Lund University

Department of Astronomy and Theoretical physics

Master Thesis - FYTM04, 60 credits

DNA damage – a novel method to measure the rate of single-stranded breaks from fragmenting dsDNA in nanochannels – theory and modeling

---

Author:
Magnus Brander

Supervisor:
Tobias Ambjörnsson
Co-Supervisor:
Jason Beech

LUND
UNIVERSITY

October 16, 2020

Abstract

Damage to DNA can cause death to an individual cell and serious harm to the host organism. Photosensitized reactions are one cause of DNA damage. It can lead to destructive chemical reactions targeted to a base of the DNA as well as a breakage along one or both of the DNA strands. Due to this, quantifying and understanding photosensitized driven DNA damage is an important topic of research.

From experimental data of fragmenting fluorescently stained linear double-stranded DNA in nanochannels, we will extract the non-observable single-stranded cleavage rate (nicking rate) from observable times of double-stranded cleaves (cuts), which lets us quantify the rate of DNA damage. To do this, we present a new probabilistic model that connects the cutting rate to the time of cuts. We present two distinct models for the cutting rate, the first one is analytical and the second is based on simulations.

We find through validation on synthetic data with known nicking rates that using the cutting rate from the simulation-based model yields more accurate estimates of the nicking rate compared to using the cutting rate from the analytical model. In addition, we manage to estimate the nicking rate for three experimental data sets with varying illumination strength. From these estimates, we conclude that the nicking rate, as expected, increases with increasing illumination strength.

We hope that this study will serve as proof of concept for our new methodology to estimate the nicking rate and provide a good starting point for other studies which want to add to the knowledge of nicking rate estimation under different conditions.

Popular science – measuring the invisible

Claiming around 8 million lives yearly, cancer is one of the deadliest diseases world-wide. Due to its negative impact on human well-being, improving our knowledge about cancer and working towards more effective treatments in the future is essential.

Accumulation of damage to our genes can lead to cell death or severe harm to the host organism. To better understand the processes that result in gene damage, it is important to study deoxyribonucleic acid (DNA) damage. DNA damage can be studied in various ways, e.g., by exposing the DNA to, cutting enzymes, high energy radiation, oxygen or visible light. In this work, we study DNA damage caused by the exposure of oxygen and visible light.

We know that harmful reactants, such as free radicals, take part in chemical reactions that can cause damage to our DNA. We also know that light and oxygen can increase the creation of free radicals. The rate of DNA damage is crucial since our bodies need to keep up with the reparation processes and avoid accumulation of mutations. It is thus important to ask how oxygen and light change the rate of DNA damage, which are both present in our daily lives.

In this work, we will look at DNA damage through experiments of DNA in nanometer-sized channels captured using a fluorescence microscope. Even though the scale of the experimental setup is very small, it is not small enough to obtain the DNA damage directly. To better understand what limits our observations of DNA damage, we begin by picturing the DNA as a spiraling ladder structure. The ladder's two side-rails and rungs, correspond to the two main strands and base-pairs of the DNA, respectively. Free radicals may chemically react with one of the strands and break it, or analogously damage one of the rails on the ladder. When such a single-stranded break (nick) happens the DNA is still held together as one molecule due to the connection of base-pairs to the other strand, just as the ladder is still in one piece held together by the rungs. This process of single-stranded breaks is not observable with the microscope. On the other hand, if yet another single-stranded break occurs close enough on the opposite strand on the DNA we obtain a double-stranded break (cut), it has the effect of damaging both side-rails between the same pairs of rungs on the ladder. In this case, the DNA molecule divides into two molecules, a constellation which can be observed with the microscope, once the DNA fragments has diffused apart within the nanochannel.

Since we want to know at which rate the nicks occur but can only observe the cuts, we are challenged to measure something non-observable. To tackle this, we here introduce a new stochastic model, which can estimate the nicking rate given a series of observed cutting times using a functional form for the rate of cuts. As the microscope has a limited resolution, we cannot observe the exact time of the cuts in the experiments and the functional form for the cutting rate must account for this fact. To do this, we have simulated all parts of the experiments to obtain the observable functional form for the cutting rate. By performing simulations for different nicking rates, we obtain different functional forms for the observed cutting rate. Each of these observed cutting rate functions, together with the observed cutting times, is then used in the stochastic model to deduce the likelihood of the current nicking rate. The observed cutting rate function that gives the highest likelihood corresponds to our estimate of the nicking rate.

With this study, we hope that our new method to measure the rate of DNA damage can find its way into the hands of more researchers who can keep expanding the knowledge of DNA damage and use our method to measure the nicking rate under different experimental conditions.

Acknowledgements

This work was made possible by many individuals and I would like to express my gratitude to all off them.

To begin with, I would like to thank my supervisor Tobias Ambjörnsson, who did not only make it possible for me to write this thesis, but also offered guidance, support and inspiration throughout the process. I am also grateful that he helped me through seemingly impossible problems, enabled my eagerness to be put to work and enlightened me about the importance of patience.

I also want to thank my co-supervisor Jason Beech from the Tegenfeldt group which has provided me with all the needed assistance when put in front of new and challenging questions about the experiments. In addition, he provided all the experimental data which made the work of this thesis possible. For that I am very thankful.

At this point I want to thank Michael Lomholt who so kindly welcomed me to Odense for discussions, help and guidance with the theoretical work. He is also the inventor of the theory that constitutes the backbone of my thesis and deserves much recognition for that.

Now it is time to thank Jens Krog for offering great assistance with the statistics parts of my thesis work in a very pedagogical way. His help has not only made contributions to my work but also taught me a great deal. He should also be thanked for providing me with his image segmentation software, without which I would not have managed.

I will not forget to thank Albertas Dvirnas for many interesting discussions throughout the year I spent at the department. His will to help, both with physics and food, was most delightful.

Lastly, I would like to thank André Nüßlein for proofreading the thesis and providing helpful critique to improve the language.

Table of contents

1 Introduction

Deoxyribonucleic acid (DNA) is of utmost importance to all living organisms. Within its double helical structure it contains the key for both current and future generations' survival. With its clever system of repeated blocks of simple information, in form of base-pairs (bp) spaced by merely 3.4 Ångstrom, it encodes the most complex proteins. The molecule we today call DNA was discovered in 1868 by F. Miescher [1, 2] although we are perhaps more familiar with the structural model proposed by Watson and Crick [3]. The model proposed by Watson and Crick remains valid for many of the key features we assign DNA with current knowledge. For example, each base is bound to the complementary base by hydrogen bond and only pair up as AT or GC. In addition, the chemical structure is such that molecules, for example proteins, can bind to the DNA molecule for various purposes, including replication.

As the central storage unit for life, damage to the DNA can be a serious threat to survival. We call a change in the structure of the DNA through chemical addition or disruption to a base, as well as cleavage of one or two strands, *DNA damage* [4]. When a base is damaged, a wrong complementary base may be inserted during replication. In addition, cleavage of both strands could result in a repair (through DNA repair enzymes) that is different from the original genome. Both of these two cases could lead to mutations in the next stage of replication if the organism does not detect the fallacious DNA. Although mutations play an important role in the biological evolution, a harmful mutation passed on to successive cells can lead to cancer. To stop this, there are multiple processes which deal with *DNA repair*, unfortunately these processes are not 100 % effective. Accordingly, the interplay between *DNA damage* and *DNA repair* is of highest importance for all living organisms, with evidence that direct repair after DNA damage is the most effective and easiest way of repair [5].

In this study, we will focus on photosensitized driven DNA damage [6]. Nanochannels are used to stretch fluorescently stained linear double-stranded DNA (dsDNA) which are subsequently imaged using fluorescence microscopy. As an already established tool used to study the mechanical properties [7] and large-scale sequence information [8] of DNA, fluorescently stained dsDNA in nanochannels will in this study instead be used to analyze single-stranded breaks (nicks) resulting in double-stranded breaks (cuts). In detail, this involves measuring the time of cuts in the experimental data as a function of illumination strength (the non-visible nicking rate is extracted through modeling). Single-stranded breaks, in whatever way they are created, do not result in a double-stranded break unless two nicks are located close enough on the opposite strands within half the *fraying distance*. Here, fraying is the notation of thermally induced DNA unzipping [9], which depends on temperature, DNA sequence and buffer composition. The relation between nicks and cuts leads to a non-trivial relation between the rate at which nicks happen and the rate at which the observable cuts happen. To quantify this relation, both in theory and in real experiments, we will in this study develop a mathematical model as well as a simulation model.

The relation between the dose of nicks and resulting cuts has earlier been subject to multiple studies. For example, formulas predicting the fraction of super-coiled, circular, linear and fragmented DNA in bulk are presented in [10]. Another example, the yield of nicks and cuts and their corresponding dependence as function of radiation dose, measured through ratios of fragmented DNA in bulk, are present in [11]. In addition, multiple studies have performed measurements of the fraying distance [12, 13, 14], with significantly varying results.

There are some main characteristics of the earlier studies we will try to improve upon in this work. Each of the earlier studies used the proportion of molecules in different stages (super-coiled, circular, linear and fragmented) or subsets of them as an observable. We intend to use another observable which can be applied to a few molecules or, in theory, a single molecule. We also note that the earlier experiments were performed in bulk, here we will instead use data from experiments of fluorescently stained linear dsDNA in nanochannels which allows us to see each molecule and exclude damaged ones. Accordingly, we hope to, in this preliminary study, give a good first order of magnitude estimate of nicking rates for DNA in nanochannels for limited amounts of data.

2 Problem statement and thesis outline

We will in this section state the problem that this work intends to solve and outline the new theoretical-experimental platform for DNA nick analysis presented in Figure 1. In Figure 1, there

