Characterization of Intracellular Transcription and Diffusion Kinetics

Research Thesis

In Partial Fulfillment of the Requirements for the Degree of Master of Science

in

Biomedical Engineering

Naor Granik

Submitted to the Senate of the Technion – Israel Institute of Technology

Tishrei 5780 Haifa October 2019

The Research Thesis was done under the supervision of Prof. Yoav Shechtman in the faculty of Biomedical Engineering and Prof. Roee Amit in the faculty of Biotechnology and Food Engineering.

The Generous Financial Help of the Technion, Israel Institute of Technology is Gratefully Acknowledged.

I would like to thank my advisers Prof. Yoav Shechtman and Prof. Roee Amit for their guidance.

I would also like to thank Dr. Lucien E. Weiss, Elias Nehme, and Noa Katz for their help on this work.

Publications

Granik, Naor, Noa Katz, Yoav Shechtman, and Roee Amit. "Live dynamical tracking of slncRNA speckles in single E. coli cells reveals bursts of fluorescence degradation." Submitted to *eLife.* (Section I of this thesis)

Granik, Naor, Lucien E. Weiss, Elias Nehme, Maayan Levin, Michael Chein, Eran Perlson, Yael Roichman, and Yoav Shechtman. "Single particle diffusion characterization by deep learning." *Biophysical Journal* (2019). (Section II of this thesis)

Table of Contents

Table of Contents

List of Figures

List of Tables

Abstract

The cellular cytoplasm is the environment in which all intracellular reactions take place. Its physical and chemical properties have a strong influence on a multitude of functions, such as signaling, transport, protein folding etc. Here, we aim to shed light on two different intracellular dynamic processes which have gained increased attention in recent years, owing to technological improvements in microscopy techniques: The formation of nuclear speckles, and the occurrence of anomalous diffusion.

Nuclear speckles are membrane-less protein-rich bodies built around a long-non-coding RNA (lncRNA) scaffold. As a means of studying the dynamics of such a formation, we investigate the dynamics of synthetic speckles in bacteria by encoding two types of synthetic lncRNAs (slncRNA), which form the basis of the bacterial speckle. The slncRNAs incorporate RNA-binding phage-coat-protein (RBP) binding sites downstream from a pT7 promoter. For both slncRNAs studied, fluorescent speckles containing dozens of RBP-bound slncRNA molecules form in cell poles. Fluorescence measurement over time reveals both positive and negative changes in intensity spaced by exponentially distributed periods of non-classified activity. We identify positive changes with transcriptional bursts, and term the negative, fluorescence degradation bursts. The data indicates that negative bursts correspond to shedding of multiple slncRNAs back to cytoplasm.

Diffusion plays a critical role in many biological processes in the cell. Direct observation of molecular movement by single-particle-tracking experiments has contributed to a growing body of evidence that many cellular systems do not exhibit classical Brownian motion, but rather anomalous diffusion. Characterization of the physical process underlying anomalous diffusion remains a challenging problem due to the fact that commonly used tools for distinguishing between these processes are based on asymptotic behavior, which is experimentally inaccessible in most cases. Additionally, an accurate analysis of the diffusion model requires the calculation of many observables since different transport modes can result in the same diffusion power-law α, which is typically obtained from the mean squared displacements (MSD). We opted to use deep learning to infer the underlying process resulting in anomalous diffusion. We implemented a neural network to classify single-particle trajectories by diffusion type, separating between Brownian motion, fractional Brownian motion (FBM) and Continuous Time Random Walk (CTRW). We demonstrate the applicability of our network architecture for estimating the Hurst exponent for FBM and the diffusion coefficient for Brownian motion on both simulated and experimental data. We show these networks achieve better accuracy than time-averaged MSD analysis on simulated trajectories while requiring fewer time-steps. Furthermore, on experimental data, both network and ensemble MSD analysis converge to similar values, with the net requiring only half the number of trajectories required for ensemble MSD to achieve the same confidence interval.

Introduction

The advent of advanced microscopy techniques has made it possible to explore cellular processes in greater spatial and temporal resolutions. This in turn, created a demand for better tools, both molecular and technical, that complement the new abilities. On the molecular side, constructs that allow for accurate labeling, novel fluorescent probes, and new model systems that enable the examination of general dynamics. On the technical side, data analysis tools that can handle the abundance of new data and interpret it correctly. Naturally, this technological advancement led to new discoveries about the dynamic nature of the cellular environment and its many components.

This work is based on two separate research projects, each designed to aid in the understanding and analysis of different phenomena discovered due to improved localization and fluorescence microscopy methods.

Section I

In recent years, naturally occurring complexes of long-non-coding RNAs (lncRNAs) that are bound by various RNA binding proteins (RBPs) have been discovered in many Eukaryotic cell types. The most prominent examples being paraspeckles[1–3] and nuclear-speckles[4]. These serve a purpose in the structure and organization of the nucleus as well as being involved in many biochemical reactions[5,6].

Synthetic systems with components of similar nature, namely lncRNAs bound by RBPs, have been in use for several decades as a means of labelling proteins of interest for tracking purposes[7–9]. These systems consist of a set of stem loops encoded either into the 5’ or 3’ end of the transcript of interest. The loops encode a binding site for the coat-protein of either MS2 or PP7 bacteriophages. In addition, the cells separately express a chimera of the particular coat-protein fused to a fluorescent protein. When co-expressed, the bound cassettes produce bright puncta or speckles which can be tracked in living cells. When viewed under a fluorescence microscope, these speckles show similar visual characteristics as other types of known nuclear bodies[3].

Motivated by this similarity we opted to study the synthetic system in bacteria as a model for the general dynamics of nuclear bodies. To this end, we tracked the expression of a pT7 promoter using two different slncRNAs as scaffold for our synthetic speckles, a cassette encoding for 24 binding sites for the PP7 bacteriophage coat protein (PP7-24x), considered as the standard in the field for labelling and tracking, as well as a new significantly shorter slncRNA encoding four PP7 and five Qβ phage-coat-protein binding sites in interlaced fashion (Qβ-5x-PP7-4x). In the new slncRNA molecule, the respective coat proteins do not recognize the other binding site family[10], and thus upon expression only half of the hairpins are bound, while the unbound half increases the stability of the entire structure.

For both slncRNAs, we show that individual speckles within single cells can be observed and dynamically tracked. Upon analyzing the resulting intensity vs. time signals, we detected not only the expected transcriptional bursts[9,11], but also for the first time, bursts of signal degradation suggesting that speckles not only can accrue new slncRNAs upon transcription but also shed them in a burst-like fashion. Further analysis revealed important differences between the PP7-24x and the Qβ-5x-PP7-4x speckles, implying that macroscopic characteristics of these compartments are dependent on the RNA cassette design. Finally, we studied a three-state random telegraph kinetic model and show that it better describes the experimental data compared to the two-state model commonly used to describe transcription dynamics[12].

Section II

Single-particle tracking (SPT) is widely used to investigate the biophysical properties of cellular membranes and other materials for extracting kinetics and other information on nanoscale processes. In recent years, rapid advances in labelling and detector sensitivity have widened the applicability for SPT to new biological systems with improved temporal and spatial resolution[13–16]. The key information gained from these obtained trajectories after analysis is a statistical model for the mode of motion and the parameters which shed light on the dominating elements of the environment governing motion[13,17,18].

An important property of SPT methods is that the list of particle positions acquired during a measurement contains temporal information. This feature can be exploited to identify transient periods of statistically-similar motion within the same trajectory, including different diffusion states[19–23], changes in diffusion type, e.g. distinguishing between Brownian, confined, and directed diffusion[24–26] and the associated kinetics of transitions and equilibrium probabilities. While identification of periods of diffusive processes gives some insight into an object behaviour, we would ideally identify a specific mathematical model that best describes the measured trajectory. The applicability of classical methods for accurately extracting the underlying parameters has been somewhat limited, thus necessitating a more reliable approach.

Here, we develop a deep-learning-based framework for both the classification of diffusion processes in long trajectories, for which it exhibits higher precision over conventional analysis methods, as well as for short and noisy trajectories, including parameter estimation from an ensemble of very short trajectories (10 time-steps). Our approach is to use a set of convolutional neural networks (CNNs) for classifying either single trajectories or a set of short trajectories as one of three selected diffusion models: Brownian motion, FBM and CTRW, with simultaneous estimation of the relevant parameters by continuous regression. This method is simple to implement and outperforms conventional approaches in terms of parameter estimation precision, convergence rate and usability of short trajectories.

Section I:
Synthetic Speckle Analysis

Abbreviations and Notations

SM-FISH	Single molecule fluorescent in-situ hybridization
slncRNA	Synthetic long noncoding RNA
RBP	RNA binding protein
AU	Arbitrary units
FP	Fluorescent protein
RTN	Random telegraph noise
RNAP	RNA polymerase
PSD	Power spectral density

Chapter 1: Background

Transcription is a complex process that depends on successive stochastic interactions between many molecular species (transcription factors, promoters, polymerases, etc.). This randomness leads to variability in gene expression levels, even in a genetically identical population of cells[27–29]. In recent years, single molecule studies of transcription in different cell types (varying from bacteria to mammalian cells), have unexpectedly revealed dynamics characterized by bursts of transcriptional events that are separated by periods of quiescence in which transcription is barely observed[9,30,31]. These observations have been predominantly obtained using two methods offering single molecule resolution. The first is single-molecule fluorescent in-situ hybridization (SM-FISH), and the second, phage-coat-protein labeling of cassettes containing multiple binding sites in living cells. SM-FISH facilitates quantitative analysis of the number of transcripts at a given time point in a population of single fixed cells by labeling mRNA molecules with fluorescently tagged DNA probes complementary to the transcript sequence. Although highly quantitative, this approach does not allow for the direct exploration of the temporal dynamics of transcription, and instead these are inferred from population statistics[32,33].

In order to directly study the dynamics of transcription, Singer and colleagues[7] introduced a second method, whereby a set of stem loops is encoded either into the 5' or 3' end of a transcript. The loops encode a binding site for the coat-protein of either MS2 or PP7 bacteriophages. In addition, the cells separately express a chimera of the particular coat-protein fused to a fluorescent protein. When co-expressed, the coat-protein-bound cassettes yield bright puncta or speckles which can be tracked in living cells. Thus, in theory, evaluation of the spot intensity allows one to interrogate the dynamics of processes at the single-cell level[7,31]. Despite extensive efforts to optimize this technology yielding a commonly used cassette developed by Tutucci and Singer[8] consisting of 12 or 24 binding sites, this approach has still not reached the single-molecule-in-single-cell threshold necessary for a direct evaluation of transcriptional dynamics. This is a result of several critical drawbacks. First, it is thought that the synthetic binding sites disrupt natural degradation, effectively artificially extending transcript lifetimes[11]. Second, sequences of multiple binding sites suffer from severe occupancy issues[34], making it impossible to accurately correlate fluorescence to transcript number. Third, puncta are often composed of multiple RNA molecules, making it difficult to disentangle signals from single molecules[34]. Finally, due to the repeating binding sites, cassettes are prone both to mutation or general instability[35]. For example, in the context of bacteria, Golding et al[9] engineered an MS2-coat protein binding site cassette containing 96 hairpin repeats. These were inserted downstream of the Plac/ara promoter, providing the first live evidence for transcriptional bursts in bacteria. However, individual transcriptional events were not resolved, and this relatively large cassette was not used in follow-up studies. Instead, later studies opted to use SM-FISH, RT-PCR, and RNA-seq to provide further proof for the pervasiveness of bursty transcription throughout the microbial genome[36,37].

Naturally occurring puncta-like complexes of have been discovered in many Eukaryotic cell types. The most well-studied examples of these natural puncta-like complexes are paraspeckles[1–3] and nuclear-speckles[4] which are composed predominantly of long-non-coding RNAs (lncRNAs) and RNA binding proteins. These particles are an example of a wider phenomenon, which has received increased attention by the research community over the past decade, of liquid-liquid phase-separated micro- and nano-compartments within cells[38]. Given the similarity in optical microscopy observations between paraspeckles (for instance) and the RNA-phage coat protein puncta (**Figure *1***), it is possible to view the latter retrospectively as studies which showed that synthetic liquid-liquid phase-separated compartments (or speckles) can be formed *in vivo* (in any cell type) with phage coat proteins and synthetic lncRNA molecules that encode cassettes of their binding sites.

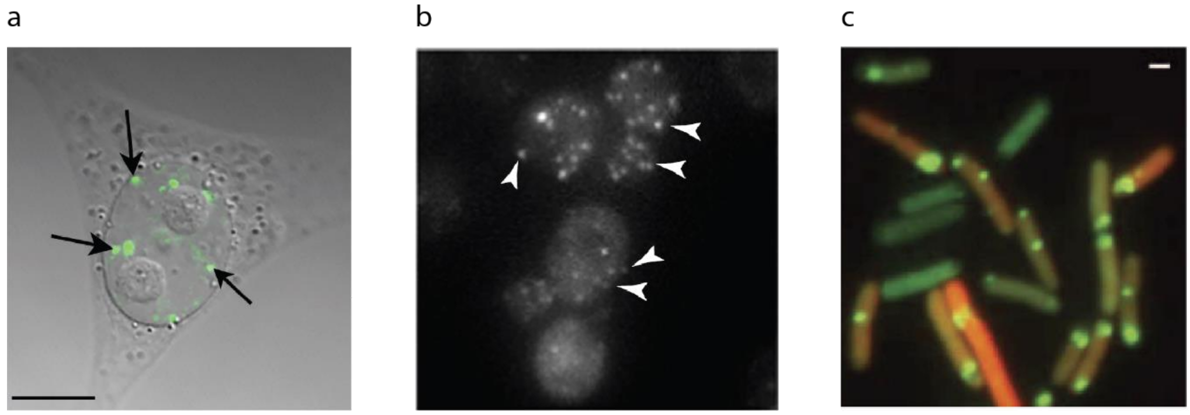

Figure 1. Fluorescence imaging of paraspeckels and RNA binding sites cassettes.

(**a**) A confocal image of a HeLa cell stained with an antibody to the paraspeckle protein; Fluorescent signal overlaid on a brightfield image of the cell. Taken from Fox et al.[3] (**b**) *S. cerevisiae* cells expressing the MDN1 mRNA tagged with 12 repeats of the MS2 binding sites (arrows mark single mRNAs). Taken from Tutucci et al.[8] (**c**) Detection of mRNA tagged with 96 repeats of MS2 binding sites in live *E. coli* cells. Taken from Golding et al.[9]

Chapter 2: Materials and Methods

Bacterial strains and plasmids

Strain	Source	Use	Genotype
TOP10	Prof. Amit lab	General cloning and storage	F- mcrA Δ(mrr-hsdRMS-mcrBC) ϕ80lacZΔM15 ΔlacX74 nupG recA1 araD139 Δ(ara-leu)7697 galE15 galK16 rpsL(StrR) endA1 λ-
BL21-DE3	Prof. Amit lab	T7 expression strain	fhuA2 [lon] ompT gal (λ DE3) [dcm] ΔhsdS λ DE3 = λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5

Table 1. *Bacterial strains*

Plasmid	Source	Resistance	Function
pCR4-24XPP7SL	Addgene # 31864	Ampicillin	Expression of 24xPP7 binding sites cassettes
pBAC-lacZ	Addgene # 13422	Chloramphenicol	BAC plasmid (maintained as single copy)
pSMART BAC	Lucigen	Chloramphenicol	BAC plasmid (maintained as single copy)
pUC57-T7-5Qβ-4PP7	GenScript	Ampicillin	Cloning of the 5Qβ-4PP7 binding sites cassette
A133-rhlr-PP7-mCherry	Prof. Amit lab	Ampicillin	Expression of PP7-mCherry fusion protein under an inducible rhlr promoter
A133-rhlr-Qβ-mCherry	Prof. Amit lab	Ampicillin	Expression of Qβ -mCherry fusion protein under an inducible rhlr promoter
pBAC-5xQβ-4PP7-lacZ	Prof. Amit lab	Chloramphenicol	Expression of the 5Qβ-4PP7 binding sites cassette under a T7 promoter.

Table 2. *Plasmids*

Construction of the pBAC-Qβ-5x -PP7-4x binding sites array

The T7 promoter followed by the binding sites sequence coding for 5Qβ-4PP7 binding sites: cctaggcgattatgacgttattctactttgattgtgatgcatgtctaagacagcatcgcctgctggtcgtgactaaggagtttatatggaaacccttacgagacaatgctaccttaccggtcgggcccacttgtttttacccatgatgcatgtctaagacagcatcgcctgctggtcgtgactaaggagtttatatggaaacccttagaaacagccgtcgccttgaagccgagaacaatgcatgtctaagacagcatatggattgcctgtctgttaaggagtttatatggaaacccttacatcaggcttcgcagtatgcaacgcttgcgatgcatgtctaagacagcatttcaccgctttcctaagtaaggagtttatatggaaacccttagtactaactcgcagatgcatgtctaagacagcatcagaaacgtcacgtcctggc.
(Qβ and PP7 binding sites marked in red and green respectively), was ordered from GenScript, Inc. (Piscataway, NJ), as part of a pUC57 plasmid, flanked by EcoRI and HindIII restriction sites. The sequence was extracted using the restriction enzymes and purified from gel. pBAC-LacZ backbone plasmid was obtained from Addgene (plasmid #13422). Both insert and vector were digested using the above restriction sites and ligated to form pBAC-Qβ-5x -PP7-4x-lacZ.

Sample preparation

BL21 cells expressing the BAC-Qβ-5x-PP7-4x and the Qβ-mCherry or PP7-mCherry expression plasmid were grown O/N in Luria Broth (LB), in 37^{o} with appropriate antibiotics (CM, AMP). O/N culture was diluted 1:100 into 3μl solution of BioAssay (BA)-LB (95%-5% v:v) with appropriate antibiotics, and induced with 1μl IPTG (final concentration 1mM) and 1.5μl *N*-butanoyl-l-homoserine lactone (C4-HSL) (final concentration 60μM) to induce expression of T7 RNA polymerase and the RBP-FP respectively. Culture was shaken for 3 hours in 37^{o} before being applied to a gel slide (3μl Dulbecco's Phosphate-Buffered Saline (Biological Industries) x1, mixed with 0.045g SeaPlaque low melting Agarose (Lonza, Switzerland), heated for 20 seconds and allowed to cool for 30 minutes).

1.5 μl cell culture was deposited on a gel slide and allowed to settle for an additional 30 minutes before imaging.

Image analysis

A single experiment was carried out by tracking a field of view for 60 minutes on Nikon Eclipse Ti-E epifluorescent microscope (Nikon, Japan) using the Andor iXon Ultra EMCCD camera at 6 frames-per-minute with a 200 msec exposure time per frame to avoid photo-bleaching and sufficient recovery of fluorescence signal. Excitation was performed at 585 [nm] wavelength by a CooLED (Andover, UK) PE excitation system. Subsequently, the brightest spots (top 10%) in the field of view were tracked using the plugin developed by Sbalazarini and Koumoutsakos[39] for imageJ[40,41]. A typical field of view usually contained dozens of cells, a portion of which were not fluorescent while others presented distinct bright speckles, localized at the cell poles as expected from literature[42].

The tracking data, (x,y,t coordinates of the bright spots centroids), together with the raw microscopy images were fed to a custom built Matlab (The Mathworks, Natick, MA) script designed to normalize the relevant spot data. Normalization was carried out as follows: for each bright spot, a 20-pixel wide sub-frame was extracted from the field of view, with the spot at its center. Each pixel in the sub-frame was classified to one of three categories according to its intensity value. The brightest pixels were classified as 'spot region' and would usually appear in a cluster, corresponding to the spot itself. The dimmest pixels were classified as 'dark background', corresponding to an empty region in the field of view. Lastly, values in between were classified as 'cell background'. Classification was done automatically using Otsu's method[43], (**Figure *2***). From each sub-frame, two values were extracted, the mean of the 'spot region' pixels and the mean of the 'cell background' pixels, corresponding to spot intensity value and cell intensity value. This was repeated for each spot from each frame in the data, resulting in sequences of intensity vs. time. Sequences were filtered for high frequency noise by a moving average filter with a window of 10 time points. Normalization was done by subtracting the cell intensity values from the spot intensity values.

