Bound2Learn: A Machine Learning Approach for Classification of DNA-Bound Proteins from Single-Molecule Tracking Experiments

Nitin Kapadia, Ziad W. El-Hajj, and Rodrigo Reyes-Lamothe*

Department of Biology, McGill University, 3649 Sir William Osler, Montreal, Qc, H3G 0B1 Canada

* Correspondence to: rodrigo.reyes@mcgill.ca

Abstract

Many proteins act on DNA for a wide range of processes, including DNA replication, DNA repair, and transcription. Their time spent on DNA can provide insight into these processes and their stability within complexes to which they may belong. Single-particle tracking allows for direct visualization of protein-DNA kinetics, however, identifying whether a molecule is bound to DNA can be non-trivial. Further complications arise with tracking molecules for extended durations in cases of processes with slow kinetics. We developed a machine learning approach, using output from a widely used tracking software, to robustly classify tracks in order to accurately estimate residence times. We validated our approach *in silico*, and in live-cell data from *Escherichia coli* and *Saccharomyces cerevisiae*. Our method has the potential for broad utility and is applicable to other organisms.

Introduction

Quantitative information regarding the kinetics of a protein provides valuable insight in the behaviour of the respective protein, as well as its relationship with other proteins in cases where it belongs to a protein complex. This in turn may inform on the activity of the protein. The residence times of DNA-bound proteins (DBP) can reveal important details on basic cellular processes such as transcription, DNA repair, and DNA replication, at the timescales at which they operate [1-7]. This is true of not only proteins that bind directly at sites on DNA such as initiator proteins, repair proteins, and chromatin remodellers, but also proteins that are part of complexes that translocate on DNA such as the DNA replication complex (replisome) and RNA polymerase (RNAP) [1, 8].

Recent advances in fluorescence microscopy have allowed us to study protein kinetics directly in living cells, with the most common techniques being single-particle tracking (SPT), fluorescence recovery after photobleaching (FRAP), and fluorescence correlation spectroscopy

(FCS) [3, 4, 8-13]. SPT has the particular advantage of being able to directly observe protein behaviour, allowing for a wealth of information to be extracted from the images, both qualitatively and quantitatively [14, 15]. Typically, SPT has been used to determine binding kinetics of DBP with very fast kinetics (hundreds of milliseconds to a few seconds), by using capture rates of few to tens of milliseconds. However, certain processes operate on much longer timescales posing issues for SPT, including photobleaching and tracking single-molecules for long durations. One way to bypass these issues is by using long-exposure times to blur out diffusing molecules, in combination with low-laser power and stroboscopic illumination to minimize photobleaching. Nonetheless, tracking single-molecules over long durations can be difficult due to intensity fluctuations caused by the molecule moving out of focus and photophysics, resulting in fragmentation of tracks [1, 16] (Figure 1A). Reducing the intensity threshold required for localizing molecules can compensate for this, although with the disadvantage of introducing false positives. In addition, even with motion blurring and tracking parameters to select for only bound molecules, it is possible to detect diffusing molecules or have false positives due to localization and tracking errors; therefore filtering steps on tracks need to be included to only isolate tracks representing DNA-bound proteins [1, 2, 17, 18]. All these complications arise from the tracking algorithms and automated analysis, since the user can typically distinguish immobile DBP when looking at the raw images when using long-exposure and stroboscopic illumination as they appear to wiggle around.

Here, we provide a user-friendly and robust method to determine residence times of DBP using machine learning and SPT in live *Escherichia coli* (*E.coli*) and *Saccharomyces cerevisiae* (budding yeast), that can be easily extended to other organisms.

Results

Random Forest for Single-molecule tracking classification

Machine learning (ML), and its branch of deep-learning, is a powerful tool for image analysis and classification, with the most common implementation being supervised learning, whereby a labelled training data set is given to the ML algorithm, which then builds a model for subsequent classification [19, 20]. Our motivation arose from recognizing the limitations associated with automated detection and tracking, and we wanted to develop a classification method to compensate for these limitations. We recognized that a known DBP (e.g. histone H3)

could be used to construct a training data set manually, for how DBP in general should move and how single-molecules should look like in images, from which we can build a ML model. We can also use this control as a photobleaching control when we estimate the residence time for our DNA-bound protein of interest (POI) [1].

We used Trackmate, a freely available plugin in Fiji, to track molecules and used variables from the tracking output that we believed to be good predictors of single-molecule DBP and help to reduce the cross-validation error (CVE) (Figure 1B) [21]. To predict whether a molecule moves like a DBP we used the following variables from the tracking output: mean speed, maximum speed, minimum speed, and median speed. For predicting whether the track represented a genuine molecule, we used the maximum quality variable – a parameter corresponding to the intensity of the molecule as well as its shape [21]. Classification of the track was done manually by looking at the raw image data, with binary classification: 0 for a diffusing molecule or noise, and 1 for a track representing a genuine immobile, DNA-bound molecule. We chose to use the random forest algorithm to construct a model from the training data, as it is accurate, tolerant of noise, and less prone to overfitting [22] (Figure 1C). The various parameters used to construct the different random forest models are listed in **Supplementary Table 1.** We used this model to classify tracks for our POI (Figure 1C), and subsequently determine their residence time after correcting for photobleaching.

We first tested our approach using computer simulations of single-molecule timelapses in *E.coli* and nuclei of budding yeast (See Materials and Methods). The main difference between the two simulations aside from cell shape, was that the budding yeast simulation had a distortion in the shape of the molecule based on its position in z, to model the point spread function (PSF) and emulate the molecule going out of the focal depth of the objective. Under our experimental conditions, *E.coli* cells have diameters of 0.7μm, which is not far from the estimated focal depth of high numerical aperture objectives commonly used in single-molecule studies (~0.4 μm), so we assumed molecules to be in focus, regardless of their z position[1]. We first constructed training data sets from simulated data with 500ms exposure, no time interval, and a mean bleach time of 10s (**Supplementary Table 2)**. We had a stable, bound fraction (Diffusion coefficient (D) of D_{bound}= 0.005μm^2/s, bound time >>> bleach time), along with a mobile fraction (D_{mobile} = 0.5μm^2/s) **(Figure 1D).**

We only considered tracks with ≥ 4 localizations, as it is difficult to discern the state of the molecule for shorter tracks. In order to make use of a single training data set, instead of constructing multiple training data sets for classifying data sets collected under different conditions, we constructed two models: ML model 1 only has the speed variables and ML model 2 has the speed variables along with maximum quality (**Figure 1D)**. We calculate the mean of mean speed as well as the mean of maximum quality from the tracks classified as being bound in the training data and use these values to scale the speed and quality variables, respectively, for data collected with different time intervals and/or illumination intensities **(Supplementary Figure S1).**