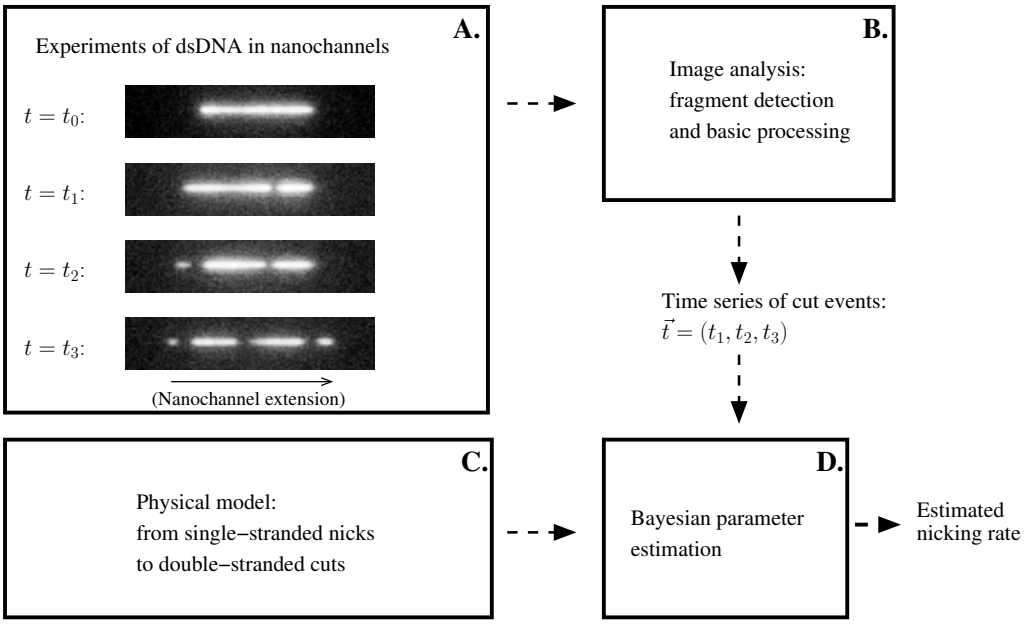


Figure 1: **A schematic overview of the theoretical-experimental platform.** The working process of this study begins by obtaining the experimental data of dsDNA stained with fluorescent molecules in nanochannels undergoing fragmentation (see box **A.**). In the four images taken at different times we show the same DNA molecule. The DNA is observed, and can be separated from the background, via the concentration of fluorescent molecules which determines the brightness of the pixels (value along z-axis in the images). Furthermore, the DNA is, as can be seen in the images, stretched out. This is due to the confinement imposed by horizontally aligned (along the x-axis) nanochannels of a width approximately equal to one pixel. We can in the images see that cuts appear along the initial intact molecule as time progress and is split into 4 observable fragments at $t = t_3$. The three last images taken at $t = t_1, t_2$ and t_3 correspond to the time at which we can observe a new cut. In addition, we see that the limited resolution extends the region of bright pixels far outside the confinement of the nanochannels. The experiments were done by Jason Beech (Jonas Tegenfeldt group, Lund University). The next stage in the work is to perform image analysis on the experimentally obtained images with the goal of detecting individual fragments. The output of the image segmentation procedure then undergoes simple processing to obtain time series of cut events (box **B.**). To describe the experimental data in box **A.** we presented a physical model which connects the rate of nicks and the observed double-stranded cut times in a probabilistic manner (box **C.**). Lastly, the time series of cuts from box **B.** and the physical model from box **C.** are combined in a probabilistic Bayesian framework which outputs an estimate of the nicking rate (box **D.**). In this thesis we will present work to cover the theoretical framework of box **B.**, **C.** and **D.**.

are 4 parts which we will now cover in some detail.

In Figure 1, box **A.**, we find a brief overview of the experiments and the type of data produced. The experiments consist of obtaining microscope images of fluorescently stained dsDNA in nanochannels undergoing fragmentation. Here we have chosen to zoom in on a single molecule to visualize the cuts better, the full image contains several molecules spread out in several nanochannels. The experiments were done by Jason Beech (Jonas Tegenfeldt group, Lund University). The details of the experiments as well as the working procedure are presented in section 3.1.

In Figure 1, box **B.**, we find the data extraction part from the experimentally obtained images. Here the goal is to observe time series of cut events. The knowledge of the time at which each image is obtained allows us to deduce a time series of cut events via the number of fragments in each image. The limited resolution of the microscope images poses a major challenge for this procedure to be accurate. The procedure of the image analysis and time series extraction is presented in section 3.2.

In Figure 1, box **C.**, we find the theoretical framework, presented in section 3.3, aiming to describe the cutting process observed in the experiments seen in box **A.**. This is the central part of this work of this thesis, and we present, in section 3.3.1, a new theoretical framework connecting

the observable cuts and the underlying nicking rate given a functional form of the cutting rate. Since the experiments have two major issues: (i) the images have limited spatial resolution and, (ii), the fragments diffuse with a limited rate, we will present two different models for the functional form of the cutting rate used in the theory presented in section 3.3.1. The first cutting rate model, model I, is analytical and ignores the experimental issues. We present model I in section 3.3.2. The second model, model II, is based on simulations and intends take the experimental shortcomings into account. To deal with issue (i) and (ii), model II attempts to simulate artificial microscope images of fragmenting DNA in nanochannels, including limited resolution and diffusive behavior, to obtain an empirical functional form of the cutting rate. We present model II in section 3.3.3.

In Figure 1, box **D.**, we find the probabilistic Bayesian framework used to estimate model parameters. The probabilistic framework, described in section 3.4, needs a physical model for observed cut times and corresponding experimental data with time series of cuts in order to estimate model parameters.

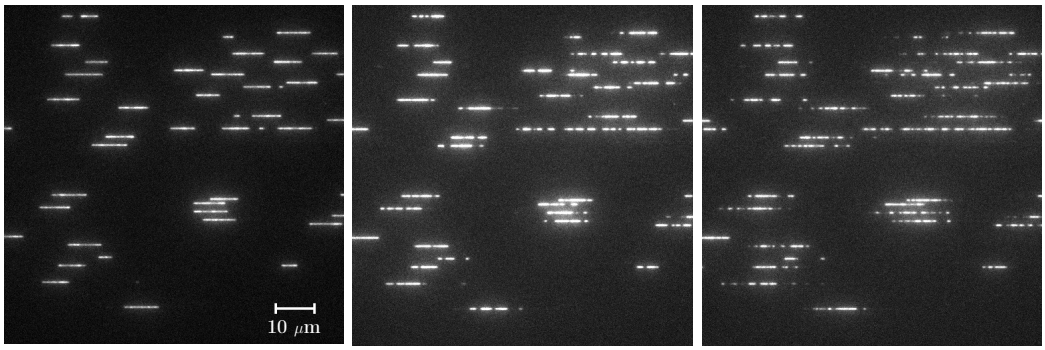
To conclude, we will in this study use the new experimental-theoretical platform presented in Figure 1 to estimate the single-stranded nicking rate. To do this, we will develop two new probabilistic models, which can connect the nicking rate with the observed times of cuts. We first test these new models on synthetic movies to check their applicability before analyzing real experimental movies.

3 Methods

In this section, we will present the methodological material of this work according to the framework in section 2 (Figure 1). The content of the material will have a brief character to keep clarity of the structure. Details and technicalities are discussed in the appendices.

3.1 Experiments

Here, we present the experimental procedures and details behind the work of obtaining videos of diffusing DNA in nanochannels performed by Jason Beech (for example images from one such video see Figure 2). All experiments were performed in a silicon device with a quartz cover-slide.



(a) $t = 0$ s (b) $t = 3$ s (c) $t = 6$ s

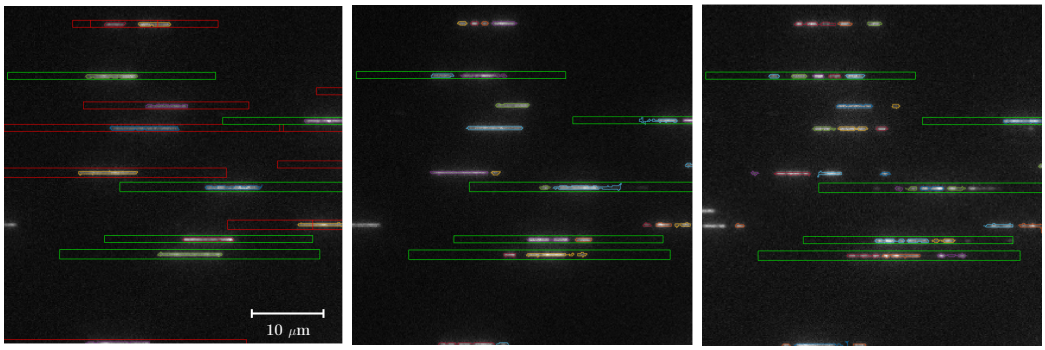
Figure 2: **Example images from one experimental movie.** Here, we show three different images taken at three different times containing fragmenting dsDNA diffusing in nanochannels. In each image, we can see ~ 40 molecules. We observe the DNA molecules as the bright regions. The images in box **A.** (Figure 1) showed a limited field of view, containing one molecule, from the movie presented here. The nanochannels (not directly visible) are horizontally aligned and each one stretches over the whole width of each image. The width of each nanochannels is 100 nm (approximately one pixel) in the vertical direction. Furthermore, we can see that the DNA molecules are stretched out in the same direction as the nanochannels due to confinement. (a): In this image we see the initial stage of the experiment where most molecules are intact and similar in length. (b): At later times, we start to notice that several molecules have been cut, giving rise to shorter fragments. (c): At this late stage in the experiment we see how the majority of molecules have undergone substantial fragmentation. The images are all 512×512 pixels wide and high and the pixel size is 0.16 µm. Experiments performed by Jason Beech (Jonas Tegenfeldt group, Lund University)

The nanochannels were 100 nm wide and 150 nm deep. To capture the needed images, a 100x

objective was used with a numerical aperture of 1.4. Data of the relationship between voltage and luminosity of the lamp used to illuminate the sample is presented in Appendix A. Two types of driving gases were used to obtain different nicking rates, oxygen and nitrogen, respectively, applied with a pressure of 250 mBar. For all experiments, λ-phage DNA was used stained with 1 dye per 5 base-pairs ($\rho = 1/5$) using YOYO1 molecules. The buffer used for all experiments was a 0.5×TBE buffer. To maximize the presence of a single gas at the time of the experiments, the device was put under pressure for 4 hours when using nitrogen and 2 hours using oxygen as driving gas before the insertion of DNA into the nanochannels was done. Once the system was saturated with a single gas and the DNA been inserted into the nanochannels, the lamp cover was removed, and the video acquisition took place until the initial molecules were highly fragmented. The exposure time, ΔT, used to capture a single image was 0.1 s, i.e., 10 images were captured every second. The acquisition of videos was performed for three different illumination strengths, 25, 50, and 75 % of maximum voltage for both nitrogen and oxygen as driving gas. For details about the fabrication of the nanofluidic system used, preparation of the DNA and the imaging system, we refer to Appendix A.

3.2 Image analysis and time series extraction

In this section, we present a working method to extract time series from fragmenting DNA diffusing in nanochannels based on image segmentation. The data consist of different stacks (videos) of gray scale images $\vec{S} = \left\{\vec{I}_1, \vec{I}_2, ..., \vec{I}_{N_{\text{im}}}\right\}$ taken with a separation ΔT in time. Each image is a two-dimensional matrix $\vec{I}_i = I(x, y)_i$ with $x \in [1, x_{\max}]$ and $y \in [1, y_{\max}]$, where $x_{\max}$ and $y_{\max}$ is the width and height of the images in terms of pixels. Three representative images from one video $\vec{S}$, at three different time points, are shown in Figure 2.



(a) $t = 0\,\text{s}$ (b) $t = 3\,\text{s}$ (c) $t = 6\,\text{s}$

Figure 3: **Illustration of fragment detection and choice of acceptable data.** Here, we have zoomed in on the top left corners of the images in Figure 2 to improve the visibility of individual molecules and allow for illustration of the data extraction procedure. The images in Figure 2 have here been segmented to detect all regions containing DNA, marked as regions with borders of various colors (not large green or red rectangles). In the first time instance, $t = 0\,\text{s}$ (Figure 3a), all molecules which are considered acceptable are enclosed with green rectangles 2 times higher and 4 times longer than the original molecule. All the non-acceptable molecules are enclosed by red rectangles. Molecules may be marked as non-acceptable due to: too long or too short (indicating merge with another fragment/molecule or already fragmented), too close to other fragment(s) or molecule(s) (enclosing rectangles overlap) and partly outside image (enclosing rectangle extends outside the image). In the second time instance, $t = 3\,\text{s}$ (Figure 3b), we have excluded all non-acceptable molecules. Here, we can also observe that we detect multiple fragments inside a single green rectangle which initially only contained one molecule. In the last time instance, $t = 6\,\text{s}$ (Figure 3c), we observe that an increasing number of fragments are detected and that most of the fragments are rather short, in some cases so short that we failed to detect them.

From a given video, we want to extract the time at which a new cut happens for each individual DNA molecule that is intact in the initial frame. To do this, we used the following procedure:

1. Locate all DNA molecules in the first frame. To do that, we segment the first image into signal and background (using the segmentation technique described in Appendix B) which returns the boundaries of all detected molecules as can be seen in Figure 3a.