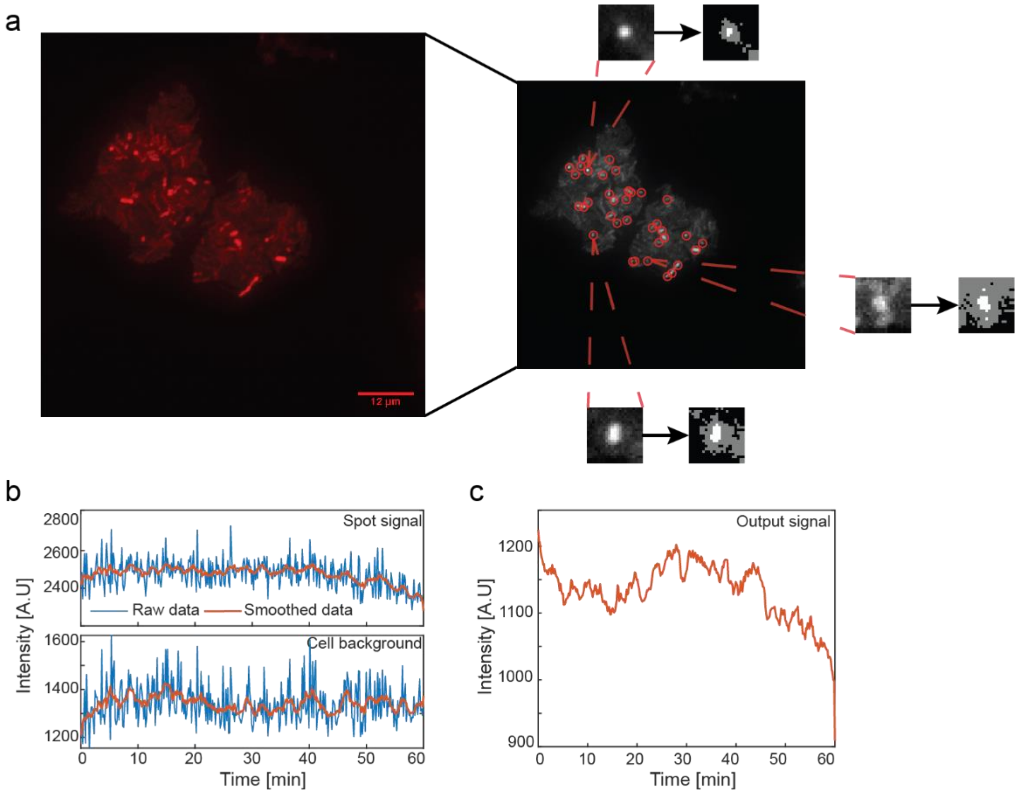

Figure 2. Image analysis and signal acquisition.

(**a**) Leftmost image – Typical field of view showing bright fluorescent spots, along with bright cells. Dim red background is cells which are not fluorescing. Middle image – Capturing top 10% of bright spots (marked in red circles). Presented are three example sub-frames, each showing a spot and its immediate surroundings. For each sub-frame, each pixel is classified to one of three intensity levels – bright spot, cell background and dark background corresponding to white, grey and black colors in the segmented images. (**b**) Sample spot signal (top), and its corresponding cell background signal. Blue line is raw data, orange line is smoothed data after a 10-point moving average. (**c**) Output signal resulting from the subtraction of the background signal from the spot signal.

Identifying burst events

We define a *burst* as a sudden change or shift in the level of the speckle's fluorescence intensity leading to either a sustainable higher or lower new signal level (**Figure *3*a- top**). To identify such shifts in the base-line fluorescence intensity, we use a moving-average window of ten points to smooth the data. The effect of such an operation is to bias the fluctuations of the smoothed noisy signal in the immediate vicinity of the bursts towards either a gradual increase or decrease in the signal (**Figure *3* a-bottom).** Random fluctuations, which do not settle on a new baseline level are not expected to generate a gradual and continuous increase over multiple time-points in a smoothed signal. As a result, we search for contiguous segments of gradual increase or decrease and record only those whose probability for occurrence are 1 in 1000 or less given a Null hypothesis of randomly fluctuating noise. As an example, we consider a constant base-line intensity amplitude with white Gaussian noise. For any particular time-point the probability that the next point will exhibit either a stronger or weaker signal is 50% respectively. Since the noise is independent and identically distributed (IID), the probability for an increase in the signal lasting 10 consecutive time points is $\frac{1}{2^{10}} = \frac{1}{1024}$. Given that our traces typically take place over ~60 minutes and include 360 frames in total, we expect <1 such random 1 in 1000 events per trace. This probability remains roughly the same even after the moving average smoothing.

For the general case, the underlying empirical signal is not constant, and may either be trending up or down. Therefore, it is necessary to normalize the probability per signal and determine a threshold – m, such that the probability for a consecutive increase of m time points is $\frac{1}{1024}$ given the underlying signal trend. For every trace, we first compute the intensity difference distribution (**Figure *3* b**). The probability that the signal increases at a random time point is calculated by summing the number of points in which the signal derivative is positive and dividing by the total length of the signal.

$$p = \frac{length\left(\frac{dS}{dt} > 0\right)}{length(S)} \tag{1}$$

This in turn allows us to compute m as follows:

$$p^m = \frac{1}{2^{10}} \Rightarrow mlog_2(p) = -10 \Rightarrow m = -\frac{10}{\log_2(p)} \tag{2}$$

The threshold is calculated for each signal separately and is usually in the range of 7-13 time points. An analogous threshold is calculated for decrements in the signal and is usually in the range $[m-1, m+1]$. We mark each trace with the number of events that exceed this threshold and define those as bursts.

While the choice of smoothing window is somewhat arbitrary, it was chosen to be sufficiently large to allow for both an identification of a gradual increase or decrease due to the burst and a stable base-line shift, without compromising our ability to properly characterize the signal on a longer time-scale. To check that our choice of smoothing parameter does not affect the interpretation of the data, we applied both shorter and longer moving-average windows showing that the over-all nature of the results remains unchanged (**Figure *3* c-d**). The main difference between the averaging windows lies in the number of significant events identified. The 5-point window results in a total of 91 positive, and 86 negative events found, while the 15-point window results in 577 positive and 424 negative events.

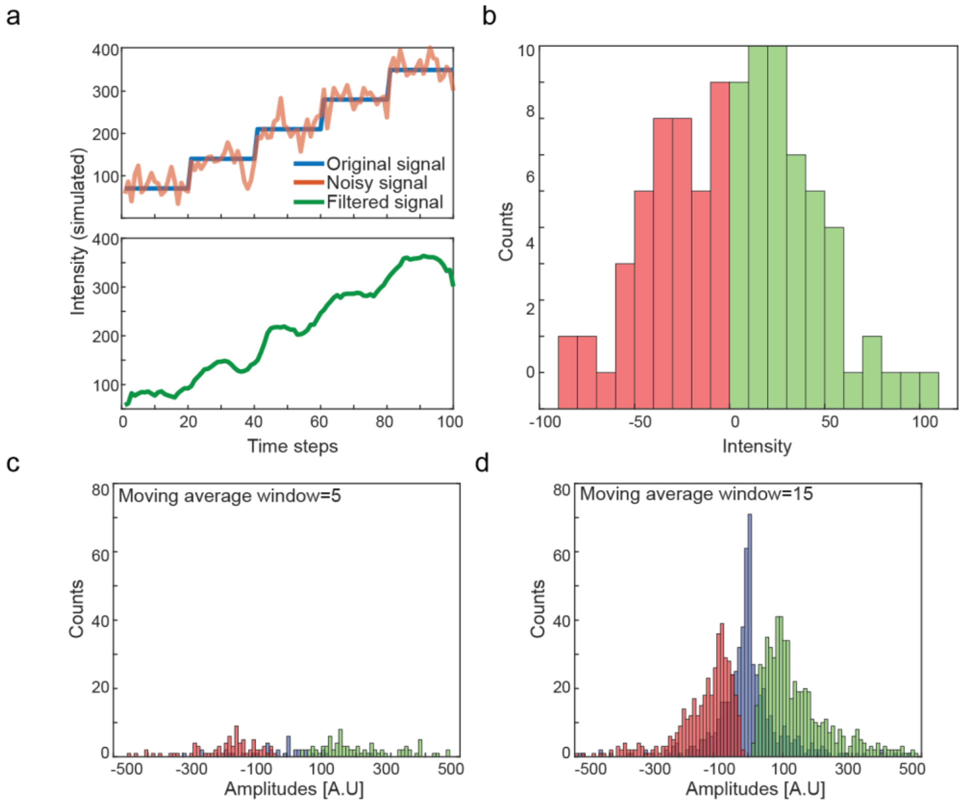

Figure 3. Identification of burst events.

(**a**) Effects of noise, and noise-filtering, on a bursty signal. Underlying simulated signal is comprised of instantaneous increases in intensity (top plot – blue line), however, this feature disappears with the addition of Gaussian noise with a standard deviation of 30 [A.U] (top plot - orange line). The noisy signal was filtered with a 10-point moving average filter and appears continuous (bottom plot). (**b**) Distribution of intensity difference between successive time points of the simulated signal appearing in (**a**), showing a slight bias towards the positive side (54% of the derivative is positive). This distribution is used to calculate the threshold for a significant event. (**c-d**) Amplitude distribution of the Qβ-5x experimental data analyzed with different moving average windows. (**c**) 5 time-points moving average (n_{pos}=605, n_{neg}=446, $n_{\text{non-classified}}$=573). (**d**) 15 time-points moving average (n_{pos}=91, n_{neg}=86, $n_{\text{non-classified}}$=36).

Chapter 3: Results

Qβ-5x-PP7-4x RBP cassette displays positive and negative intensity bursts

In order to study synthetic speckle formation with slncRNA-RBP complexes, we designed a short mRNA binding-site cassette, consisting of four native PP7, and five native Qβ binding sites in an interlaced manner (**Figure *4* a**). The cassette was cloned downstream to a pT7 promoter on a BAC (Addgene plasmid # 13422), and transformed, together with a plasmid encoding for Qβ-mCherry from a pRhlR inducible promoter, to BL21 *E. coli* cells. Single cells expressing the Qβ-5x-PP7-4x together with Qβ-mCherry (data gathered from these cells is denoted as Qβ-5x data) were imaged every 10 seconds for 60 minutes under constant conditions (200 msec integration time, 37$^{\text{o}}$ c), and subsequently the intensity of bright speckles (**Figure *4* b**) resulting from the bound cassette in each cell were analyzed for every timepoint resulting in a trace of intensity vs. time. During processing each trace was smoothed by a 10-point moving average and subsequentially normalized by subtracting the background of the cellular environment surrounding the speckle (which was smoothed in similar manner). The resulting signals are either decreasing or increasing in overall intensity, and occasionally signals that initially show an increase and subsequent decrease are observed (**Figure *4* c**)

To determine whether the "sharp" increases or decreases in intensity correspond to a distinct signal, and are not part of the underlying noise, we employed a scheme, whereby statistically significant changes in intensity are identified as burst events. In brief, we assume the total fluorescence is comprised of three distinct signal processes: total transcript fluorescence, background fluorescence and noise. We further assume that background fluorescence is slowly changing, as compared with total transcript fluorescence which depends on the dynamic and frequent processes of transcription and degradation. Finally, we consider noise to be a symmetric, memory-less process. We define a "burst" as a change or shift in the level of signal intensity leading to either a higher or lower new sustainable signal intensity level. To identify such shifts in the baseline intensity we search for continuous segments of gradual increase or decrease whose probability of occurrence, under our assumption of random symmetric noise, is 1 in 1000. From this probability we set a threshold for the minimum length of a gradual shift, where events lasting longer than this threshold are classified as burst events. Segments within the signal that are not classified as either a negative or positive burst event are considered unclassified. Unclassified segments are typically signal elements whose noise profile does not allow us to make a classification into one or the other event-type. Such segments can consist of multiple event types, for example: bursts that do not pass the false positive threshold that we set, or events where no transcriptional or degradation processes are recorded. We mark the classifications with positive "burst", negative "burst", and non-classified events in green, red, and blue respectively (**Figure *4* c-right**). We confine our segment analysis between the first and last significant segments identified in a given signal, since we cannot correctly classify signal sections that extend beyond the observed trace. These unprocessed segments (before the first significant event, and after the last) are marked in a dashed black line.

Next, using this classification criteria for bursts, we annotated the three features of our signal (increasing bursts, decreasing bursts, and non-classified events) for ~1000 speckle traces. From this data we aggregated the amplitude (*ΔI*) distribution and rate of intensity change (*ΔI/ Δt)* for all three event types (**Figure *4* d-e**). The plots show distributions with three separated populations of non-classified, increasing, and decreasing bursts, with the number of positive and negative burst events being approximately equal (<15% difference regardless of moving average window and the statistical threshold set for identification).

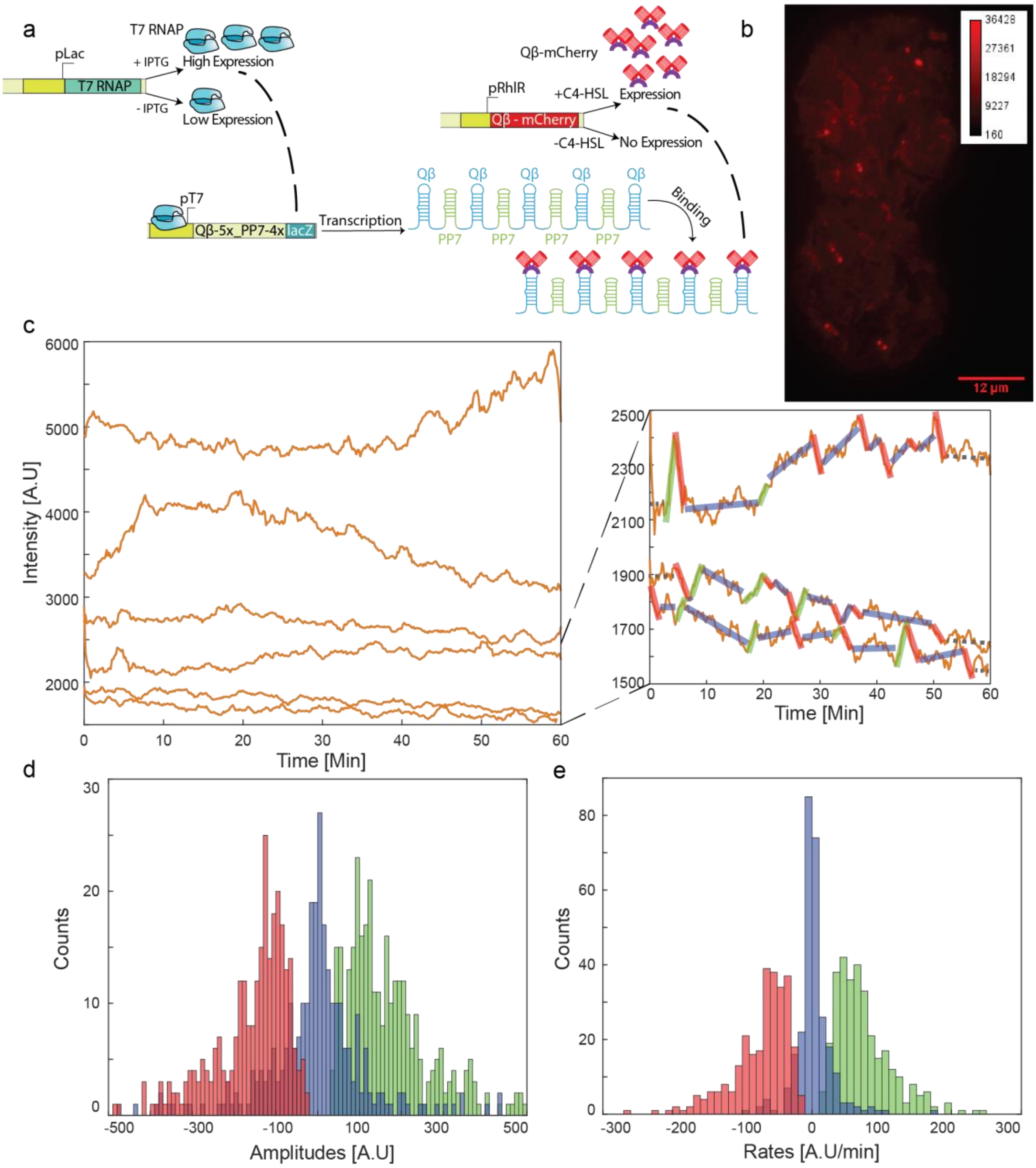

Figure 4. Biological setup, sample signals and amplitude distributions.

(**a**) Biological scheme of the experimental system comprised of the slncRNA with binding site configuration Qβ-5x-PP7-4x and the RBP-FP fusion protein Qβ-mCherry. (**b**) A region of interest (ROI) from a field of view of an experiment showing dark background, cells (dim red) and bright speckles (bright red), resulting in a high dynamic range as shown by the color bar presenting intensity values in A.U. (**c**) Sample intensity signals taken from different speckles from different experiments on separate days showing a range of behaviors. Zoom-in shows the three lower signals with overlaid segments presenting three signal states: strong increase (green), strong decrease (red), non-classified (blue). The black dashed lines mark data that was not analyzed (segments extend beyond signal range). (**d**) Empirical amplitude distributions gathered from 1200 signals. Green, blue and red correspond to increasing, decreasing and non-classified signal segments. (**e**) Rate distributions gathered from 1200 signals.

Negative and positive burst distributions indicate 1-3 molecules per burst

To further explore the burst distributions, we first take into consideration the assumed nature of burst events. To a first approximation, molecular bursting events are thought to vary over a small range of integer values (e.g. 1-10 as for transcriptional bursting[9]), and should thus exhibit a Poisson-like distribution. Consequently, the accumulation of the fluorescent amplitudes of a large number of burst events should behave according to a Poisson distribution as well. Since each burst amplitude should be directly proportional to the number of events which comprise it, the mean number of burst events and fluorescence signature per molecule can then be extracted from such distributions.

We plot separately the distributions for the positive and negative burst amplitudes as blue dots (**Figure *5* a-b**). Each plot is overlaid with three Poisson distribution fits with parameter λ = 1 (red), 2 (green), and 3 (black) respectively, corresponding to a mean of 1-3 transcripts per burst. Given the fact that we cannot directly infer the fluorescence intensity associated with a single RNA cassette, we fitted the distributions with a modified Poisson function of the form:

$$p(I) = \frac{\lambda^{\frac{I}{k_0}} e^{-\lambda}}{\left(\frac{I}{k_0}\right)!} \qquad (3)$$

where *I* is the experimental fluorescence amplitude, λ is the Poisson parameter (rate), and K_0 is a fitting parameter whose value corresponds to the amplitude associated with a single RBP-bound slncRNA molecule within the burst. For each rate we chose the fit to K_0 that minimizes the deviation from the experimental data. The fits show that while the λ=3 distribution provides the best fit to the data (corresponding to a mean of three transcriptional slncRNA's per burst), the λ=1 distribution provides the best fit to the tail of the distribution, but fails at lower amplitude values which may be due to our analysis threshold that treats many of these small amplitude events as unclassified. Since higher rate distributions provide a progressively worse fit (as Poisson distribution resembles a Gaussian curve in higher rates), we conclude that both the transcriptional and degradation distributions provide a reasonable match to Poisson distributions with λ=1 to 3. This match suggests that the range of fluorescence amplitude spanning ~40-95 (A.U) most likely corresponds to the amplitude generated by the addition or subtraction of a single Qβ-5x-PP7-4x slncRNA into the speckle under our experimental conditions. Finally, given the match with the lower rate Poisson distributions, it seems likely that the number of cassettes involved in both the positive and negative bursts varies between 1 and 3 molecules per burst.