The entire procedure for the classification of tracks for a POI is as follows:

1) We performed a two-component Gaussian mixture model (GMM) fit on the log of mean speed, of all the tracks, and the component with the lowest mean is selected as representing the immobile molecules. This step does not need to be robust as it will be used to do some initial filtering for subsequent steps.
2) Using this value, we scale the speed variables accordingly, using the mean of mean speed calculated from the training data, as follows: scale $\text{factor}_{\text{speed}}$ = mean (mean $\text{speed}_{\text{training}}$)/mean(mean $\text{speed}_{\text{boundPOI}}$). This helps to ensure the ML models can classify tracks obtained from different time intervals than the training data. We then run the tracks through ML model 1 for initial filtering.
3) From the resulting tracks, we calculate the mean of maximum quality using a two-component GMM, similar to step 1 except selecting the component with the higher mean and use the same variable from the training data to scale the quality variable, similar to the previous step. We then run the tracks through ML model 2 for final classification.

To assess how well the procedure worked, we quantified both the accuracy (proportion of tracks accurately predicted to be bound), and the recovery error (RE) (the fraction of tracks known to be representing bound molecules recovered by the classification procedure). While high classification accuracy is important to ensure accurate estimation of residence time, we also wanted to make sure that the recovery error is small, as single-molecule studies are often plagued by low sample sizes thereby reducing confidence in the estimate [1, 14]. In addition, the out-of-bag (OOB) errors, equivalent to the CVE for random forests [22], were estimated for each ML model **(Supplementary Table 1)**. On test data, we obtained an accuracy of 0.96 and RE of 0.10

for the *E.coli* simulation, and an accuracy of 0.97 and RE of 0.06, for the budding yeast simulation (**Supplementary Table 3)**.

Accurate estimation *in silico* of residence times under different conditions

Next, we tested how well the ML models would work on simulated data with different time intervals and spot intensities (**Supplementary Table 2)**. First, we tested on data with a 1s time interval and spot intensities the same as the training data set. The residence time was set to 8s, while the bound fraction was set to 0.5. As **Figure 2A** (top) illustrates, we first calculate the mean of mean speed of the bound population, in order to scale the speed variables. After the final classification step, for both the *E.coli* simulation and the budding yeast, we were able to obtain an estimate for the residence time that was within error of 8s (**Figure 2A** (bottom)). In addition, the accuracy values were 0.93 and 0.96, with recovery errors of 0.11 and 0, respectively, for the *Ecoli* and budding yeast simulations (**Supplementary Table 4)**. We then tested whether we can obtain an accurate estimate in a situation representing poorer image quality (e.g. lower spot intensity). As **Figure 2B** (top) illustrates, we use the GMM fit on the max quality values to find an appropriate scaling factor. After classification, we find that we can still recover an estimate of the residence that is within error of the known bound time, for both simulations, with accuracy values and REs similar to the previous condition (**Figure 2B, Supplementary Table 4)**.

While a 500ms exposure helps to blur out diffusing molecules, it might be too long for faster processes. Therefore, we constructed a training data set with 100ms exposure and tested on data with a mean residence time set to 1s while the bleach time for this condition was set to 2s (20 frames) (**Figure 2C)**. Once again, we were able to obtain an estimate within error of 1s, although with lower accuracy and higher recovery errors than with 500ms, as one would expect under conditions where the separation between two diffusive states is more difficult to discern. (**Figure 2C)**. We tested this by changing D_{mobile} to 5 $\mu m^2/s$, and found the errors were drastically reduced **(Supplementary Table 4)**. We then asked whether the 500ms training data can be used to classify the 100ms data, and surprisingly, we found that it still performed well (**Figure 2C)**, with comparable errors to the 100ms training data (**Supplementary Table 4)**. We also found that we could get accurate estimates of the residence times of a heterogenous population of bound molecules, where two distinct binding regimes are present (**Figure 2D).** We note that for the simulation representing a heterogenous population of bound molecules, we had to change the

bleach time to 10s, in order to recover the two binding times. As others have alluded to, detecting multiple populations is highly dependent on acquisition settings which is why we had to change the bleach time to recover binding times of 1s and 7s [23]. These results suggest that one can use ML models constructed from a single training data set and use them for classifying tracks obtained from data collected under widely different conditions, along with obtaining accurate residence times of homogenous and heterogenous bound populations.

Experimental validation in *E.coli*

We then wanted to determine if our approach could work with estimating residence times in *E.coli*. For our DNA-bound control we used LacI, a transcriptional repressor, fused to the photoconvertible mMaple, which upon illumination with 405nm light, converts to the red form [24]. LacI has been reported to bind stably to the *lacO* array site on DNA (~5 minutes), thus making it suitable as a photobleaching control, under our acquisition times [25] **(Figure 3A)**. Although we had initially used this strain for photobleaching correction in [1], we decided to use it for constructing a training data set as well. The LacI data used for the training data was collected with 500ms exposure, as fast as possible.

Once the ML models were constructed, we asked whether we could get an accurate estimate of the residence time for *E.coli* DNA polymerase, PolIII, for which we had previously reported to have a residence time of around ~10s, [1]. We reanalyzed a dataset from [1] taken of the PolIII subunit, ε, also tagged with mMaple (dnaQ-mMaple), to determine if the ML approach could give identical results (**Figure 3B)**. We found that we were able to recover an estimate of the residence time that was very similar to the previously reported estimate **(Figure 3C),** and with small confidence intervals, confirming the robustness of our approach in estimating residence times of DBP in *E.coli*.

Estimating Residence Time of Topoisomerase II in Budding Yeast

We asked if our approach could be used in live haploid budding yeast. For our photobleaching control and training data construction, we used histone H3 fused to HaloTag (H3-HaloTag), due to its expected long residence time and high bound fraction [26] (**Figure 4A)**. In order to detect the protein, we incubated the cells with the cell-permeable photoactivatable (PA)

dye, JF-PA549, and the experimental protocol was very similar to that described in *E.coli* [1], including highly-inclined and laminated optical (HILO) sheet stroboscopic illumination with 500ms exposure, but with cycles of low-dose 405nm activation every 40 frames **(Materials and Methods)**. To improve image quality with budding yeast, we also used the refractive-index matching media, Optiprep, to minimize light refraction due to the cell wall [27]. The mean photobleaching duration estimated, with 1s interval acquisition, was ~23 seconds **(Materials and Methods)**.

