Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing*, 378:112–119, 2020.
- [22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [23] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [24] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
- [25] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023.
- [26] Christopher M Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:1122–1128, 2006.
- [27] Orsolya K Wade, Johannes B Woehrstein, Philipp C Nickels, Sebastian Strauss, Florian Stehr, Johannes Stein, Florian Schueder, Maximilian T Strauss, Mahipal Ganji, Joerg Schnitzbauer, et al. 124-color super-resolution imaging by engineering dna-paint blinking kinetics. *Nano letters*, 19(4):2641–2646, 2019.
- [28] Mukta Singh-Zocchi, Sanhita Dixit, Vassili Ivanov, and Giovanni Zocchi. Single-molecule detection of dna hybridization. *Proceedings of the National Academy of Sciences*, 100(13):7605–7610, 2003.
- [29] Asif Iqbal, Sinan Arslan, Burak Okumus, Timothy J Wilson, Gerard Giraud, David G Norman, Taekjip Ha, and David MJ Lilley. Orientation dependence in fluorescent energy transfer between cy3 and cy5 terminally attached to double-stranded nucleic acids. *Proceedings of the National Academy of Sciences*, 105(32):11176–11181, 2008.
- [30] Shizhao Sun, Wei Chen, Liwei Wang, Xiaoguang Liu, and Tie-Yan Liu. On the depth of deep neural networks: A theoretical view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

- [31] X Li. Exploring the effect of depth and width of cnn models on binary classification of dogs and cats. *Applied and Computational Engineering*, 47(1):147–158, 2024.
- [32] David M Jameson. *Introduction to fluorescence*. CRC press, 2025.
- [33] Rukiye Savran Kızıltepe and John Q Gan. 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*, pages 77–87. Springer, 2020.
- [34] Tim Pearce, Alexandra Brintrup, and Jun Zhu. Understanding softmax confidence and uncertainty. *arXiv preprint arXiv:2106.04972*, 2021.
- [35] Laura C Zanetti-Domingues, Christopher J Tynan, Daniel J Rolfe, David T Clarke, and Marisa Martin-Fernandez. Hydrophobic fluorescent probes introduce artifacts into single molecule tracking experiments due to non-specific binding. *PloS one*, 8(9):e74200, 2013.
- [36] Chandrasekhar Roychoudhuri, Katherine Creath, and A Kracklauer. The nature of light: what is a photon? SPIE, 2005.
- [37] Amnon Yariv and Pochi Yeh. Optical waves in crystal propagation and control of laser radiation. 1983.
- [38] Jerrold T Bushberg and John M Boone. *The essential physics of medical imaging*. Lippincott Williams & Wilkins, Philadelphia, PA, 2011.
- [39] SA Haider, A Cameron, P Siva, D Lui, MJ Shafiee, A Boroomand, N Haider, and A Wong. Fluorescence microscopy image noise reduction using a stochastically-connected random field model. *Scientific reports*, 6(1):20640, 2016.
- [40] Jozsef Dudas, Linda M Wu, Cory Jung, Glenn H Chapman, Zahava Koren, and Israel Koren. Identification of in-field defect development in digital image sensors. In *Digital Photography III*, volume 6502, pages 319–330. SPIE, 2007.
- [41] Yan Shan Ang and Lin-Yue Lanry Yung. Protein–dna conjugates with a discrete number of oligonucleotide strands for highly reproducible protein quantification by the dna proximity assay. *Analytical Chemistry*, 95(32):12071–12079, 2023.
- [42] Yiwei Hou, Wenyi Wang, Yunzhe Fu, Xichuan Ge, Meiqi Li, and Peng Xi. Multi-resolution analysis enables fidelity-ensured deconvolution for fluorescence microscopy. *eLight*, 4(1):14, 2024.

Response Letter

Journal: Nature Communications

Ref No.: NCOMMS-24-77281A

Title: “Discriminating Single-Molecule Binding Event from Diffraction-Limited Fluorescence”

Authors: Yueming Yin, Nithin Pathoor, Kamal Kant Sharma, Shiwen Zhu, Iong Ying Loh, Yan Shan Ang, Shao Ren Sim, Lanry Lin Yue Yung, Thorsten Wohland, Lipo Wang

Dear Reviewers,

Thank you for reviewing. Following your feedback, we have made revisions to improve both the scientific depth and clarity of the manuscript. All changes are highlighted in blue in the revised manuscript and this document. A summary of the revisions and a point-by-point response to your comments are attached below.

Sincerely,

Yueming Yin, Nithin Pathoor, Kamal Kant Sharma, Shiwen Zhu, Iong Ying Loh, Yan Shan Ang, Shao Ren Sim, Lanry Lin Yue Yung, Thorsten Wohland, Lipo Wang

We have made revisions to address all the remaining issues raised by the reviewers. A summary of the changes is provided below:

Additional Experiments:

1. We conducted experiments using scrambled sequences of image frames to demonstrate the presence of temporal information within binding events in both the DNA origami and cell experiments (Comment 2.1).
2. We tracked the convolutional outputs for fluorescence spots accumulated over time to investigate the contribution of temporal information, explore why peripheral regions contribute more than central intensities (Comment 2.1), and interpret domain-discriminative features in both DNA origami and cell experiments (Comment 2.3).

Presentation Improvements:

1. We added connecting lines and used abbreviated annotations in the revised Figure 8 to make performance patterns easier to see (Comment 1.1).
2. We plotted classification accuracies for each event length (similar to Fig. 6d), and visualized one-frame binding events (similar to Fig. 6e) for the cell experiments (Comment 2.2).
3. We reviewed relevant literature to provide a more informed speculation of the underlying mechanisms behind the different binding types (Comment 2.4).
4. We reviewed additional literature to identify plausible physical and geometric mechanisms that could explain the distinct inter-frame correlation patterns observed in Domain 1 (Comment 2.5).
5. We re-examined binding durations on a linear scale and attributed the apparent peak at 11 frames to the blinking correction threshold. This is now clarified with a zoomed figure and explanation for Fig. 6d (Comment 2.6).
6. We updated Figure 1a with annotations to clarify that it was sampled from single frames at a specific exposure time. We also explicitly stated the imager concentration in the Methods section (Comment 3.1).
7. We clarified the distinction between “length-based” and “image-based” methods, added dwell time estimates, and explained axis offsets in the revised Figure 1b (Comment 3.2).

1 Response to the Reviewer #1's comments.

We would like to thank Reviewer #1 for the suggestion to improve the quality of the manuscript. The following changes have been made to address Reviewer #1's comment, where the changes are highlighted in blue in the manuscript.

1.1 The Reviewer comments that

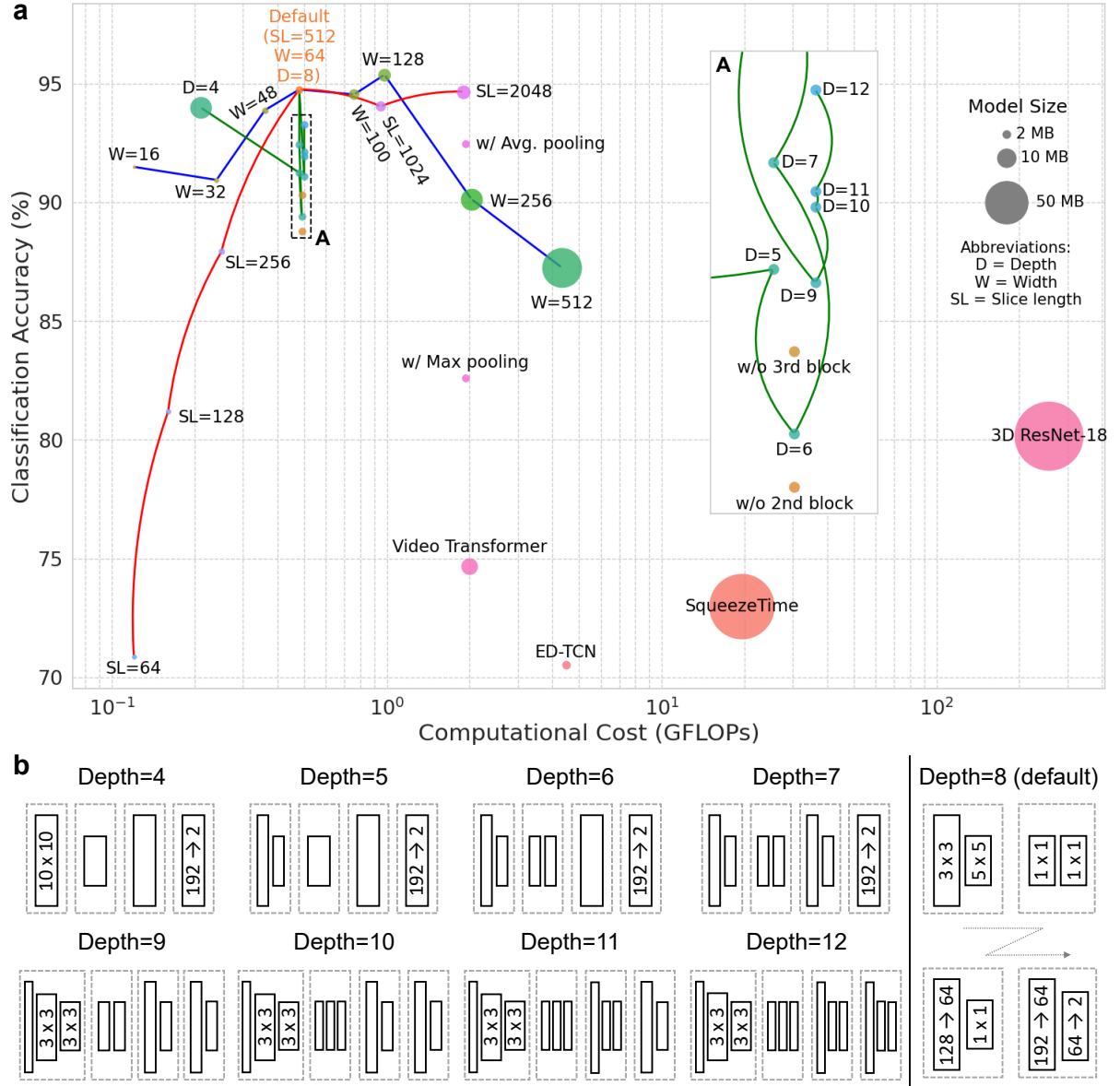
It is particularly hard to understand Figure 8, where too many configurations aggregate on the same page.

Response: We sincerely thank the reviewer for this valuable comment. We agree that the original figure could be challenging to interpret due to the number of configurations presented. To improve clarity while preserving the comparative advantages of displaying all configurations together, we have revised Figure 8 by adding lines that link models with the same hyperparameter type (i.e., width, depth, or slice length). For visual simplicity, hyperparameter types are abbreviated in the figure. These annotations facilitate the reader's visual tracing and comparison across different configurations.

The revised Figure 8 is attached below. We hope this revised figure now better serves its purpose of enabling comparison while being easy to follow.