To provide additional confirmation that we are able to detect a signal from a single slncRNA molecule, we repeated the experiment with a strain expressing PP7-mCherry with a 24x PP7 binding site-cassette (sequence obtained from Addgene plasmid #31864 [44]). For this cassette, denoted as PP7-24x, the plots show similar distributions as to the one observed for Qβ-5x (**Figure *5* c-d**). The distributions appear to be well described by Poisson distributions with λ=1-3 as well, but with an increase of 62-64% in the fitted K_0 per λ. This result is consistent with past observations, which have shown that these cassettes are only occupied by 8-14 proteins, resulting in a reduced intensity relative to the intensity expected from the number of designed binding sites[34]. The consistency of the Poisson fit results for both cassettes indicate that in the PP7-24x case, we are also observing 1-3 mRNAs per burst, and that the nature of the reporting cassette does not have a large effect on the outcome of the experiment.

Given the values extracted for the fluorescence intensity that is associated with a single reporter cassette, we computed a lower estimate for the average number of RBP-bound slncRNAs that make-up a single speckle. To do so we took the average value of the Poisson 1- K0 value (93 and 151 A.U. for the Qβ-5x and PP7-24x respectively) and computed the average number of cassettes per trace by dividing the average trace intensity with the appropriate K0. The results (**Figure *5* e-f**) show that a single speckle can be estimated to be made up of at least 10-30 slncRNA molecules on average.

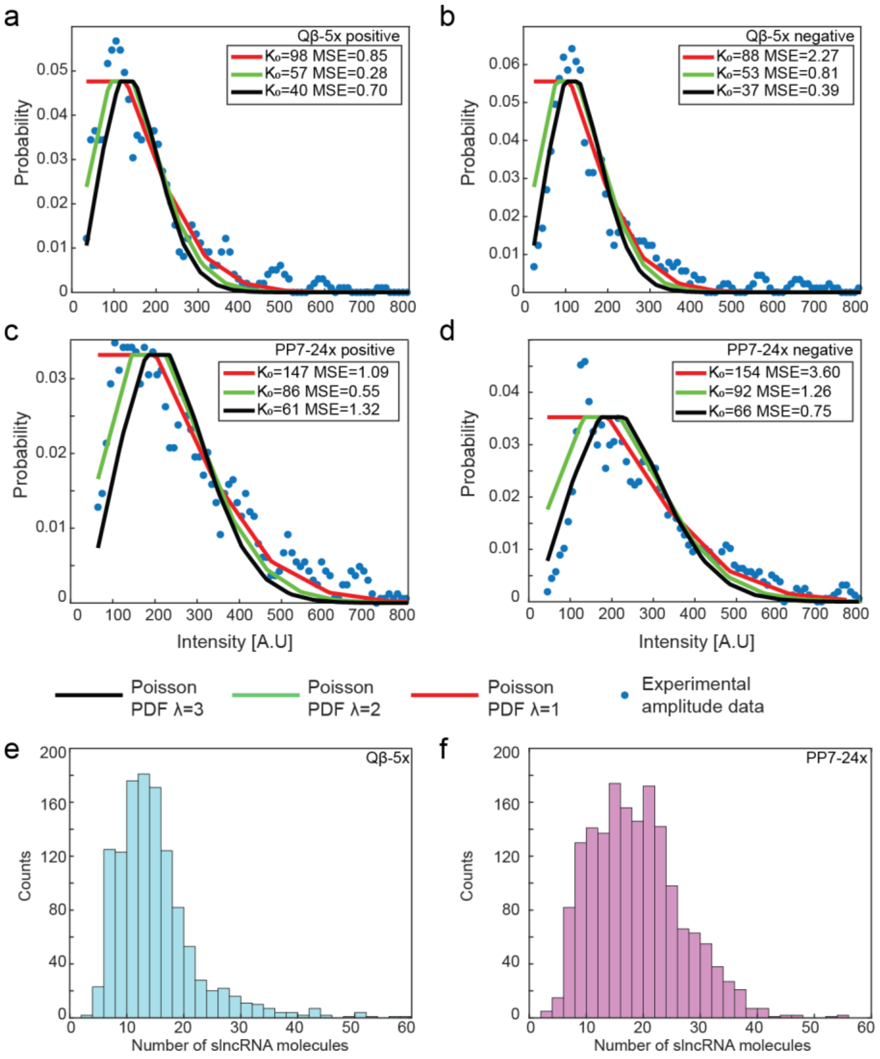

Figure 5. Poisson distribution fitting of empirical amplitude data.

(**a-d**) Experimental data presented by a scatter plot, overlaid by theoretical Poisson probability distribution functions (PDFs) with parameter values 1-3, presented in red, green, and black lines, respectively. Theoretical fits normalized to correct x-axis by the factor K_0 (**a**) Qβ-5x positive amplitudes (*n* = 333). (**b**) Qβ-5x negative amplitudes (*n* = 298). (**c**) PP7-24x positive amplitudes (*n* = 549). (**d**) PP7-24x negative amplitudes (*n* = 533). (**e-f**) Average number of binding sites cassettes per signal, evaluated by dividing the average signal intensity by the value of K_0 calculated from the Poisson PDF with λ=1 fit. (**e**) Qβ-5x data. (**f**) PP7-24x data.

Measuring the effect of background choice on observed signal

To check that our analysis is independent of our choice of image background and segmentation results, we repeated the analysis on all experimental traces using the same definition for a "burst", but with an alternative selection of the sub-frame size used to calculate the background intensity for normalization. A large sub-frame would undoubtedly include other cells, with possibly different spots of themselves, inserting a bias to the background intensity signal. On the other hand, a small sub-frame might not have a sufficient spot-to-background area ratio, resulting in an underestimated background signal.

To test whether this has a meaningful effect on the data, we repeated the entire data analysis process with a sub-frame length of 10 pixels (**Figure *6***), i.e. a 10-pixel-wide sub frame was extracted from the FOV, with the spot at its center, in contrast to the 20-pixel wide used for analysis. Smaller lengths would cause a loss of spot intensity data and larger lengths would bring about biases originating from other cells and therefore were not considered.

Overall, this control comes to show that the statistical results presented in the text, and the conclusions drawn from them do not suffer any change upon altering this step in the data processing. That being said, this control does show that absolute quantitative results, i.e. intensity value per binding site, will be difficult to correctly calculate without further work and calibrations.

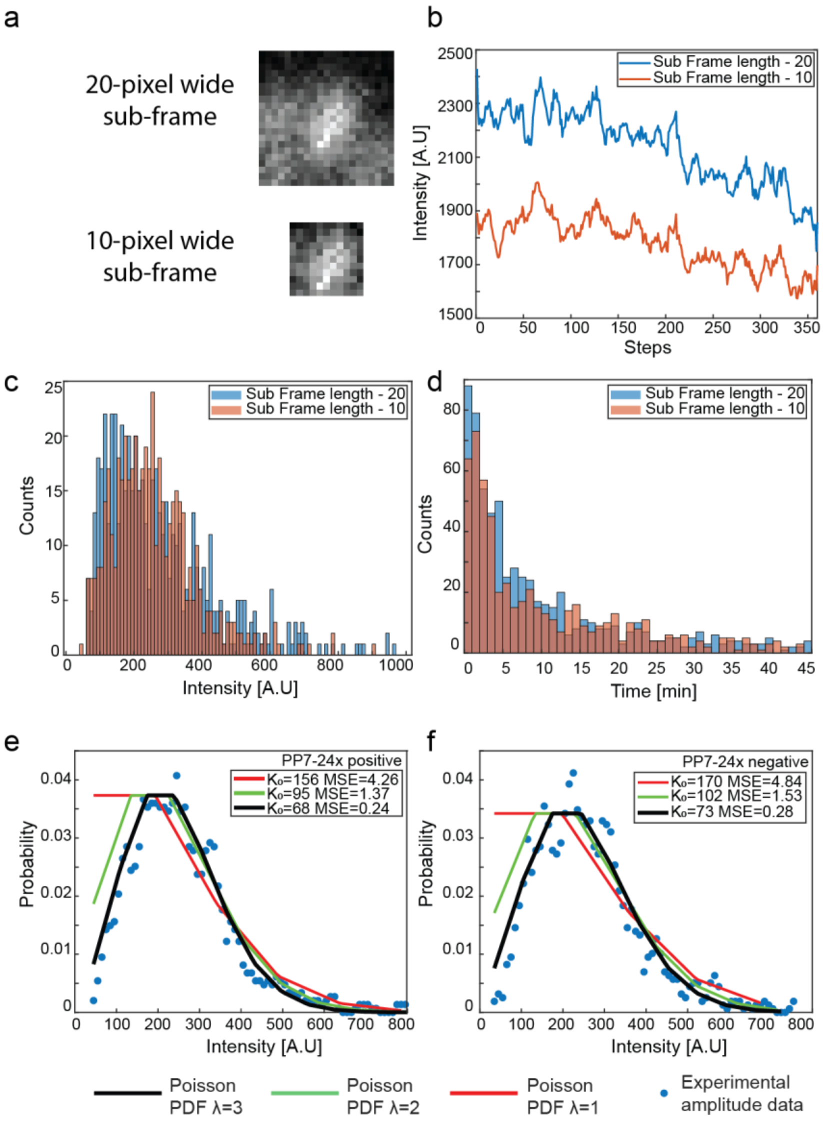

Figure 6. Background selection controls.

(**a**) Sample comparison between two sub-frames with 20-pixel width (top) and 10-pixel width (bottom) (**b**) sample intensity signal (after moving average and background normalization), demonstrating the effects of a small sub-pixel, resulting from the number of pixels being classified as 'spot' being smaller in the 10-pixel wide sub-frame. Despite this difference, both signals show similar behavior (i.e. similar 'hills' and 'valleys'). (**c**) Comparison between positive amplitude statistics for the PP7-24x experiments calculated with the two sub-frame sizes. (**d**) Quiescent segments durations of the PP7-24x showing no discernible difference in the durations of quiescent periods between the two variations. (**e-f**) fittings of the amplitude data gathered using 10-pixel wide sub-frames, to theoretical Poisson probability density functions with rates 1,2,3, similarly to the process described in eq. 3. The optimal K_0 values in this case differ by no more than 10% in all cases, compared to their 20-pixel-wide counterparts presented in Figure 4.

(**e**), and negative amplitudes (**f**)

Signal simulations support a multi-amplitude bursty model

To check that our analysis is consistent with an underlying random burst signal, we simulated three types of base signals with an added white Gaussian noise of magnitude 30 [A.U] peak-to-peak amplitude, matching the value calculated from the experimental traces. For each simulation type, 1000 signals of 360 time-points were simulated and analyzed using the same data analysis process used for the experimental signals.

We simulated a flat, constant signal with noise (**Figure *7* a - top**), a gradually ascending signal with noise (**Figure *7* b - top**), and a three-state random telegraph signal with noise (**Figure *7* c - top**). We then applied our burst-detection algorithm described above and found that for the flat signal (**Figure *7* d**) positive and negative bursts (green and red respectively) and non-classified events are detected. However, a closer examination of the results reveals that the burst amplitude width is smaller by a factor of ~3 as compared with the experimental data bursts, and the total number of events observed is significantly smaller than the experimental data (i.e. 371 positive, 439 negative, and 274 non-classified segments found), indicating less than one event per signal, as expected from our base assumption that a rare noise event occurs once in a thousand time points. For the gradually increasing signal with additional noise, (**Figure *7* e**) a negligible number of burst-like events was detected by our algorithm, with a pronounced bias towards positive events. The scarcity of events can be explained by the positive bias in the signal which results in a steep increase in the statistical threshold for event identification.

Finally, a signal designed to mimic our interpretation of the experimental data containing randomly distributed instantaneous bursts, both increasing and decreasing with multiple possible amplitudes was analyzed (**Figure *7* f**). Our simulated signals resulted in a symmetric amplitude distribution, comprising of non-Gaussian or skewed amplitude distributions. Additionally, the range of amplitudes observed is 2x larger as compared with the case for the constant signal, with the non-classified amplitudes presenting a wider distribution. A total of 2297 positive, 2221 negative and 2981 non-classified segments were found, which is approximately an order of magnitude larger than the number of events observed for the constant signal, and similarly to the density of events observed experimentally, is larger than 1 event per 1000 time-points (2.4 and 6.2 events per 1000 time-steps for the experimental Qβ-5x-PP7-4x and bursty simulated signal respectively).

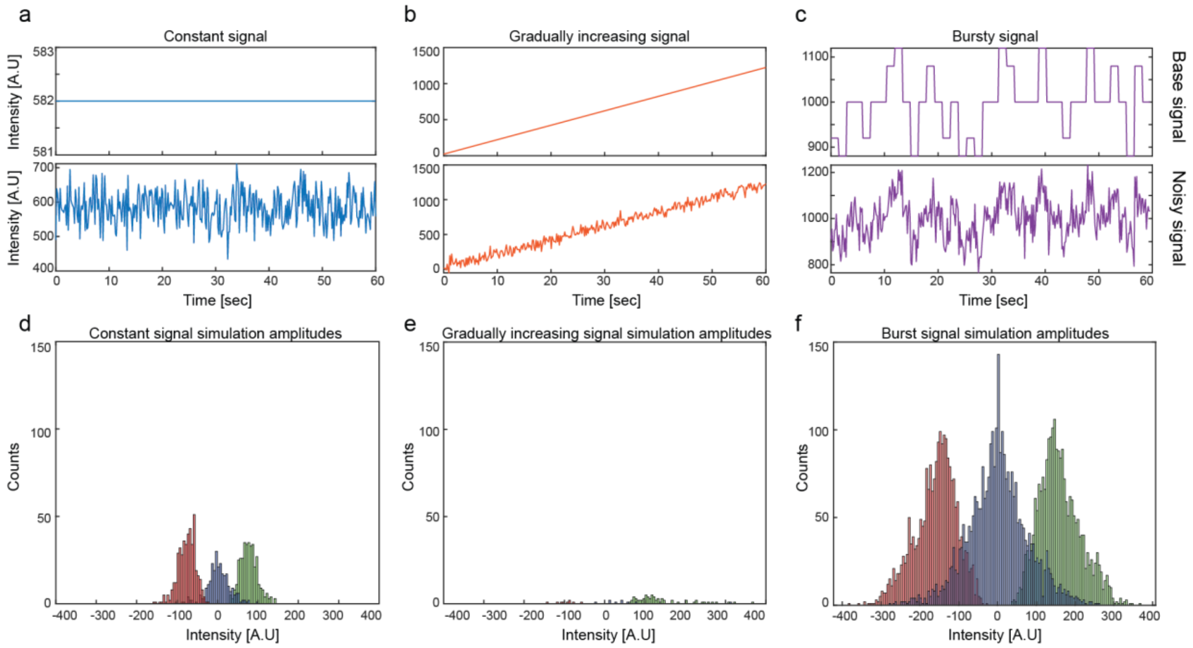

Figure 7. Numerical simulations of potential signals.

(**a-c**) Simulated signal without noise (top) and with Gaussian noise (bottom) for a constant (**a**), gradually increasing (**b**), and intermittent/bursty (**c**) signals. (**d-f**) Amplitude distributions computed with our signal analysis algorithm for the constant (n_{pos}=371, n_{neg}=439, $n_{\text{non-classified}}$=274) (**d**), gradually increasing (n_{pos}=76, n_{neg}=9, $n_{\text{non-classified}}$=9) (**e**), and intermittent/bursty (n_{pos}=2297, n_{neg}=2221, $n_{\text{non-classified}}$=2981) (**f**) signal simulations. In all three panels red, blue, and green bars correspond to strongly decreasing, non-classified, and strongly increasing events, respectively. Given the close match of the intermittent signal simulation to the experimental data, this result indicates that in our experimental data, no more than ~10% of the called "events" are false positives.

Duration of events further supports the three-state random telegraph model

We next compared frequency and duration of burst and non-classified states (**Figure *8***). Burst states (**Figure *8* a-b**) for both the Qβ-5x-PP7-4x and PP7-24x cassettes appear to last approximately 2.5 minutes, irrespective of cassette size or burst-type, indicating a possible temporal resolution issue arising from the statistical analysis process wherein any event shorter than the temporal threshold will be missed. In contrast to these narrow duration distributions, non-classified state durations (**Figure 8c**) are exponentially distributed, with an average decay rate of about 10 minutes for both Qβ-5x-PP7-4x and PP7-24x. The non-classified amplitude distribution (**Figure *8* d**) for the PP7-24x cassette shows a slight preference for slow signal degradation trends (sample skewness of -0.55), that may be consistent with increased mRNA stability that has been previously attributed to these cassettes[34] or to the underlying structural characteristic of the speckle which differ from the Qβ-5x-PP7-4x example. Specifically, the binding sites cassette either slows down, or entirely halts the signal degradation process, resulting in negative amplitudes that do not meet our statistical criteria, but instead appears in the non-classified distribution.

In order to provide context to the information generated by the duration data, we studied the duration of events in our simulated signals. Interestingly, both positive and negative durations for the burst events in each signal bear a striking resemblance to the experimental data (**Figure *8* g - top**). This result is consistent with an interpretation that burst duration measurements are limited by the resolution of the experiment (1 frame every 10 seconds) and choice of smoothing algorithm (ten experimental frames or simulated time points). Together, these constraints result in a lower bound of 150 seconds, or 15 frames on the temporal resolution, in which a burst can be detected. Any process occurring faster would be obscured by the smoothing algorithm and missed. By contrast, important information can be deduced from the simulated duration of the non-classified events (**Figure *8* e-g – bottom**). Here, the three different signals generate visually distinguishable duration signals, which correspond to a distinct fingerprint for each signal type in this case. While both the increasing and constant signal generate a gradually declining and spread-out non-classified duration distributions, the random telegraph signal generated an exponentially distributed duration distribution. This is consistent with the experimental observations and provides further evidence that the underlying signal in our experimental data is a multi-state random telegraph noise.

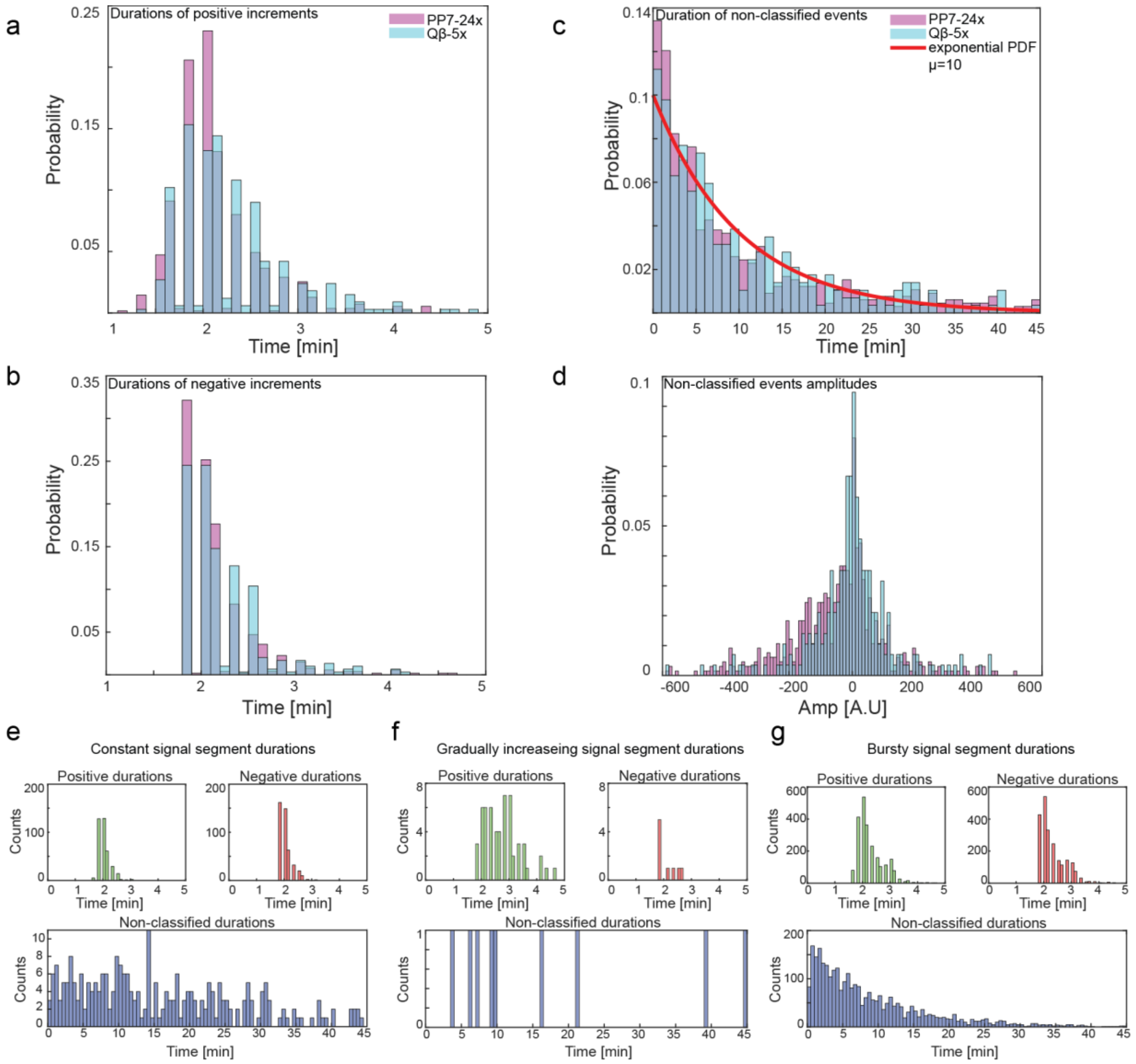

Figure 8. Temporal statistics and non-classified amplitudes for experimental and simulated signals.