For our experimental test, we tagged Topoisomerase II (Top2) with HaloTag (Top2-HaloTag) and to segment nuclei in order to isolate tracks found only in nuclei, we tagged proliferating nuclear cell antigen (PCNA) with mNeonGreen and acquired a z-stack in the green channel prior to acquisition (**Figure 4B)**. The resulting z-stack went through a smooth manifold extraction (SME) process for better image quality of the nuclei [28]. We chose Top2 as mammalian Top2 enzymes have been reported to be dynamic and can complement Top2 function in budding yeast *Δtop2* strains, suggesting they function similarly, albeit with the possibility of slightly different dynamics [13]. We collected the data under the same acquisition settings as H3-HaloTag. As indicated in **Figure 4B** with the red arrows, we could visualize dissociation of the protein under our acquisition settings, suggesting it is dynamic. Following tracking analysis and ML classification, we obtained an estimate for the residence time of Top2 of ~30s, consistent with dynamic behaviour reported previously (**Figure 4C)**. As one can notice, the spot quality in live budding yeast, is of slightly poorer quality compared to *E.coli* (compare **Figure 3B vs Figure 4B)**, but we surprisingly found our ML approach still worked quite well in this system, suggesting our approach can be used in a range of conditions and organisms.

Discussion

Here we have provided a robust, easy-to-use classification approach that can be used on long-capture single-molecule data, to isolate tracks of DBP in order to accurately estimate their residence times. While there are alternatives to classify bound molecules such as estimating their diffusion coefficient, step-size per frame of DBP, PSF, etc., and applying a threshold for acceptance, they often suffer from not being precise estimators or require additional analysis steps and/or experiments, and setting hard thresholds may discard too many tracks [1, 2, 18, 29].

We do acknowledge the limitations with our approach, namely that in cases where the diffusing fraction has similar mobility to the bound fraction, the error in classification will increase, although this would likely be the case for most classifiers. Also, constructing training data sets manually can take significant time and choosing the right parameter values for building random forest models does require some expertise; however, our results suggest that once the initial models have been created, they can be used across a range of conditions whether that be different exposure times, time intervals, and data quality, not to mention the potential to estimate residence times in a high-throughput manner for cases where one is interested in an entire protein complex. Finally, we did have to adjust our simulation parameters (akin to adjusting experimental conditions) to detect two binding regimes in cases where there was a heterogenous population of bound molecules, similar to what a recent study has reported [23]. Nonetheless, this is not an issue with our classification approach but rather the acquisition settings used (or in our case, simulation parameters), and our results suggest simulations can help identifying the optimal acquisition conditions to use for detecting multiple binding states.

Finally, although we have shown our ML approach, along with the experimental protocol, works well in small organisms such as *E.coli* and budding yeast, it should be easily extendable to other systems such as mammalian cells.

References

1. Beattie, T.R., et al., *Frequent exchange of the DNA polymerase during bacterial chromosome replication.* Elife, 2017. **6**.
2. Uphoff, S., et al., *Single-molecule DNA repair in live bacteria.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(20): p. 8063-8068.
3. Mazza, D., et al., *A benchmark for chromatin binding measurements in live cells.* Nucleic Acids Research, 2012. **40**(15).
4. Hager, G.L., J.G. McNally, and T. Misteli, *Transcription Dynamics.* Molecular Cell, 2009. **35**(6): p. 741-753.
5. Mehta, G.D., et al., *Single-Molecule Analysis Reveals Linked Cycles of RSC Chromatin Remodeling and Ace1p Transcription Factor Binding in Yeast.* Mol Cell, 2018.
6. Stracy, M., et al., *Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid.* Proceedings of the National Academy of Sciences of the United States of America, 2015. **112**(32): p. E4390-E4399.
7. Xie, X.S., et al., *Single-molecule approach to molecular biology in living bacterial cells.* Annual Review of Biophysics, 2008. **37**: p. 417-444.
8. Sprouse, R.O., et al., *Regulation of TATA-binding protein dynamics in living yeast cells.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(36): p. 13304-13308.

9. Gebhardt, J.C.M., et al., *Single-molecule imaging of transcription factor binding to DNA in live mammalian cells.* Nature Methods, 2013. **10**(5): p. 421-+.
10. Mueller, F., et al., *FRAP and kinetic modeling in the analysis of nuclear protein dynamics: what do we really know?* Current Opinion in Cell Biology, 2010. **22**(3): p. 403-411.
11. Manley, S., et al., *High-density mapping of single-molecule trajectories with photoactivated localization microscopy.* Nature Methods, 2008. **5**(2): p. 155-157.
12. Sprague, B.L., et al., *Analysis of binding reactions by fluorescence recovery after photobleaching.* Biophysical Journal, 2004. **86**(6): p. 3473-3495.
13. Christensen, M.O., et al., *Dynamics of human DNA topoisomerases II alpha and II beta in living cells.* Journal of Cell Biology, 2002. **157**(1): p. 31-44.
14. Elf, J. and I. Barkefors, *Single-Molecule Kinetics in Living Cells.* Annu Rev Biochem, 2018.
15. Yu, J., *Single-Molecule Studies in Live Cells.* Annual Review of Physical Chemistry, Vol 67, 2016. **67**: p. 565-585.
16. Etheridge, T.J., et al., *Quantification of DNA-associated proteins inside eukaryotic cells using single-molecule localization microscopy.* Nucleic Acids Research, 2014. **42**(19).
17. Callegari, A., et al., *Single-molecule dynamics and genome-wide transcriptomics reveal that NF-kB (p65)-DNA binding times can be decoupled from transcriptional activation.* Plos Genetics, 2019. **15**(1).
18. Ball, D.A., et al., *Single molecule tracking of Ace1p in Saccharomyces cerevisiae defines a characteristic residence time for non-specific interactions of transcription factors with chromatin.* Nucleic Acids Res, 2016. **44**(21): p. e160.
19. Moen, E., et al., *Deep learning for cellular image analysis.* Nat Methods, 2019. **16**(12): p. 1233-1246.
20. Angermueller, C., et al., *Deep learning for computational biology.* Mol Syst Biol, 2016. **12**(7): p. 878.
21. Tinevez, J.Y., et al., *TrackMate: An open and extensible platform for single-particle tracking.* Methods, 2017. **115**: p. 80-90.
22. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
23. Ho, H.N., et al., *Identification of Multiple Kinetic Populations of DNA-Binding Proteins in Live Cells.* Biophysical Journal, 2019. **117**(5): p. 950-961.
24. McEvoy, A.L., et al., *mMaple: A Photoconvertible Fluorescent Protein for Use in Multiple Imaging Modalities.* Plos One, 2012. **7**(12).
25. Hammar, P., et al., *Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation.* Nature Genetics, 2014. **46**(4): p. 405-+.
26. Kimura, H. and P.R. Cook, *Kinetics of core histones in living human cells: Little exchange of H3 and H4 and some rapid exchange of H2B.* Journal of Cell Biology, 2001. **153**(7): p. 1341-1353.
27. Boothe, T., et al., *A tunable refractive index matching medium for live imaging cells, tissues and model organisms.* Elife, 2017. **6**.
28. Shihavuddin, A., et al., *Smooth 2D manifold extraction from 3D image stack.* Nature Communications, 2017. **8**.
29. Michalet, X. and A.J. Berglund, *Optimal diffusion coefficient estimation in single-particle tracking.* Phys Rev E Stat Nonlin Soft Matter Phys, 2012. **85**(6 Pt 1): p. 061916.
30. Javer, A., et al., *Persistent super-diffusive motion of Escherichia coli chromosomal loci.* Nature Communications, 2014. **5**.
31. Hajjoul, H., et al., *High-throughput chromatin motion tracking in living yeast reveals the flexibility of the fiber throughout the genome.* Genome Research, 2013. **23**(11): p. 1829-1838.
32. Weber, S.C., A.J. Spakowitz, and J.A. Theriot, *Bacterial Chromosomal Loci Move Subdiffusively through a Viscoelastic Cytoplasm.* Physical Review Letters, 2010. **104**(23).