(continued on next page)

In Main Text Figure 8, on Page 14,



Main Text Figure 8. a) Comparison of classification accuracy, computational cost, and model sizes for variants of the proposed T2C CNN. Baseline methods, including 3D ResNet-18 [1], Video Transformer [2], ED-TCN [3], and SqueezeTime [4], 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. Lines connect T2C CNN variants sharing the same hyperparameter type (i.e., width, depth, or slice length) to visualize performance trends. **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 $\rightarrow$ M.” Rectangles without labeled kernel sizes or dimension transformations follow the default hidden layer configuration for an 8-layer depth.

2 Response to the Reviewer #2's comments.

We would like to thank Reviewer #2 for the suggestion to improve the quality of the manuscript. The following changes have been made to address Reviewer #2's comments, where all the changes are highlighted in blue in the manuscript and the supplementary material.

2.1 The Reviewer comments that

Throughout the paper, the importance of using the temporal convolution is emphasized, but the contribution of this component in the classification is uncertain. Although this was used in training, Fig. 6d shows that the analysis of a single frame already reaches a 95% classification accuracy. There is no temporal information involved here. This may indicate that the temporal information (i.e., binding kinetics) does not contribute much to the classification. In other words, just a large number of diffraction-limited images may be sufficient for training.

It may be interesting to check if training performed with the scrambled sequence of image frames (i.e., no kinetic information) or training with background images excluded would give a similar classification accuracy. This would assess how important it is to include the temporal information in the analysis.

Response: Thank you for the insightful comment. As shown in Supplementary Table 1 (also attached below), we conducted additional experiments using scrambled sequences of image frames. The results demonstrate that scrambling—both within and across binding events—significantly reduces the classification accuracy, confirming that temporal information contributes meaningfully to the model's performance.

Regarding the high classification accuracy achieved for single-frame events, we would like to clarify that this does not imply the absence of temporal information. Even for a single-frame event, two forms of temporal meta-information are available to the model: (1) the event duration is one frame, and (2) the image corresponds to the first (and only) frame of that binding event. Furthermore, each frame represents a 100 ms accumulation of photons by the camera. While sub-frame temporal dynamics are not explicitly available, the integrated signal over this period still captures binding-type characteristics, contributing to correct classification, as validated by our model.

To further investigate how temporal information and single-frame characteristics contribute to classification, we analyzed the outputs of the most discriminative convolutional kernels in the T2C CNN model, as presented in Supplementary Note 4 and Supplementary Fig. 4–6 (also attached below). For single-frame events, the convolutional responses are weak but still distinguishable, suggesting that even minimal dynamics or structural cues within one frame can inform the model. In contrast, longer binding events generate stronger but more variable responses across frames—contributing more temporal features but also introducing higher variability, which can lead to occasional misclassifications.

As for background content, background images were not used during training. Instead, fluorescence spots were cropped from the background and padded to a fixed size (e.g., 10×10 pixels) using a constant value. All training experiments were thus designed to exclude background influence on classification outcomes.

Below, we've attached all the revised parts related to this comment.

Added parts in the revised manuscript:

1) In the “Binding-type information in diffraction-limited fluorescent spots”, on Page 7,

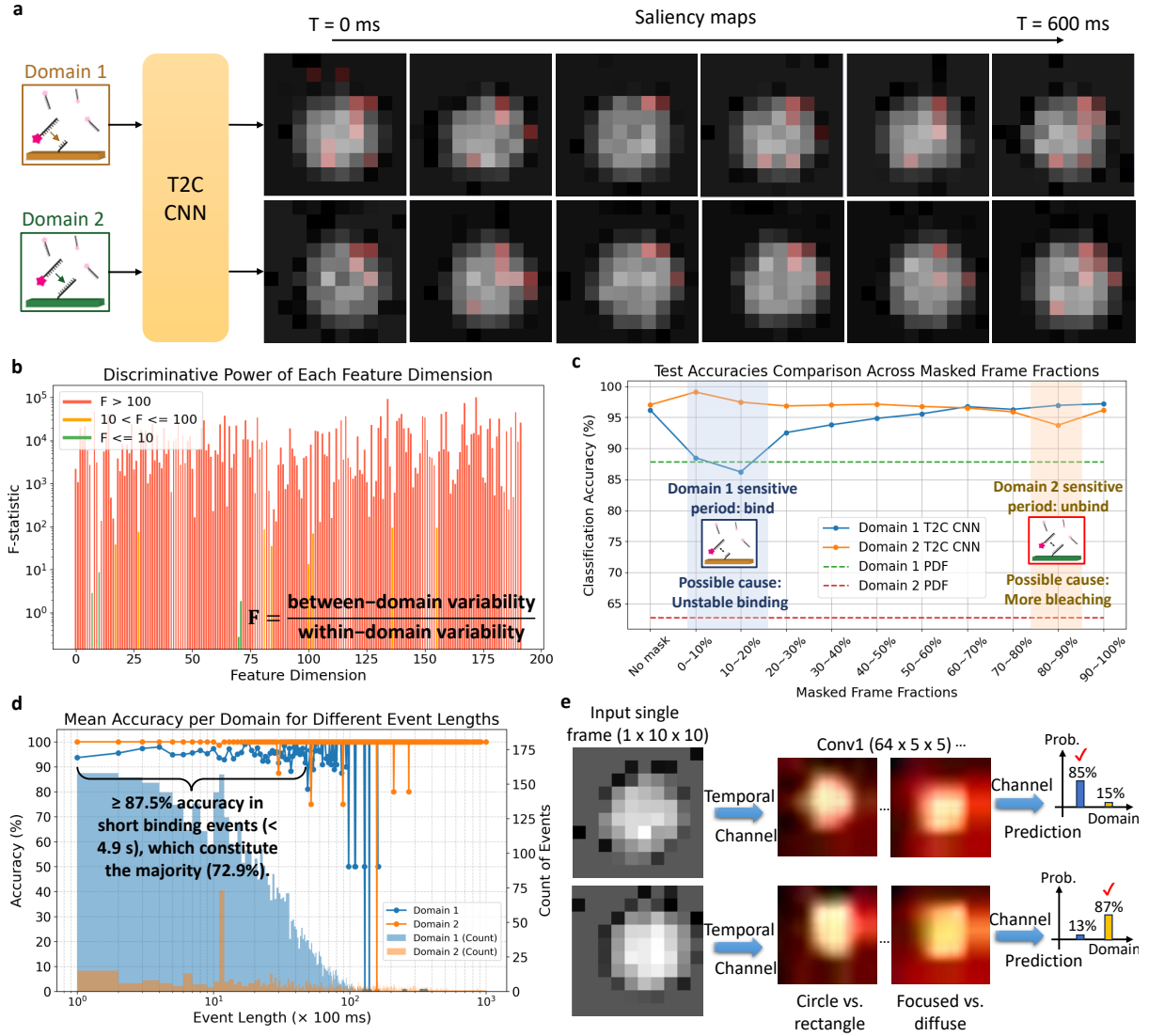
...

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% $\rightarrow$ 89.47%; Cell: 74.09% $\rightarrow$ 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 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.

(continued on next page)

2) In Main Text Figure 6, on Page 12,



Main Text Figure 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 seconds arises from merged short events due to blinking correction thresholds (1 second). **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 $\times$ 3) with a (2 $\times$ 2) stride and (1,1,1,1) padding, converting the (1 $\times$ 10 $\times$ 10) input images into 64 feature maps with spatial dimensions of (5 $\times$ 5). More examples and analysis are provided in Supplementary Fig. 4, 5, 6 and Supplementary Note 4.

(continued on next page)

Added parts of the revised supplementary material:

1) In Supplementary Table 1: Classification accuracy on scrambled fluorescence image sequences under varying information constraints, on Page 23,

Supplementary Table 1. Classification accuracy on scrambled fluorescence image sequences under varying information constraints.

Experiment	Method type ¹	Scramble type	Contained information ²				Accuracy (%) ³
			Lengths	Images	Images in event	Image order	
DNA Origami	Length-based	–	✓	–	–	–	83.88±0.21
	Image-based	Random-length	–	✓	–	–	78.98±2.04
		Cross-event	✓	✓	–	–	87.66±4.73
		In-event	✓	✓	✓	–	89.47±1.41
		–	✓	✓	✓	✓	94.76±0.47
HER2 Cell	Length-based	–	✓	–	–	–	66.34±0.37
	Image-based	Random-length	–	✓	–	–	62.64±1.18
		Cross-event	✓	✓	–	–	70.25±3.31
		In-event	✓	✓	✓	–	71.59±1.96
		–	✓	✓	✓	✓	74.09±4.54

¹ We take the proposed T2C CNN and probability density function (PDF) as representatives of image-based and length-based methods, respectively. ² This indicates the type of information retained or lost in different scrambling strategies. “In-event”: Scrambling image sequences within an event destroys temporal order but retains event integrity. “Cross-event”: Cross-event scrambling mixes frames from different events, disrupting both temporal order and event-level coherence, while preserving individual event lengths. “Random-length”: scrambling random-length sequences groups frames arbitrarily, eliminating both temporal and event-level information. The length-based method uses only the lengths of multiple events as input. ³ For image-based methods, we report 5-fold class-wise accuracy for single-event classification (requires ~5 s of measurement). For length-based methods, we report 5-fold class-wise accuracy for single-molecule classification (requires ~10 min of measurement), using all observed events per molecule. Accuracy is reported as mean ± standard deviation across 5 folds.

2) In the Supplementary Note 4: Analysis on fluorescence spots and T2C CNN convolutional outputs from distinct binding events, on Page 5,

Main Text Fig. 6d and 6e show that T2C CNN can accurately classify short binding events, including single-frame events. Does this imply that temporal information contributes little to the classification? To explore the answer, we tracked the response of T2C CNN’s top kernels to fluorescence spots over time in Fig. 4 and 5, corresponding to the DNA origami and cell experiments, respectively. Overall, we found that the accumulated convolutional output becomes more pronounced and distinguishable as more frames are included, indicating that temporal information does contribute to the prediction. However, accumulating outputs across frames within a binding event does not always enhance previously observed patterns. For example, in the cell experiment (lower right of Fig. 5b), the Top-3 kernel’s accumulated output at frame 3 for Domain 2 weakens the pattern accumulated in frames 1 and 2. This is consistent with the input fluorescence spot at frame 3, which appears slightly smaller and dimmer, with a noticeable change in shape and intensity distribution compared to the first two frames.

When comparing across domains, we found that the accumulated convolutional outputs generally share a similar overall style—due to similar input spot characteristics and the use of the same convolutional kernels—but differ in detail. For instance, in the Top-1 kernel’s accumulated output in the DNA origami experiments (second row in Fig. 4a and 4b), both domains show low values at

the center. However, Domain 1 exhibits a horizontally extended low-value region, while Domain 2 shows a lower-right clustered pattern, especially as more frames are accumulated. These distinct patterns are not readily discernible from the raw fluorescence spots. This observation supports the hypothesis that domain-specific differences may stem from subtle variations in fluorophore positioning, orientation, or environmental constraints, although the exact physical origin remains unclear. Such differences may lead to local mismatches between kernel patterns and specific regions of the fluorescence spots, contributing to the observed convolutional differences.