(**a**) Duration of positive segments (transcription bursts) showing a Poisson-like distribution shifted by 1 minutes, in both cases. (**b**) Duration of negative segments (degradation bursts), showing an exponential-like distribution shifted by ~2 minutes for both cases. (**c**) Duration of non-classified segments, showing an exponential distribution decay time τ≈10 (min) for both cases. (**d**) Amplitude distribution of non-classified segments. The PP7-24x data is skewed toward negative values (sample skewness of -0.55 for PP7-24x, and 0.83 for Qβ-5x data). For all cases, Qβ-5x-PP7-4x, and PP7-24x cassettes are presented in cyan and magenta, respectively. (**e-g**) Temporal statistics for the simulated signals for the (**e**) constant, (**f**) gradually increasing, (**g**) and bursty signals respectively.

Finally, we checked if the bursts occurred at random or whether there was some bias in the order of the bursts. To do so we examined whether after a non-classified period that lasted more than 2.5 minutes there was a bias for one type of burst or the other. The data shows that no such bias seems to exist, i.e. either a positive or negative burst seems to occur after non-classified events with equal probability (**Figure *9***). This observation is consistent with the fact that we measured fluorescence from bright speckles, which appear after accumulation of multiple binding-sites cassettes, meaning the transcript levels in the cell are at a steady state.

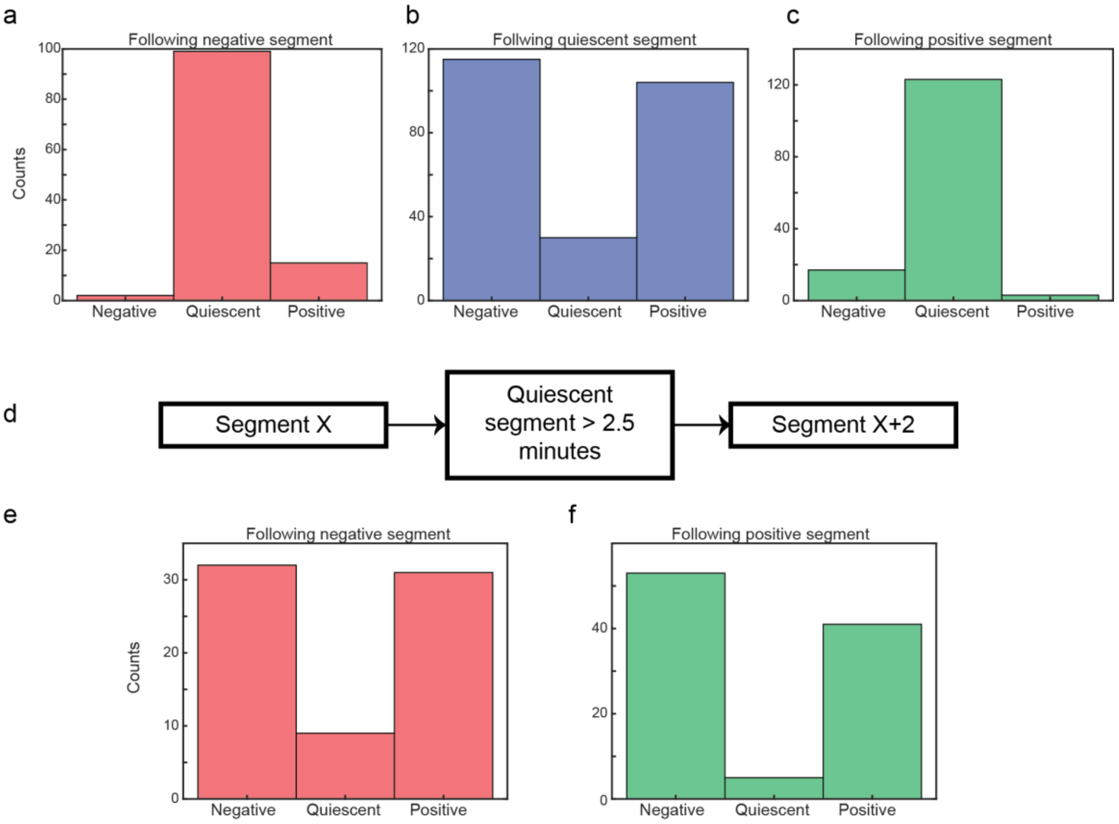

Figure 9. Distribution of burst sequence.

(**a-c**) Distribution of successive bursts according to current burst type. (**a**) negative, (**b**) quiescent, (**c**) positive. (**d**) Schematic of the data presented in (**e,f**). Distribution of bursts following a quiescent event longer than 2.5 minutes, which follows a burst-type. (**e**) Following a negative segment. (**f**) Following a positive segment.

A double random telegraph noise model describes steady state speckle intensity

The existence of negative bursts that appear to be independent of labelling cassette implies that the well-studied two-state model for transcription dynamics[11,12,45], which incorporates one state with a strong transcriptional rate and another state with a weak rate, must be modified, since the kinetics of this model cannot generate two isolated sets of bursts. We therefore add another degradation rate, extending the two-state to a four state-model.

The mathematical derivations presented here are based on the work done by Sanchez et al[45]. The general form of the model contains two stochastic variables: the number of mRNA molecules *n*, and the state of the system *S*, with the system being the two major processes controlling speckle fluorescence, transcription and degradation of molecules. The general model (**Figure *10***), is comprised of four possible states for the system, strong transcription and strong degradation; strong transcription and weak degradation; weak transcription and strong degradation; weak transcription and weak degradation. The rate of transcription is determined by the state of the promoter (open/closed), and the rate of degradation is similarly determined by the accessibility of mRNA molecules to RNase binding or the availability of the degradation enzymes.

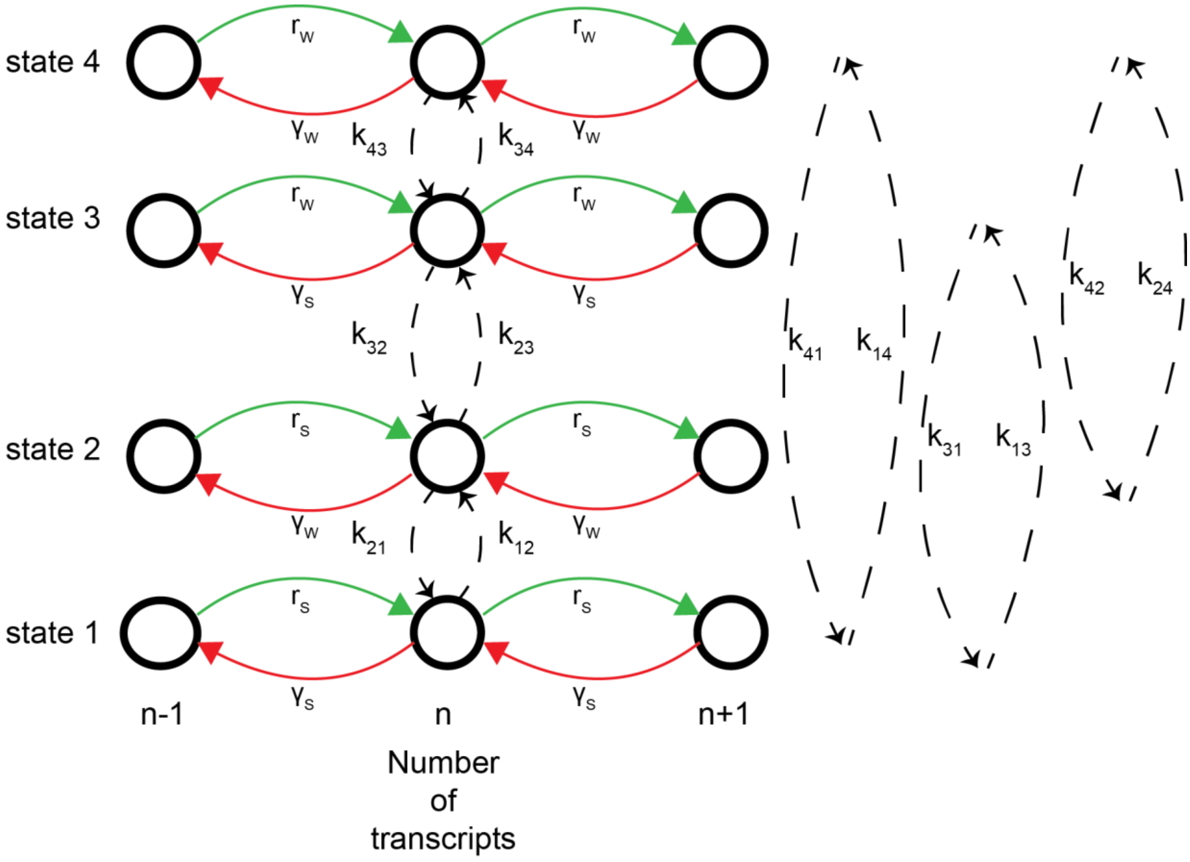

Figure 10. Scheme of the general four-state model.

Black circles represent the possible states of the system for any *n>1* number of RNA molecules. Red and green arrows indicate degradation and transcription processes, respectively, governing the number of molecules. Dashed arrows stand for transitions between system states.

The temporal dynamics of this stochastic system are given by the master equation, which can be derived by listing all possible reactions leading to a change in *S* or in *n*.

$$\begin{aligned}
\frac{d}{dt}P_{1,n} &= r_S P_{1,n-1} + \gamma_S(n+1)P_{1,n+1} + k_{21}P_{2,n} + k_{31}P_{3,n} + k_{41}P_{4,n} - P_{1,n}[\gamma_S n + r_S + k_{12} + k_{13} + k_{14}] \\
\frac{d}{dt}P_{2,n} &= r_S P_{2,n-1} + \gamma_W(n+1)P_{2,n+1} + k_{12}P_{1,n} + k_{32}P_{3,n} + k_{42}P_{4,n} - P_{2,n}[\gamma_W n + r_S + k_{21} + k_{23} + k_{24}] \\
\frac{d}{dt}P_{3,n} &= r_W P_{3,n-1} + \gamma_S(n+1)P_{3,n+1} + k_{13}P_{1,n} + k_{23}P_{2,n} + k_{43}P_{4,n} - P_{3,n}[\gamma_S n + r_W + k_{31} + k_{32} + k_{34}] \\
\frac{d}{dt}P_{4,n} &= r_W P_{4,n-1} + \gamma_W(n+1)P_{4,n+1} + k_{14}P_{1,n} + k_{24}P_{2,n} + k_{34}P_{3,n} - P_{4,n}[\gamma_W n + r_W + k_{41} + k_{42} + k_{43}]
\end{aligned} \tag{4}$$

Here the subscripts ‘S’ and ‘W’, stand for strong and weak respectively, indicating the strength of the biochemical process.

Using the following definitions:

$$\begin{aligned}
&\vec{P}(n) = \begin{pmatrix} P_{1,n} \\ P_{2,n} \\ P_{3,n} \\ P_{4,n} \end{pmatrix} \\
&\hat{K} = \begin{pmatrix} -[k_{12}+k_{13}+k_{14}] & k_{21} & k_{31} & k_{41} \\ k_{12} & -[k_{21}+k_{23}+k_{24}] & k_{32} & k_{42} \\ k_{13} & k_{23} & -[k_{31}+k_{32}+k_{34}] & k_{43} \\ k_{14} & k_{24} & k_{34} & -[k_{41}+k_{42}+k_{43}] \end{pmatrix} \\
&\hat{R} = \begin{pmatrix} r_S & 0 & 0 & 0 \\ 0 & r_S & 0 & 0 \\ 0 & 0 & r_W & 0 \\ 0 & 0 & 0 & r_W \end{pmatrix} ; \ \hat{\Gamma} = \begin{pmatrix} \gamma_S & 0 & 0 & 0 \\ 0 & \gamma_W & 0 & 0 \\ 0 & 0 & \gamma_S & 0 \\ 0 & 0 & 0 & \gamma_W \end{pmatrix}
\end{aligned} \tag{5}$$

We can write the above equations in matrix form.

$$\frac{d}{dt}\vec{P}(n) = \left[\hat{K} - \hat{R} - n\hat{\Gamma}\right]\vec{P}(n) + \hat{R}\vec{P}(n-1) + (n+1)\hat{\Gamma}\vec{P}(n+1) \tag{6}$$

The matrix $\hat{K}$ describes the transition rates between system states, the matrices $\hat{R}$ and $\hat{\Gamma}$ describe the rates of transcription initiation rates, and degradation rates, respectively.

At this point, we can derive equations from which the first and second moments of the mRNA distribution in steady state can be computed. The derivation is similar to the one presented by Sanchez et al[45] and therefore will not be repeated here. Since it is convenient to write the resulting equations in terms of partial moments of the mRNA distribution, they will be defined here as:

$$\vec{n}_0 = \sum_{n=0}^{\infty} n^0\, \vec{P}(n) = \begin{pmatrix} \sum_{n=0}^{\infty} n^0 P_{1,n} \\ \sum_{n=0}^{\infty} n^0 P_{2,n} \\ \sum_{n=0}^{\infty} n^0 P_{3,n} \\ \sum_{n=0}^{\infty} n^0 P_{4,n} \end{pmatrix} = \begin{pmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{pmatrix} \;;\; \vec{n}_1 = \sum_{n=0}^{\infty} n\, \vec{P}(n) = \begin{pmatrix} \sum_{n=0}^{\infty} nP_{1,n} \\ \sum_{n=0}^{\infty} nP_{2,n} \\ \sum_{n=0}^{\infty} nP_{3,n} \\ \sum_{n=0}^{\infty} nP_{4,n} \end{pmatrix};$$

$$\vec{n}_2 = \sum_{n=0}^{\infty} n^2\, \vec{P}(n) = \begin{pmatrix} \sum_{n=0}^{\infty} n^2 P_{1,n} \\ \sum_{n=0}^{\infty} n^2 P_{2,n} \\ \sum_{n=0}^{\infty} n^2 P_{3,n} \\ \sum_{n=0}^{\infty} n^2 P_{4,n} \end{pmatrix} \tag{7}$$

$\vec{n}_0$ can be found from the equation:

$$\hat{K}\vec{n}_0 = 0 \tag{8}$$

Together with the normalization condition:

$$P_1 + P_2 + P_3 + P_4 = 1 \tag{9}$$

$\vec{n}_1$ from the equation:

$$\left(\hat{K} - \hat{\Gamma}\right)\vec{n}_1 + \hat{R}\vec{n}_0 = 0 \tag{10}$$

$\vec{n}_2$ from the equation:

$$\hat{K}\vec{n}_2 + \hat{R}(2\vec{n}_1 + \vec{n}_0) + \hat{\Gamma}(\vec{n}_1 - 2\vec{n}_2) = 0 \tag{11}$$

Using these equations, the Fano factor can be extracted and computed numerically.

Biological assumptions can be made to simplify the model in order to simulate it using Monte Carlo methods. To this end, we give the different states biological meaning.

Model State	Biological state
1	Strong transcription; strong degradation
2	Strong transcription; weak degradation
3	Weak transcription; strong degradation
4	Weak transcription; weak degradation

Table 3. *Biological meaning of model states*

First, we assume that a simultaneous transition of both transcription and degradation is unlikely, i.e. $k_{14}, k_{41}, k_{23}, k_{32} = 0$. Second, transition rates in a specific system do not change between the different states, meaning that, switching of a promoter from ‘open’ to ‘closed’ will have the same rate regardless of degradation and vice versa. In mathematical terms:

$k_{12} = k_{34} = k_{off}^{d}$	Strong degradation → Weak degradation
$k_{21} = k_{43} = k_{on}^{d}$	Weak degradation → Strong degradation
$k_{13} = k_{24} = k_{off}^{t}$	Strong transcription → Weak transcription
$k_{31} = k_{42} = k_{on}^{t}$	Weak transcription → Strong transcription

Table 4. Parameter designations for biological model

Finally, for simplicity we set $\gamma_W = 0$, as we believe this value to be insignificant compared to γ_L

The matrices then take the following form:

$$\hat{K} = \begin{pmatrix} -\left[k_{off}^{d} + k_{off}^{t}\right] & k_{on}^{d} & k_{on}^{t} & 0 \\ k_{off}^{d} & -\left[k_{on}^{d} + k_{off}^{t}\right] & 0 & k_{on}^{t} \\ k_{off}^{t} & 0 & -\left[k_{on}^{t} + k_{off}^{d}\right] & k_{on}^{d} \\ 0 & k_{off}^{t} & k_{off}^{d} & -[k_{on}^{t} + k_{on}^{d}] \end{pmatrix}$$

$$\hat{R} = \begin{pmatrix} r_S & 0 & 0 & 0 \\ 0 & r_S & 0 & 0 \\ 0 & 0 & r_W & 0 \\ 0 & 0 & 0 & r_W \end{pmatrix} \; ; \; \hat{\Gamma} = \begin{pmatrix} \gamma_S & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma_S & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{12}$$

We note that we cannot differentiate between all four possible states experimentally, as two of these result in similar mRNA behavior. Depending on parameter choices, the indistinguishable pair could be states 1 and 2 (assuming the effect of transcription is stronger regardless of degradation state), or states 1 and 4 (assuming the effect of transcription to be similar to that of degradation). Since this is a living biological system, it is safe to assume the former case.

In practical terms, the first and second states result in transcriptional bursts and cannot be discerned neither in simulation nor experiments. The third state results in negative bursts when the degradation rate is larger than the weak transcriptional rate, while the fourth state results in a quiescent state of apparent inactivity when both transcription and degradation rates are sufficiently small. Therefore, in order to account for the negative-burst finding, we are left with three discernible states, forming a model similar to the three-state random telegraph noise (RTN) model[46] (**Figure *11* a**).

To test this new model, we implemented both it, and the traditional two-state model, using kinetic Monte Carlo simulations[47]. We simulated time-lapse sequences, modelling the number of transcripts as a function of time. Both simulations were run for 3600 steps (corresponding to 1 hour, simulated at 1 sec intervals), with sampling every 10 steps to emulate the sampling conditions of the experimental data. The transition parameters between states for the two-state model, and for the transcriptional burst part of the three-state model are given in table 5.