Acknowledgements

We thank Tatiana Karpova (NCI) for providing the yeast strains carrying H3-Halo and pdr5Δ;Luke Lavis (Janelia Farms) for providing the Halo substrate coupled to PA-JF549 dye. **Funding:** The experiments were done using equipment from the Integrated Quantitative Bioscience Initiative (IQBI), funded by CFI 9. This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC# 435521-2013), the Canadian Institutes for Health Research (CIHR MOP# 142473), the Canada Foundation for Innovation (CFI# 228994), and the Canada Research Chairs program. **Author contributions:** NK, Conceptualization, Formal analysis, Investigation, and Writing. ZWEH, Formal analysis, Investigation, RR-L, Supervision, Funding acquisition. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** All data, code, and materials used in this work are available by request.

Figures

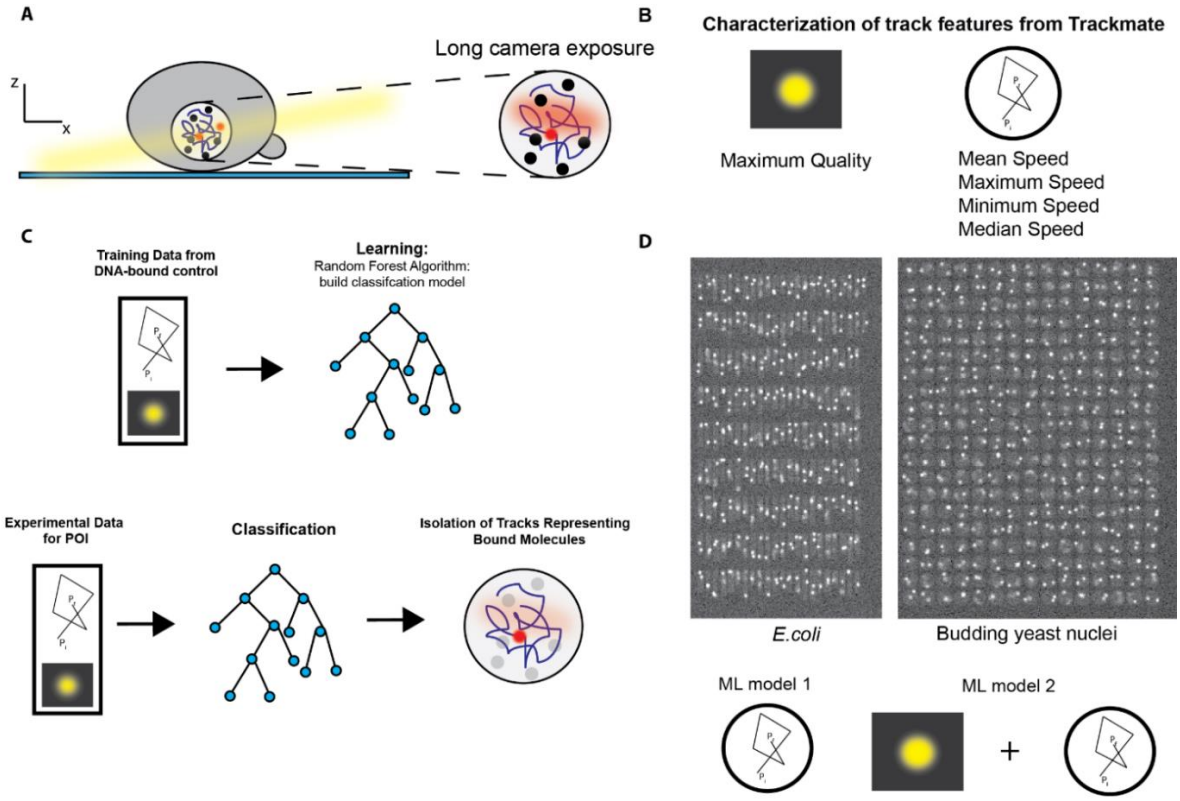


Figure 1- Approach to isolate DNA-bound molecules. A) Diagram of experimental setup to do single-molecule tracking with photoactivatable/photoconvertible fluorophores to calculate residence times on DNA. B) Variables from Trackmate used to predict if a track represents a genuine bound molecule. C) Illustration of the general procedure for ML and how it can be used to predict tracks from experimental data for POI. D) Top - Example images from simulations of single-molecule movies representing *E.coli* cells and budding yeast nuclei. Bottom – Illustration of the variables used in ML models 1 and 2.