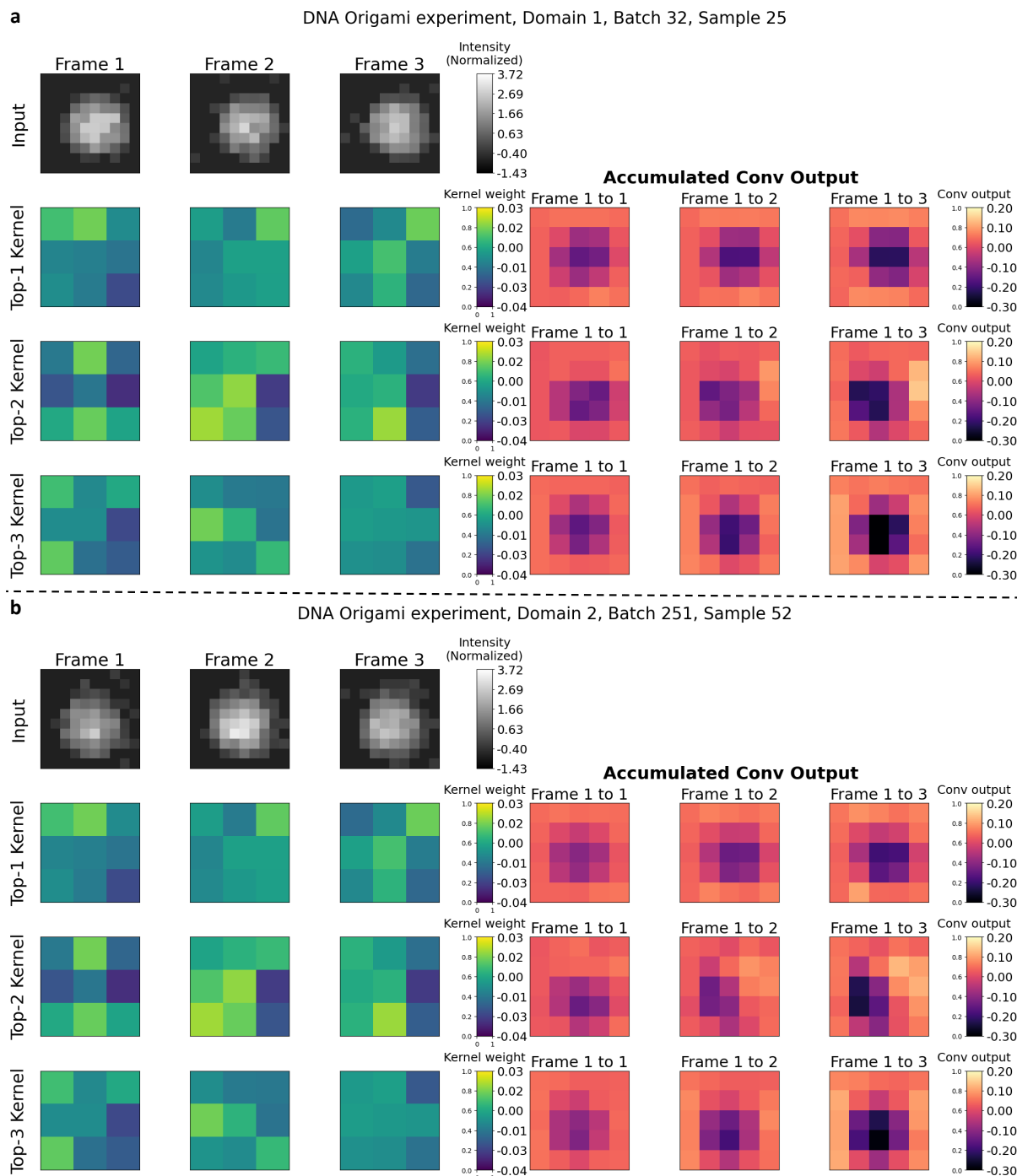
When comparing across experiments, we observed that convolutional kernels in the DNA origami experiments (first three columns in Fig. 4) contain more negative values than those in the cell experiments (first three columns in Fig. 5). Kernels with clustered negative values can produce stronger responses to certain peripheral patterns in fluorescence spots. In the DNA origami experiments, these peripheral regions are typically cleaner and more circular. When these regions exhibit domain-specific differences, T2C CNN is likely to focus on them rather than the more random center intensity. This observation may help explain the phenomenon in Main Text Fig. 6a, where the periphery appears to contribute more to T2C CNN’s predictions. In contrast, the fluorescence spots in the cell experiments display more irregular or complex periphery and center intensity, leading to more variability. In this case, the Top-1 kernel tends to emphasize peripheral regions (more negative values), while the Top-2 and Top-3 kernels focus more on center intensities (more positive values).

When analyzing single-frame binding events (Fig. 6), we found that their convolutional outputs are generally as weak as the first frame of longer events in Fig. 4 and 5. Despite being weak, these outputs can still differ between domains, e.g., through differences in value distributions. The output patterns are driven by the fluorescence inputs and do not form consistent intra-domain features. However, the trained top kernels act as fixed templates that highlight differences in domain-specific spot characteristics. For example, the Top-1 kernel in the DNA origami experiments (second row in Fig. 6a and 6b) shows a peripheral gradient pattern, such as a medium–high–low intensity transition in the first row of the kernel. In the cell experiments, the Top-1 kernel (second row in Fig. 6c and 6d) exhibits a center-focused pattern, characterized by a high-intensity pixel located at the upper-right of two adjacent high-intensity pixels within the same row. These kernel patterns serve as templates to measure the degree of alignment in the input fluorescence spots. Higher convolutional layers and fully connected layers then use these measurements to abstract more high-level representations for final classification.

From the above analyses, we found that temporal information plays an important role in T2C CNN’s decision-making by changing the final convolutional patterns. Misclassifications of rare long events may arise from diluted or inconsistent temporal patterns. Single-frame events, while weaker in response, can still be distinguished by domain-specific convolutional patterns. The increased complexity and variability of fluorescence spots in cell experiments contribute to lower classification accuracy. In contrast, the cleaner and more regular peripheral regions of DNA origami spots encourage the model to focus on periphery-based patterns for more reliable classification.

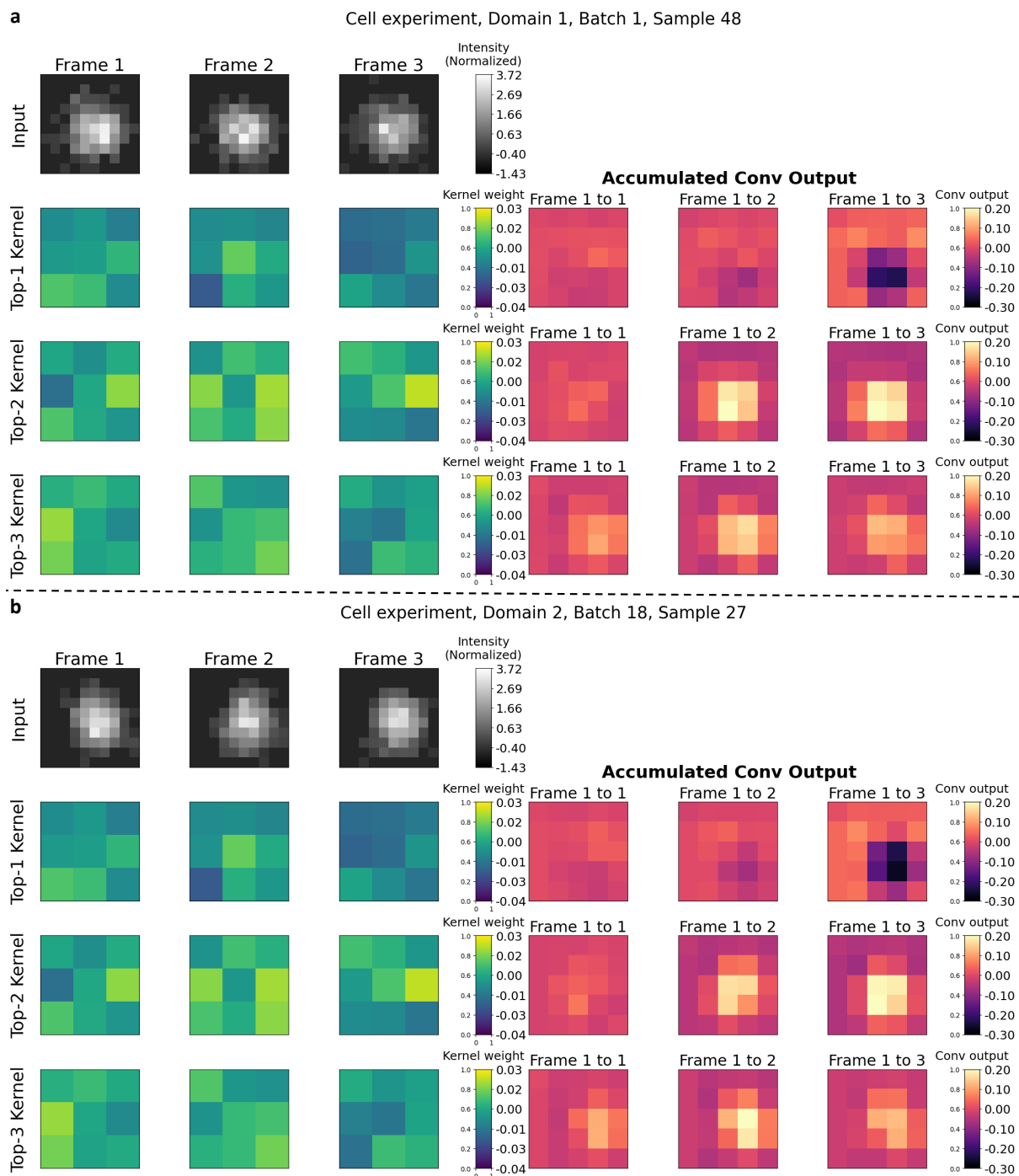
(continued on next page)

3) In Supplementary Fig. 4: Example fluorescence spots and temporally accumulated T2C-CNN convolutional outputs corresponding to distinct binding events in DNA origami experiments, on Page 17,