2. Find a rectangular area which is likely to contain a given DNA molecule at all time frames. To do this, we define, for each intact molecule, a rectangle which is 2 times higher, 4 times longer and centered around the initial molecule.

3. Exclude all non-acceptable molecules. To do this, we exclude all molecule which larger rectangles overlap or extend outside the image, as well as all molecules which observed length does not satisfy the criterion

$$L_{\text{est}} - 3\sigma \quad < L_{\text{obs}} < \quad L_{\text{est}} + 3\sigma. \tag{3.1}$$

   Here, L_{obs} is the observed length, L_{est} the estimated expected observed length and σ the standard deviation of L_{est}. We use Eq. (G.1) to compute L_{est} and Eq. (G.2) to compute σ. Non-acceptable molecules are surrounded by larger red rectangles in Figure 3a.

4. Count the *observed* number of cuts at each time frame, $\tilde{N}_{\text{cuts}}(k)$, for all molecules ($k$ labels different time frames). We do this by segmenting each image and count the number of detected fragments in each rectangle (example segmentation in Figure 3b and 3c). The number of cuts is one less than the number of fragments.

5. Estimate the *actual* number of cuts, $N_{\text{cuts}}(k)$, at each time frame k from the observed number of cuts. To do this we define $N_{\text{cuts}}(k) = \max\left(\tilde{N}_{\text{cuts}}(j)\right)$ with $j \in [1, k]$, i.e., the *cumulative max* in time of the observed number of cuts.

After completion of steps 1-5 above, we have obtained a time series with the actual number of cuts, $N_{\text{cuts}}(k)$, for a specific molecule at time frame k.

Let us here comment on two things in the data extraction procedure just described. Firstly, we take the cumulative max in time of the observed number of cuts to minimize the effects of missed cuts. Missed cuts are caused by two or more fragments located closer to each other than the resolution limit given by our segmentation technique. We can estimate the resolution limit to be approximately equal to σ_{PSF} given in Eq. (F.2). In this study $\sigma_{\text{PSF}} \approx 222$ nm or equivalently 1.4 pixels. Thus, missed cuts can be present when a cut recently happened and the two resulting fragments have not yet diffused apart a distance longer than the resolution limit, or when already separated fragments diffuse together a distance closer than the resolution limit. Secondly, the choice of height and width in step 2 is based on observations of multiple image sequences and aims to produce rectangles of a size such that we balance the degree of fragment enclosure with data loss.

We now seek the *time series of cuts* $\vec{t}$. For a specific molecule we can identify the time frames when a cut happened, $\vec{k}_{\text{cut}}$, by satisfying the following condition

$$N_{\text{cut}}(k+1) - N_{\text{cut}}(k) > 0 \quad \text{for} \quad k \in [1, N_{\text{im}}]. \tag{3.2}$$

We note that in Eq. (3.2) we need to account for the fact that multiple cuts could have happened from time frame k to $k+1$. Therefore, one value of k appears equally many times as the increase in cutting number between consecutive images. The time series of cuts for a single molecule is thus given by

$$\vec{t} = \vec{k}_{\text{cut}} \Delta t. \tag{3.3}$$

We have here assumed that the first image in $\vec{S}$ was taken at $t = 0$, if the recording of images started before the experiment itself we translate the indices of the images as $i = i - i_{t=0}$. Note that, for a given video, we will obtain a set of individual time series $\vec{t}$ equal to the number of accepted molecules in the first image N_{mol}, i.e., $\vec{\tau} = \left\{\vec{t}_1, \vec{t}_2, ..., \vec{t}_{N_{\text{mol}}}\right\}$.

3.3 The physical models

In this section, we present the theoretical framwork of box **C.** in Figure 1. This work consists of three main parts. Firstly, we present the new probabilistic model for the observed times of cuts given a functional form for the cutting rate, $r(t)$, in section 3.3.1. After this, we present an analytical model for $r(t)$ (model I) in section 3.3.2. Lastly, we present a simulation-based model for $r(t)$ (model II) in section 3.3.3.

3.3.1 From nicks to cuts

In this section, we introduce a new probabilistic model for the observed times of cuts given a function for the rate of cuts $r(t)$. To tackle this problem, we begin by approaching it from a general perspective utilizing a state like waiting time model for the cuts presented in [15], chapter 15.2.3.

The model proceeds as followed: Consider a time interval $t \in [0, T]$. Divide this time interval into N shorter time intervals where time interval i is given by $[t_{i-1}, t_i]$. We assign our system two possible states $\sigma_i = 0$ and $\sigma_i = 1$ for having no cut and a cut in time interval i, respectively. The evolution of our system in time can now be described by a set of σ values representing the state in each time interval, $\vec{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_N)$. We now ask: assuming that each state is independent of any other state, what is the probability of observing $\vec{\sigma}$? We can write this as product of individual probabilities

$$\tilde{P}\left(\vec{\sigma}\right) = \prod_{i=1}^{N} \hat{P}(\sigma_i). \tag{3.4}$$

To deduce the probability of $\sigma = 0$ and $\sigma = 1$, respectively, we need to confront the physical world of rate equations. We realize that the probability of observing a cut in time interval i can be related to the rate of cuts at that time. We write this probability as $\hat{P}(\sigma_i = 1) = r(t_i)\delta t$. Where $r(t)$ is the cutting rate at time t and δt is the length of the time interval. The probability of not having a cut in time interval i is then given by $\hat{P}(\sigma_i = 0) = 1 - \hat{P}(\sigma_i = 1) = 1 - r(t_i)\delta t$. Assuming that $1 \gg r(t_i)\delta t$, we can, using the exponential approximation $e^{-x} = 1 - x + \mathcal{O}(x^2)$, rewrite $\hat{P}(\sigma_i = 0) \approx e^{-r(t_i)\delta t}$. With this approximation the total probability for a set of cuts at times $\vec{t} = (t_1, t_2, ...t_n)$, i.e., a time series of cuts, can be written as

$$P\left(\vec{t}\right) \approx \left(\prod_{i=1}^{n} r(t_i)\,\delta t\right) e^{-\sum_{j=1}^{N} r(t_j)\delta t}. \tag{3.5}$$

We note that the probabilities $\hat{P}(\sigma_i = 1)$ for all time intervals in which a cut happened are still included in the product above. Although not correct, the error will vanish later when we let the length of the time intervals approach zero.

At this point we need to realize that $P(\vec{t}) \approx p(t_1)\delta t \cdot p(t_2)\delta t \cdot ... \cdot p(t_n)\delta t$, where $p(t)$ is the probability density function for observing a cut at time t. Dividing both sides of Eq. (3.5) with $(\delta t)^n$ and letting δt approach zero, we obtain

$$p(\vec{t}) = \prod_{i=1}^{n} p(t_i) \approx r(t_1)\, r(t_2) \ldots r(t_n)\, e^{-\int_0^T r(t')dt'}. \tag{3.6}$$

We point out that $p(\vec{t})$ is a joint probability density function for all observed cuts and has dimension time^{-n}. Furthermore, we observe that Eq. (3.6) is valid for any choice of model for $r(t)$. This observation will prove itself useful later as we want Eq. (3.6) to be valid for $r(t)$ from both model I and model II. With that, we will now continue and search for a model, and an associated functional form for $r(t)$, which satisfactorily connects the rate of nicks to the rate of observable cuts.

3.3.2 Model I for $r(t)$

To deduce the relationship between single-stranded nicks and double-stranded cuts we will now introduce a physical model that attempts to predict $r(t)$.

Considering the cutting rate $r(t)$, we realize that it can be represented by the probability of having a cut at time t given that a new nick is introduced, $P_{\text{cut}}(t)$, multiplied by the number of attempts at that time κ. With that in mind, the cutting rate becomes

$$r(t) = \kappa P_{\text{cut}}(t). \tag{3.7}$$

The number of attempts to cut at each time, κ, is nothing but the nicking rate per length, which we denote as α, multiplied by the length of the molecule that can be nicked

$$\kappa = 2L\alpha. \tag{3.8}$$

Here the factor 2 enters the equation since we have two strands in the DNA that are susceptible to nicking. Note that the unit of L determines the unit of α. In this study L is given in base-pairs which means we measure α as the nicking rate per base-pair. In Appendix C, we perform, as additional material, a calculation which links α to the concentration of molecules that nick the DNA in the surrounding buffer.

We now need to find the quantity $P_{\text{cut}}(t)$ in order to recover the expression for the cutting rate. Instead of going directly for $P_{\text{cut}}(t)$ we ask: given one nick at one of the strands, what is the probability Q_{nick} for a new nick to be placed within a distance of $\xi/2$ in both directions on the opposite strand from the original nick, assuming it to be uniformly likely along the whole strand? (In Appendix D we cover the *fraying distance*, ξ, in more detail and theoretically estimate it to approximately 1 base-pair). The answer is

$$Q_{\text{nick}} = \frac{\xi}{L}, \tag{3.9}$$

where L is the total length of one of the strands. Equivalently we can say that the probability of not placing the new nick within a distance of $\xi/2$ in both directions from the original nick on the opposite strand is

$$Q_{\text{no nick}} = 1 - Q_{\text{nick}} = 1 - \frac{\xi}{L}. \tag{3.10}$$

We here point out that once we place a new nick on one of the strands we have to reconsider Eq. (3.10) the same number of times as there are nicks on the opposite strand. Therefore, the probability of having no new cut at some instance is given by

$$P_{\text{no cut}} = (Q_{\text{no nick}})^{nL} = (1 - \frac{\xi}{L})^{nL}. \tag{3.11}$$

Note that we rewrote the number of nicks as nL, where n is the density of nicks. After inspection of Eq. (3.11) and invoking the assumption that $L \gg \xi$ ($L \sim 10^4$), we can make the approximation

$$P_{\text{no cut}} \approx e^{-\xi n}. \tag{3.12}$$

It is here important to note that $n = n(t)$ as the density of nicks changes with time. Now, we can finally return to specifying the probability of having a cut for each new nick and find, using Eq. (3.12), that

$$P_{\text{cut}}(t) = 1 - P_{\text{no cut}} = 1 - e^{-\xi n(t)}. \tag{3.13}$$

With this, we can finally write down the cutting rate given in Eq. (3.7) as

$$r(t) = 2\alpha L(1 - e^{-\xi n(t)}). \tag{3.14}$$

We have so far deduced an expression for the cutting rate which can be used in Eq. (3.6) to perform parameters estimation given a time series of cuts but one thing remains to be clarified, the nicking density $n(t)$. We assume, in accordance with [10], that the nicking process can be described as a Poisson process resulting in a linear increase of the nicking density equal to

$$n(t) = \alpha t + n_0, \tag{3.15}$$

where n_0 is the initial nicking density.

We will now analyze some properties of $r(t)$ in Eq. (3.14). For short times, we have $1 \gg \xi n(t)$, which allows us to Taylor expand to first order, yielding

$$r(t) \approx 2\alpha L\xi(\alpha t + n_0). \tag{3.16}$$

We can now choose to look at the case when $n_0 = 0$. Integration of $r(t)$ in Eq. (3.16) with $n_0 = 0$ from 0 to T gives the expected number of cuts

$$\langle N_{\text{cuts}} \rangle = \alpha^2 L\xi T^2, \tag{3.17}$$

and shows that the nicking process, with a *linear* increase of nicks in time, results in a cutting process where the number of cuts is *quadratic* in time.