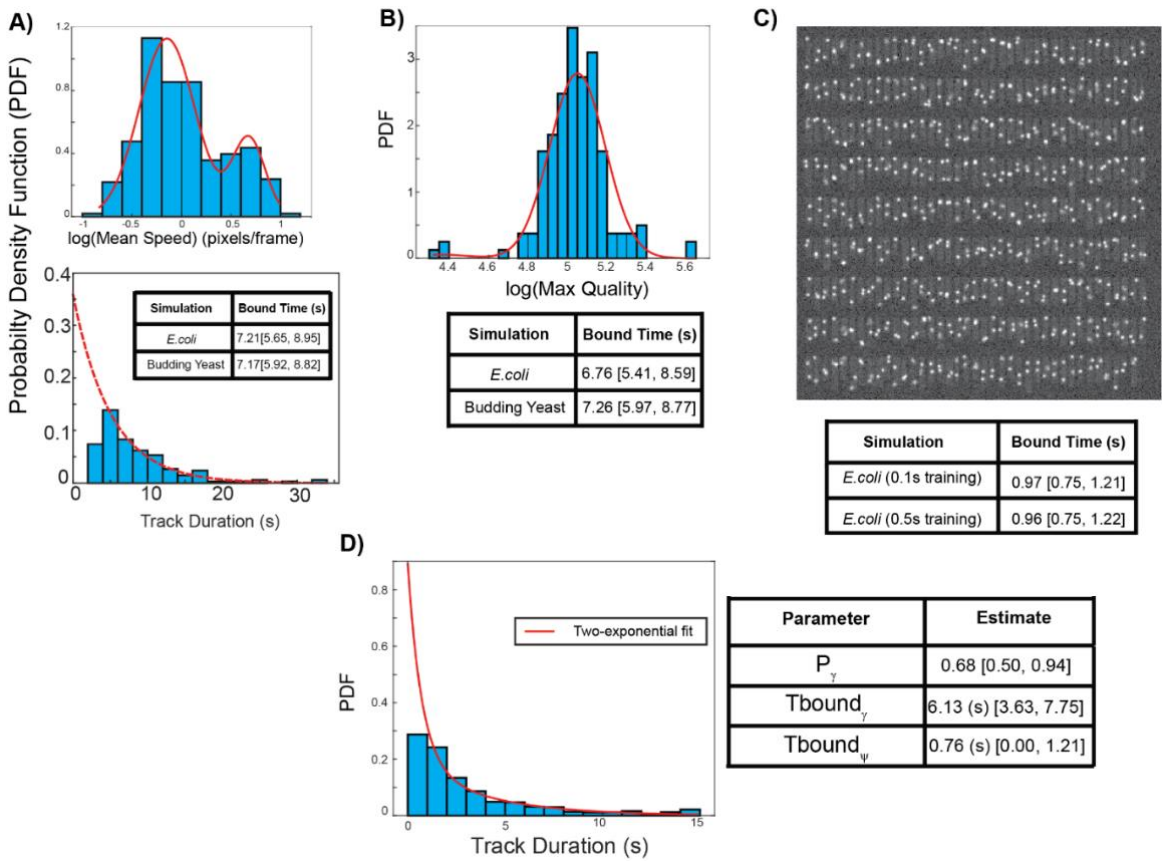
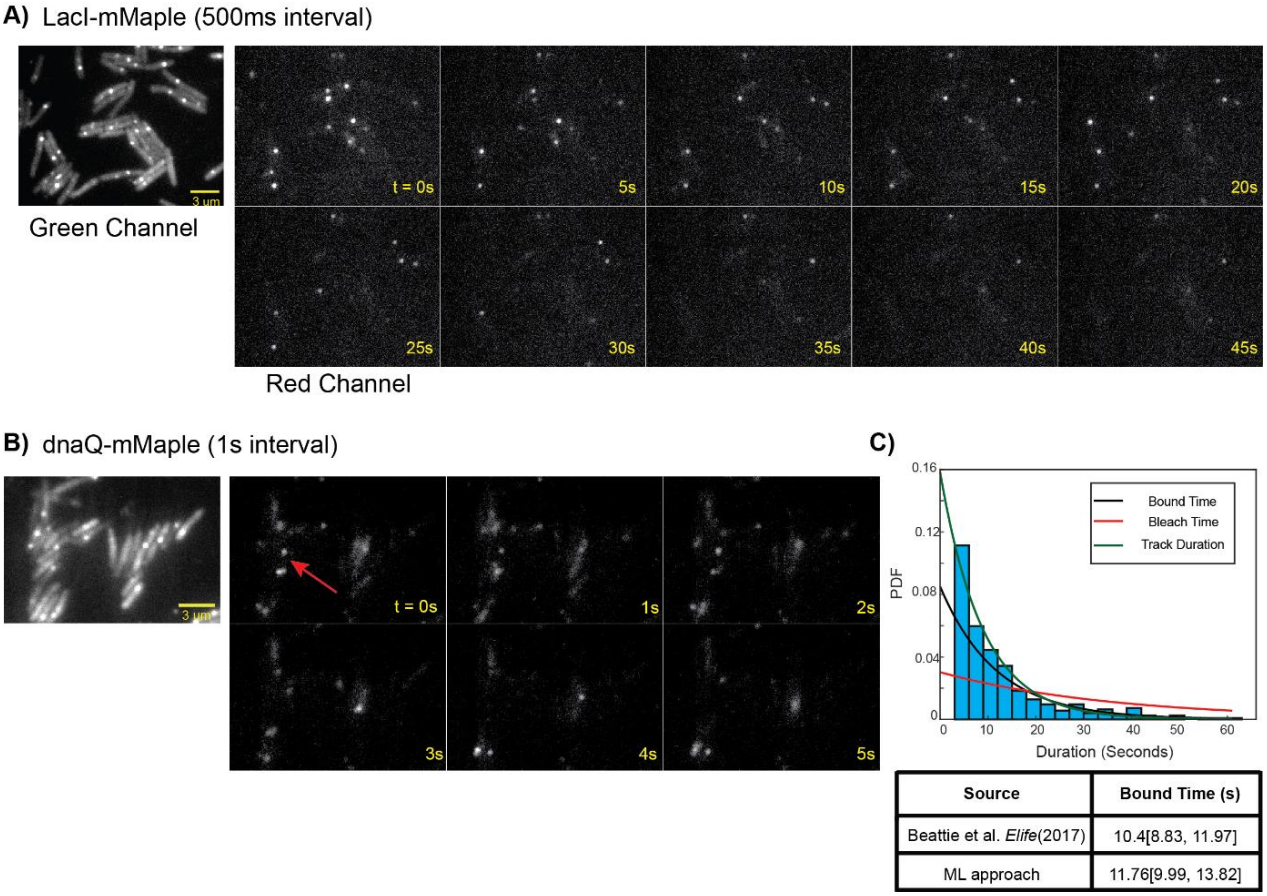


Figure 2 – Estimating residence times in a range of conditions. A) Top – Diagram illustrating the GMM fit to determine the mean of mean speed for 1s interval data, which can be used to rescale the speed variables. Bottom – Estimates of residence times in both simulations. B) Top – Diagram illustrating the GMM fit to maximum quality values from a data set with lower fluorescent intensities. Bottom – Estimates of residence times from simulated data representing poorer image quality. C) Top – Example of image from simulated 100ms timelapse in *E.coli*. D) Two-exponential fitting to a data set with heterogenous population of bound molecules. For all estimates, 95% confidence intervals are shown.