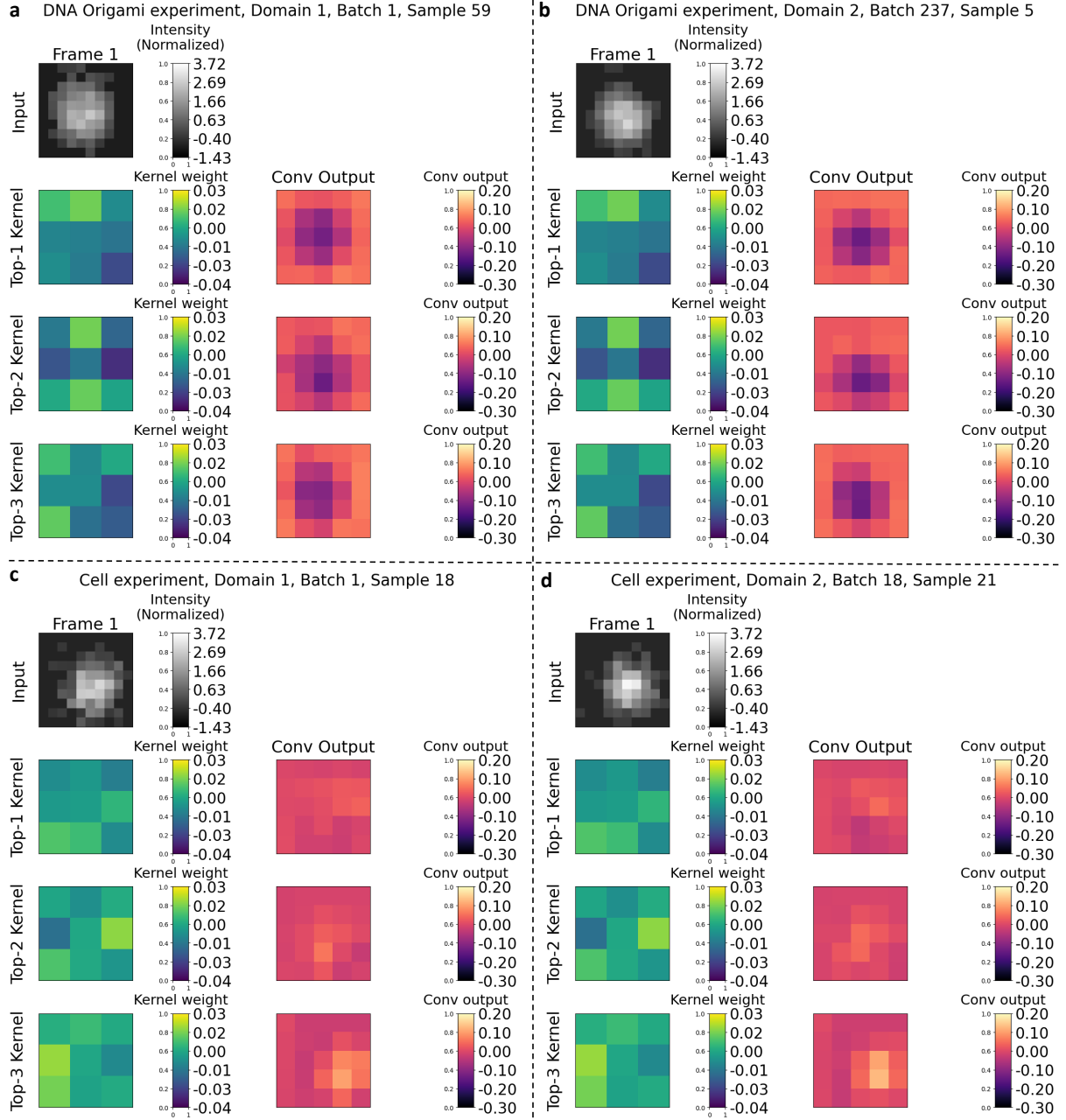
Supplementary Figure 4. Example fluorescence spots and temporally accumulated T2C-CNN convolutional outputs corresponding to distinct binding events in DNA origami experiments. The T2C CNN applies distinct convolutional kernel groups synchronously to each frame within a predefined slice length. Convolutional outputs from kernels within the same group are summed across accumulated frames of an event. The top-3 kernel groups were selected globally for visualization, each showing distinct responses to fluorescence spots associated with different domain events. Subfigures **a** and **b** present representative examples from Domain 1 and Domain 2, respectively. The right and left legend labels indicate the original and normalized value ranges, respectively, which are consistent across both domains for comparability. Fluorescence spot intensities were mean–standard deviation normalized within each event. For clearer visualization, a global range truncation was applied to both fluorescence spot intensities and convolutional output values.

4) In Supplementary Fig. 5: Example fluorescence spots and temporally accumulated T2C-CNN convolutional outputs corresponding to distinct binding events in cell experiments, on Page 18,



Supplementary Figure 5. Example fluorescence spots and temporally accumulated T2C-CNN convolutional outputs corresponding to distinct binding events in cell experiments. The T2C CNN applies distinct convolutional kernel groups synchronously to each frame within a predefined slice length. Convolutional outputs from kernels within the same group are summed across accumulated frames of an event. The top-3 kernel groups were selected globally for visualization, each showing distinct responses to fluorescence spots associated with different domain events. Subfigures **a** and **b** present representative examples from Domain 1 and Domain 2, respectively. The right and left legend labels indicate the original and normalized value ranges, respectively, which are consistent across both domains for comparability. Fluorescence spot intensities were mean–standard deviation normalized within each event. For clearer visualization, a global range truncation was applied to both fluorescence spot intensities and convolutional output values.

5) In Supplementary Fig. 6: Example fluorescence spots and T2C CNN convolutional outputs from distinct single-frame binding events, on Page 19,



Supplementary Figure 6. Example fluorescence spots and T2C CNN convolutional outputs from distinct single-frame binding events. The T2C CNN applies distinct convolutional kernel groups synchronously to each frame within a predefined slice length. Convolutional outputs from kernels within the same group are summed across all frames of an event. The top-3 kernel groups were selected globally for visualization, each showing distinct responses to fluorescence spots associated with different domain events. Subfigures **a** and **b** present representative examples from Domain 1 and Domain 2 in DNA origami experiments, respectively, while subfigures **c** and **d** show examples from cell experiments. The right and left legend labels indicate the original and normalized value ranges, respectively, which are consistent across both domains for comparability. Fluorescence spot intensities were mean–standard deviation normalized within each event. For clearer visualization, a global range truncation was applied to both fluorescence spot intensities and convolutional output values.

2.2 The Reviewer comments that

From the above perspective, the relatively lower accuracy in the cell image data analysis (Figs. 9 and 10) may not result from the overlapping binding duration, but from a relatively small difference in the diffraction-limited images of the two domains. Is there a plot similar to Fig. 6d and e for the cell image data?

Response: Thank you for the insightful comment. In the newly added Supplementary Note 4 and Supplementary Fig. 6 (also attached above), we visualized and analyzed single-frame binding events in the cell experiments, in a manner similar to Fig. 6e. We observed that their fluorescence spot patterns were indeed more complex and variable, which likely contributed to the relatively lower classification accuracies.

In addition, we plotted the classification accuracies of the cell experiments for each event length (similar to Fig. 6d), as shown in Fig. 1 below. We found that the accuracies fluctuated more substantially across event lengths, particularly for the relatively rare long-duration events. The average class-wise accuracy reached 79.64%.

This performance outperformed both the best-performing state-of-the-art deep learning method we evaluated, SqueezeTime [4] (63.12%), and the length-based method (PDF, 54.08%). The relatively low performance of the PDF method reflects the high degree of overlap in binding durations, as shown in Fig. 9b of the main text (Page 17). However, there remains considerable room for improvement due to the greater overlap in binding durations and the increased complexity of the cell fluorescence images.

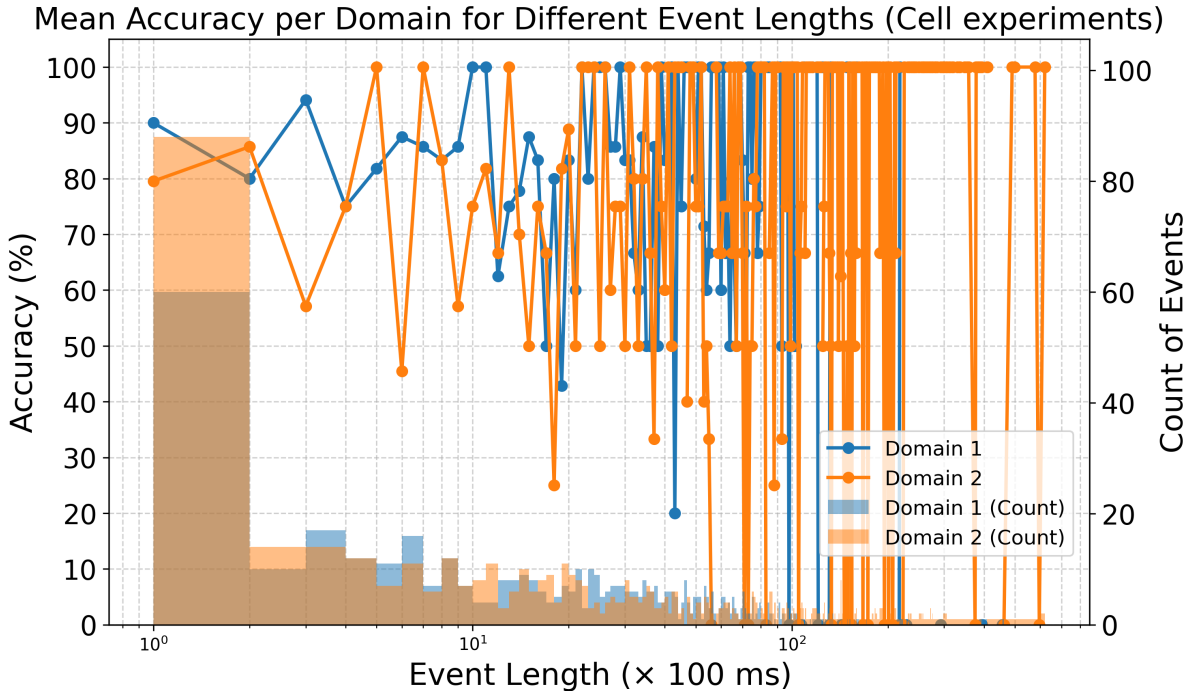


Figure 1: Classification accuracies of the T2C CNN for binding events of varying lengths in the cell experiments. The plot shows average classification accuracy per event length. Accuracy varies more across lengths compared to the DNA origami dataset, particularly for longer and rarer events.

2.3 The Reviewer comments that

Fig. 6a shows that the peripheral region of the diffraction-limited image is the most critical part for classification. This is quite counterintuitive. The images in Fig. 6a clearly show intensity variations by shot noise (fluctuation by a finite number of photons in each pixel). The relative shot noise (i.e., signal-to-noise ratio) should be the largest in the peripheral region due to the low intensity (small number of photons). Then, how can this intrinsically noisy part discriminate between the two domains? In particular, Fig. 6d shows 95% accuracy in the single frame analysis. There is one example in Fig. 6e for explanation, but it is not clear why this intrinsic noise is not transferred to the convoluted images.

In fact, this was one of the reasons that I suspected the average intensity per frame may be different for the two domains in my first round comments. (The intensity of the two input images in Fig. 6e actually looks different.) However, Fig. 5b shows that they are overall almost the same. It may be one of the magics of deep learning, but the high performance of the T2C network in a single frame image is indeed puzzling.

Response: Thank you for the insightful and thought-provoking comment. We agree that the apparent importance of the peripheral region in the saliency maps (Fig. 6a) is counterintuitive at first glance, especially given the lower signal-to-noise ratio due to shot noise in these regions. However, we would like to clarify a few key aspects that may help reconcile this observation.

First, the saliency map in Fig. 6a highlights regions where the model's output is most sensitive to perturbations in the input, but it does not imply that these regions have high raw intensity or are visually discriminative to the human eye. The T2C CNN may have learned subtle yet consistent spatial features—such as differences in the radial decay of intensity, diffraction ring shape, or local asymmetries—that are more distinguishable in the peripheral region and vary systematically between the two domains. Although these features may be masked by shot noise at the individual-pixel level, the model's convolutional layers aggregate information over local spatial neighborhoods and across time. Since T2C views a long temporal dimension as additional convolutional channels, it can integrate spatiotemporal context to enhance signal extraction, enabling discrimination even in low-intensity or noisy regions.