At this point, we can also note that it is possible to estimate the time and how many nicks we need in order to observe the first cut. Solving Eq. (3.17) for T and making the substitution $N_{\text{nicks}} = T\alpha L$ gives, when solved for N_{nicks},

$$\langle N_{\text{nicks}} \rangle = \sqrt{\frac{L\langle N_{\text{cuts}} \rangle}{\xi}}. \tag{3.18}$$

We have so far not only derived a functional form for the cutting rate, but also realized that the relation between nicks and cuts is related in the non-obvious way found in Eq. (3.17).

3.3.3 Model II for $r(t)$

In this section, we will outline the procedure used to obtain a simulation-based cutting rate function, $r(t)$. To that end, we perform simulations that intend to mimic the data presented in Figure 2. For a more detailed algorithm, we refer to Appendix F. In addition, we present, in Appendix G, theory which allows us to estimate the observed length of dsDNA in nanochannels. On top of that, we cover the basics of DNA diffusion in nanochannels in Appendix H.

The problem at hand here is that we are, from the experiments presented in section 3.2, presented with *blurred* time series of cuts, i.e., time series of cuts obtained under the influence of limited spatial resolution due to camera physics and limited diffusion rate of fragments. These two properties give rise to hidden, delayed and disappearing cuts. Thus, model I for $r(t)$ may be inaccurate and biased for experimentally obtained data. Luckily, Eq. (3.6) is valid for all $r(t)$, which allows us to choose any $r(t)$ that better corresponds to the physical world. To do this, we intend to simulate $\langle N_{\text{cuts}} \rangle(t)$, corresponding to Eq. (3.17), and obtain a realistic version of $r(t)$ for experimental data by taking the time derivative of $\langle N_{\text{cuts}} \rangle(t)$. By simulating $r(t)$ for different values of the nicking rate α, we can then estimate the correct α value using the theory in section 3.4 for a given set of time series of cuts.

To simulate a single version of $r(t)$ for a specific set of parameters, we use a four-step procedure, including theory from the Gillespie simulation method [16, 17, 18]: (i) Simulate the single-stranded nicking process, (ii) simulate the diffusion process of the resulting fragments (iii) generate a synthetic movie from (i) and (ii), (iv) segment the synthetic movie and estimate the number of actual cuts, using the procedure described in section 3.2. Finally, we repeat step (i)-(iv) M times to obtain reliable statistics. To do this, we assume that all four steps are independent of each other such that we can perform them separately. Below we describe, in brief, the four steps.

Single-stranded nicking

We begin with the simulation of the single-stranded nicking process. This process is primarily governed by the nicking rate per nicking site denoted as α and the number of available nicking sites $N_{\text{ns}} = 2N_{\text{bp}}$, where N_{bp} is the number of base-pair along one strand of the DNA. The nicking process is closely related to the chemical decay process [17]. Let us assume that each nicking site can only be nicked once. The simulation of the nicking process is given by:

1. Initialize.
2. Sample a Gillespie time to the next nick.
3. Place a nick at a randomly chosen site and update running time.
4. Check if cut happened by locating nicks on the opposite strand within half the fraying distance, if yes: save the current time of cut, split the fragment which was cut and save the fragment constellation at the time after the cut.
5. Repeat from step 2 until stopping time is reached or all sites are nicked.

With the method just described we have simulated a unique stochastic series of single-stranded nicks on a DNA molecule of length N_{bp} in base-pairs and obtained all positions and times of resulting cuts.

Diffusion of fragmenting DNA molecules

At this point we will deal with the second step, outlining the stochastic diffusion simulation process for several fragments simultaneously diffusing in one dimension. Using the output from the single-stranded nicking process the diffusion simulation process is:

1. Initialize using the cutting times from the nicking process.
2. Compute observed length and position of all fragments.
   - First instance: place the initial molecule at its start position.
   - Second or later cut instance: place all fragments according to the previous positions except for the fragment that was cut, whose children fragments are to be placed side by side conserving the center of mass of the parent fragment.
3. Sample a Gillespie time to the next move.
4. Randomly sample a fragment with weights according to their individual diffusion constants (the diffusion constant is inversely proportion to the contour length, see Appendix H). Also sample, with equal probabilities, a direction to move.
5. Move the fragment if it will not result in an overlap with another fragment.
6. Save the current positions of all fragments together with the current time.
7. Repeat from step 3 until stopping time of current instance, if we passed current cut time return to step 1, or quit if we passed last cut time.

Let us here point out some details regarding the simulation procedure just described. Firstly, we note that the simulation of the diffusion process is made between the cut instances produced by the nicking simulation such that the lengths of the fragments are constant in all simulations. Furthermore, we update the simulation time also for failed attempts to move due to resulting overlap with another fragment [19].

Generating a synthetic movie from the diffusion simulation

Now, we will present an algorithm that translates the previously simulated trajectory of the diffusing DNA fragments into a synthetic movie mimicking the experimental fluorescent images captured using a microscope and a EMCCD camera. We do this as followed:

1. Divide the simulation time of the diffusion process into intervals equal to the acquisition time (see section 3.1).
2. Create an artificial background image of photon counts.
3. Create an empty signal image of same size as the background image.
4. While the simulation time is smaller than the start time of the next time interval of image acquisition:
   (a) Compute the exposure time before move for current fragment positions.
   (b) Sample the number of photons emitted in each pixel based on the exposure time.
   (c) Add photon numbers to signal image.
5. Add background and signal image to get *total photon count image* and convert this image to the *final image* with correct pixel counts by applying the experimental setups optical and camera specific properties.
6. Save the *final image* and repeat from step 2 for all time intervals of image acquisition.

We will now mention a few important parts when simulating a synthetic movie from diffusion trajectories. Firstly, it is important to take the optical setup and camera parameters used in the experiments into account. This includes the gain setting, the quantum yield, noise contributions, the numerical aperture and the photon to electron conversion factor. Furthermore, the separation in time between images Δt and acquisition time of each image plays a crucial role. In this study these two quantities are the same.

In addition, we present in Appendix I synthetic images generated according to the procedure described above and compare these to real experimental images.

Estimating $r(t)$ from the synthetic movie

In order to estimate $N_{\text{cuts}}(t)$ from the synthetic movie, we use the method presented in section 3.2 modified to only include steps 4 - 5 from the list.

In order to get $\langle N_{\text{cuts}} \rangle(t)$ found in Eq. (3.17), we need to redo the simulations, starting from the single-stranded nicking process up until this point, several times such that we can take the average of many $N_{\text{cuts}}(t)$ for all time instances. With $\langle N_{\text{cuts}} \rangle(t)$ at hand we obtain the cutting rate for model II as

$$r(t) = \frac{\mathrm{d}\langle N_{\text{cuts}} \rangle(t)}{\mathrm{d}t}. \tag{3.19}$$

Since only a limited number of $N_{\text{cuts}}(t)$ can be simulated, we expect statistical fluctuations in both $\langle N_{\text{cuts}} \rangle(t)$ and $r(t)$, respectively. To decrease the fluctuations, we choose to apply a mean filter in time on $r(t)$ of length $10\Delta t$, based on empirical grounds, which in real conditions would correspond to an averaging over one second.

We make a final note on the structure and division of the four steps of the simulation procedure described in this section. In the detailed algorithm of the complete simulation presented in Appendix F, we still make the separation between the single-stranded nicking process and the diffusion process. On the other hand, we perform, due to high memory requirements of saving the trajectories of all diffusing fragments and all synthetic movies, the diffusion simulation, the generation of synthetic images and the image segmentation simultaneously. This allows us to only save the previous positions of all fragments and instead update the photon count in the signal image as we proceed through the simulation. Although different, the two approaches will produce equivalent results.

We can here note that, in addition to $r(t)$, the simulations can create blurred as well as *non-blurred* time series of cuts. To obtain a blurred time series of cuts we utilize the same strategy already outlined in section 3.2 with Eq. (3.2) and Eq. (3.3) using $N_{\text{cuts}}(t)$ obtained from a single synthetic movie. The non-blurred time series of cuts can be obtained directly as an output from the single-stranded nicking simulation.

Furthermore, we point out that it is possible to obtain a third version of $r(t)$ from the single-stranded nicking simulation. This cutting rate resembles the cutting rate we would observe without the influence of hidden, delayed or disappearing cuts and should be consistent with $r(t)$ from model I if the assumptions we made to derive it are true. To obtain $r(t)$ from the single-stranded nicking simulation, we do the following: (i) simulate M individual nicking processes (ii) for each simulation save $N_{\text{cuts}}(t)$ (with time discretized into intervals using Δt) (iii) take the average of all $N_{\text{cuts}}(t)$ to obtain $\langle N_{\text{cuts}} \rangle(t)$ (iv) use Eq. (3.19) with $\langle N_{\text{cuts}} \rangle(t)$ to obtain $r(t)$. In Appendix K, we show that $r(t)$ from the nicking simulation and $r(t)$ from model I yield consistent results.

3.4 Bayesian parameter estimation

In this section, we cover the basics of parameter estimations using Bayesian data analysis. Equipped with a stochastic model, we want to estimate its parameters $\vec{\theta}$ given some *data*. To do this, we use Bayes theorem [20]

$$p(\vec{\theta} \,|\, \text{data}) = \frac{p(\text{data} \,|\, \vec{\theta}) p(\vec{\theta})}{p(\text{data})}. \tag{3.20}$$

Here, $p(\vec{\theta} \,|\, \text{data})$ is the conditional probability density of $\vec{\theta}$ given some data, $p(\text{data} \,|\, \vec{\theta})$ is the conditional probability density of the data given $\vec{\theta}$, $p(\vec{\theta})$ the prior probability density of observing

$\vec{\theta}$ and $p(\text{data})$ the probability density of observing the data. The factor $p(\text{data})$ in Eq. (3.20) is referred to as the *evidence* and can for purposes of parameter estimations be omitted an Eq. (3.20) becomes

$$p(\vec{\theta}\,|\,\text{data}) \propto p(\text{data}\,|\,\vec{\theta})p(\vec{\theta}). \tag{3.21}$$

Conveniently, Bayes rule thus lets us relate $p(\text{data}\,|\,\vec{\theta})$ (what we have) to $p(\vec{\theta}\,|\,\text{data})$ (what we want), given a prior probability density, through proportionality.

In this study, will use two different stochastic models to estimate two different parameters. Firstly, we will use Eq. (3.6) to estimate the nicking rate, this means that $\vec{\theta} = \alpha$ and the data is given by a number of time series of cuts. Secondly, we will use Eq. (H.2) to estimate the diffusion constant. In this case, $\vec{\theta} = D$, where D is the diffusion constant, and the data consists of molecule positions.