Kinetic transition	**Mathematical Notation**	**Biological meaning**	**Value [s^{-1}]**
Closed promoter -> Open promoter	k_{on}^{t}	RNAP association	0.0027
Open promoter -> Closed promoter	k_{off}^{t}	RNAP disassociation	0.0023
Open promoter -> mRNA	r_S	Strong transcription	0.04
Close promoter -> mRNA	r_W	Weak transcription	0.003
mRNA -> Quiescent state	k_{off}^{d}	Halted degradation	0.2
Quiescent state -> mRNA	k_{on}^{d}	Resumed degradation	0.4
mRNA degradation	γ_S	Degradation rate	0.011

Table 5. *Parameters used for kinetic modelling*

The association/disassociation rates were gathered from the work by Sanchez et al[45] , and are valid for the pLac promoter in *E. coli*. These rates do not represent empirical biological kinetic data, as no *in-vivo* association/disassociation rates could be found for the T7 RNAP in literature. Instead, these should be regarded as qualitative only, representing the order of magnitude to the association of a protein (e.g. RNAP, transcription factor) to a promoter, in a crowded cellular environment.

The transition parameters into and out of the degradation burst state were set such that the distribution of number of mRNA molecules would be similar between the two models, meaning that there should be no apparent change in the number of transcripts (**Figure *11* b**).

To confirm that a three-state RTN model is better suited for describing our experimental data, we used two different methods to compare its performance to that of the two-state model. The first test is rooted in telegraph processes theory, which provided the basis for the historic two-state model[48,49]. A common analysis of signals generated from such processes is based on the power spectral density (PSD), which is proportional to the square of the Fourier transform of the signal[50]. To generate a suitable signal from our data, we concatenated the slopes of all identified segments from all signals (under an assumption of memoryless noise), generating a signal resembling a telegraph process. We calculated the PSD for both simulations and experimental data (**Figure *11* c**). From a heuristic viewpoint, the experimental PSD bears a closer similarity to that of the three-state model in the low frequency range, as compared with the PSD of the two-state model, providing further evidence that the three-state model provides a better description for the experimental data.

For the second test, we analyzed the simulated sequences in the same manner used for the experimental data, to generate amplitude histograms (**Figure *11* d-e**) (3 state and 2-state respectively). The most notable difference between the two models lies in the distribution of negative amplitudes. While the two-state model produces a Gaussian distribution centered at -1, representing constant degradation, the three-state simulation generates a more flattened, spread-out distribution, similar to the experimental data (**Figure *4* d**).

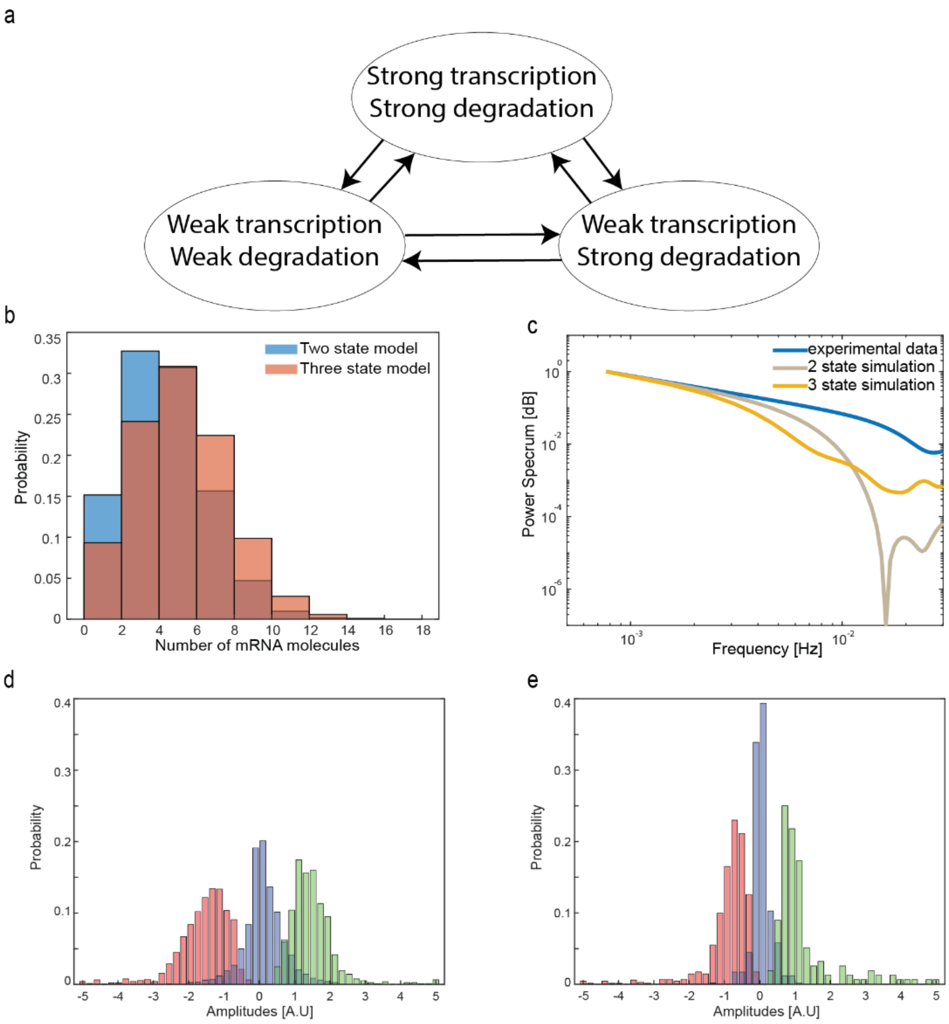

Figure 11. Three state model analysis.

(**a**) Model schematic showing transition possibilities between the three experimentally discernible states. (**b**) Number of mRNA molecules arising from simulations of three-state (orange) and two-state (blue) models (n = 300). (**c**) Power spectrum density of experimental Qβ-5x, three-state model and two-state model. (**d-e**) Intensity amplitude analysis for the three-state model (n_{pos}=769, n_{neg}=1266, $n_{non\text{-}classified}$=1480) (**d**) and two-state model (n_{pos}=156, n_{neg}=691, $n_{non\text{-}classified}$=584) (**e**).

Chapter 4: Section Summary

In the present study, we aimed to learn the kinetic features of paraspeckles and other, similar nuclear bodies via a synthetic bacterial system. We used the pT7 promoter in bacteria to express synthetic lncRNA (slncRNA) molecules that incorporated several RBP binding sites, which together with their cognate RBP formed fluorescent speckles in the cell poles. Our speckles were composed of two different types of slncRNA molecules or binding-site cassettes: a new Qβ-5x-PP7-4x and the common-place PP7-24x cassette. Our findings reveal that at steady-state, speckles are likely composed of 15-30 RBP bound slncRNA molecules. Changes in speckle fluorescence is characterized by exponentially distributed random bursts of either positive or negative amplitudes, spaced by periods where no classifiable change in speckle fluorescence can be made. To our knowledge there was one previous study in bacteria that tracked synthetic fluorescent speckles in single cells[9]. In that study, the RNA component was an mRNA gene that was labelled at its 5'UTR by a ~5 kbp cassette of 96 binding sites for the phage coat protein for MS2, which in turn self-assembled into speckles that only exhibited bursts of increasing fluorescence intensity that were attributed to bursts of transcriptional activity. In this study, we identified both positive and negative bursts in approximately equal proportions in synthetic speckles composed of two significantly shorter RNA molecules. As in the Golding *et al.*[9] paper, we also interpret positive bursts as a near simultaneous addition of multiple fluorescent slncRNA molecules to the speckle resulting from bursts of transcriptional activity. By contrast, a negative burst can be interpreted as a simultaneous removal of slncRNA molecules from the speckle. It is therefore reasonable to conclude that negative bursts may correspond to bursts of degradation.

Unlike transcriptional bursts, which have been attributed to kinetic and structural nucleoid-related processes, there is no particular reason why degradation of speckle signal in bacteria should also present itself in bursts. There seem to be two possible kinetic scenarios; either slncRNA molecules are actively degraded in an intermittent fashion directly within a given speckle, or the speckle itself sheds intermittently a number of slncRNA molecules which are then further degraded in the cytoplasm. The implications of the first scenario are that RNase and potentially other major degradation enzymes may also be expressed in bursts themselves leading to an intermittent set of degradation events. The second scenario implies that speckles are nano-particles which are comprised of entangled protein-RNA complexes that are effectively phase-separated from the rest of the cytoplasm and thus do not allow access to other molecular species such as the ribosome or degradation enzymes. When considering the increased negative skewedness observed in the non-classified set of events for the 24x lncRNA as compared with its 5x shorter counterpart (**Figure *8* d**), the latter scenario offers a more compelling explanation. Namely, synthetic speckles that are composed of increasingly complex slncRNA molecules (i.e. more binding sites) are likely to be more entangled leading to a slower release of the molecules from the biomolecular complex as compared with their shorter counter parts. An increased entanglement due to binding site number is also consistent with the wider distribution for the estimated "number of slncRNAs" within speckles observed for the 24x as compared with the shorter example (**Figure *5* e-f**), and the lack of negative bursts observed for the 96x speckle by Golding *et al.*[9].

Observing both positive and negative fluorescent bursts and considering the presence of a phase separated compartment which stores RNA and proteins led us to a new modelling scheme for the regulation of gene expression. While a two-state RTN kinetic model suffices for describing the mRNA distribution resulting from bursty gene expression process, in order to take the effects of speckles into account a multi-state RTN model is needed. Here, the presence of the speckles was accounted for by adapting the two-state model to one which included an additional state where degradation is a rare event. This, in turn, led to the desired three-state RTN prediction. Consequently, this prediction may be relevant to not only this synthetic context, but to natural systems as well. This, therefore, become particularly pertinent in lieu of many recent observations in Eukaryotes, which suggests that intracellular liquid-liquid phase-separation into molecularly isolated compartments[51] (e.g. paraspeckles[1–3], nuclear speckles[4], etc.) may be a common and generic cellular phenomenon that affects the regulation of gene expression. Thus, bursting may not just be a process that characterizes transcription but may be characteristic of other molecular processes related to the regulation of gene expression as well.

Finally, we believe that the ability to reach the negative burst finding is due in large part to our new binding-site cassette design. Our cassette is ~450 bps in length, which is about 5 times shorter than the 24x, and an order of magnitude shorter than the cassette used by Golding *et al.*[9]. Comparison of our cassette to the results generated by the PP7-24x indicates that our cassette is likely to be fully occupied by RBPs, as compared with <50% occupancy for the PP7-24x, as has been previously reported[34].

Section II: Anomalous Diffusion Analysis

Abbreviations and Notations

SPT	Single Particle Tracking
FBM	Fractional Brownian Motion
CTRW	Continuous Time Random Walk
(TA)MSD	(Time Averaged) Mean Square Displacement
CNN	Convolutional Neural Network
H	Hurst exponent
D	Diffusion coefficient
MT-network	Multi-Track network
ST-network	Single-Track network
STD	Standard deviation

Due to the stochastic nature of diffusive processes, a single trajectory does not uniquely correspond to a type of diffusion, unless long enough to show asymptotic behaviour. For example, Fractional Brownian Motion (FBM) and Continuous Time Random Walk (CTRW) can both produce highly similar motion characteristics[15,17], yet arise from two completely different underlying physical mechanisms. It is therefore usually only possible to assign a probable mechanism by examining either very long trajectories or multiple, intermediate-length trajectories. This analysis is most frequently based on the mean squared displacements (MSD) curves, which describe the average spatial distances, $\langle r^2 \rangle$, within a trajectory measured between increasingly long time lags, Δt or τ[52]. This analysis has the benefit of being simple to apply, and the function scales with time in the form of a power law $\langle r^2 \rangle \propto \tau^{\alpha}$. For normal diffusion (*i.e.*, pure Brownian motion), $\alpha = 1$; whereas, anomalous diffusion can be subdiffusive ($\alpha < 1$), or superdiffusive ($\alpha > 1$), (**Figure *12***)[53,54]. While MSD is the most commonly used approach, it is worth noting that other methods can also distinguish between diffusion types and extract the power law parameter, α; e.g. mean maximal excursion[55], and power spectral density analysis[56].

Several families of frequently-encountered processes leading to anomalous diffusion include: FBM, CTRW, random walk on a fractal, and Lévy flights[15]. Classifying a trajectory to a diffusion model can reveal a great deal about the physical environment being investigated; for example, FBM might imply a crowded cellular environment, while CTRW can indicate an environment containing traps[57].

Existing methods for identification of the best-fitting diffusion model fall into two categories: qualitative (searching for ergodicity-breaking behaviour or spatial constrains during motion), and quantitative (p-variation test, Gaussian-breaking parameters). Problematically, many of these criteria concern asymptotic behaviour, and therefore require long single trajectories or hundreds of moderate-length trajectories from the same environment[57,58]. Additionally, these methods are ill-suited in cases of subordinated diffusion, where more than one mechanism determines the type of motion[59,60].

Yet another issue is how to best determine model-specific parameters from a dataset for a given theoretical model. For example, using only the first two time lags of the MSD curve to determine the diffusion coefficient in pure Brownian trajectories was shown to be preferable under most conditions because it was more robust to noise[61]. Identifying the optimal approach for the relevant parameters describing subdiffusive motion has not yet been investigated, despite known inaccuracies in applying traditional methods to deficient datasets[53,54]. A particular interest is how to make use of very short trajectories to accurately extract subdiffusive parameters, i.e. those produced by single-molecule fluorescence microscopy experiments using genetically encoded labels. While these probes are seemingly ideal for reporting on the biological environments[62], their limited photostability and brightness typically produces many short trajectories. The key challenge presented by these experiments is how to infer material properties of biological systems and transport properties of single biomolecules from very short trajectories in which the asymptotic behaviour is experimentally inaccessible.

The applicability of classical methods for accurately extracting the underlying parameters in this regime has been somewhat limited, thus necessitating a more reliable approach. Machine-learning algorithms, and in particular deep learning[63], excel at extracting concealed correlations in large datasets which can then be used to create a predictive tool for analysis of similar data[64]. This makes the problem of characterizing single-particle diffusion well suited for deep-learning analysis.

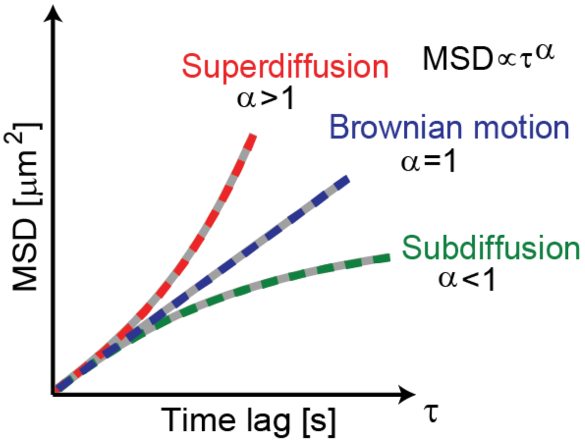

Figure 12. Mean Square Displacement curves.

Mean Square Displacement (MSD) curves for sub-, super- and Brownian diffusion.

Chapter 6: Methods

Diffusion trajectories

We focus on three diffusion models which are often encountered in SPT experiments: Brownian motion, FBM and CTRW (**Figure *13***). Training a neural network generally requires tens of thousands of data samples to achieve high performance. To answer this requirement, we simulated trajectories governed by each of the three motion models.

Brownian Motion

Brownian motion was generated as a random walk process with independent identically distributed (IID) Gaussian steps (eq 13) with $\{s_i, i \geq 1\}$ a zero-mean Gaussian process, and N_t the total number of steps taken up to time t.

$$x(t) = \sum_{i=1}^{N_t} s_i \qquad (13)$$

Fractional Brownian Motion

$x(t)$ is an FBM process if it is a continuous Gaussian process with a mean of zero, and it satisfies the following covariance function (eq. 14):

$$cov(x_t, x_s) = \frac{1}{2}(|t|^\alpha + |s|^\alpha - |t-s|^\alpha), t, s \geq 0 \tag{14}$$

Where x_t, x_s are positions in the trajectory and α is the anomalous exponent (twice the Hurst exponent). Simulated FBM trajectories were generated using the circulant embedding algorithm[65] which is a fast simulation method for stationary Gaussian processes on uniformly spaced grids which is based on Fourier transformations.

Continuous Time Random Walk

CTRW can be regarded as a combination of random walks in both time and space, with temporal 'steps' (waiting times) drawn from a heavy-tailed distribution with an asymptotic power-law behavior (eq. 15).

$$h(t) \sim \frac{1}{t^{\alpha+1}}, t \rightarrow \infty \tag{15}$$

This general formulation has led to many variations over the years[66]; we chose to implement the uncoupled CTRW simulation presented by Fulger and coworkers[67] due to its relative simplicity and speed. In short, spatial increments are drawn from a symmetric α-stable Lévy distribution (in our case, $\alpha = 2$, corresponding to a Gaussian distribution), and temporal increments are drawn from a Mittag-Leffler distribution. The two distributions are then matched to produce a spatio-temporal trajectory comprising of a power-law distribution in dwell times.

Experimental trajectories

For experimental validation of our method, we performed three experiments, each illustrating a different diffusion mechanism: diffusion of fluorescent beads in actin networks, demonstrating FBM, where the Hurst exponent, H, varies as a function of gel-mesh density[68,69]; diffusion of fluorescent beads in a water-glycerol solution, demonstrating pure Brownian motion[70]; and diffusion of TrkB and p75 receptors along the plasma membrane of HEK293T cells, demonstrating FBM subordinate to CTRW[71]. The experimental data is available online[72]

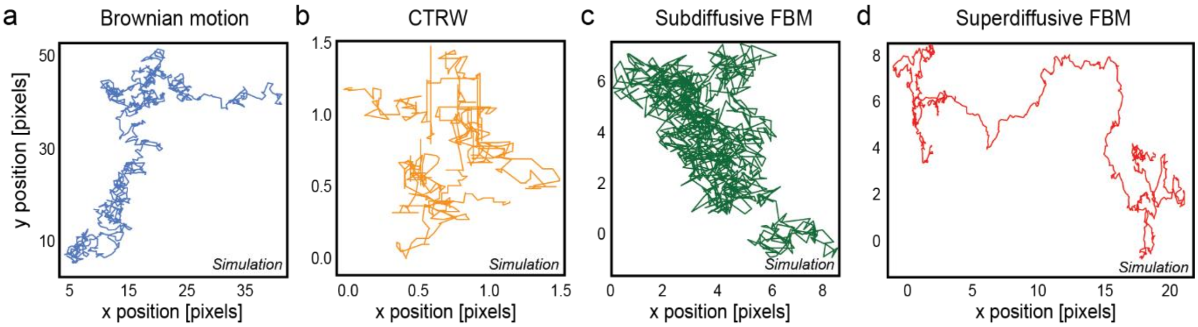

Figure 13. Diffusion models.

(**a-d**) Simulated trajectories illustrating diffusion processes discussed in this work. (**a**) Brownian motion. (**b**) Continuous time random walk. (**c**) Subdiffusive FBM. (**d**) Superdiffusive FBM

SNR definition

We define the signal-to-noise ratio (SNR) of a simulated trajectory x_{sim} as the standard deviation of signal increments divided by the standard deviation of the Gaussian noise added to the signal.

$$SNR_{sim} \triangleq \frac{std\left(\frac{dx_{sim}}{dt}\right)}{std(N)} \tag{16}$$

where N is a zero-mean Gaussian process. An example for SNR=1,5 can be seen in (**Figure *14***).

Experimental SNR is defined as the standard deviation of signal increments of a diffusing agent x_{ex} divided by the standard deviation of signal increments of an immobile agent x_{fixed} from the same environment.

$$SNR_{exp} \triangleq \frac{std\left(\frac{dx_{ex}}{dt}\right)}{std\left(\frac{dx_{fixed}}{dt}\right)} \tag{17}$$

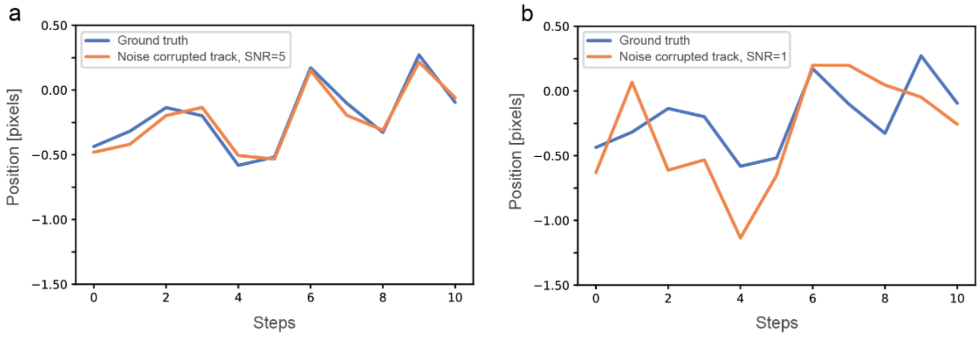

Figure 14. Simulated trajectories containing noise.