Second, as the reviewer correctly noted, we carefully examined and ruled out average intensity differences as the driving factor behind the high classification accuracy. As shown in Fig. 5b, the average intensity distributions across the two domains are nearly identical. This suggests that the model is not relying on global brightness cues, but rather on spatial and structural features within the fluorescence profiles.

Finally, the example shown in Fig. 6e is intended to illustrate representative visual differences between one-frame events. However, we agree that intensity differences in single examples can be misleading. To address this concern, we have included a new Supplementary Fig. 6 in which the input intensities are normalized for visualization, thereby mitigating potential bias from brightness variations. We observed that the corresponding convolutional responses remain weak but still distinguishable, suggesting that the model may rely on subtle structural cues or minimal dynamics—even within a single frame—to inform its predictions.

We appreciate the reviewer highlighting this apparent contradiction, and we agree that this highlights the nontrivial capacity of deep learning models to exploit subtle patterns beyond what is visually intuitive. We have now added a clarifying discussion of this point in Supplementary Note 4 to assist the reader in interpreting these results.

2.4 The Reviewer comments that

Regarding possible reasons for the formation of the different diffraction-limited images, it was stated on page 9 that “This design captures frequency variations across spatial regions, effectively encoding DNA binding dynamics that can affect fluorescence intensity, angle, and density. For instance, hybridization-induced changes in fluorophore orientation can alter emission polarization or angular distribution [33, 34]. Leveraging these insights greatly improves the efficiency of binding type classification.”

→ This explanation is not sufficient. A fluorophore was not used in Ref. 33. The experiment in Ref. 34 was single-molecule two-color FRET, which can effectively detect the stacking of dyes at the end of DNA duplex because the relative orientation of the two fluorophores is fixed. In one-color experiment performed in this study, however, even if the dye stacks at the end of DNA, due to the rapid movement of DNA strand, the orientation would not be fixed, but averaged in the measurement time scale of 100 ms (one frame). The dye can stick to a rigid part such as the surface, which may fix the orientation or hinder its rotation. However, if this happens, there must be large intensity variations, which doesn’t seem to be the case.

Response: We thank the Reviewer for highlighting the limitations of our original explanation and for clarifying the differences in experimental conditions relative to Refs. 33 and 34. We agree that under our current imaging conditions, rapid motion of the fluorophore-tethered DNA would likely average out orientation-dependent effects over the 100 ms integration time, making polarization-related explanations less relevant.

In the revised manuscript, we now state more cautiously that the precise mechanisms remain speculative and likely involve a combination of geometric and dynamic factors. While we cannot draw firm conclusions under the current experimental conditions, we speculate that differences in linker length or binding strength could, in principle, influence fluorophore mobility and slightly alter the PSF shape even within a 100 ms frame. These points are intended as speculative possibilities rather than definitive explanations, and their actual contributions remain to be determined.

To clarify, we have updated the relevant section as follows:

Revised parts of the manuscript:

In the “Cross-experiment evaluation of existing methods for discriminating single-dye binding types”, on Page 8,

... 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. We acknowledge that the precise mechanisms remain speculative and likely involve a combination of geometric and dynamic factors. Future experimental studies using polarization-resolved imaging or higher temporal resolution may help clarify these contributions. We speculate that differences in linker length and binding strength may influence, albeit subtly, the mobility of the fluorophore bound to docking strands [5], and that slight changes in the PSF [6] averaged over 100 ms may contribute to the enhanced classification accuracy observed with diffraction-limited images. ...

2.5 The Reviewer comments that

In the explanation of the correlation plot of Fig. 3, it was stated on page 5, “These intermittent low-correlation segments in Domain 1 likely stem from peak shifts, spot shape changes, or transient fluctuations—more prevalent due to its weaker binding affinity. Such frame-to-frame correlation patterns provide useful signals for distinguishing binding types in fluorescence videos.”

→ It is unlikely that weaker binding affinity would cause those fluctuations unless binding duration is shorter than the frame time of 100 ms as in the case of the new 6-bp data. However, in this case, the average intensity per frame is lower, which is not the case of 8-bp construct (domain 1). Peak shift clearly indicates the construct is stuck to a rigid part.

Response: We thank the Reviewer for this careful analysis and agree that lower binding affinity alone cannot explain the observed fluctuations unless the average binding durations are comparable to or shorter than the camera’s integration time (100 ms). For the 8-bp construct (Domain 1), the comparable mean intensity with Domain 2 suggests that binding events are typically longer than one frame, supporting the Reviewer’s observation.

Nevertheless, Domain 1 exhibits more frequent low-correlation segments, which we speculate may result from a combination of relatively shorter binding durations and physical/geometric factors that increase fluctuations within individual binding events. These include:

- **Asymmetric duplex geometry:** The 8nt–10nt hybrid contains a single-stranded overhang that disrupts duplex symmetry and increases rotational freedom. Such asymmetry has been shown to amplify thermal motion of the docking-imager complex. As discussed in [7]: “A potential issue deriving from the extension of the docking strands is the loss of spatial resolution, as the flexible docking-imager complexes undergo rapid thermal fluctuations during binding events (see Supplementary Note 2).”
- **Subpixel-scale lateral motion:** Fluorophores tethered via more flexible linkers (e.g., single-stranded overhangs) may drift slightly between frames, altering spatial intensity patterns even if the fluorophore remains bound. In [8], the authors note: “...the localization precision is limited by molecular thermal fluctuations, which is a few nanometers depending on the molecular size and flexibility.” This effect is more pronounced in Domain 1.

Thus, we speculate that both the temporal dynamics (relatively shorter binding durations) and the enhanced structural flexibility of Domain 1 contribute to the observed correlation fluctuations. We thank the Reviewer again for encouraging this deeper examination, and have revised the main text accordingly.

Revised parts of the manuscript:

In the “Binding-type information in diffraction-limited fluorescent spots”, on Page 5,

Frame correlation quantifies the similarity between frames based on pixel intensity patterns, revealing temporal and structural consistency (definition in Supplementary Note 1). Fig. 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, likely reflecting a combination of fewer long-duration events and enhanced structural fluctuations. *These intermittent low-correlation segments in Domain 1 may arise from increased torsional and lateral fluctuations enabled by its asymmetric partial duplex geometry [7] and flexible overhangs [8], 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.

2.6 The Reviewer comments that

In Fig. 6d, why were an order of magnitude more binding events (~ 70 vs several) observed at the duration of 2 s (20×100 ms) for Domain 2?

Response: Thank you for the thoughtful comment. We re-examined Fig. 6d by zooming into the linear (non-log) scale and found a notable peak at a binding duration of 11 frames, especially in Domain 2. Upon reviewing our data processing pipeline, we identified the cause to be related to the blinking correction threshold, which was set at 10 frames (1 s).

This threshold merges two fluorescence events detected at the same binding site within 1 second, under the assumption that they arise from the same binding event with an intermediate dark period (blinking). However, in some cases, two separate short binding events (each lasting one frame, i.e., 100 ms) may occur at the same site with a 9-frame (900 ms) dark gap in between. These would be interpreted as a single 11-frame binding event under our correction rule. A typical example would show 1 frame on, 9 frames off, 1 frame on, followed by a longer dark interval.

From the zoomed figure below (Fig. 2), we estimate there are approximately 27 such events in Domain 1 (representing 0.1% of 19,457 events) and 36 in Domain 2 (0.6% of 6,073 events). These are more apparent in Domain 2 due to its lower overall number of short binding events.

Since these events constitute only a minor fraction of the total, and it is unclear whether they truly originate from single binding events, we grouped them together and treated them as a single binding event for simplicity.

We have added a brief explanation in Fig. 6d to clarify this phenomenon for the reader. Revised Fig. 6d is attached below.

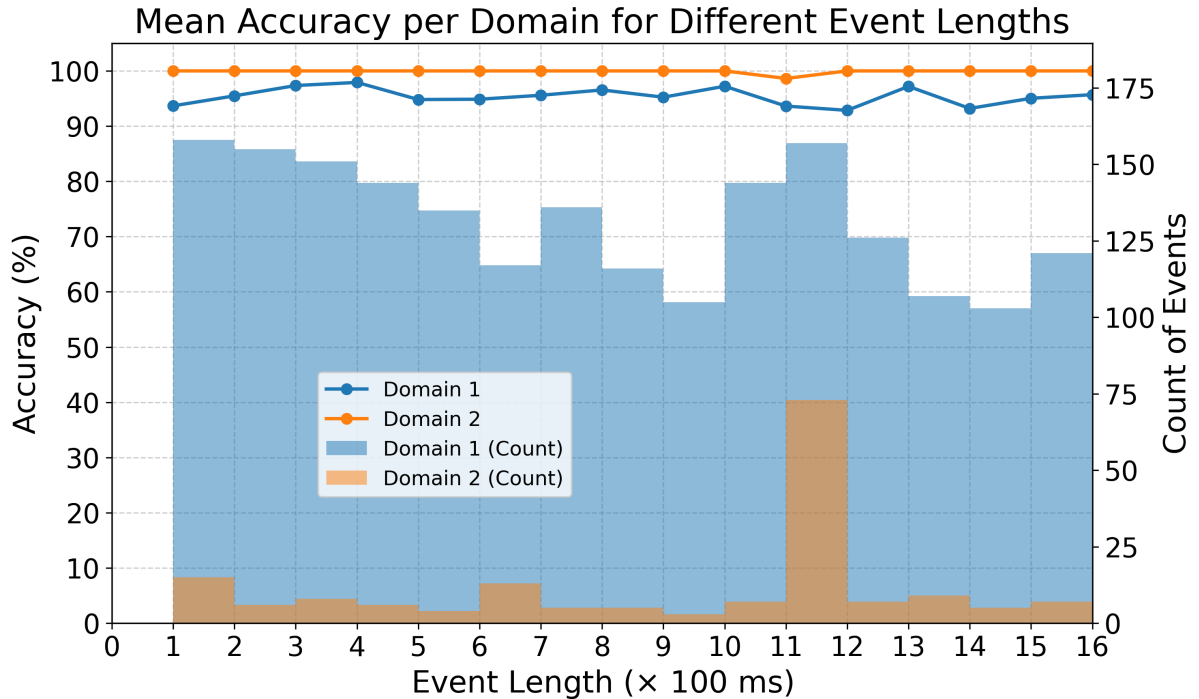
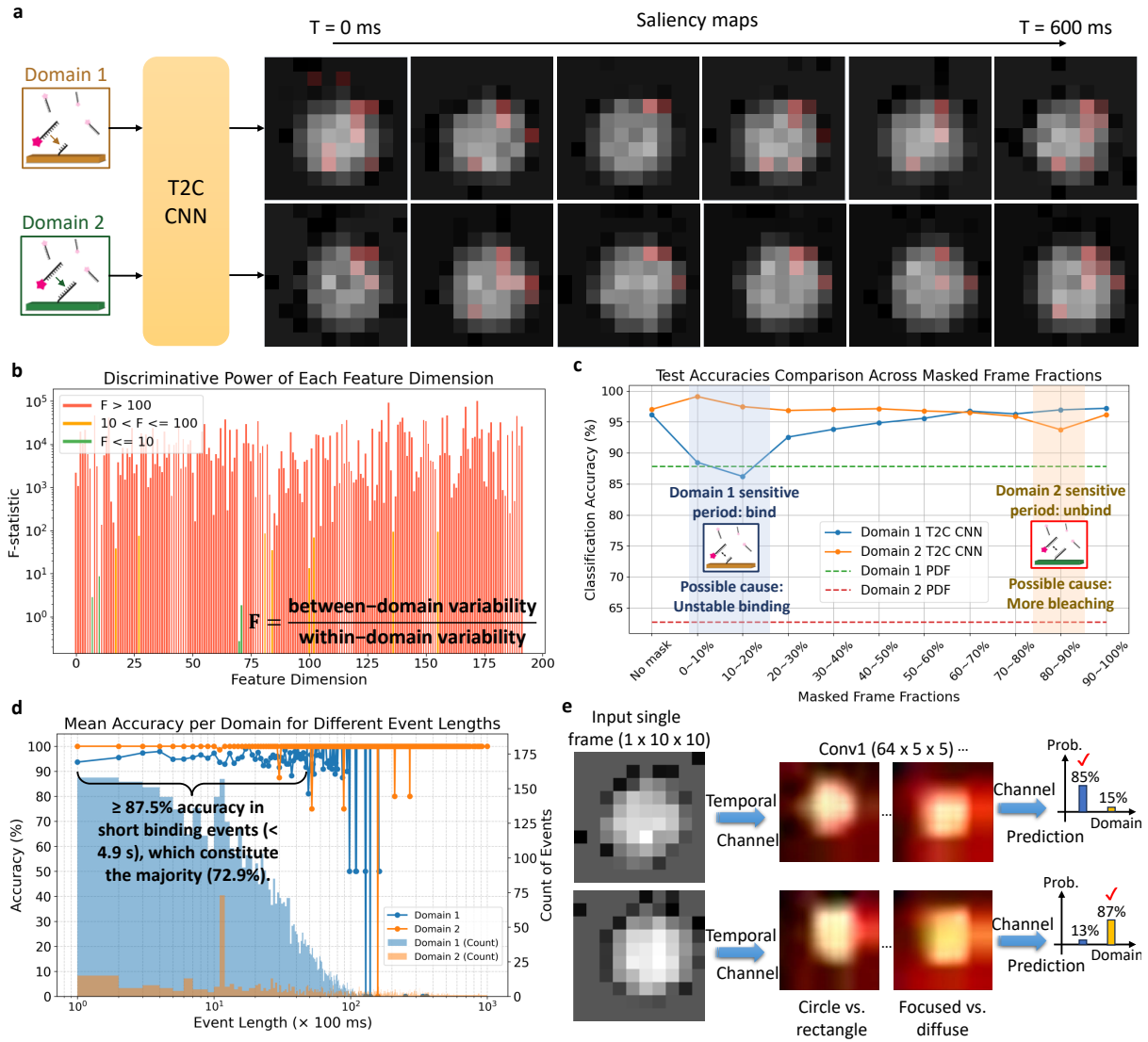