Source	Bound Time (s)
Beattie et al. *Elife*(2017)	10.4[8.83, 11.97]
ML approach	11.76[9.99, 13.82]

Figure 3- Estimating residence times for PolIII subunit, ε - A) Example of LacI-mMaple data collected with 500ms exposure. Red channel represents image after 405 induced photoconversion of mMaple. B) Example of dnaQ-mMaple timelapse. Red arrow indicates molecule that we observed dissociating. C) Estimation of residence times and comparison to estimate from [1], with 95% confidence intervals.

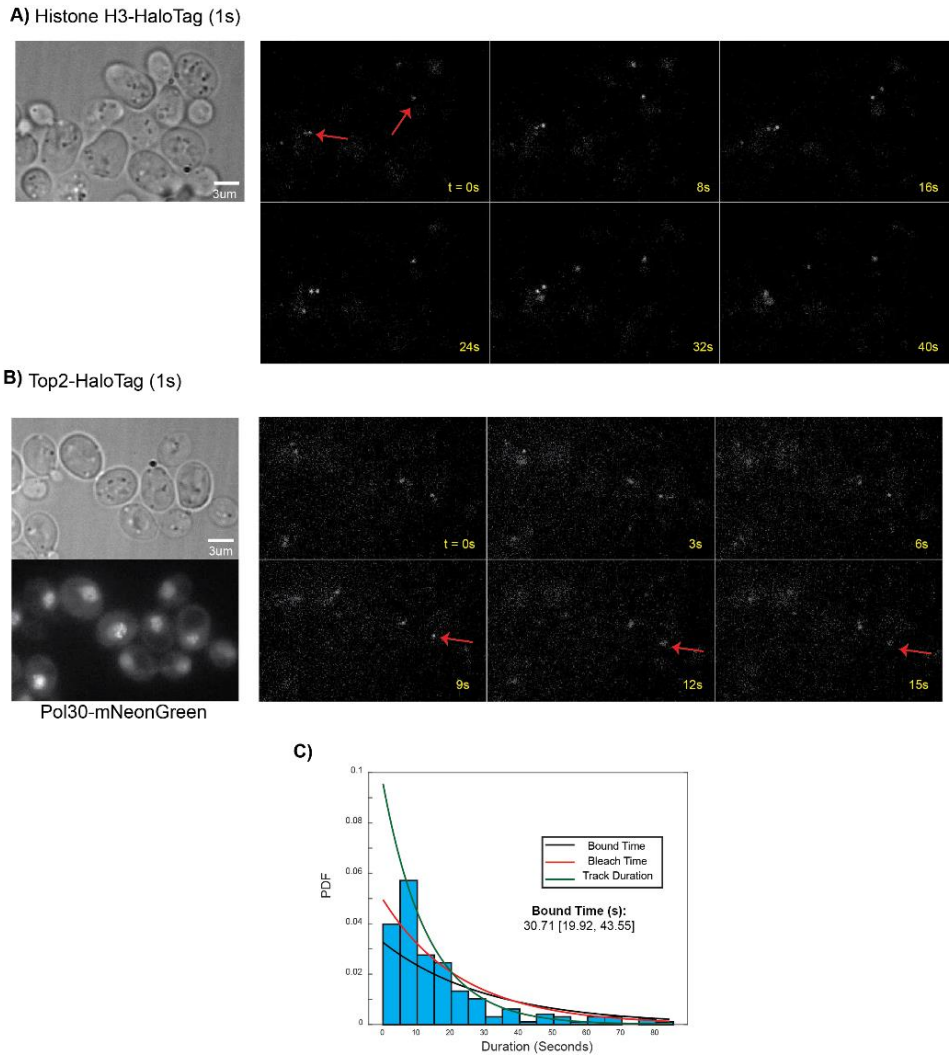


Figure 4 -Estimating residence time of Top2. A) Example of H3-HaloTag images, with timelapse shown after photoactivation with 405. B) Example of Top2-HaloTag timelapse after photoactivation. Red arrows indicate molecule we observed dissociating. Also shown is the SME of a Pol30-mNeonGreen z-stack. C) Estimation of residence time with 95% confidence intervals.

Materials and Methods

Computer Simulations of Single-Molecule Timelapses

Computer simulations of images were written in Python 3.6.

First, a 5000x5000 array was constructed as an image. Each array element was 10nm, therefore the total size was meant to represent 50µm x 50µm. Cells were placed in a grid-like pattern to prevent overlap and given a low intensity value of around 0.2 per array element to represent cell autofluorescence. *E.coli* cells were modelled as 3D rectangles, with average width of 0.7µm, and an average length of 3µm., while budding yeast nuclei were modelled as spheres of average diameter 2µm. For each cell, unless stated otherwise, two fluorescent molecules were assigned such that their initial locations were confined to the interior of their respective cell. A weighted sampling method was used to assign their initial diffusive state.

To model fluorescent spots, a spot intensity was assigned at the center of the molecules and using a standard deviation of 130nm, the intensity was spread across the region using a Gaussian filter. In the case of the spots in the budding yeast simulation, the standard deviation used was increased by 5% for every 150nm increase/decrease in z position of the molecule's position. For example, from +/- 150nm from the z-origin of 0, the molecule was assumed to be in focus. However, if the molecule moved to a |z position| of >150nm but less than 300nm, its standard deviation was increased by 5%. After the Gaussian filter, poisson shot noise at each element was added by sampling from a poisson distribution with mean equal to the initial intensity value at that element.