(**a-b**) Ground truth trajectory (blue) with added localization precision (orange) (**a**) SNR = 5 (**b**) SNR = 1.

Mean square displacement calculation

Time averaged mean squared displacement (TAMSD) was calculated as:

$$\delta^2(\tau) = \frac{1}{\frac{L}{\Delta t}} \cdot \sum_{m=1}^{\frac{L}{\Delta t}} \big(x(m\Delta t + \tau) - x(m\Delta t)\big)^2 \tag{18}$$

Where $x(t)$ is a trajectory of length L, taken at time intervals Δt.

Ensemble MSD was calculated as the average of TAMSDs obtained from different trajectories. For fractional Brownian motion, $\delta^2(\tau) \propto \tau^{2H}$, therefore the Hurst exponent (H) was estimated by fitting log $(\delta^2\big(log(\tau)\big)$ to the linear function $a + bx$, where $b = 2H$. For Brownian motion, $\delta^2(\tau) \propto 2D\tau$ (for the one-dimensional case), therefore the diffusion coefficient was estimated by fitting the first 5 time-lags of log $(\delta^2\big(log(\tau)\big)$ to the linear function $a + bx$, where $a = 2D$.

CNN network architecture

Network architecture is based on the design proposed by Bai and co-workers[73]. In brief, four sets of convolution blocks with different filter sizes [2, 3, 4, & 10], operate in parallel (**Figure *15* a**). Each block consists of 1D dilated causal convolution layers with increasing dilation factors (**Figure *15* b**). This setup is meant to find correlations spanning multiple time scales of unknown length. The specific architecture used was selected in a process of trial and error, based on classification and regression performance on simulated data.

Each net was constructed for a specific input trajectory length, namely, by the inclusion of additional convolution layers. For example, the network designed for 1000-step trajectories, relative to the 100-step network has an additional convolution block with a filter size of 20.

For the Multi-Track (MT) networks, the 1D convolution layers were replaced by 2D convolution layers with dilation factors operating on the temporal axis only (i.e. for an input matrix, M, with the shape [Number of tracks × Number of steps] the dilation factor will be [1 × d]).

Classification networks for differentiating between discrete motion types were trained with categorical cross entropy loss, with the following mathematical formula:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} 1_{y_i \in C_c} log P_{model}[y_i \in C_c] \tag{19}$$

The double sum is over the observations i, and the categories *c* whose number is C. The term $1_{y_i \in C_c}$ is an indicator function of observation *i* belonging to the category *c*. The term $P_{model}[y_i \in C_c]$ is the predicted probability that observation *i* belongs to the category *c*.

Regression networks, which estimate a continuous variable, were trained with mean squared error loss, with the following mathematical formula:

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^p)^2 \tag{20}$$

where y_i is the ground truth and y_i^p the predicted value of observation i.

Networks were implemented and trained using Keras (Version 2.2.4) with TensorFlow backend version (1.8.0) in Python (version 3.5). Other packages used: NumPy (version 1.14.5), SciPy (version 1.2.1), Stochastic (version 0.4). Training was done on NVIDIA GeForce Titan GTX graphics card in a Windows environment.

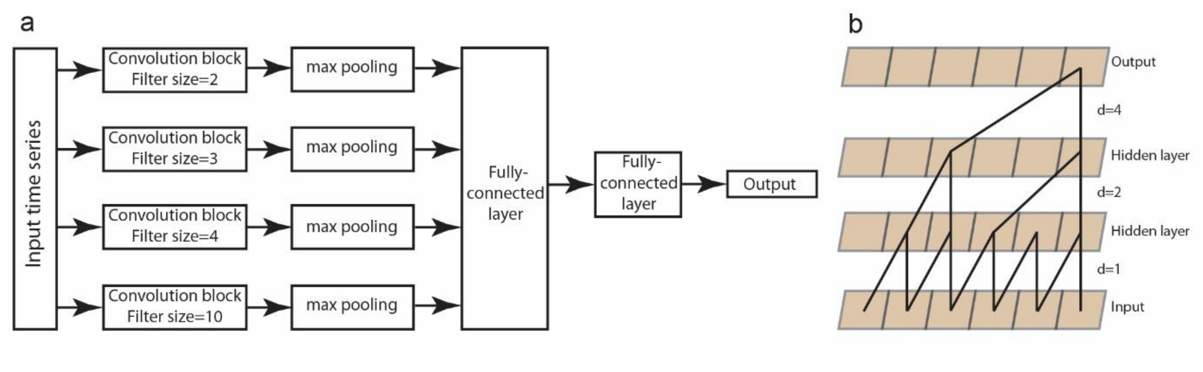

Figure 15. Neural network architecture.

(**a**) Schematic of basic network structure. (**b**) An example convolution block with filter size = 2 and dilation factors $d = 1,2,4$.

Fluorescent beads in Glycerol solution

100 nm green and 200 nm red fluorescently labelled microspheres (Life Technology) were diluted into 40% glycerol in water (v/v). From the mixture 1 µL was pipetted onto a glass coverslip and pressed onto a glass slide and sealed with clear nail polish. Both surfaces were pretreated with a ~20 mg/mL casein solution to decrease the propensity for sticking of the fluorescent beads to the glass. Three fluorescent filters were used in combination with different excitation-laser combinations to image the green, red, and both beads at once. Green-bead images were recorded with a 488 nm excitation; red beads were imaged with a 650 nm laser. Imaging was done using a multi-bandpass filter set (ZT405/588/561/647rpc, ZET405/488/561/647m, Chroma). All imaging was done using a standard inverted microscope system (TI Eclipse, Nikon) with a 20X objective (Plan Apo NA .75, Nikon) using an sCMOS detector (Photometrics). Movies were recorded with 50 ms frames using NIS Elements software (Nikon) and analyzed using the Mosaic plugin[39] for FIJI[41].

Beads in glycerol experiment theoretical calculations

Theoretical diffusion coefficient values for the diffusion-in-glycerol experiment were calculated using the Stokes-Einstein equation for diffusion of spherical particles through a liquid with low Reynolds number[74].

$$D = \frac{K_B T}{6\pi\eta r} \tag{21}$$

Where:
D – Diffusion coefficient $\left[\frac{\mu m^2}{s}\right]$
K_B – Boltzmann constant $\left[\frac{m^2 kg}{k\cdot s^2}\right]$
T – Temperature $[k]$
η – Viscosity $\left[\frac{kg}{m\cdot s}\right]$
r – Particle radius $[nm]$

The experiments were conducted in room temperature with beads of two different sizes, $100{,}200$ $[nm]$ in a solution of 40% glycerol in water, giving a viscosity coefficient of 0.00372 $\left[\frac{kg}{m\cdot s}\right]$ [75].

$$D_{100[nm]} = \frac{K_B T}{6\pi\eta r} = \frac{1.3806\cdot 10^{-23}\cdot 293}{6\pi\cdot 0.00372\cdot 100\cdot 10^{-9}} = 5.722\cdot 10^{-13}\left[\frac{m^2}{s}\right] = 0.57\left[\frac{\mu m^2}{s}\right] \tag{22}$$
$$D_{200[nm]} = \frac{K_B T}{6\pi\eta r} = \frac{1.3806\cdot 10^{-23}\cdot 293}{6\pi\cdot 0.00372\cdot 200\cdot 10^{-9}} = 2.884\cdot 10^{-13}\left[\frac{m^2}{s}\right] = 0.28\left[\frac{\mu m^2}{s}\right] \tag{23}$$

Fluorescent beads in F-actin networks

Experiments were performed by the lab of Prof. Yael Roichman, Tel Aviv University[68].

F-actin gels are described as networks of semiflexible polymers. We determine the mesh size from C_A, the concentration of the actin monomer (G-actin), according to $\xi_s = 0.3\sqrt{C_A}$ [76]. G-actin was purified from rabbit skeletal muscle acetone powder[77], with a gel filtration step, stored on ice in G-buffer (5 mM Tris HCl, 0.1 mM CaCl2, 0.2mM ATP, 1 mM DTT, 0.01% NaN3, pH 7.8) and used within two weeks.

The concentration of the G-actin was determined by absorbance, using a UV/Visible spectrophotometer (Ultraspec 2100 pro, Pharmacia) in a cuvette with a 1 cm path length and extinction coefficient of ϵ_{290} = 26, 460 $M^{-1}cm^{-1}$. Polystyrene colloids with radii of α = 0.55 μm (Invitrogen, lot #742530) were pre-incubated with a 10 mg/ml BSA (bovine serum albumin, Sigma) solution to prevent nonspecific binding of protein to the bead surface.

Glass samples were prepared from glass coverslips (diameter, 40mm) coated with methoxy-terminated PEG (Polyethylene glycol, Mw=5000 g/mol, Nanocs) to prevent F-actin filaments from sticking to the chamber walls. After polymerization sample was loaded into a glass cell and left to equilibrate for 30 min at room temperature.

Tracking of fluorescent transmembrane receptors in HEK cells

Experiments were performed by the labs of Prof. Yael Roichman and Prof. Eran Perlson, Tel Aviv University[71].

Preparation of expression plasmids for the study of membrane receptors

TrkB-GFP plasmid, encoding rat full-length TrkB fused to EGFP at the C'-terminus under a CMV promoter was gifted by Rosalind Segal (Harvard University). p75-GFP plasmid, encoding rat p75NTR fused to EGFP was a gift by Francisca C Bronfman (Pontifical Catholic University of Chile). The pLL3.7- CMV-EGFP 3rd generation lentiviral vector (Addgene #11795) was gifted by Uri Ashery (Tel Aviv University). LV-TrkB-GFP and LV-p75-GFP were cloned by inserting TrkB and p75 from TrkB-GFP and p75-GFP into pLL3.7-CMV-EGFP downstream of the CMV promoter.

Total Internal Reflection Fluorescence (TIRF) microscopy

Live cell TIRF imaging was done on a FEI-Munich iMic-42 digital microscope equipped with fast 360° spinning beam scanner to allow even illumination of the entire diameter of the back focal plane of the objective. A 100x Olympus 1.49 numerical aperture TIRF objective was used for objective based TIR. As illumination source, 4 solid-state laser lines at 405, 488, 561 and 640nm were used with maximum output power of 50mW each. Control of stage, excitation and acquisition parameters were via Live Acquisition 2 software. Images were captured using Ixon897 EMCCD camera (Andor). In all live imaging experiments, a 37°C, 5% CO2 and humidity conditions were kept using a custom environmental control system (Live Imaging Services). For SPT experiments, exposure time was 25 milliseconds with 1 millisecond delay. Laser intensities used were 40% and 70% for 561 and 488 respectively. TIRF angle varied for each plate but was between 2.480-2.500. all movies acquired were 1500 frames long.

Chapter 7: Results

Classification network

Simulated data

We trained a neural network to classify each trajectory as one of the three diffusion models described above. The network was trained on ~300,000 simulated trajectories of 100 steps, with a normally distributed localization error and a signal-to-noise ratio (SNR) of four. For FBM and CTRW, which are parameter dependent, each simulated trajectory was generated with a random parameter drawn from a uniform distribution in the range [0,1]. Both models converge to Brownian motion for specific parameter values (0.5 for FBM and 1 for CTRW); to account for this, trajectories generated with a parameter in the range [0.4,0.6] for FBM, and [0.9,1] for CTRW, were labeled as Brownian motion for network training purposes.

During testing, the network receives the derivatives of trajectories in the form of two vectors (i.e., dx and dy data), and outputs the probabilities of each being associated with one of the included diffusion models (**Figure *16***). This configuration can be used to uncover or evaluate motion characteristics incurred in orthogonal directions, such as those incurred for diffusion in a constraining geometry[22].

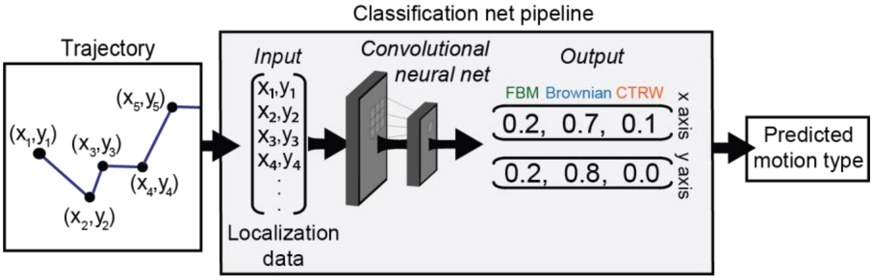

Figure 16. Classification network schematic representation.

The net input is the position coordinates of a single trajectory, and outputs the model-classification.

The performance of the trained network was evaluated using realistic trajectory simulations with various levels of localization precisions.

In terms of accuracy, the classification net performs well on simulated data over a large range of SNRs and diffusion-model parameters (**Figure** *17*), including cases in which the diffusion approaches pure Brownian motion, and are recognized as such by the network.

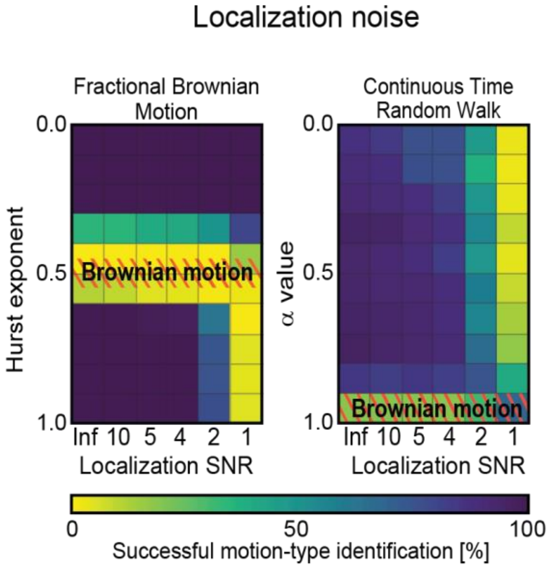

Figure 17. Classification network results.

Heat maps representing the percentage of correct classification as a function of model parameter and SNR. Each pixel corresponds to 200 simulated trajectories. Left – FBM trajectories, right – CTRW trajectories.

In addition to the above tests, we quantified network performance using confusion matrices which show specifically when the network errs. The tables were produced by simulating a set of 300 trajectories, 100 for each considered diffusion model. Parameters for CTRW and FBM were selected at random from the range of values that should not result in Brownian motion ($\alpha \in [0.05,0.9], H \in [0.05,0.45] \cup [0.55,0.95]$), in order to maintain correct statistics in the data set. This dataset was then analyzed by the net under various levels of localization noise.

			Ground truth		
			FBM	Brownian	CTRW
Network prediction	**SNR = ∞**	FBM	84	14	2
		Brownian	0	99	1
		CTRW	5	2	93
	SNR = 10	FBM	82	16	2
		Brownian	0	99	1
		CTRW	6	3	91
	SNR = 5	FBM	80	17	3
		Brownian	0	99	1
		CTRW	9	5	86
	SNR = 2	FBM	77	20	3
		Brownian	26	74	0
		CTRW	28	12	60
	SNR = 1	FBM	67	31	2
		Brownian	99	1	0
		CTRW	69	23	8

Table 6. *Confusion percentages for classification by single-track network*

The confusion matrices show the identification network is accurate even at relatively low SNR levels, beginning to falter at SNR=2. Another important result is the uncertainty between FBM and Brownian motion, even with no addition noise. This is likely due to the fact that FBM is a generalization of Brownian motion, with certain parameter choices causing the network to err between the two. This occurs despite the fact that the Brownian motion parameter range of the Hurst exponent, [0.45-0.55], was not used during generation of the dataset. To illustrate this, we show below the confusion table for SNR=∞, but for a data set wherein H was selected from the parameter range – [0.05,0.35] ∪ [0.65,0.95].

		Ground truth		
		FBM	Brownian	CTRW
Network prediction **SNR = ∞**	FBM	93	7	0
	Brownian	0	100	0
	CTRW	8	1	91

Table 7. *Confusion percentages for classification by the single-track network for a subset of Hurst parameters*

Interestingly, the net finds no ambiguity between CTRW and Brownian motion. This possibly has less to do with parameter choices, but rather with the extracted features learned by the net for classification. During the training phase, each filter learns different features of the signal, CTRW is characterized by long waiting periods between jumps, which results in the diffusing particle being transiently 'stuck'. It seems most likely that the network found this significant, as trajectories displaying some apparent 'sticking' being most frequently classified as CTRW. A further indication can be found in the confusion table for SNR=1, where the high noise masks the signal itself, thus appearing more FBM in character to the net.

Experimental data

On experimentally-obtained trajectories (**Figure** ***18***), the net's classification agrees with the expected diffusion models: FBM for bead in a crowded actin gel (250 nm mesh size), Brownian motion for bead in 40% glycerol solution, and FBM combined with CTRW for transmembrane protein diffusion (TrkB and p75 receptors).

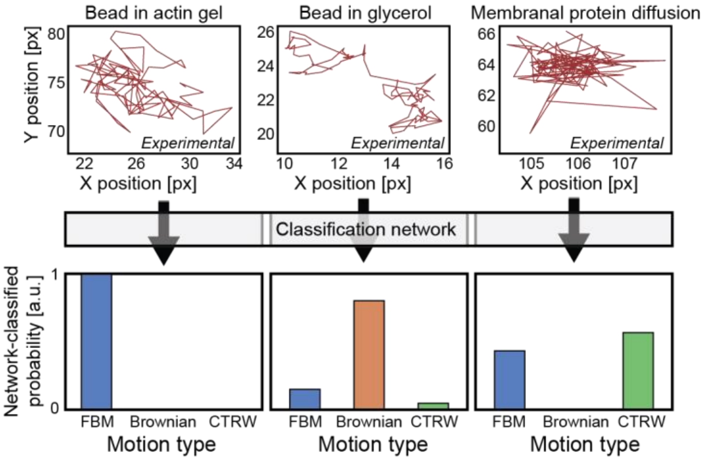

Figure 18. Sample of experimental results.

Network analysis for three experimentally-measured trajectories; left to right: a fluorescent bead diffusing in an actin gel demonstrating FBM; a fluorescent bead diffusing in a glycerol-water mixture, demonstrating Brownian motion; and protein diffusion on a live-cell membrane exhibiting a combination of FBM and CTRW.

Analysis of the individual datasets reveals a more complex image of the diffusion modes. For beads diffusing in glycerol solution (***Figure 19*** **a**), the classification is not perfect, showing nearly similar numbers of FBM and Brownian motion (the minor CTRW population represents beads stuck to the surface unable to move, these do not appear in the H-estimation analysis). The fault most likely lies in a combination of precision errors and other unknown factors relating to the experiment (e.g. effects of fluid dynamics). Analysis using the Hurst exponent regression network shows a population centered around H = 0.6 with standard deviation of 0.07 (***Figure 19*** **b**), in agreement with the classification results (i.e. approximately half classified as FBM, and half as Brownian motion).

The transmembrane protein diffusion experiment presents a unique challenge in that the motion does not fit into any one anomalous diffusion model. For this reason, we cannot simply set the highest probability in the network output as the selected model, but instead must look at probabilities themselves (***Figure 19*** **c**). X, Y axes represent probabilities of being assigned to FBM and CTRW models, respectively. The data closely follows a $y = -x$ trend line, with clusters of tracks being scattered around (x,y) = (0.5,0.5), or (x,y) = (1,0) From this we can conclude: The network identifies features from both models, while almost entirely disregarding the Brownian motion model (otherwise the sum of P_{fbm} and P_{CTRW} would not be one); The network shows a bias towards FBM as was previously shown on simulated data (confusion tables).