Figure 2: Zoomed-in average classification accuracies of T2C CNN for binding events of varying lengths within 1.6 s. This figure corresponds to a magnified view of Fig. 6d in the main text.

In Main Text Figure 6, on Page 12,



Main Text Figure 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 seconds arises from merged short events due to blinking correction thresholds (1 second).** **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 Fig. 4, 5, 6 and Supplementary Note 4.

3 Response to the Reviewer #3's comments.

We would like to thank Reviewer #3 for the suggestion to improve the quality of the manuscript. The following changes have been made to address Reviewer #3's comments, where all the changes are highlighted in blue in the manuscript and the supplementary material.

3.1 The Reviewer comments that

It is unclear whether the raw images shown in Figure 1a represent single frames at a specific exposure time or cumulative images compiled from multiple binding events. Please clarify and provide additional details. If the surface-anchored docker strands are loaded at similar densities, the number and intensity of the observed imager strand spots should be comparable, differing only in their dwell durations. Also, please specify the concentrations of the imager strands in the Materials and Methods section.

Response: Thank you for the thoughtful comment. We have revised Figure 1a to clarify that it shows a representative single frame (100 ms exposure) from the raw DNA-PAINT video, not a cumulative image. This is now directly annotated on the figure, and additional details have been added to the main text where Figure 1a is introduced.

To reduce spot overlap, the two domains were loaded at different docker strand concentrations: 1 nM for Domain 1 (8 nt) and 400 pM for Domain 2 (10 nt). Despite the lower loading for Domain 2, its longer binding dwell time leads to a higher chance of temporal overlap, resulting in more observable spots per frame.

The same 1 nM imager strand solution was used for both domains. This information has been explicitly added to the revised Methods section. Please see the updated text and Figure 1a below.

Revised parts of the manuscript:

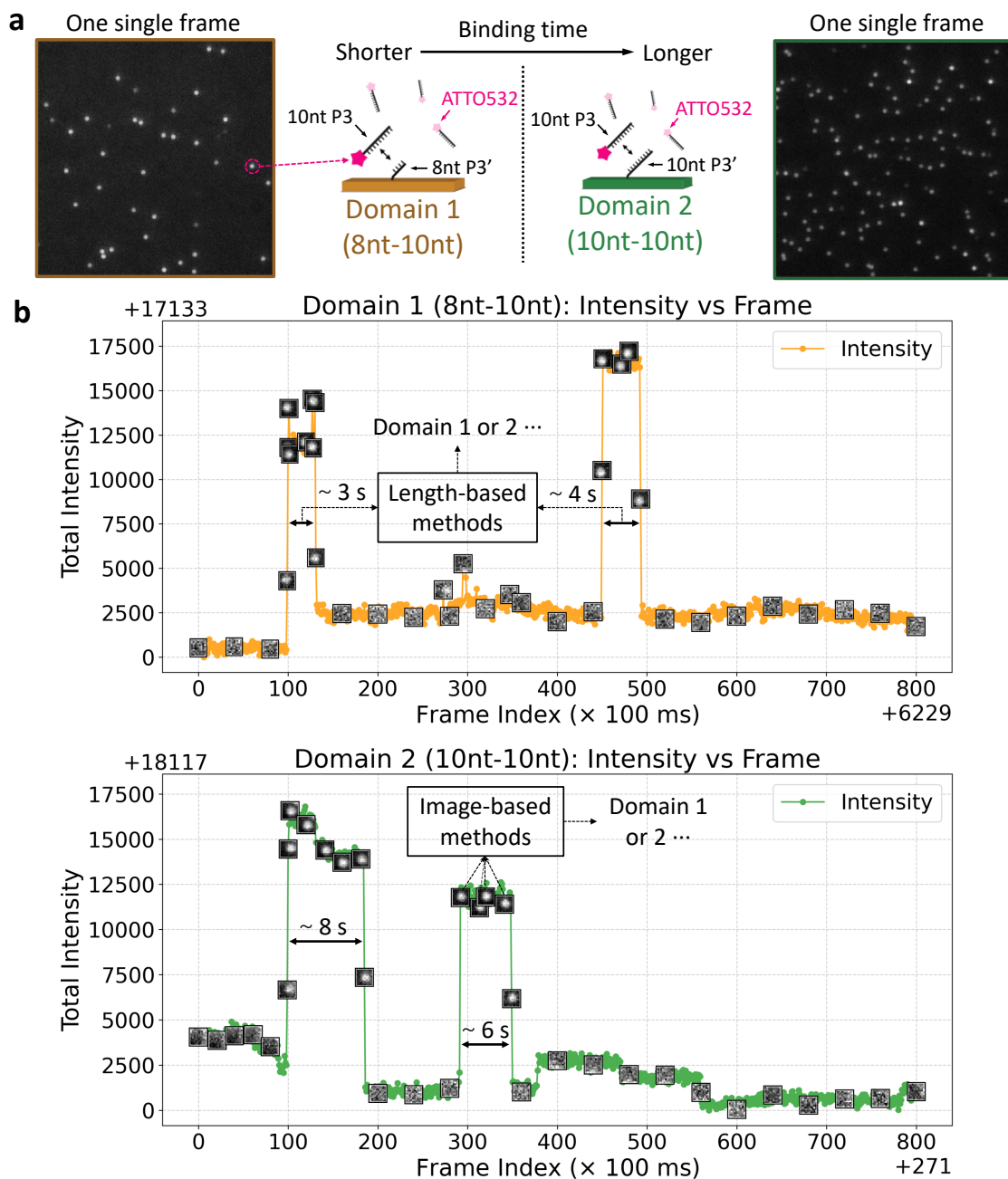
1) In the Overview of raw data and methods, on Page 3,

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. [Fig. 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 [9], 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).

...

(continued on next page)

2) In Main Text Figure 1, on Page 4,



Main Text Figure 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 [10]). Both domains share the same dye (“ATTO532”) 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 estimates t_{on} as 2.0 ± 1.2 s for Domain 1 and 8.0 ± 6.5 s for Domain 2 via NUPACK [11]), 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.

3) In the Sample preparation, in Methods, on Page 4,

...

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 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). All experiments were conducted on separate samples.

3.2 The Reviewer comments that

Figure 1b: The distinction between “length-based methods” and “image-based methods” is difficult to discern from the current illustration. Secondly, as stated in the text and figure caption, “Domain 2 exhibits longer binding durations than Domain 1,” but this is not readily apparent. A dwell time analysis is recommended to better demonstrate this point. Finally, the meaning of the (+17133/+18117) annotation on the y-axis and (+6229/+271) on the x-axis is unclear and should be clarified.

Response: Thank you for the helpful comment. We revised Figure 1b to clarify the difference between the “length-based” and “image-based” methods by adding labels. To support the claim about binding durations, we performed a dwell time analysis using NUPACK [11], which estimates the t_{on} for Domain 1 (8nt–10nt P3) as 2.0 ± 1.2 s and for Domain 2 (10nt–10nt P3) as 8.0 ± 6.5 s. We also clarified that the axis offsets (+17133/+18117 for the y-axis and +6229/+271 for the x-axis) indicate the base intensity and event start frame, respectively. These updates are included in the revised caption. The updated Figure 1b is shown in the previous response (Figure 1).