Here, we present the procedure used to estimate the nicking rate. To begin with, we note that the cutting rate is dependent on the nicking rate, i.e., $r(t) = r(t, \alpha)$. We want to estimate the most likely value of $\vec{\theta} = \alpha$ using the stochastic model in Eq. (3.6) combined with Eq. (3.21) given data in form of N times series of cuts $\vec{t}^{(n)} = (t_1^{(n)}, t_2^{(n)}, ..., t_T^{(n)})$, where $n \in [1, N]$ labels different time series. To do this, we firstly define a function which is proportional to the joint probability density function for a single time series of cuts, $\vec{t}^{(j)}$, using Eq. (3.6) as

$$y(\vec{t}^{(j)}\,|\,\alpha) = \prod_{i=1}^{T} r(t_i^{(j)}, \alpha) e^{-\int_0^{t_T^{(j)}} r(t',\alpha)\,dt'} \left[U(t_i^{(j)} - t_{i-1}^{(j)})\right] \tag{3.22}$$

where $t_0^{(j)} = 0$. In addition, $U(x)$ is defined to be 0 if $x < 0$ and 1 if $x \geq 0$. Extending Eq. 3.22 to include all time series of cuts as well as a prior probability we obtain

$$p(\vec{t}^{(n)}\,|\,\alpha)p(\alpha) \propto \prod_{j=1}^{N} y(\vec{t}^{(j)}\,|\,\alpha)\ \left[H(\alpha - \alpha_{\min}) - H(\alpha - \alpha_{\max})\right]. \tag{3.23}$$

Note that we chose a uniform prior probability density $p(\alpha) \propto H(\alpha - \alpha_{\min}) - H(\alpha - \alpha_{\max})$, where $H(\cdot)$ is the Heaviside step function [21]. Here, $\alpha_{\min}$ and $\alpha_{\max}$ denotes the lower and upper bounds of α, respectively. We also note that the right-hand side of Eq. (3.23) is proportional to $p(\alpha\,|\,\vec{t}^{(n)})$. This allows us to estimate the most likely nicking rate by varying α as to maximize the right-hand side of Eq. (3.23).

Let us now present the procedure used to estimate the diffusion constant, D. To begin with, in this case $\vec{\theta} = D$ and the data is now given by N molecules center of mass trajectories $\vec{x}^{(n)} = (x_1^{(n)}, x_2^{(n)}, ..., x_T^{(n)})$, where $n \in [1, N]$ labels different trajectories. The positions in the trajectories are sampled with time intervals Δt and the center of mass of a molecule is the midpoint between its edge pixels located furthers to the left and right in the horizontal direction of the image, respectively. To estimate the most likely value of D we vary it as to maximize $\tilde{p}(D\,|\,\vec{x}^{(n)})$, which is proportional to

$$\tilde{p}(\vec{x}^{(n)}\,|\,D)\tilde{p}(D) \propto \prod_{j=1}^{N}\prod_{i=2}^{T} \frac{1}{\sqrt{4\pi D\Delta t}} e^{-\frac{\left(x_i^{(j)} - x_{i-1}^{(j)}\right)^2}{4D\Delta t}} \left[H(D - D_{\min}) - H(D - D_{\max})\right]. \tag{3.24}$$

Here, we chose a uniform prior probability density $\tilde{p}(D) \propto H(D - D_{\min}) - H(D - D_{\max})$, with $D_{\min}$ and $D_{\max}$ begin the lower and upper bounds for D, respectively. We note that Δt must be chosen sufficiently large such that the difference in displacement between the sampling times resembles a normal distribution. To ensure this, we systematically increase Δt until the estimated D has reached a plateau. The value of D at this plateau is our final estimate of D.

4 Results

Here, we present the results of this study according to the following structure: Firstly, we present an estimation of the diffusion constant for an intact DNA molecule diffusing in a nanochannel (used in model II). After that, we will compare $r(t)$ from model I with $r(t)$ from model II for

different nicking rates. Next, we will estimate the nicking rate for synthetic movies with known ground truth values of the nicking rate using both model I and model II. Lastly, we will estimate the nicking rate for experimental movies using model II. More elaborate discussions of the results will be saved to section 5.

For all the results presented in this section, we used some fixed values of parameters and constants (if nothing else is mentioned) listed in Table 1.

Table 1: Fixed numerical values of parameters used in this study.

Parameter	Numerical value	Unit	Description	Reference
D	0.0635	$\mu m^2/s$	Diffusion constant	section 4.1
L	48490	bp	Contour length of λ-DNA	[22]
ρ	1/5		YOYO1 dye to base-pair ratio	Appendix A
ξ	1	bp	Fraying distance	Appendix D
n_0	0	1/bp	Initial nicking density	
$\langle \bar{N}_{\text{cuts}} \rangle$	50		Wanted number of cuts in simulation	
M	600		Number of simulations	
$T_{\min}$	13	s	Minimum simulation time	
λ_{bg}	100	s^{-1}	Average photon number – background	
λ_{sig}	1300	s^{-1}	Average photon number – signal	
λ_{YOYO1}	509	nm	Emitted wavelength from YOYO1	[23]
ΔT	0.1	s	Image acquisition time	section 3.1
$\alpha_{\min}$	0.0001	$1/s \cdot bp$	Lower boundary for nicking rate	
$\alpha_{\max}$	0.005	$1/s \cdot bp$	Upper boundary for nicking rate	
$D_{\min}$	0.01	$\mu m^2/s$	Lower boundary for diffusion constant	
$D_{\max}$	0.5	$\mu m^2/s$	Upper boundary for diffusion constant	

4.1 Estimation of diffusion constant

In this section, we present an estimate of the diffusion constant D as a function of Δt by maximizing the right-hand side of Eq. (3.24). We measured the trajectories of the DNA molecules up until the first cut appeared for low rates of nicking, i.e., data acquired with nitrogen as driving gas. We have here assumed that 3 different data sets can be combined even though the illumination strength varies between them. All experiments were performed using a buffer of strength 0.5×TBE. The result can be seen in Figure 4.

In Figure 4 the estimated value of the diffusion constant decreases rapidly the first second, followed by a slower decrease after that. For $\Delta t > 5\,\text{s}$ the decrease is not observable and the fluctuations between individual points becomes insignificant. Fitting a constant to the estimated values for the diffusion constant for $\Delta t > 5\,\text{s}$ yields $D_{\text{est}} = 0.0635\ \mu m^2/s$. To show that the obtained negative log-likelihood functions for each value of Δt are well behaved and yield a robust estimate of D, we show for three different values of Δt, in Appendix J, the corresponding negative log-likelihood functions.

4.2 Comparison between $r(t)$ from model I and model II

Here, we present how $r(t)$ from model I and model II compare for three different values of the nicking rate. The results can be seen in Figure 5.

In Figure 5 we can see that $r(t)$ from model I and $r(t)$ from model II are significantly different at most times for all three nicking rates. The only correspondence is found for the lowest nicking rate at short times seen in Figure 5a. We conclude from these results that including the experimental properties and limitations as well as the limited accuracy of the image segmentation technique makes a difference for the functional form of $r(t)$. Furthermore, we can also see that the discrepancy between the two models increases for an increasing nicking rate.

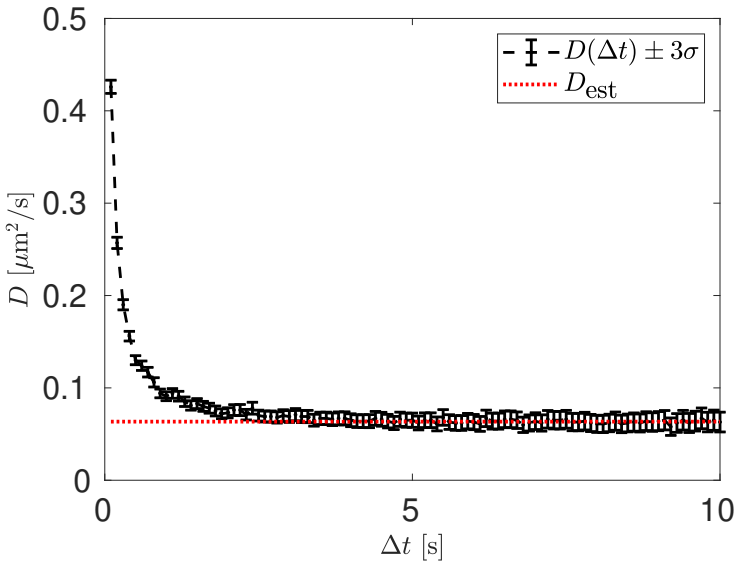
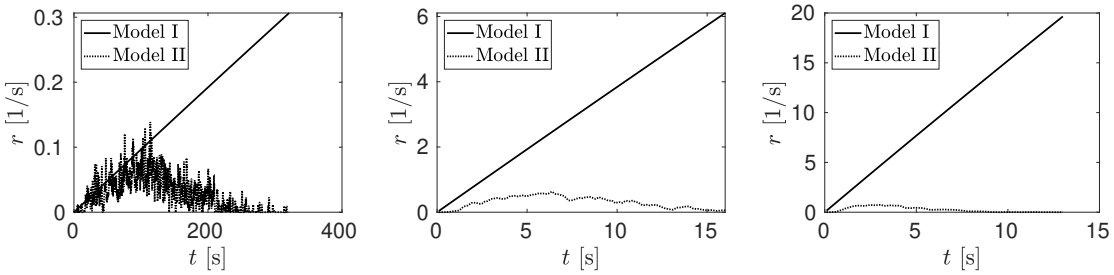


Figure 4: **Estimation of diffusion constant.** Plot of the diffusion constant estimation as a function of time between position samples Δt and final estimate of the diffusion constant, D_{est}. We note that the drop in D is fast up until $\Delta t \sim 1$ s, after that it slows down to transition into almost zero after $\Delta t = 5$ s. The estimate of D appears to plateau for $\Delta t > 5$ s, which indicates that $D(\Delta t > 5)$ are good data points for estimating the diffusion constant. The fitted constant of the values of D for $\Delta t \geq 5$ s gives $D_{\text{est}} = 0.0635$ µm^2/s. We point out that the number of samples decrease with increasing Δt, so the variance in D increases when Δt increases.



(a) $\alpha = 0.0001$ 1/s · bp. (b) $\alpha = 0.002$ 1/s · bp (c) $\alpha = 0.004$ 1/s · bp

Figure 5: **Comparison between $r(t)$ from model I and model II.** Here we have plotted $r(t)$ from model I and model II for three different nicking rates separately. We can see that $r(t)$ from model II tends to deviate more from $r(t)$ from model I with increasing nicking rate. For $\alpha = 0.0001$ 1/s · bp found in Figure 5a we observe that $r(t)$ from model I and model II agree rather well for short times but starts to deviate significantly from each other after $t > 100$ s. The correspondence for short times between $r(t)$ from model I and model II can not be observed in Figure 5b or 5c. This indicates that model II, where we included experimental constraints of limited spatial resolution and finite diffusion rate for fragments, may be increasingly important for increasing nicking rates.

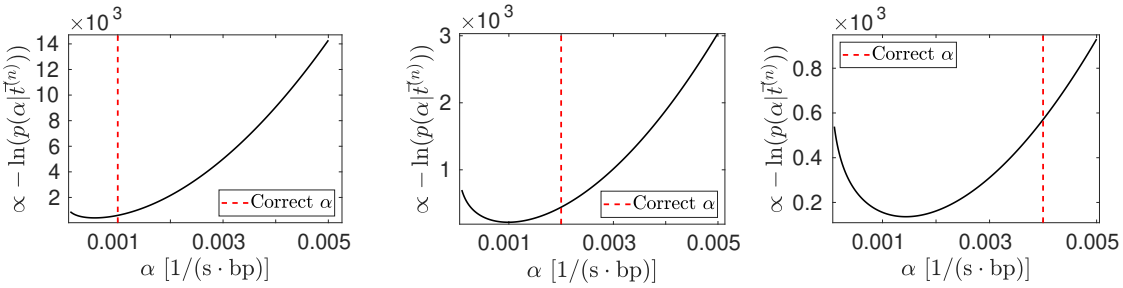
4.3 Nicking rate estimation on synthetic movies

In this section, we present estimations of the nicking rate for blurred time series of cuts from synthetic movies with three different ground truth nicking rates using $r(t)$ from model I and model II. The ground truth nicking rates for the synthetic movies are: $\alpha = 0.001$, $\alpha = 0.002$ and $\alpha = 0.004$ 1/s · bp, respectively. For each value of α we produced 30 individual synthetic movies from which 30 blurred time series of cuts were extracted.