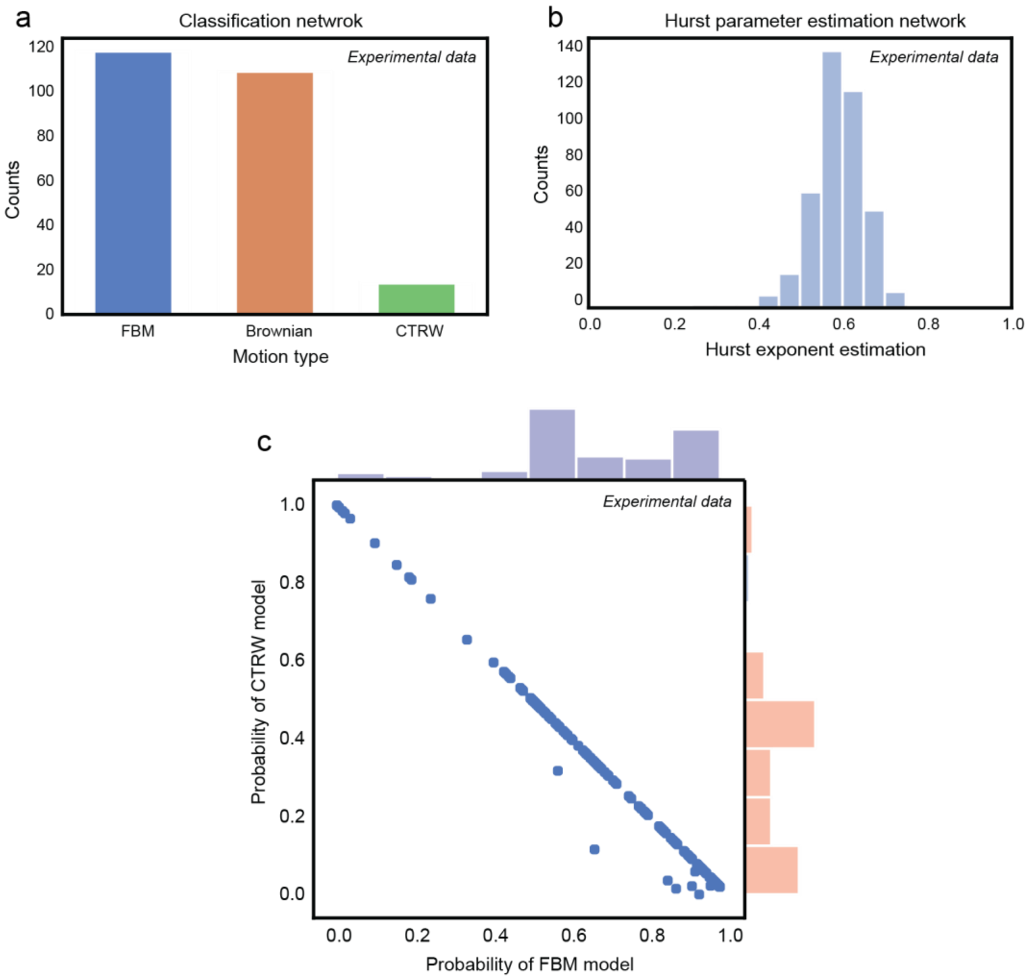

Figure 19. Classification of experimental datasets.

(**a-b**). Classification of bead trajectories diffusing in a glycerol-water mixture. (**a**) Classification. (**b**) Determination of the Hurst parameter for the subset of trajectories classified as FBM. (**c**) Classification probability of protein trajectories diffusing on membrane surface. Results presented are the probabilities of being identified as FBM model or CTRW model, where $P_{Brownian} = 1 - P_{FBM} + P_{CTRW}$.

Single-track Hurst exponent regression network

Simulated data

For estimation of the Hurst exponent, the key parameter in FBM[78], we trained a set of neural networks to estimate a continuous variable. When estimating a continuous variable, we need not alter the training process or network architecture, but rather only the loss function to fit a continuous estimate. Training was performed on ~150,000 simulated tracks with randomly selected Hurst exponents in the range [0.05,0.95]. Two separate net types were trained: first, an array of single-track networks (ST-networks), which receive as input the autocorrelation of the derivative of a single 1D trajectory, known as the velocity autocorrelation. Each network in the set was optimized for different track lengths: from

25 to 1000 steps. The networks were tested on simulated data with various levels of added noise and compared to time averaged MSD (TAMSD) estimation. On simulated data, the ST-networks outperforms TAMSD in estimation accuracy, the number of steps required to achieve the same confidence interval, and robustness to noise (**Figure *20***). The distribution and mean of estimated Hurst exponents for simulated 100 step-tracks with H = 0.2 is shown in Figure 20a, and for a range of parameters in Figure 20b. At finite track lengths, the TAMSD is negatively biased away from the true parameter; however, at long track lengths (>2000 steps), converges to the true value.

Much like MSD analysis, the net's performance improves when more data is available, i.e. additional steps are used for analysis (**Figure *20* c**). Thus, when a 500-step trajectory is available, it is best to use a net designed for that many steps, although sub-trajectories can be given to nets generated for shorter track lengths.

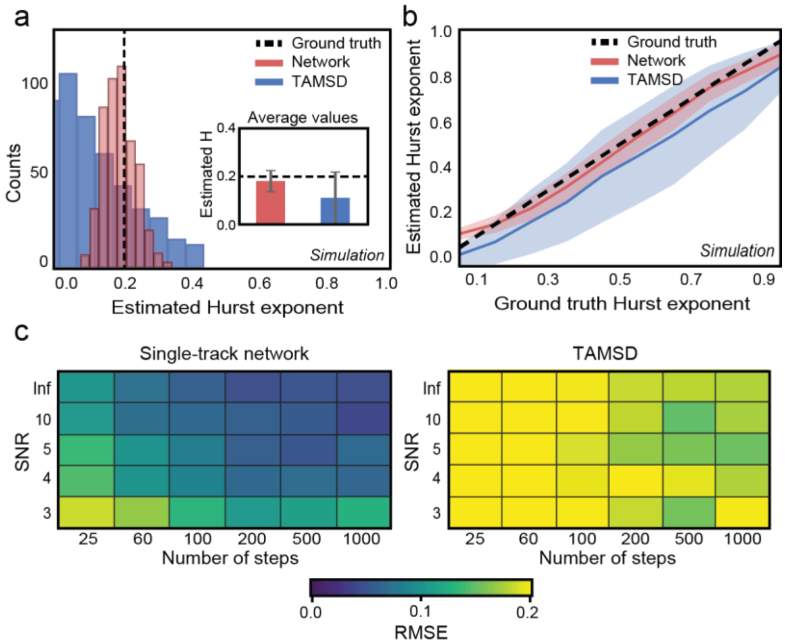

Figure 20. Hurst exponent estimation network – simulated data.

(**a**). Comparison of the 100-step, single-track network to Time Averaged MSD (TAMSD): estimation of the Hurst exponent (0.2) for 200 simulated trajectories. Inset shows the average estimated values and standard deviations. (**b**) Comparison over a range of H values [0.05,0.15,..0.95], using 1000 tracks generated with SNR = 4, and evaluated by the 100-step network and TAMSD, displayed as the mean and standard deviation. (**c**). Root mean squared error (RMSE) heat maps between estimated H and ground truth per simulated trajectory as a function of SNR and track length. Each pixel represents 100 trajectories with random H values in the range [0,1].

Experimental data

Performance of the Hurst exponent-estimation network was tested on experimentally obtained trajectories of fluorescent beads diffusing in entangled F-actin network gels with various mesh sizes[68], allowing control of the Hurst exponent by changing the crowding of the environment (**Figure *21* a**).

In this case, due to the lack of a ground truth, the net's estimation was compared to that of TAMSD and ensemble MSD. Relative to TAMSD (**Figure *21* b**), the network-estimated values were typically slightly higher. For all data points, network estimation is well within the TAMSD standard deviation (STD), with its own STD being less than half that of MSD (~0.06 vs 0.15). Compared to ensemble MSD (**Figure *21* c**), network and MSD estimations converge to relatively equal mean values (within ±0.01), with the network exhibiting lower standard deviation of the mean (0.001 vs 0.005).

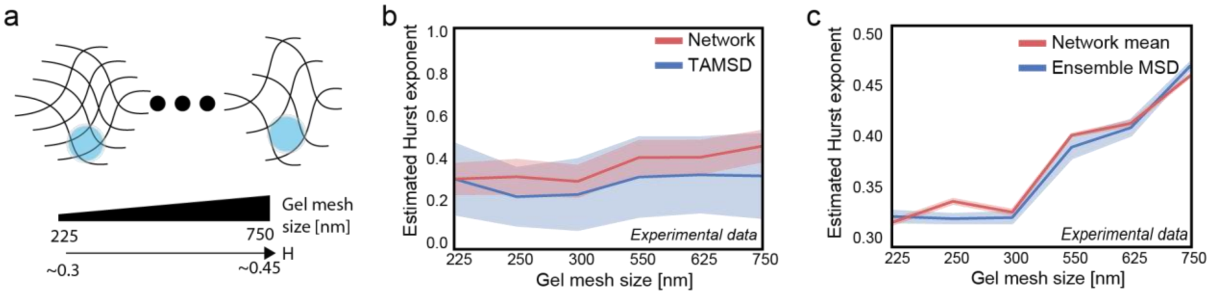

Figure 21. Hurst exponent estimation network – experimental data.

(**a**) Schematic of beads moving through a crowded actin network, where the ratio of bead size to gel-mesh size determines the value of Hurst exponent. (**b**) Estimated Hurst exponents for experimental data using the net and time averaged MSD for truncated 100-step long trajectories displayed as the mean and standard deviation. (**c**) The network and MSD averages for the full-length trajectories shown in (b), ~1000 steps, where the standard deviation is estimated by bootstrapping.

Multi-track Hurst exponent regression network

The second version of the Hurst exponent estimation network aims to tackle a problem of high practical importance of experiments in which only numerous very short trajectories are available, i.e. ~10 step, rather than a single long trajectory. This is often the case when tracking fluorescent proteins that are quick to photobleach[79]. To this end, we trained a set of multi-track networks (MT-networks), that receive an array of 1D-velocity autocorrelations obtained from 10-step trajectories (**Figure *22* a**). The MT-networks were compared to ensemble MSD estimation on simulated data, where the net exhibited better accuracy, standard deviation, and convergence of the mean (**Figure *22* b, c**). At longer track lengths (>100 steps), ensemble MSD analysis surpasses the performance of MT-networks in terms of RMSE.

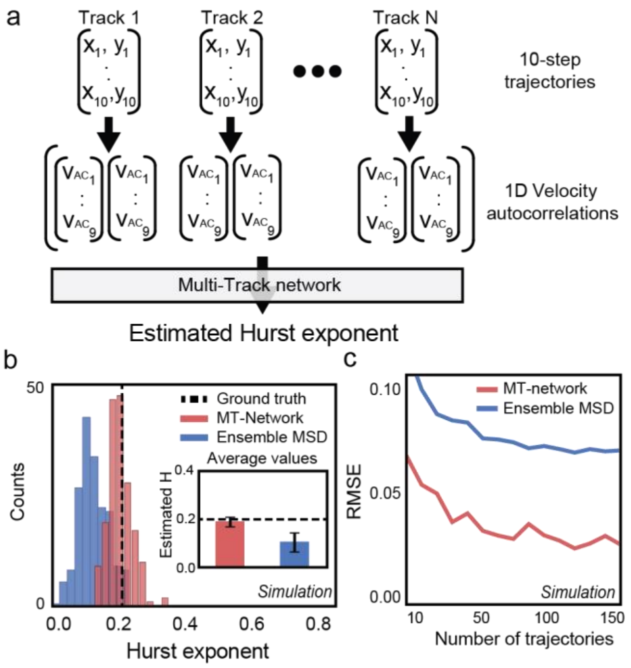

Figure 22. Hurst exponent multi-track network.

(**a**) Schematic of the MT-network (**b**) Estimation of single H value from 50 10-step trajectories, analyzed by ensemble MSD and the MT-network. (**c**) RMSE estimation per trajectory set for varying numbers of 10-step trajectories.

Diffusion coefficient regression network

Simulated data

To estimate the diffusion coefficient from pure Brownian motion trajectories, we trained a single net which receives two inputs: the mean and standard deviation of the absolute value of the single-frame displacements, note that the MSD operates on the squared displacements of the first few time lags.

Training was performed on ~100,000 tracks of 1000 steps, with diffusion coefficients randomly drawn from a uniform distribution in the range [0.1,10] $\frac{\mu m^2}{s}$. For a single value, the network was found to have a lower standard deviation than TAMSD on tracks of only 50 steps (**Figure *23* a**). When considering the range of possible values, the network was discovered to be precise as TAMSD analysis on the low to medium range ([0,5] $\frac{\mu m^2}{s}$), where estimation results converge with those of TAMSD (**Figure *23* b**) and exhibit a similar variance. For this problem, no MT-networks were trained; Multi-track analysis is unnecessary in this case, due to the fact that Brownian motion is a memoryless process, and therefore different tracks from the same population can be concatenated and analyzed in the same manner. Using this concatenation method, the network was compared to ensemble MSD estimation on collections of 10-steps tracks and was found to be slightly more accurate regardless to the number of tracks (**Figure *23* c**).

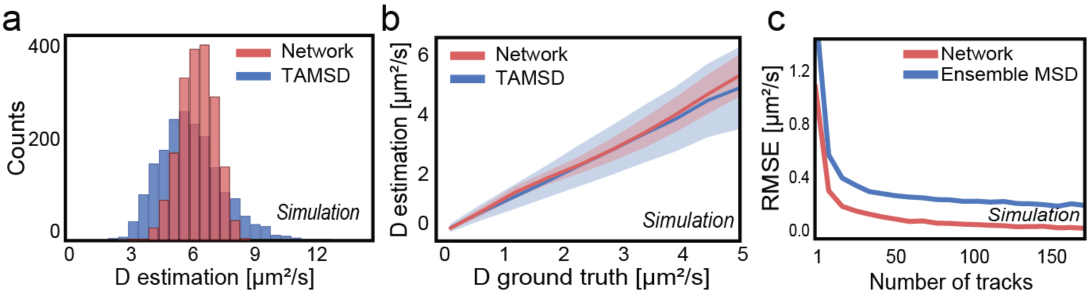

Figure 23. Diffusion coefficient (D) Network – simulated data.

(**a**) Per-trajectory estimation of 2000 simulated 50-step trajectories with D = 6 $\left[\frac{\mu m^2}{s}\right]$. (**b**) Mean and standard deviation for various D values. For each D value, 1000 simulated trajectories with 100 steps were analyzed by the network and a linear fit of the first five time lags of the MSD. (**c**) RMSE for sets of 10-step, D=3 $\left[\frac{\mu m^2}{s}\right]$ trajectories.

Experimental data

Performance was tested on experimental data of fluorescent beads of two sizes (100 nm and 200 nm) diffusing in 40% glycerol solution (see work by Hershko et. al[70] for complete details of preparation). Network estimation shows two different populations, with mean values of 0.29 and 0.54 $\frac{\mu m^2}{s}$, which are similar to predicted theoretical diffusion coefficients calculated from the Stokes-Einstein equation – 0.27 and 0.58 $\frac{\mu m^2}{s}$ (**Figure *24***). Time-averaged MSD estimation of the same data shows relatively close values; however the existence of two populations cannot be distinguished with 100-steps (inset shows the TAMSD estimation using the full 500-step trajectories).

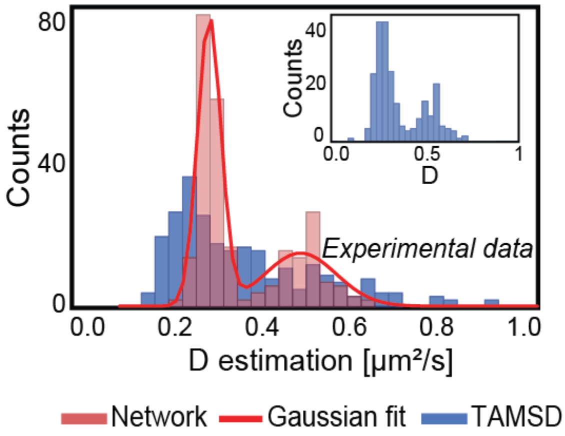

Figure 24. Diffusion coefficient (D) Network – experimental data.

Analysis of two experimentally measured populations of beads diffusing in a glycerol-water mixture were analyzed by the network and TAMSD using 100-step truncated segments of the full trajectories. The inset shows the result of TAMSD using the entire 600-step trajectories.

Chapter 8: Section Summary

Deep learning is revolutionizing signal analysis owing to its ability to identify complex models in large quantities of data, relative simplicity of implementation, (once trained) inference speed, and robustness. This revolution is providing tools enabling the extrapolation of biological conclusions from seemingly unintelligible measurements. In this work, we take a first step in strengthening single-particle-diffusion analysis using a set of neural networks for model selection and parameter estimations. We have shown these to be more precise than current standard methods on both simulated and experimental data, while requiring a smaller number of steps, with increased robustness to noise and the advantage of being parameter free.

The framework we have developed here enables concatenation of different neural nets, providing end-to-end localization to classification and parameter estimation. The networks presented are computationally inexpensive and can be trained in the span of minutes to several hours to process different experimental conditions on a standard GPU-accelerated personal computer. The duration of the training process is generally determined by the implementation of the diffusion simulations and by the complexity of the diffusion processes themselves. Generally, the amount of training data required increases with model complexity. For example, training on CTRW data would require a significant number of trajectories to capture rare events, i.e. the long tail of the temporal dwell time distribution. Regardless of training complexity, the trained network can analyze hundreds of trajectories in a manner of seconds.

Although easy to use train and use, the networks presented are limited by the data they are trained on. The simulations used to build the training set require as input the localization precision levels (SNR), total time of the experiment, number of steps, and for the D-network, the pixel size of the experimental system. A grossly incorrect selection of one of these can result in erroneous estimation by the networks.

In addition, the parameter estimation networks are constrained to the theoretical model they were trained on. For example, attempting to estimate the Hurst exponent for a more complex case of diffusion (e.g. and FBM process subordinated to a second FBM process), will result in incorrect estimations.

Discussion

In section I we presented an experimental synthetic model system used to as a means to study the dynamics of nuclear bodies. Fluorescence tracking of this system revealed both positive and negative sharp changes in intensity which we termed bursts. The positive most likely corresponding to transcription of new slncRNA molecules and their accumulation in the synthetic speckle, and the negative to shedding of molecules back to the cytoplasm to be degraded naturally. We have shown these results repeat themselves when using the standard 24x cassette, and our new much shorter design of Qβ-5x-PP7-4x.

We believe that continued exploration of additional designs of such synthetic long non-coding RNA molecules has the potential to provide important biophysical insight into both the assembly and characteristics of natural membrane-free intracellular compartments in all cell-types. Given the increased importance that these compartments are now thought to have in many biological processes, constructing and studying such objects synthetically has the potential to provide important biophysical insight for this new class of intracellular compartments.

In section II we presented a deep-learning based system for the analysis of single particle trajectories originating from diffusing molecules. The system classifies the input trajectories to the most probable theoretical anomalous diffusion model, separating between Brownian motion, CTRW and FBM, and estimates the anomalous exponent for the case of FBM, or the diffusion coefficient for the case of Brownian motion. We have shown this system to be more accurate than the mean squared displacement method which is the standard used today for estimation anomalous diffusion parameters.

Future work in this area will extend the set of networks to incorporate other motion models, e.g. CTRW (estimation of model parameters with an in-depth analysis of the different implementations and the effects of non-ergodicity on inference capabilities), motion on a fractal, Lévy flights, and more complex cases of subordinate processes. Additionally, other values of interest can be estimated from short trajectories such as fluctuations in TAMSD amplitudes, and the shape of the displacement probability density function[80,81].