References

- [1] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.
- [2] Javier Selva, Anders S Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B Moeslund, and Albert Clapés. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12922–12943, 2023.
- [3] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.
- [4] Yingjie Zhai, Wenshuo Li, Yehui Tang, Xinghao Chen, and Yunhe Wang. No time to waste: Squeeze time into channel for mobile video understanding. *arXiv preprint arXiv:2405.08344*, 2024.
- [5] Matthew D Lew, Mikael P Backlund, and WE Moerner. Rotational mobility of single molecules affects localization accuracy in super-resolution fluorescence microscopy. *Nano letters*, 13(9):3967–3972, 2013.
- [6] Christiaan N Hulleman, Rasmus Ø Thorsen, Eugene Kim, Cees Dekker, Sjoerd Stallinga, and Bernd Rieger. Simultaneous orientation and 3d localization microscopy with a vortex point spread function. *Nature Communications*, 12(1):5934, 2021.
- [7] Alexander H Clowsley, William T Kaufhold, Tobias Lutz, Anna Meletiou, Lorenzo Di Michele, and Christian Soeller. Repeat dna-paint suppresses background and non-specific signals in optical nanoscopy. *Nature communications*, 12(1):501, 2021.
- [8] Guangzhong Ma, Zijian Wan, Yunze Yang, Wenwen Jing, and Shaopeng Wang. Three-dimensional tracking of tethered particles for probing nanometer-scale single-molecule dynamics using a plasmonic microscope. *ACS sensors*, 6(11):4234–4243, 2021.
- [9] Joerg Schnitzbauer, Maximilian T Strauss, Thomas Schlichthaerle, Florian Schueder, and Ralf Jungmann. Super-resolution microscopy with dna-paint. *Nature protocols*, 12(6):1198–1228, 2017.
- [10] Orsolya K Wade, Johannes B Woehrstein, Philipp C Nickels, Sebastian Strauss, Florian Stehr, Johannes Stein, Florian Schueder, Maximilian T Strauss, Mahipal Ganji, Joerg Schnitzbauer, et al. 124-color super-resolution imaging by engineering dna-paint blinking kinetics. *Nano letters*, 19(4):2641–2646, 2019.
- [11] Joseph N Zadeh, Conrad D Steenberg, Justin S Bois, Brian R Wolfe, Marshall B Pierce, Asif R Khan, Robert M Dirks, and Niles A Pierce. Nupack: Analysis and design of nucleic acid systems. *Journal of computational chemistry*, 32(1):170–173, 2011.

Response Letter

Journal: Nature Communications

Ref No.: NCOMMS-24-77281B

Title: “Discriminating single-molecule binding events from diffraction-limited fluorescence”

Authors: Yueming Yin, Nithin Pathoor, Kamal Kant Sharma, Shiwen Zhu, Iong Ying Loh, Yan Shan Ang, Shao Ren Sim, Lanry Lin Yue Yung, Thorsten Wohland, Lipo Wang

Dear Reviewers,

Thank you for reviewing. Following your feedback, we have made revisions to improve both the scientific depth and clarity of the manuscript. All changes are highlighted in blue in the revised manuscript and this document. A summary of the revisions and a point-by-point response to your comments are attached below.

Sincerely,

Yueming Yin, Nithin Pathoor, Kamal Kant Sharma, Shiwen Zhu, Iong Ying Loh, Yan Shan Ang, Shao Ren Sim, Lanry Lin Yue Yung, Thorsten Wohland, Lipo Wang

We have made revisions to address all the remaining issues raised by the reviewers. A summary of the changes is provided below:

1. We provided clear explanations on why the classification accuracy of the single-frame analysis is also high (Comment 1.1).
2. We added the survival curves of dwell times as Supplementary Figure 1 and included the corresponding experimental data in the Source Data file alongside other figures. In the caption of Main Text Figure 1, the t_{on} values have been updated to reflect experimental measurements rather than NUPACK estimations. (Comment 2.1).

1 Response to the Reviewer#1’s comments.

We would like to thank Reviewer#1 for the suggestion to improve the quality of the manuscript. The following changes have been made to address Reviewer#1’s remaining comment, where the changes are highlighted in blue in the manuscript.

1.1 The Reviewer comments that

Overall, it is not still clear to me why the classification accuracy of the single-frame analysis is so high, but it may not be possible to understand this completely.

Response: We appreciate the Reviewer’s comment and provide the following clarification to help explain the high classification accuracy observed in the single-frame analysis:

First, we would like to clarify that the single-frame classification accuracy is not the highest among short binding events. As shown in Figure 1 below, the classification accuracy of Domain 1 (8nt-10nt) increases from single-frame to 4-frame binding events, indicating that additional frames up to this point provide more useful information.

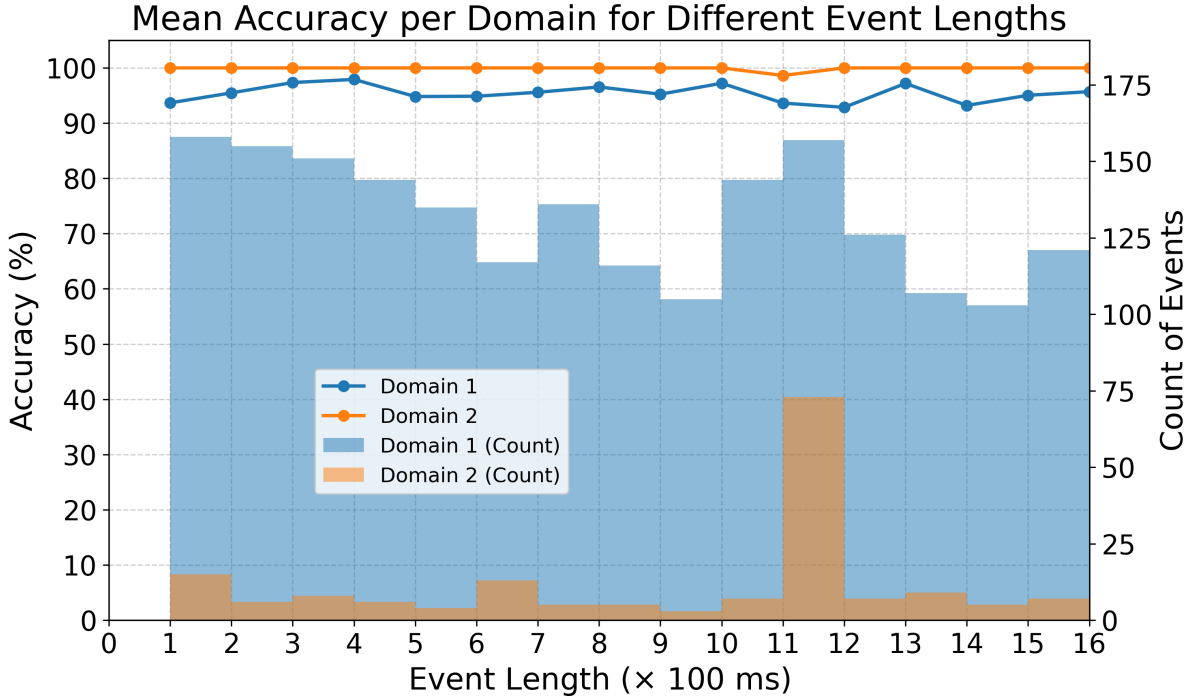
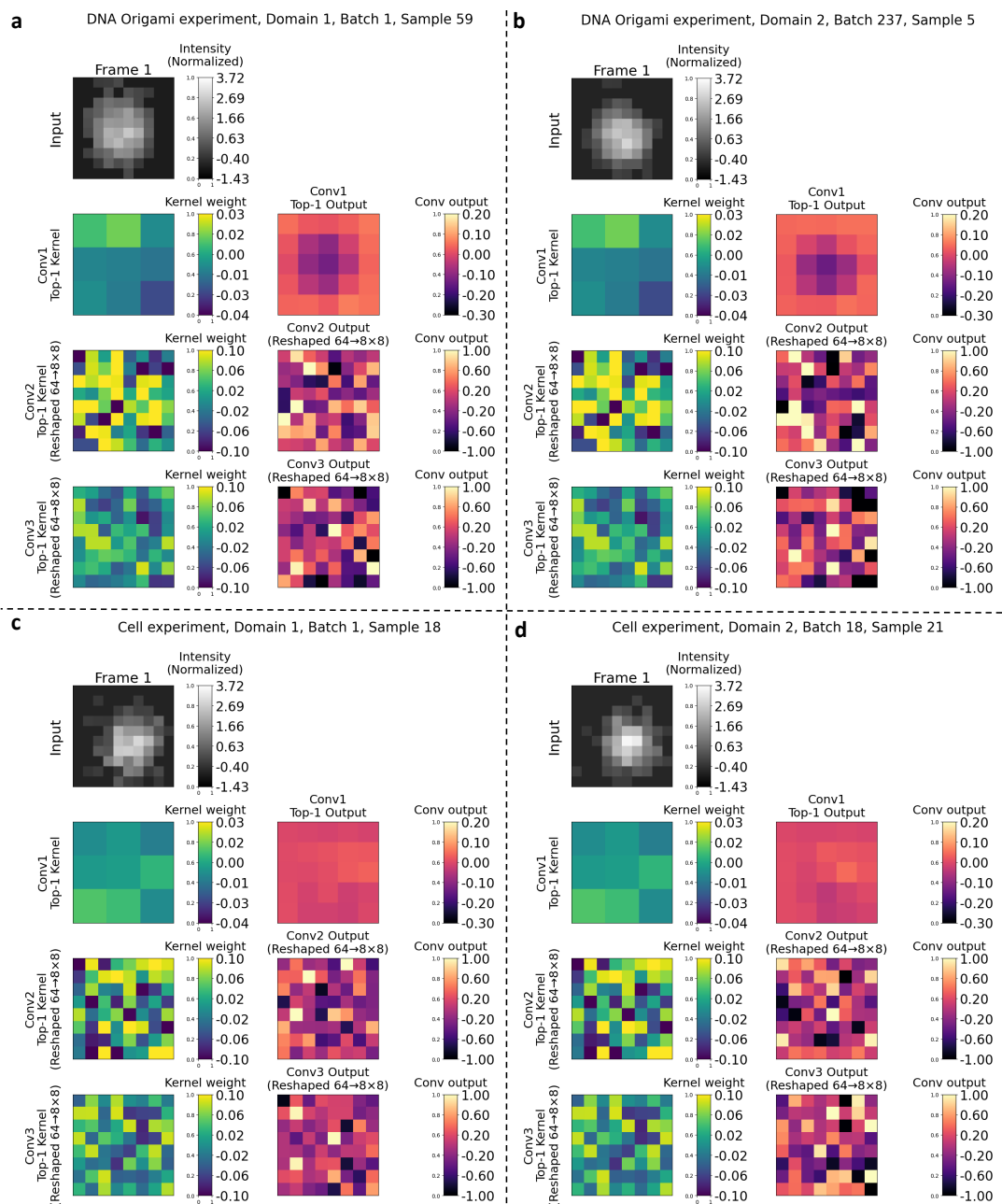


Figure 1: Zoomed-in average classification accuracies of T2C CNN for binding events of varying lengths within 1.6 s. This figure corresponds to a magnified view of Fig. 6d in the main text.

Second, the single-frame events represent an at most 100 ms accumulation of photons within a complete (from dwell to free) but very short binding events. These single frames accumulate, in principle, more binding dynamics than every later single frame in longer binding events with extra binding stability. This trait enables them to be distinguishable as longer binding events.

Last but not least, the proposed T2C CNN is inherently capable of distinguishing single-frame binding events accurately. By reshaping the temporal dimension into the channel axis, even a single accumulated frame provides a sufficiently rich representation for convolutional filters to extract discriminative spatial-frequency features. Combined with cross-connected blocks and a pooling-free design, the model preserves fine-grained intensity patterns and amplifies domain differences (as shown in Supplementary Figure 7 and Supplementary Note 4 below), allowing accurate classification even when temporal context is minimal.