4.3.1 Model I

We start by presenting estimations of the nicking rate using $r(t)$ from model I. The results can be seen in Figure 6.

In Figure 6, we can see that the ground truth nicking rate is not recovered in any of the three



(a) $\alpha = 0.001$ 1/s · bp.
$\alpha_{\text{est}} = 0.0056 \pm 0.00002$ 1/s · bp

(b) $\alpha = 0.002$ 1/s · bp.
$\alpha_{\text{est}} = 0.0099 \pm 0.00004$ 1/s · bp

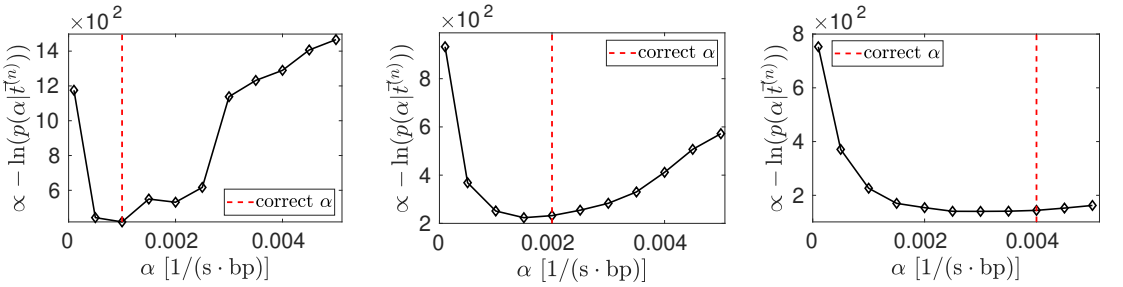
(c) $\alpha = 0.004$ 1/s · bp.
$\alpha_{\text{est}} = 0.0014 \pm 0.00007$ 1/s · bp

Figure 6: **Nicking rate estimation on synthetic movies using model I.** Here, we have plotted the negative log of the right-hand side of Eq. (3.23), which is proportional to $p(\alpha \,|\, \vec{t}^{(n)})$, using $r(t)$ from model I for three sets of blurred time series of cuts. The blurred time series of cuts are obtained from synthetic movies with three different values of ground truth nicking rates (marked with a dashed red vertical line). We observe that the estimated nicking rates do not correspond to the ground truth nicking rate, within the error margins, for any of the nicking rates. Furthermore, we observe that the estimated nicking rate tends to increasingly deviate from the ground truth for increasing values of the ground truth nicking rate. On the other hand, the model returns estimates in the right order of magnitude for all values of the nicking rate.

cases. For the lowest nicking rate the estimation is only a factor 2 wrong. However, for the two higher values of the nicking rate, we see that the estimate deviates increasingly from the ground truth values.

4.3.2 Model II

Here, we present estimations of the nicking rate using $r(t)$ from model II. The results can be seen in Figure 7.



(a) $\alpha = 0.001$ 1/s · bp.
$\alpha_{\text{est}} = 1 \cdot 10^{-3}$ 1/s · bp

(b) $\alpha = 0.002$ 1/s · bp.
$\alpha_{\text{est}} = 1.5 \cdot 10^{-3}$ 1/s · bp

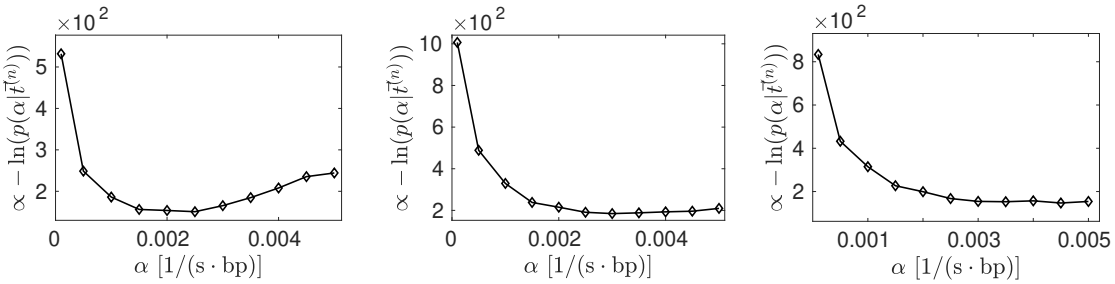
(c) $\alpha = 0.004$ 1/s · bp.
$\alpha_{\text{est}} = 3 \cdot 10^{-3}$ 1/s · bp

Figure 7: **Nicking rate estimation on synthetic movies using model II.** Here, we have plotted the negative log of the right-hand side of Eq. (3.23), which is proportional to $p(\alpha \,|\, \vec{t}^{(n)})$, using $r(t)$ from model II. The data consists of the same three sets of blurred time series of cuts from synthetic movies used to obtain the results in Figure 6. The ground truth nicking rates are marked with dashed red vertical lines. We observe that the estimate of the nicking rate, α_{est}, corresponding to the value of α that gives lowest value of negative log-likelihood function, is accurate for $\alpha = 0.001$ and also rather good for $\alpha = 0.002$ 1/s · bp. The estimate for $\alpha = 0.004$ 1/s · bp does not recover the ground truth value but the variance in the estimate is large.

We observe in Figure 7 that the estimated nicking rate is accurate for the lowest ground truth value. Furthermore, the estimated nicking rate for $\alpha = 0.002$ 1/s · bp (seen in Figure 7b) is also rather accurate. Regarding the estimate of the highest ground truth nicking rate (Figure 7c) we see that it deviates from the correct answer, but the uncertainty in the estimate is large as we can see from the flat negative log-likelihood function. From these observations, we see that model II provides good estimates of α, but where the variance in the parameter increases with an increasing nicking rate.

4.4 Nicking rate estimation on experimental movies

In this section, we present estimations of the nicking rate for three sets of blurred time series of cuts obtained from experimental data with oxygen as driving gas using $r(t)$ from model II. The experimental data sets are obtained at three different illumination strengths: 25, 50 and 75 % applied voltage out of maximum voltage for the lamp. For each illumination strength, 25, 50 and 75 %, the corresponding set of blurred time series of cuts contained 21, 33 and 30 individual blurred time series of cuts of various length, respectively. The results can be seen in Figure 8. For the interested reader, we present corresponding results using $r(t)$ from model I in Appendix L.



(a) 25 % illumination strength. $\alpha_{\text{est}} = 2.5 \cdot 10^{-3}$ 1/s · bp

(b) 50 % illumination strength. $\alpha_{\text{est}} = 3 \cdot 10^{-3}$ 1/s · bp

(c) 75 % illumination strength. $\alpha_{\text{est}} = 4.5 \cdot 10^{-3}$ 1/s · bp

Figure 8: **Estimation of nicking rate for experimental movies using model II.** Here, we have plotted the negative log of the right-hand side of Eq. 3.23, which is proportional to $p(\alpha \,|\, \vec{t}^{(n)})$, using $r(t)$ from model II for three sets of experimentally obtained blurred time series of cuts. The three sets of blurred time series of cuts have been obtained with oxygen as driving gas for three different illumination strengths, 25, 50 and 75 %, respectively. We can observe that the curvature of the negative log-likelihood curves are small around their minimum values, especially for the case in Figure 8c, indicating a large uncertainty in the estimated nicking rates. The estimated values of the nicking rate, α_{est}, correspond to the nicking rate which gives the lowest value of the negative log-likelihood function.

The results in Figure 8 show that the estimated nicking rate is increasing for increasing illumination strength. Furthermore, we can see in Figure 15a and 15b that the negative log-likelihood functions both have well defined minimums. Regarding the negative log-likelihood function in Figure 15c, we can also observe a minimum, but it is less apparent compared to those from the negative log-likelihood functions in Figure 15a and 15b. Lastly, we point out that the negative log-likelihood functions in Figure 8, in difference to the estimated nicking rates, are the final results as they represent the distributions of α in our model for each given data set.

5 Discussion

Based on the results presented in this study, we are optimistic about the possibility of measuring the nicking rate from experimental data of dsDNA in nanochannels. It appears that model II can satisfactorily estimate the nicking rate even if the variance in the parameter estimate increase with increasing values of the nicking rate. Furthermore, we can observe that using $r(t)$ from model I leads to a consistently larger underestimation of the nicking rate compared to using $r(t)$ from model II. This difference seems to increase for increasing values of the ground truth nicking rate. This conclusion is supported by the notion that cuts will take time before becoming visible due to the dynamics of diffusion and limited resolution in the images, which in turn delay the observed cut time leading to a smaller estimate of the nicking rate. The results presented in section 4.3 with known values of the ground truth nicking rate allow us to observe the magnitude of this error as well as whether or not we can satisfactorily eliminate it with more elaborate models for $r(t)$.

We should point out that this study does not provide any insight into how good we have, through the synthetic movies from model II, managed to mimic the experimental movies. This study only provides an alternative way, which is sufficiently consistent and physically more plausible compared to model I for experimental data, to deduce the functional form of the cutting rate function. To this end, let us add a more detailed discussion regarding some parts of this work.

5.1 The diffusion constant

Let us here look at the result obtained in section 4.1 where we estimated the diffusion constant. In the result we observe that the estimate of the diffusion constant is dependent on Δt. We have during the estimation chosen a range of Δt such as to maintain a statistically large enough sample without excluding too much of the plateau region. The estimated value might be significantly different from the true value, but it appears to be at the same order of magnitude as obtained by [7] and we have, for the purpose of this study, considered it to be satisfactory taking into account other plausible sources of error. To change the diffusion constant, away from the right value, appears to be a plausible extension for further work in order to test how sensitive the estimate of the nicking rate is to the estimate of the diffusion constant.

5.2 The fraying distance

Here we will consider ξ, the fraying distance, and how we dealt with it in this study. In this study, we choose to set $\xi = 1$ bp, meaning that a double-stranded cut can only appear when two nicks are located between the same two base-pairs. The estimation in appendix D shows that indeed ξ should be close to 1. It is true that the probability of opening a single base-pair is low, but we did not specify during how long time it attempted to open as well as if there are multiple chances of opening. To clarify, our simple model does not consider that base-pairs may attempt to open with a certain rate for all times. This could be the reason our estimation of the fraying distance is significantly smaller than those presented by [12, 13, 14]. We can here see the possibility of improving upon the methodology of this study by incorporating proper simulations utilizing the dynamics of the base-pair opening where time scales matter. In addition to this, it might also be of value to take into account the difference in energy required to open up a specific base-pair combined with the specific sequence of base-pairs for the λ-DNA used in this study. Lastly, we see the possibility of not fixing ξ. This would mean that we instead estimate both ξ and α.

5.3 Initial nicking density

Yet another quantity, which we did not devote a substantial amount of time to is the initial nicking density n_0. Likewise, our study does not provide any insight into whether this quantity might be important for the estimation of the nicking rate. Setting $n_0 = 0$, as we did in this study, assumes that the DNA has no single-stranded nicks when the experiment starts. This is most likely not the case in many experiments, but the importance of this effect would require some further analysis. Using Eq. (3.18) with $L = 48490$ base-pairs and $\xi = 1$ base-pair, we find that approximately 200 single-stranded nicks are required before one cut appears. In case we can guarantee that no double-stranded cuts exist before we start the experiment this would lead to a rough estimate of the initial nicking density equal to $n_0 \sim 0.002$. We conclude that n_0 is most likely small but the overall nicking density is probably also small, thus the interplay would have to be further investigated.