Another significant problem to be addressed in future work is how to best identify transient behavior in a trajectory, i.e. switching between diffusion models during motion. For this problem, neural networks have shown promising capabilities in their ability to infer data from short trajectories, an ability which will be vital for handling transient trajectories. Such an intricate problem might require several neural networks working sequentially to identify diffusion types along a trajectory, and extract parameters from each part separately in an approach analogous to those already used for object classification within images[82,83].

References

1. Clemson, C. M. *et al.* An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol. Cell* **33**, 717–726 (2009).
2. Staněk, D. & Fox, A. H. Nuclear bodies: news insights into structure and function. *Curr. Opin. Cell Biol.* **46**, 94–101 (2017).
3. Fox, A. H., Nakagawa, S., Hirose, T. & Bond, C. S. Paraspeckles: Where Long Noncoding RNA Meets Phase Separation. *Trends in Biochemical Sciences* **43**, 124–135 (2018).
4. Spector, D. L. & Lamond, A. I. Nuclear speckles. *Cold Spring Harb Perspect Biol* **3**, (2011).
5. Bond, C. S. & Fox, A. H. Paraspeckles: nuclear bodies built on long noncoding RNA. *J Cell Biol* **186**, 637–644 (2009).
6. Gomes, E. & Shorter, J. The molecular language of membraneless organelles. *J. Biol. Chem.* **294**, 7115–7127 (2019).
7. Bertrand, E. *et al.* Localization of ASH1 mRNA particles in living yeast. *Molecular cell* **2**, 437–445 (1998).
8. Tutucci, E. *et al.* An improved MS2 system for accurate reporting of the mRNA life cycle. *Nature Methods* **15**, 81–89 (2018).
9. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell* **123**, 1025–1036 (2005).
10. Katz, N. *et al.* An in Vivo Binding Assay for RNA-Binding Proteins Based on Repression of a Reporter Gene. *ACS Synth. Biol.* **7**, 2765–2774 (2018).
11. Jones, D. & Elf, J. Bursting onto the scene? Exploring stochastic mRNA production in bacteria. *Current Opinion in Microbiology* **45**, 124–130 (2018).
12. Paulsson, J. Models of stochastic gene expression. *Physics of Life Reviews* **2**, 157–175 (2005).
13. Manzo, C. & Garcia-Parajo, M. F. A review of progress in single particle tracking: from methods to biophysical insights. *Reports on Progress in Physics* **78**, 124601 (2015).

14. Golding, I. & Cox, E. C. Physical nature of bacterial cytoplasm. *Phys. Rev. Lett.* **96**, 098102 (2006).

15. Metzler, R., Jeon, J.-H., Cherstvy, A. G. & Barkai, E. Anomalous diffusion models and their properties: non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. *Phys. Chem. Chem. Phys.* **16**, 24128–24164 (2014).

16. Shen, H. *et al.* Single Particle Tracking: From Theory to Biophysical Applications. *Chemical Reviews* **117**, 7331–7376 (2017).

17. Sokolov, I. M. Models of anomalous diffusion in crowded environments. *Soft Matter* **8**, 9043 (2012).

18. Metzler, R., Jeon, J.-H. & Cherstvy, A. G. Non-Brownian diffusion in lipid membranes: Experiments and simulations. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1858**, 2451–2467 (2016).

19. Matsuda, Y., Hanasaki, I., Iwao, R., Yamaguchi, H. & Niimi, T. Estimation of diffusive states from single-particle trajectory in heterogeneous medium using machine-learning methods. *Physical Chemistry Chemical Physics* **20**, 24099–24108 (2018).

20. Hansen, A. S. *et al.* Robust model-based analysis of single-particle tracking experiments with Spot-On. *eLife* **7**, e33125 (2018).

21. Persson, F., Lindén, M., Unoson, C. & Elf, J. Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nature Methods* **10**, 265–269 (2013).

22. Weiss, L. E., Milenkovic, L., Yoon, J., Stearns, T. & Moerner, W. E. Motional dynamics of single Patched1 molecules in cilia are controlled by Hedgehog and cholesterol. *PNAS* **116**, 5550–5557 (2019).

23. Calderon, C. P., Weiss, L. E. & Moerner, W. E. Robust hypothesis tests for detecting statistical evidence of two-dimensional and three-dimensional interactions in single-molecule measurements. *Phys Rev E Stat Nonlin Soft Matter Phys* **89**, 052705 (2014).

24. Briane, V., Vimond, M. & Kervrann, C. An adaptive statistical test to detect non Brownian diffusion from particle trajectories. in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* 972–975 (2016). doi:10.1109/ISBI.2016.7493427.

25. Dosset, P. *et al.* Automatic detection of diffusion modes within biological membranes using back-propagation neural network. *BMC Bioinformatics* **17**, (2016).

26. Kowalek, P., Loch-Olszewska, H. & Szwabiński, J. Classification of diffusion modes in single particle tracking data: feature based vs. deep learning approach. *arXiv:1902.07942 [q-bio]* (2019).

27. Blake, W. J. *et al.* Phenotypic Consequences of Promoter-Mediated Transcriptional Noise. *Molecular Cell* **24**, 853–865 (2006).

28. Elowitz, M. B. Stochastic Gene Expression in a Single Cell. *Science* **297**, 1183–1186 (2002).

29. McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences* **94**, 814–819 (1997).

30. Chong, S., Chen, C., Ge, H. & Xie, X. S. Mechanism of Transcriptional Bursting in Bacteria. *Cell* **158**, 314–326 (2014).

31. Tutucci, E., Livingston, N. M., Singer, R. H. & Wu, B. Imaging mRNA In Vivo, from Birth to Death. *Annual Review of Biophysics* **47**, 85–106 (2018).

32. Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of Single RNA Transcripts in Situ. *Science* **280**, 585–590 (1998).

33. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods* **5**, 877–879 (2008).

34. Fusco, D. *et al.* Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. *Current Biology* **13**, 161–167 (2003).

35. Wells, A. L., Condeelis, J. S., Singer, R. H. & Zenklusen, D. Imaging Real-Time Gene Expression in Living Systems with Single-Transcript Resolution: Construct Design and Imaging System Setup. *Cold Spring Harbor Protocols* **2007**, pdb.top28 (2007).

36. So, L. *et al.* General properties of transcriptional time series in Escherichia coli. *Nature Genetics* **43**, 554–560 (2011).

37. Taniguchi, Y. *et al.* Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science* **329**, 533–538 (2010).

38. Li, X. & Fu, X.-D. Chromatin-associated RNAs as facilitators of functional genomic interactions. *Nature Reviews Genetics* 1 (2019) doi:10.1038/s41576-019-0135-1.

39. Sbalzarini, I. F. & Koumoutsakos, P. Feature point tracking and trajectory analysis for video imaging in cell biology. *J. Struct. Biol.* **151**, 182–195 (2005).

40. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**, 671–675 (2012).

41. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9**, 676–682 (2012).

42. Fei, J. & Sharma, C. M. RNA localization in bacteria. *Microbiol Spectr* **6**, (2018).

43. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**, 62–66 (1979).

44. Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A. & Singer, R. H. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science* **332**, 475–478 (2011).

45. Sanchez, A., Garcia, H. G., Jones, D., Phillips, R. & Kondev, J. Effect of Promoter Architecture on the Cell-to-Cell Variability in Gene Expression. *PLoS Computational Biology* **7**, e1001100 (2011).

46. Gong, T. *et al.* Classification of Three-Level Random Telegraph Noise and Its Application in Accurate Extraction of Trap Profiles in Oxide-Based Resistive Switching Memory. *IEEE Electron Device Letters* **39**, 1302–1305 (2018).

47. Battaile, C. C. The Kinetic Monte Carlo method: Foundation, implementation, and application. *Computer Methods in Applied Mechanics and Engineering* **197**, 3386–3398 (2008).

48. Larson, D. R., Singer, R. H. & Zenklusen, D. A Single Molecule View of Gene Expression. *Trends Cell Biol* **19**, 630–637 (2009).

49. Lionnet, T. & Singer, R. H. Transcription goes digital. *EMBO Rep* **13**, 313–321 (2012).

50. Miller, S. & Childers, D. *Probability and Random Processes With Applications to Signal Processing and Communications*. (Elsevier Science & Technology Books, 2012).

51. Hyman, A. A., Weber, C. A. & Jülicher, F. Liquid-Liquid Phase Separation in Biology. *Annual Review of Cell and Developmental Biology* **30**, 39–58 (2014).

52. Bressloff, P. C. *Stochastic Processes in Cell Biology*. (Springer International Publishing, 2014).

53. Kepten, E., Weron, A., Sikora, G., Burnecki, K. & Garini, Y. Guidelines for the fitting of anomalous diffusion mean square displacement graphs from single particle tracking experiments. *PLoS One* **10**, e0117722 (2015).

54. Burnecki, K., Kepten, E., Garini, Y., Sikora, G. & Weron, A. Estimating the anomalous diffusion exponent for single particle tracking data with measurement errors - An alternative approach. *Sci Rep* **5**, 11306 (2015).

55. Tejedor, V. *et al.* Quantitative Analysis of Single Particle Trajectories: Mean Maximal Excursion Method. *Biophysical Journal* **98**, 1364–1372 (2010).

56. Krapf, D. *et al.* Spectral Content of a Single Non-Brownian Trajectory. *Phys. Rev. X* **9**, 011019 (2019).

57. Meroz, Y. & Sokolov, I. M. A toolbox for determining subdiffusive mechanisms. *Physics Reports* **573**, 1–29 (2015).

58. Magdziarz, M., Weron, A., Burnecki, K. & Klafter, J. Fractional Brownian Motion Versus the Continuous-Time Random Walk: A Simple Test for Subdiffusive Dynamics. *Phys. Rev. Lett.* **103**, 180602 (2009).

59. Weigel, A. V., Simon, B., Tamkun, M. M. & Krapf, D. Ergodic and nonergodic processes coexist in the plasma membrane as observed by single-molecule tracking. *PNAS* **108**, 6438–6443 (2011).

60. Tabei, S. M. A. *et al.* Intracellular transport of insulin granules is a subordinated random walk. *PNAS* **110**, 4911–4916 (2013).

61. Michalet, X. & Berglund, A. J. Optimal diffusion coefficient estimation in single-particle tracking. *Phys. Rev. E* **85**, 061916 (2012).

62. Elf, J. & Barkefors, I. Single-Molecule Kinetics in Living Cells. *Annual Review of Biochemistry* **88**, 013118–110801 (2019).

63. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L. & Muller, P.-A. Deep learning for time series classification: a review. *Data Min Knowl Disc* (2019) doi:10.1007/s10618-019-00619-1.

64. Webb, S. Deep learning for biology. *Nature* **554**, 555 (2018).

65. *Stochastic Geometry, Spatial Statistics and Random Fields: Models and Algorithms*. (Springer International Publishing, 2015).

66. Liang, Y. & Chen, W. Continuous time random walk model with asymptotical probability density of waiting times via inverse Mittag-Leffler function. *Communications in Nonlinear Science and Numerical Simulation* **57**, 439–448 (2018).

67. Fulger, D., Scalas, E. & Germano, G. Monte Carlo simulation of uncoupled continuous-time random walks yielding a stochastic solution of the space-time fractional diffusion equation. *Physical Review E* **77**, 021122 (2008).

68. Sonn-Segev, A., Bernheim-Groswasser, A. & Roichman, Y. Extracting the dynamic correlation length of actin networks from microrheology experiments. *Soft Matter* **10**, 8324–8329 (2014).

69. Wong, I. Y. *et al.* Anomalous Diffusion Probes Microstructure Dynamics of Entangled F-Actin Networks. *Phys. Rev. Lett.* **92**, 178101 (2004).

70. Hershko, E., Weiss, L. E., Michaeli, T. & Shechtman, Y. Multicolor localization microscopy and point-spread-function engineering by deep learning. *Opt. Express, OE* **27**, 6158–6183 (2019).

71. Chein, M., Perlson, E. & Roichman, Y. Flow arrest in the plasma membrane. *bioRxiv* 575456 (2019) doi:10.1101/575456.

72. Software | Nano-bio-optics lab – Yoav Shechtman. https://nanobiooptics.net.technion.ac.il/software/.

73. Bai, S., Kolter, J. Z. & Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv:1803.01271 [cs]* (2018).

74. Miller, C. C. The Stokes-Einstein Law for Diffusion in Solution. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **106**, 724–749 (1924).

75. Segur, J. B. & Oberstar, H. E. Viscosity of Glycerol and Its Aqueous Solutions. *Industrial & Engineering Chemistry* **43**, 2117–2120 (1951).

76. Schmidt, C. F., Baermann, M., Isenberg, G. & Sackmann, E. Chain dynamics, mesh size, and diffusive transport in networks of polymerized actin: a quasielastic light scattering and microfluorescence study. *Macromolecules* **22**, 3638–3649 (1989).

77. Spudich, J. A. & Watt, S. The regulation of rabbit skeletal muscle contraction. I. Biochemical studies of the interaction of the tropomyosin-troponin complex with actin and the proteolytic fragments of myosin. *J. Biol. Chem.* **246**, 4866–4871 (1971).

78. Mandelbrot, B. & Van Ness, J. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Rev.* **10**, 422–437 (1968).

79. P. Clausen, M. & Christoffer Lagerholm, B. The Probe Rules in Single Particle Tracking. *Current Protein & Peptide Science* **12**, 699–713 (2011).

80. Cherstvy, A. G. & Metzler, R. Anomalous diffusion in time-fluctuating non-stationary diffusivity landscapes. *Phys. Chem. Chem. Phys.* **18**, 23840–23852 (2016).

81. Wang, B., Anthony, S. M., Bae, S. C. & Granick, S. Anomalous yet Brownian. *Proceedings of the National Academy of Sciences* **106**, 15160–15164 (2009).

82. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv:1506.02640 [cs]* (2015).

83. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv:1311.2524 [cs]* (2013).

תקציר

הציטופלסמה התאית היא הסביבה בה כל הריאקציות התאיות מתרחשות. תכונותיה הפיזיקליות והכימיות בעלות השפעה רבה על מגוון פעולות תאיות דוגמת תקשורת בין תאית, מנגנוני תנועה, קיפול חלבונים ועוד. בעבודה זו אנו מנסים לשפוך אור על שני תהליכים דינמיים תוך-תאיים שזכו לתשומת לב מוגברת בשנים האחרונות, הודות להתקדמות טכנולוגית רבה בתחום המיקרוסקופיה: יצירתם של גופים גרעיניים (Paraspeckles) בתוך תאים, ותופעת הדיפוזיה האנומלית (Anomalous Diffusion).

ראשית, Paraspeckles הם גופים תוך-תאיים עתירי חלבונים הבנויים סביב מולקולת lncRNA המשמשת כבסיס. כאמצעי לחקירת הדינמיקה של מבנה כזה, החלטנו לחקור בחיידקים גופים סינתטיים בעלי הרכב דומה על ידי הכנסת שני סוגים של lncRNA סינתטיים היוצרים את הבסיס המתאים. ה- lncRNA מקודדים ליצירת אתרי קישור עבור חלבוני מעטפת של בקטריופאג' (RNA-binding phage-coat-protein (RBP)), תחת בקרה של פרומוטר חיידקי מסוג T7. בנוסף התא החיידקי מבטא את חלבוני המעטפת עצמם מחוברים לחלבון פלורסנטי. עבור שני סוגי ה- lncRNA שנבדקו בתור בסיס, כאשר מתבוננים על החיידקים תחת המיקרוסקופ, ניתן לראות בבירור נקודות פלורסנטיות המכילות עשרות מולקולות lncRNA הקשורות לחלבוני המעטפת התואמים. במעקב אחרי עוצמת הפלורסנציה לאורך זמן מתגלים מקטעים בהם עוצמת הפלורסנציה עולה או יורדת בצורה משמעותית, המרווחים על ידי מקטעי זמן המפולגים בצורה אקספוננציאלית בהם אין שינוי בולט בעוצמה, הנמשכים כ-10 דקות בממוצע. אנחנו משייכים את המקטעים החיוביים לפרצים של שעתוק (Transcriptional bursts), ומכנים את המקטעים השליליים, פרצים של ירידה בעוצמה. המידע הוביל אותנו למסקנה כי הפרצים השליליים מסמנים השרה של מספר מולקולות lncRNA חזרה לציטופלסמה, וכי הגופים הגרעיניים מהווים הגנה על מולקולות הlncRNA מפני פירוק.

שנית, לדיפוזיה יש תפקיד משמעותי בתהליכים ביולוגיים רבים. תצפיות ישירות של תנועות מולקולריות על ידי ניסויי עקיבה ברזולוציה של מולקולה בודדת (single-particle-tracking) הניבו ראיות לכך שמערכות תוך-תאיות רבות אינן נעות בתנועת דיפוזיה רגילה (תנועה בראונית), אלא בצורה של דיפוזיה אנומלית (anomalous diffusion). אפיון של התהליך הפיזיקלי הגורם לדיפוזיה אנומלית נותר אתגר קשה בתחום, זאת עקב העובדה שהכלים האנליטיים שבהם משתמשים לרוב כדי לאפיין תהליכים אלו תלויים בהתנהגות אסימפטוטית שאינה נגישה לנו תחת תנאי ניסוי. הכלי המרכזי המשמש לאפיון תהליכים אלו כיום הוא שיטת ממוצע ההפרשים המרובעים (mean squared displacement). שיטה זו מאפשרת למצוא את הפרמטר האנומלי α שמכיל מידע רב על סוג הדיפוזיה האנומלית, אבל שיטה זו בעייתית מכיוון שמודלים דיפוזיביים שונים בתכלית יכולים לספק את אותו פרמטר אנומלי, דבר המקשה של זיהוי נכון של סוג התנועה. בדרך כלל אפיון מדויק של המודל הדיפוזיבי דורש חישוב של מספר רב של פרמטרים אחרים הדורשים שימוש בשיטות אחרות.

אנחנו בחרנו להשתמש בלמידה עמוקה על מנת למצוא את המודל הפיזיקלי המתאים הגורם לדיפוזיה אנומלית. לצורך זאת יישמנו רשת נוירונים אשר מסוגלת לסווג מסלול תנועה של חלקיק לפי מודל הדיפוזיה האנומלית המתאים, כאשר היא מבדילה בין תנועה בראונית, תנועה בראונית פרקטלית (Fractional Brownian motion), והילוך רנדומלי רציף (Continuous time random walk). אנו מדגימים את יישומיות הרשת שלנו גם עבור שיערוך של פרמטר הורסט (Hurst exponent) עבור תנועה בראונית פרקטלית, ועבור שיערוך של קבוע הדיפוזיה (Diffusion coefficient) עבור תנועה בראונית, על מסלולים שנוצרו בסימולציה, ועל מסלולים שנמדדו בצורה ניסיונית. אנו מראים שעל מסלולים שנוצרו בסימולציה רשתות אלו מדויקות יותר מאשר שיטת ממוצע ההפרשים המרובעים ודורשות מסלולים קצרים יותר על מנת לעבוד. בנוסף, על מידע שנאסף באופן ניסיוני, אנו מראים שהרשתות מספקות את אותן תוצאות כמו שיטת ממוצע ההפרשים המרובעים.

פרסומים

Granik, Naor, Noa Katz, Yoav Shechtman, and Roee Amit. "Live dynamical tracking of slncRNA speckles in single E. coli cells reveals bursts of fluorescence degradation." Submitted to *eLife.* (Section I of this thesis)

Granik, Naor, Lucien E. Weiss, Elias Nehme, Maayan Levin, Michael Chein, Eran Perlson, Yael Roichman, and Yoav Shechtman. "Single particle diffusion characterization by deep learning." *Biophysical Journal* (2019). (Section II of this thesis)

המחקר נעשה בהנחייתו של פרופסור יואב שכטמן בפקולטה להנדסה ביו-רפואית ופרופסור רועי עמית בפקולטה להנדסת ביוטכנולוגיה ומזון.

אני מודה לטכניון – מכון טכנולוגי לישראל על התמיכה הכספית הנדיבה בהשתלמותי.

אפיון תהליכי שעתוק ודיפוזיה תוך תאיים

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר

מגיסטר למדעים בהנדסה ביורפואית

נאור גרניק

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

תשרי תש"פ   חיפה   אוקטובר 2019