In the Supplementary Figure 7, on Page 19,



Supplementary Figure 7. Example fluorescence spots and T2C CNN convolutional outputs from distinct single-frame binding events. The T2C CNN applies distinct convolutional kernel groups synchronously to each frame within a predefined slice length. Convolutional outputs from kernels within the same group are summed across all frames of an event. Visualizations are shown for the first convolutional layer in each of the first three blocks, referred to as Conv1, Conv2, and Conv3. The top-1 kernel groups were selected globally for visualization, each showing distinct responses to fluorescence spots associated with different domain events. Subfigures **a** and **b** present representative examples from Domain 1 and Domain 2 in DNA origami experiments, respectively, while subfigures **c** and **d** show examples from cell experiments. The right and left legend labels indicate the original and normalized value ranges, respectively, which are consistent across both domains for comparability. Fluorescence spot intensities were mean–standard deviation normalized within each event. For clearer visualization, a global range truncation was applied to both fluorescence spot intensities and convolutional output values. Similar results were observed in 10 randomly selected sets of fluorescence spots.

In the Supplementary Note 4, on Page 5,

...

When analyzing single-frame binding events (Supplementary Fig. 7), we found that their convolutional outputs are generally as weak as the first frame of longer events shown in Supplementary Fig. 5 and Supplementary Fig. 6. Despite this weakness, each of the 64 convolutional outputs from the first convolutional layer (Conv1) contributes collectively to the subsequent convolutional layers (Conv2 and Conv3), resulting in distinct output patterns in Conv2 and Conv3 (last two rows of each subfigure in Supplementary Fig. 7). For example, in the last row of Supplementary Fig. 7a and Supplementary Fig. 7b, the Conv3 output of the Domain-1 sample exhibits three major high-value regions, while the Domain-2 sample shows four, located at different positions. They also differ significantly at specific positions, such as the last two columns of the 4th and 5th rows, and the first three columns of the 4th to 7th rows of the reshaped Conv2 output map. These apparent differences, abstracted from similar fluorescence spots using the same convolutional kernels, enable accurate classification between domains.

...

2 Response to the Reviewer#2's comments.

We would like to thank Reviewer#2 for the suggestion to improve the quality of the manuscript. The following changes have been made to address Reviewer#2's comments, where all the changes are highlighted in blue in the manuscript and the supplementary material.

2.1 The Reviewer comments that

Regarding the response to reviewer 3.1 comment, please include the data (such as survival curves of dwell times) to support the t_{on} estimates (as stated in figure caption 1: 2.0 ± 1.2 s for Domain 1 and 8.0 ± 6.5 s for Domain 2 via NUPACK).

Response: We have added the survival curves of dwell times as Supplementary Figure 1 and included the corresponding experimental data in the Source Data file alongside other figures. Their file names are:

- Supplementary Figure 1 - Dwell times (Domain 1).csv
- Supplementary Figure 1 - Dwell times (Domain 2).csv
- Supplementary Figure 1 - Survival probabilities (Domain 1).csv
- Supplementary Figure 1 - Survival probabilities (Domain 2).csv

In the caption of Main Text Figure 1, the t_{on} values have been updated to reflect experimental measurements rather than NUPACK estimations. The revised parts are attached below.

(continued on next page)

Revised Figure 1 caption in the manuscript:

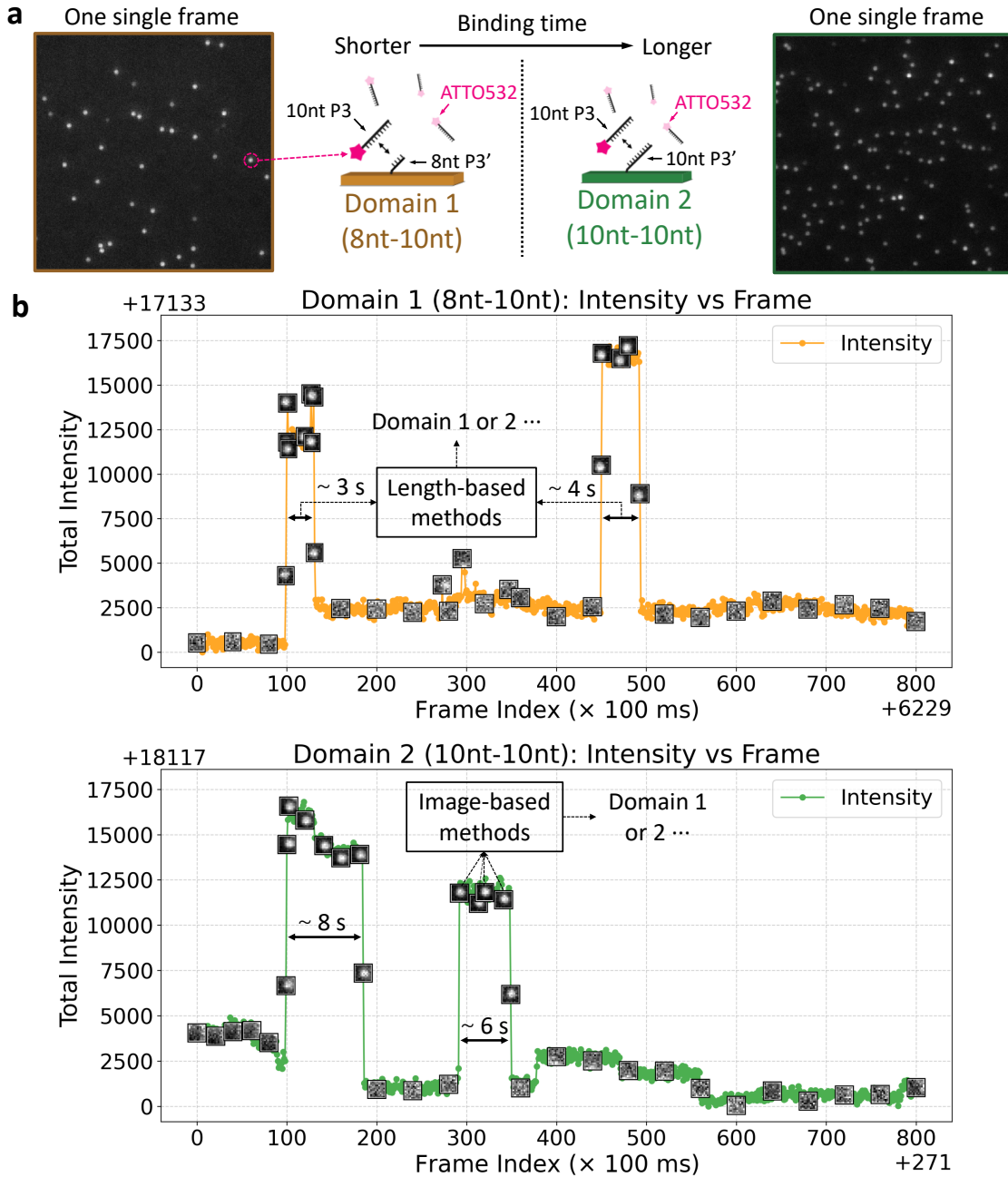
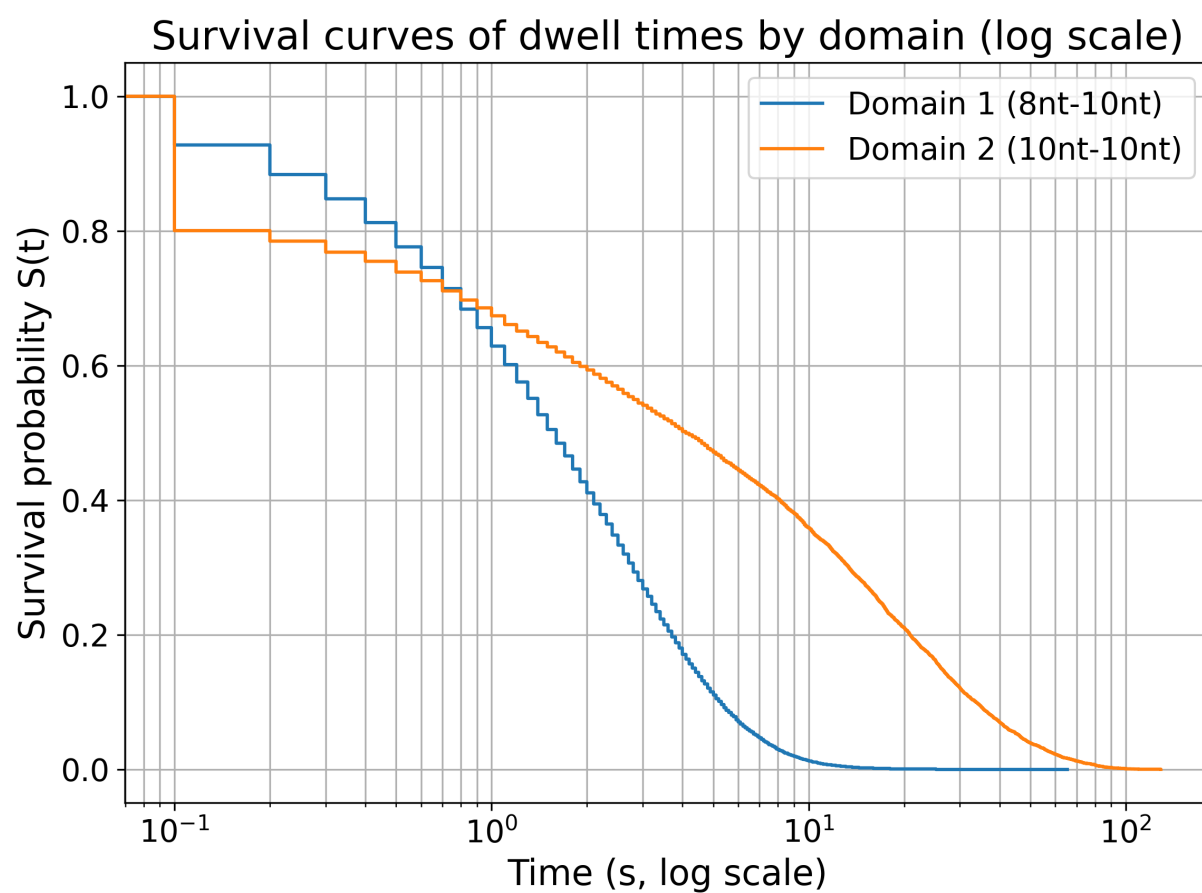


Figure 2: An example of raw images and time courses from the DNA-PAINT experiments used in this study. **a)** Schematic of the two DNA-binding domains used to collect diffraction-limited fluorescent spots (adapted from [1]). Both domains share the same dye (“ATTO532”) and strand (“10nt P3”), but differ 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.

Added Supplementary Figure 1 in the revised supplementary material:



Supplementary Figure 1. Survival curves of dwell times by domain. Survival curves were derived from the raw experimental data corresponding to Main Text Figure 1. Source data are provided as a [Source Data](#) file.

References

- [1] Orsolya K Wade, Johannes B Woehrstein, Philipp C Nickels, Sebastian Strauss, Florian Stehr, Johannes Stein, Florian Schueder, Maximilian T Strauss, Mahipal Ganji, Joerg Schnitzbauer, et al. 124-color super-resolution imaging by engineering dna-paint blinking kinetics. *Nano letters*, 19(4):2641–2646, 2019.