5.4 Model comparison

Regarding the comparison between $r(t)$ from model I and model II in section 4.2 we found a large discrepancy. Here we discuss three possible reasons to this discrepancy.

Firstly, we discuss the observed discrepancy as a possible effect of the diffusion rate and the resolution limit. The limited resolution in the experimental movies results in fragments located closer than the resolution limit to be observed as a single fragment. A limited diffusion rate, in turn, limits the distance a fragment, on average, diffuses during a certain time. The combination between a non-zero resolution limit and a finite diffusion rate yields a delay between the time at which a cut happens and when we can observe it. It is thus reasonable that the effects on the observed cutting rate is small when this time delay is much shorter than the time between new cuts, and conversely large when the time delay is much longer than the time between new cuts. Since we showed that the cutting rate is linearly dependent on the nicking rate (Eq. (3.14)) this reasoning should also hold for the nicking rate. This partly motivates why the discrepancy between $r(t)$ for model I and model II increases for an increasing nicking rate.

Secondly, we discuss the observed discrepancy as a potential effect of the limited size of the synthetic images. Recall that we set the width of the synthetic images to 4 times the length of the initial molecule. This means that fragments can diffuse out of the image, as is the case for the fragments in the real images, and the number of observable fragments thus decrease. From this, we can conclude that there should exist a draining effect on the number of cuts due to a finite image size. To this we need to add that a finite image size, together with a non-zero resolution limit, results in a maximum number of fragments we can observe.

Lastly, we discuss the discrepancy as possible effect of disappearing cuts. Here, we note that fragments can visually merge with each other when they are located closer to each other than the resolution limit. This effectively results in a smaller number of observable cuts. It is reasonable to assume that the number of cuts that we miss due to this effect is proportional to the total number of cuts, thus the effect should be important for long times and high nicking rates as we observed in the comparison between $r(t)$ from model I and model II.

5.5 Model consistency

Let us first note that we have shown, in Appendix K, that $r(t)$ from model I is consistent with $r(t)$ obtained from the nicking simulation. This indicates that the assumptions made to derive $r(t)$ in model I hold true in the parameter regime we used in this study.

Regarding model II and its capability of recovering the ground truth nicking rate, we observe that it is capable to do this for the lower values of nicking rates and deviate slightly for the highest value of the nicking rate. Here, we need to point out that the variance in estimated nicking rate for the highest ground truth nicking rate is large since the curvature around the estimate nicking rate value in the negative log-likelihood function is small. This leads us to believe that estimating nicking rates for $\gtrapprox 0.004\ 1/\mathrm{s} \cdot \mathrm{bp}$ might be challenging for data similar to that used in this study. One reason for this might be the fact that the diffusive rate should set an upper bound for the maximum rate of new cuts that we can observe.

5.6 Estimation of nicking rate from experimental data

Let us here discuss the estimation of the nicking rate from the experimental movies. We begin by pointing out that the estimated nicking rate increases for increasing illumination strength. This is expected since we know that photosensitized driven DNA damage increases for increasing illumination strength. We also point out that the estimated nicking rate is only a part of the result presented in section 4.4. The more complete results are the distributions of the nicking rates given by the negative log-likelihood functions for the different illumination strengths. Furthermore, comparing the estimates of the nicking rate from the synthetic movies and experimental movies we observe that the estimated nicking rates are all found in the same order of magnitude. This consistency indicates that the synthetic movies should describe the experimental movies reasonably accurate. In addition, we observe that the negative log-likelihood function in Figure 8c shows no distinct minimum, which leads us to believe that the nicking rate for the experimental data obtained with 75 % illumination strength is situated around the upper limit of what we can estimate with the current method and data. To that end, larger amounts of experimental data would most likely allow for more accurate estimates of the nicking rate.

5.7 Model applicability

Here we will cover some aspects of the applicability of model I and model II. When looking at the applicability of model I compared to model II, we can deduce that model I is easier to use and apply for a given set of time series. To that end, there are several benefits with using model I: it is computationally fast, it is easy to implement, it offers the practical possibility of uncertainty estimation and it can be supported by a theoretical reasoning. On the other hand, these benefits are of little value if the model cannot recover the ground truth nicking rate correctly. From the results in this study this seems to be the case when the ground truth nicking rate grows large.

Regarding model II, we conclude that it improves upon some shortcomings of model I but creates new, practical, problems. Model II is computationally intense, and the time needed to

obtain a few instances (~ 10) of the cutting rate function for different values of the nicking rate is on the order of hours or days depending on the available computational resources. This makes model II unpractical for purposes of scanning through a large set of nicking rate values when there are no good prior guesses. On the other hand, already simulated cutting rate functions can be stored and reused for experiments with the same conditions.

6 Summary

In this study we have presented a new way to quantify photosensitized driven DNA damage from experiments of fragmenting fluorescently stained dsDNA in nanochannels. In detail, we have developed a new methodology based on both analytical and simulation-based models to measure the non-visible nicking rate from observable times of cuts in the data. We found, during validation of nicking rate estimations on synthetic data, that the simulation-based model (model II) outperformed the analytical model (model I) in terms of accuracy. In addition, we have successfully managed to estimate the nicking rate for experimental data. From the estimations, we observe that the nicking rate increases with increasing illumination strength.

7 Outlook

We hope that this work can come to use in future studies and provide a good base from which to improve the estimation of nicking rates from DNA in nanochannels. With that said, there are a number of possible extensions and improvements that can be made to the work presented in this study.

To begin with, we would like to estimate the nicking rate for arbitrary experimental conditions. Thus, we would determine the nicking rate as a function of, e.g., buffer strength, temperature, oxygen concentration and illumination strength. With that, we hope to obtain a better knowledge of the nicking rate and with theory explain the empirically obtained behaviors.

To enable accurate estimations of the nicking rate for arbitrary experimental conditions, there are several things which the current method needs to improve upon. Firstly, we see a need to improve the theory describing the probability of base-pair opening by including its dependency on the DNA sequence as well taking time scales into account. Furthermore, we would most likely need to improve the current method used to detect cuts with, in order to estimate nicking rates for high illumination strengths in combination with high levels of oxygen. Lastly, we may also want to investigate the implications of using a non-zero initial nicking density.

In case satisfactory improvements can be made to increase the accuracy of the nicking rate estimation, we would still require additional experiments in order to measure the nicking rate for arbitrary experimental conditions. The new experiments would have to assure that only one parameter was varied at a time. Additionally, it would be much beneficiary to acquire more data for each set of experimental conditions than we had access to in this study. This would allow for nicking rate estimates with lower variance, a necessity if it turns out that the variations within one of the experimental condition are small.

References

[1] R. Dahm and M. Banerjee, "How we forgot who discovered dna: Why it matters how you communicate your results.," *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 41, no. 4, p. e1900029, 2019.

[2] R. Dahm, "Discovering dna: Friedrich miescher and the early years of nucleic acid research," *Human genetics*, vol. 122, no. 6, pp. 565–581, 2008.

[3] J. D. Watson and F. H. C. Crick, "A structure for deoxyribose nucleic acid.," *Nature*, vol. 421, no. 6921, pp. 397 – 398, 2003.

[4] C. Bernstein, A. R. Prasad, V. Nfonsam, and H. Bernstein, "Dna damage, dna repair and cancer," in *New Research Directions in DNA Repair* (C. Chen, ed.), ch. 16, Rijeka: IntechOpen, 2013.

[5] Z. Yu, J. Chen, B. N. Ford, M. E. Brackley, and B. W. Glickman, "Human dna repair systems: an overview," *Environmental and molecular mutagenesis*, vol. 33, no. 1, pp. 3–20, 1999.

[6] N. Paillous and P. Vicendo, "Mechanisms of photosensitized dna cleavage," *Journal of Photochemistry and Photobiology B: Biology*, vol. 20, no. 2-3, pp. 203–209, 1993.

[7] D. Gupta, A. B. Bhandari, and K. D. Dorfman, "Evaluation of blob theory for the diffusion of dna in nanochannels," *Macromolecules*, vol. 51, no. 5, pp. 1748–1755, 2018.

[8] F. Persson and J. O. Tegenfeldt, "Dna in nanochannelsdirectly visualizing genomic information," *Chemical Society Reviews*, vol. 39, no. 3, p. 985, 2010.

[9] M. Zgarbová, M. Otyepka, J. Sponer, F. Lankas, and P. Jurecka, "Base pair fraying in molecular dynamics simulations of dna and rna," *Journal of chemical theory and computation*, vol. 10, no. 8, pp. 3177–3189, 2014.

[10] R. Cowan, C. M. Collis, and G. W. Grigg, "Breakage of double-stranded dna due to single-stranded nicking," *Journal of theoretical biology*, vol. 127, no. 2, pp. 229–245, 1987.

[11] M. A. Siddiqi and E. Bothe, "Single-and double-strand break formation in dna irradiated in aqueous solution: dependence on dose and oh radical scavenger concentration," *Radiation research*, vol. 112, no. 3, pp. 449–463, 1987.

[12] G. Van Der Schans, "Gamma-ray induced double-strand breaks in dna resulting from randomly-inflicted single-strand breaks: temporal local denaturation, a new radiation phenomenon?," *International Journal of Radiation Biology and Related Studies in Physics, Chemistry and Medicine*, vol. 33, no. 2, pp. 105–120, 1978.

[13] D. Freifelder and B. Trumbo, "Matching of single-strand breaks to form double-strand breaks in dna," *Biopolymers: Original Research on Biomolecules*, vol. 7, no. 5, pp. 681–693, 1969.

[14] J. Van Touw, J. Verberne, J. Retel, and H. Loman, "Radiation-induced strand breaks in ϕx174 replicative form dna: An improved experimental and theoretical approach," *International Journal of Radiation Biology and Related Studies in Physics, Chemistry and Medicine*, vol. 48, no. 4, pp. 567–578, 1985.

[15] R. Phillips, J. Kondev, J. Theriot, and H. Garcia, *Physical biology of the cell*. Garland Science, 2012.

[16] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of computational physics*, vol. 22, no. 4, pp. 403–434, 1976.

[17] R. Erban, J. Chapman, and P. Maini, "A practical guide to stochastic simulations of reaction-diffusion processes," *arXiv preprint arXiv:0704.1908*, 2007.

[18] D. T. Gillespie, "Stochastic simulation of chemical kinetics," *Annu. Rev. Phys. Chem.*, vol. 58, pp. 35–55, 2007.

[19] T. Ambjörnsson, L. Lizana, M. A. Lomholt, and R. J. Silbey, "Single-file dynamics with different diffusion constants," *The Journal of chemical physics*, vol. 129, no. 18, p. 11B612, 2008.

[20] D. Sivia and J. Skilling, *Data analysis: a Bayesian tutorial*. OUP Oxford, 2006.

[21] R. de Lennart and B. Westergren, *Mathematics handbook for science and engineering*. Springer, 5 ed., 2011.

[22] S. R. Casjens and R. W. Hendrix, "Bacteriophage lambda: early pioneer and still relevant," *Virology*, vol. 479, pp. 310–330, 2015.

[23] H. S. Rye, S. Yue, D. E. Wemmer, M. A. Quesada, R. P. Haugland, R. A. Mathies, and A. N. Glazer, "Stable fluorescent complexes of double-stranded dna with bis-intercalating asymmetric cyanine dyes: properties and applications," *Nucleic acids research*, vol. 20, no. 11, pp. 2803–2812, 1992.

[24] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Parson, 2008.

[25] J. Crank, *The mathematics of diffusion*. Oxford university press, 1979.