Kinetics of the molecule were determined using transition matrices: one for the photophysics (e.g. photobleaching), and the other for transitions to different diffusive states. We determined the probability of transition using a time step (τ) of 5ms. We omitted photoblinking as the lower laser powers used in long-exposure experiments typically do not cause it. This simplified the photophysics transition matrix to:

	$F_{state}(t+\tau)$	$B_{state}(t+\tau)$
$F_{state}(t)$	$1-CDF_{bleach}$	CDF_{bleach}
$B_{state}(t)$	0	1

Where the F_{state} is the fluorescent state, B_{state} is the irreversible photobleached state, and CDF_{bleach} is the probability of photobleaching within τ, calculated from the cumulative distribution function (CDF) of an exponential distribution with a specified mean bleach time. If the molecule went to the $B_{state,}$ it was removed from subsequent iterations.

The transition matrix for transitions to different diffusive states was:

	$Mobile_{state}$ $(t+\tau)$	$Bound_{state}(t+\tau)$
$Mobile_{state}(t)$	$1\text{-}CDF_{search}$	CDF_{search}
$Bound_{state}(t)$	CDF_{bound}	$1\text{-}\ CDF_{bound}$

Where $Mobile_{state}$ represents the state with $D = D_{mobile}$, and $Bound_{state}$ represents the state with $D = D_{bound,}$ CDF_{search} is probability that the molecule switches to the $Bound_{state}$, based on an exponential distribution with a specified mean search time, and CDF_{bound}, is the probability that the molecule switches to the $Mobile_{state}$ based on an exponential distribution with a specified mean bound time. In the case of the heterogenous population of bound molecules, a second $Bound_{state}$ was added with a different bound time.

From the transition matrices and the respective probabilities, we would select the next state using sampling from a multinomial distribution with the weights given by the probabilities.

To simulate molecule movement, the step size in each x,y,z direction was picked from a Gaussian distribution with variance $2D\tau^{\alpha}$, and $\alpha = 1$ for the mobile fraction, and 0.4 for the bound fraction to represent subdiffusive behaviour of genomic loci, due to the Rouse model of DNA polymer motion, although we do acknowledge that loci in *E.coli* can undergo more ballistic motions [30-32].

After each time step, the sum of the values of 10x10 array elements was taken to simulate a 100nm pixel size and get a 500x500 pixel image. Subsequently, camera noise was added by sampling from gaussian distribution with mean = 150 and standard deviation 20. To obtain images of different exposure times, time steps were integrated (e.g. 100ms = 5ms * 20 steps of integration). In the case of time intervals, we allowed molecule movement but no image formation until the next image is taken. E.g. for 1s interval, a 500ms image was followed by 500ms of only molecule movement but no image formation, followed by another 500ms image.

Construction of Top2-Halo

The Top2-Halo used in this study is from a BY4741 background. Plasmids used in this study are derivatives of pUC18 (ColE1 origin, Ampicillin resistance); pTB16 carries mNeonGreen and a downstream NatMX marker, while pSJW01 carries the HaloTag gene with a HygB marker. Both mNeonGreen and HaloTag genes encode an 8 amino acid linker at the 5' end (sequence: GGTGACGGTGCTGGTTTAATTAAC). Plasmids were maintained in *E. coli* DH5α and were extracted by growing in LB with 100 μg/ml ampicillin then using the Presto Mini Plasmid Kit (Geneaid).

All PCRs were made using either Phusion or Q5 (NEB). PCR reaction were in a volume of 50 μl and included water, 3% DMSO, the reaction buffer, 2.5 mM of each dNTP, 0.2 μM of each primer, either 1ng of plasmid DNA (for insertions) or 1 μl of genomic DNA (for screening insertions), and 0.5 μl of polymerase.

Fluorescent fusions were made by PCR amplification from pTB16 or pSJW01 using the primers listed in Table S2. PCR products were transformed into wild-type diploid BY4743. A single colony was grown at 30°C in 5 ml yeast peptone dextrose (YPD) overnight, then diluted to a final OD_{600} of 0.1 in 10ml of YPD. Cells were taken at OD_{600}=0.5-0.6 and centrifuged at 4000 rpm for 5 min. The pellet was washed twice with 25 ml of sterile deionized water, then once with 1ml of 100mM lithium acetate. Cells were then resuspended in 240 μl of 50% PEG, then 50 μl of salmon sperm DNA (thawed at 95°C for 5min, then incubated on ice for at least 10min), 50 μl of the PCR product and 36 μl of 1M lithium acetate were added in this order. The mixture was thoroughly mixed by pipeting and incubated on a rotator at 25°C for 45min, followed by heat shock at 42°C for 30min. Cells were pelleted in a microcentrifuge, washed in 500 μl of sterile water, then resuspended in 200 μl of YPD and plated on YPD agar. After growing at 30°C overnight, the cell lawn was replica-plated onto selective YPD agar, either with 100 μg/ml cloNAT (Werner) for mNeonGreen, or with 200 μg/ml Hygromycin B (Life Technologies) for HaloTag. Transformants were screened for the presence of an insert by PCR using the indicated primers in Table S2.

Confirmed clones were then sporulated by taking 750 μl of a YPD overnight cultures, washing 4 times with 1 ml sterile deionized water, then washing once with 1 ml of potassium acetate sporulation medium (Kac), and finally resuspending in 2 ml of KAc and incubating at 25°C

wish shaking. After 5 days the sporulating cultures were checked by microscopy for the appearance of numerous tetrads, then 750 µl was taken and washed 3 times in sterile water before final resuspension in 1ml water and storage at 4°C. For dissection, 45 µl of spores were treated with 5 µl of zymolase for 10 min, then tetrads were dissected on YPD plates to isolate haploids with the tagged fusion. Genomic DNA was isolated from the haploid by vortexing the cells in the presence of zirconia/silica beads, followed by phenol extraction and ethanol precipitation. The insertion site was once again amplified using the same screening primers as above, and the PCR product was sequenced to confirm that the tag and linker were both mutation-free.

The HaloTag haploids were combined with PCNA-mNeonGreen (from YTB31) and the pdr5Δ::KanMX deletion (from a haploid sporulated from YTK1414) by mating. To do this, 10 µl of water was spotted on a YPD plate, and colonies from the Mat a and Mat α haploids to be combined were mixed together into the water drop and incubated at 30°C overnight. Cells were then restreaked on an auxotrophic -Met-Lys plate on which only the mated diploid could grow. Diploids resulting from a mating were dissected as above, and eventually haploids with all three markers were isolated and used for imaging.

Single Molecule Imaging in Budding Yeast

A single colony from a YPD plate was placed in 5mL synthetic complete (SC) medium and grown with shaking at 30 degrees Celsius for ~5-6 hours. This culture was diluted by transferring ~50uL into 5mL of fresh S.C and grown overnight at 30 degrees. The overnight culture was diluted to 0.15 the next day and grown until the optical density (OD) reached 0.30. 1mL of this culture was spun down for 1 min @ 13000RPM, and the pellet was resuspended in 500uL of fresh S.C. Janelia Farms photoactivatable 549 (JF-PA549) was added to the 500uL culture for a final dye concentration of 50nM, except for YTK1434-Halo (Histone H3), where a concentration of 10nM was used to compensate for the higher copy number. This culture was placed in a thermomixer at 30 degrees and 500RPM for 40 minutes. After incubation, 3 wash cycles using fresh S.C were done to wash away unbound dye. After the final wash step, the pellet was resuspended in 50uL of S.C, and 3uL of the culture was placed on an agarose pad consisting of SC and Optiprep (Sigma), within a Gene Frame (Thermo Scientific). The pad was made by taking a 2% agarose Optiprep mixture (0.02g in 1mL Optiprep) - that was heated to 90 degrees - and mixing 500uL with 500uL 2xSC, resulting in a 1% agarose 30% Optiprep SC mixture.

Approximately 140uL of this mixture was placed within the Gene Frame, with excess being removed with a KimWipe. Prior to imaging, we waited ~15 minutes to let any unbound dye be released.

Coverslips were cleaned with the following steps: 1) Place in 2% VersaClean detergent solution overnight. 2) Wash with MilliQ water 3x. 3) Sonicate in acetone for 30 minutes. 4) Wash with MilliQ water 3x. 5) Place in methanol and flame coverslips using Bunsen burner. 6) Place in Plasma Etch plasma oven for 10 minutes.

Microscopy was done at 23 degrees, on a Leica DMi8 inverted microscope with a Roper Scientific iLasV2 (capable of ring total internal reflection fluorescence (TIRF)), and an Andor iXON Ultra EMCCD camera. An Andor ILE combiner was used, and the maximum power from the optical fiber was 100mW for the 405nm wavelength, and 150mW for the 488nm and 561nm wavelengths. The iLasV2 was configured to do ring highly inclined and laminated optical sheet (HILO), for selective illumination and single-molecule sensitivity. Metamorph was used to control acquisition. A Leica HCX PL APO 100x/1.47 oil immersion objective was used, with 100nm pixel size. Any z-stacks were doing using a PInano piezo Z controller**.**

Single-particle photoactivated localization microscopy (sptPALM) experiments were performed by activating molecules with low power (0.5% in software) 405nm light to photoactivate ~1 molecule/cell, followed by stroboscopic, long-exposure (500ms) illumination with 561nm light (5% in software) to image primarily bound molecules. A brightfield image and a z-stack of 6um (0.3um step size) in the 488nm channel, was taken before and after each timelapse, to ensure normal cell health and to find nuclei.

Tracking Analysis

Tracking was done with Trackmate [21]. Spots were localized using the Laplacian of Gaussian (LoG) method, with an estimated spot size of 2.5 pixels, with the exception of the *E.coli* experimental data where it was set to 5 pixels. The intensity threshold was set a bit lower to prevent track fragmentation due to intensity fluctuations. The linear assignment problem (LAP) algorithm was used to form tracks with costs on quality ranging from 0.1-0.5. We set a gap frame of 1 to allow temporary disappearance of the molecule, and track merging and splitting was allowed in cases where multiple molecules crossed paths with one another.

To isolate tracks found only in cells/nuclei, we used the binary images to locate tracks whose mean positions coincided with values of 1 in the binary image.

Machine Learning and Tracking Analysis

All machine learning and subsequent analysis for estimation of residence times was done using Matlab.

To construct training data sets, we had binary classification, with a value of 0 assigned to false positive/diffusing molecule, and a value of 1 to a track representing a bound molecule. We manually looked at the raw image data to determine if the molecule appeared immobile.

For the learning procedure, the "TreeBagger" function in Matlab was used, representing the random forest algorithm. The hyperparameters that were adjusted were: InBagFraction (representing the fraction of the training data given to each tree), MinLeafSize (minimum leaf size), NumPredictorstoSample (number of predictors to sample at each node), and NumTrees (the number of trees to construct). The hyperparameters were adjusted until the best OOB error was achieved.

For GMM fitting, the expectation-maximization (EM) algorithm was used.

After the final classification, we analyzed the tracks to extract residence times.

We then fit the track durations of the remaining tracks with a truncated exponential model, to compensate for discarding short duration tracks, using Maximum Likelihood Estimation (MLE) through Matlab's "mle" function, to calculate the mean track duration.

2)

$$PDF = \left(\frac{1}{\tau}\right) e^{\frac{-(x-L)}{\tau}}$$

where τ is the mean track duration, and L is the truncation point. For photobleaching controls, this was equivalent to estimating the mean bleach time.

The 95% confidence intervals were calculated by bootstrapping 1000 samples. Bound times were calculated using the following equation:

3)

$$Tbound = Ttrack * Tbleach/(Tbleach - Ttrack)$$

To calculate the errors on the estimate, we performed bootstrap sampling on the track durations to the following equation:

4)

$$(\frac{1}{T_{track}})e^{-t/T_{track}} = \left((\frac{1}{T_{bound}}) + (\frac{1}{T_{bleach}})\right)e^{-((\frac{1}{T_{bound}})+(\frac{1}{T_{bleach}}))t}$$

With 10% variation allowed for the T_{bleach} estimate, in order to obtain biologically sensible results.

To check for two-exponential mixtures, the track durations were fit with the following two-exponential model:

5)

$$p\left(\frac{1}{\tau_1}\right)e^{\frac{-(x-L)}{\tau_1}} + (1-p)\left(\frac{1}{\tau_2}\right)e^{\frac{-(x-L)}{\tau_2}}$$

where τ_1 = (*Tbleach* + *Tbound*$_\gamma$)/(*Tbleach* **Tbound*$_\gamma$), τ_2 = (*Tbleach* + *Tbound*$_\psi$)/(*Tbleach* * *Tbound*$_\psi$), p is the mixture proportion, and L is the truncation point.

The lower and upper bounds on the two binding timescales were 0.0001s and 6000s, respectively, while allowing for a 10% variation in the bleaching estimate.

To check for overfitting and to identify whether the two-exponential model significantly fit the data better, we used the BIC test and the Loglikelihood ratio (LLR) test, as described in [1]. We looked for cases when the two-exponential model estimates did not simply return the lower and/or upper bounds as this would indicate that no physically sensible solution was found.