[26] M. Guéron, M. Kochoyan, and J.-L. Leroy, "A single mode of dna base-pair opening drives imino proton exchange," *Nature*, vol. 328, no. 6125, pp. 89–92, 1987.

[27] T. Ambjörnsson, S. K. Banik, O. Krichevsky, and R. Metzler, "Breathing dynamics in heteropolymer dna," *Biophysical journal*, vol. 92, no. 8, pp. 2674–2684, 2007.

[28] M. Peyrard, S. Cuesta-Lopez, and G. James, "Nonlinear analysis of the dynamics of dna breathing," *Journal of biological physics*, vol. 35, no. 1, p. 73, 2009.

[29] A. Krueger, E. Protozanova, and M. D. Frank-Kamenetskii, "Sequence-dependent basepair opening in dna double helix," *Biophysical Journal*, vol. 90, no. 9, pp. 3091–3099, 2006.

[30] D. Lide, *CRC Handbook of Chemistry and Physics, 84th Edition*. CRC HANDBOOK OF CHEMISTRY AND PHYSICS, Taylor & Francis, 2003.

[31] A. Vologodskii and M. D. Frank-Kamenetskii, "Dna melting and energetics of the double helix," *Physics of life reviews*, vol. 25, pp. 1–21, 2018.

[32] P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii, "Base-stacking and base-pairing contributions into thermal stability of the dna double helix," *Nucleic acids research*, vol. 34, no. 2, pp. 564–574, 2006.

[33] E. Protozanova, P. Yakovchuk, and M. D. Frank-Kamenetskii, "Stacked–unstacked equilibrium at the nick site of dna," *Journal of molecular biology*, vol. 342, no. 3, pp. 775–785, 2004.

[34] K. I. Mortensen, L. S. Churchman, J. A. Spudich, and H. Flyvbjerg, "Optimized localization analysis for single-molecule tracking and super-resolution microscopy," *Nature methods*, vol. 7, no. 5, p. 377, 2010.

[35] *Biomaterials science an introduction to materials in medicine*. Academic/Elsevier, 2013.

[36] W. Reisner, J. P. Beech, N. B. Larsen, H. Flyvbjerg, A. Kristensen, and J. O. Tegenfeldt, "Nanoconfinement-enhanced conformational response of single dna molecules to changes in ionic environment," *Physical review letters*, vol. 99, no. 5, p. 058302, 2007.

[37] V. Iarko, E. Werner, L. Nyberg, V. Müller, J. Fritzsche, T. Ambjörnsson, J. Beech, J. Tegenfeldt, K. Mehlig, F. Westerlund, *et al.*, "Extension of nanoconfined dna: Quantitative comparison between experiment and theory," *Physical review E*, vol. 92, no. 6, p. 062701, 2015.

[38] A. B. Bhandari, J. G. Reifenberger, H.-M. Chuang, H. Cao, and K. D. Dorfman, "Measuring the wall depletion length of nanoconfined dna," *The Journal of Chemical Physics*, vol. 149, no. 10, p. 104901, 2018.

[39] W. Reisner, J. N. Pedersen, and R. H. Austin, "Dna confinement in nanochannels: physics and biological applications," *Reports on Progress in Physics*, vol. 75, no. 10, p. 106601, 2012.

[40] R. Van Der Hofstad, F. Den Hollander, W. König, *et al.*, "Large deviations for the one-dimensional edwards model," *The Annals of Probability*, vol. 31, no. 4, 2003.

[41] E. Werner and B. Mehlig, "Confined polymers in the extended de gennes regime," *Physical review E*, vol. 90, no. 6, p. 062602, 2014.

[42] C. U. Murade, V. Subramaniam, C. Otto, and M. L. Bennink, "Force spectroscopy and fluorescence microscopy of dsdnayoyo-1 complexes: implications for the structure of dsdna in the overstretching region," *Nucleic Acids Research*, vol. 38, pp. 3423–3431, 01 2010.

[43] K. Gnther, M. Mertig, and R. Seidel, "Mechanical and structural properties of yoyo-1 complexed dna," *Nucleic Acids Research*, vol. 38, pp. 6526–6532, 05 2010.

[44] X. Bian, C. Kim, and G. E. Karniadakis, "111 years of brownian motion," *Soft Matter*, vol. 12, no. 30, pp. 6331–6346, 2016.

[45] M. Doi, S. F. Edwards, and S. F. Edwards, *The theory of polymer dynamics*, vol. 73. oxford university press, 1988.

[46] D. Sage, T.-A. Pham, H. Babcock, T. Lukes, T. Pengo, J. Chao, R. Velmurugan, A. Herbert, A. Agrawal, S. Colabrese, *et al.*, "Super-resolution fight club: assessment of 2d and 3d single-molecule localization microscopy software," *Nature methods*, vol. 16, no. 5, pp. 387–395, 2019.

Appendices

Appendix A Details of experimental setup

A.1 Properties of the illumination source

Here, we have plotted the relationship between applied voltage and luminosity of the lamp used in the experimental setup. The result can be seen in Figure 9.

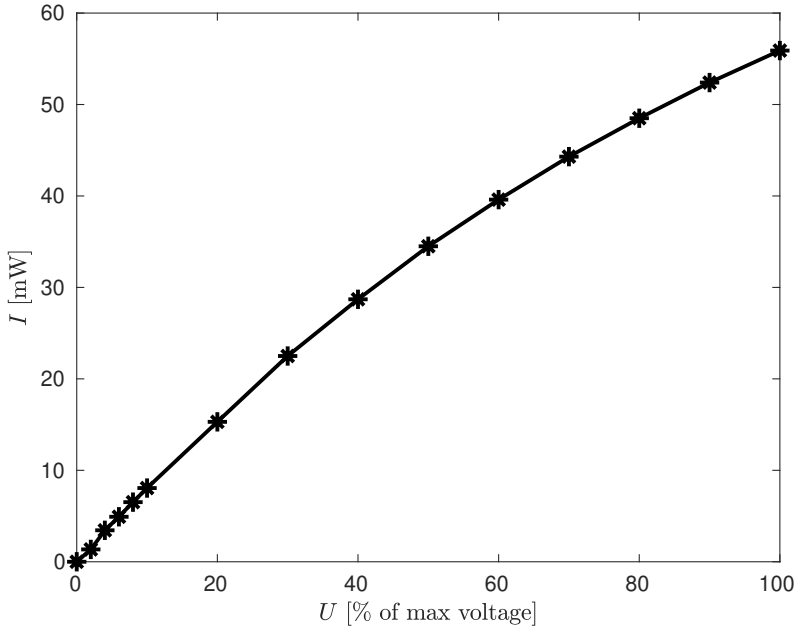


Figure 9: **Relationship between applied voltage and illumination strength.** Here, we have plotted the intensity of the lamp used to illuminate the sample as a function of applied voltage. We see how the intensity is increasing with increasing voltage throughout the voltage range. The trend seems to be non-linear and dI/dU appears to decrease with increasing voltage.

In Figure 9 we see how the illumination intensity response to variations in voltage is well-behaved and increases slightly slower than a linear response, especially for the higher values of the voltage.

A.2 Fabrication of nanofluidic systems

Here, we present, for reproducibility, the details of the procedure used to fabricate the nanofluidic systems. The compilation of the material here presented describing this procedure is done by Jason Beech.

The fabrication was performed by Joachim Fritzsche at Chalmers. Fabrication of the nanofluidic systems was carried out in cleanroom facilities of Fed. Std.209 E Class 10 100, using electron-beam lithography (JBX-9300FS / JEOL Ltd), optical lithography (MA 6 / Suss MicroTec), reactive-ion etching (Plasmalab 100 ICP180 / Oxford Plasma Technology), electron-beam evaporation (PVD 225 / Lesker), magnetron sputtering (MS150 / FHR), deep reactive-ion etching (STS ICP / STS) and wet oxidation (wet oxidation / Centrotherm), fusion bonding (AWF 12/65 / Lenton), and dicing (DAD3350 / Disco). In particular, the fabrication comprised the following processing steps of a 4"-silicon (p-type) wafer: Thermal oxidation: Wet oxidation in water atmosphere for 780 minutes (min) at 1050 °C (2 µm thermal oxide).

Reactive-ion etching of alignment marks: (a) Spin coating HMDS at 3000 rpm for 30 s and soft baking (HP) at 115 °C for 2 min. Spin coating S1813 (Shipley) at 3000 rpm for 30 s and soft baking (HP) at 115 °C for 2 min. (b) Expose alignment marks for 8 s in contact aligner at 6 mW/cm^2 intensity. (c) Development in MF-319 (Microposit) for 60 s, rinsing in water and drying under N_2-stream. (d) RIE for 20 min at 30 mTorr chamber pressure, 150 W RF-power, 50 sccm Ar-flow, 50 sccm CHF3-flow (800 nm etch depth in silicon). (f) Removal of resist in S1165 (Microposit) at 75 °C, rinsing in water and drying under N_2-stream. Reactive-ion etching of nanochannels: (a) Electron-beam evaporation of 20 nm Cr (hard mask). (b) Spin coating ZEP520A : anisole (2:1) (ZEONREX Electronic Chemicals) at 2000 rpm for 60 s and soft baking (HP) at 180 °C for 10 min. (c) Electron-beam exposure of lines (110 nm width, 20 µm pitch) at 2 nA with a shot pitch of 4 nm and 280 µC/cm^2 exposure dose. (d) Development in n-amyl acetate for 120 s, rinsing in isopropanol and drying under N_2-stream. (e) RIE for 10 s at 40 mTorr chamber pressure, 40 W RF-power, 40 sccm O_2-flow (descum). RIE for 90 s at 20 mTorr chamber pressure, 50 W RF-power, 200 W ICP-power, 20 sccm O_2-flow, 50 sccm Cl_2-flow (selective Cr hard-mask etch). RIE for 100 s at 8 mTorr chamber pressure, 50 W RF-power, 50 sccm NF_3-flow (30 nm etch depth in thermal oxide).

Reactive-ion etching of microchannels: (a) Spin coating HMDS at 3000 rpm for 30 s and soft baking (HP) at 115 °C for 2 min. Spin coating S1813 (Shipley) at 3000 rpm for 30 s and soft baking (HP) at 115 °C for 2 min. (b) Expose microchannels for 8 s in contact aligner at 6 mW/cm^2 intensity. (c) Development in MF-319 (Microposit) for 60 s, rinsing in water and drying under N_2-stream. (d) RIE for 30 min at 30 mTorr chamber pressure, 150 W RF-power, 50 sccm Ar-flow, 50 sccm CHF3-flow (1200 nm etch depth in silicon). (f) Removal of resist in S1165 (Microposit) at 75 °C, rinsing in water and drying under $N_2$2-stream.

Deep reactive-ion etching of inlets: (a) Magnetron-sputtering of 200 nm Al (hard mask). (b) Spin coating S1813 at 3000 rpm for 30 s and soft baking (HP) at 115 °C for 2 min. (c) Expose inlets for 10 s in contact aligner at 6 mW/cm^2 intensity. (d) Development in MF-319 for 60 s, rinsing in water and drying under N_2-stream. (e) Aluminum wet etch (H3PO4:CH3COOH:HNO3:H2O (4:4:1:1)) for 10 min to clear the hard mask at inlet positions. (f) Deep reactive-ion etching (SF6 / C4F8 based Bosch process) of inlets through the substrate. (g) Removal of Al-hard mask in aluminum wet etch (see above) for 60 min. Fusion bonding: (a) Cleaning of the substrate together with a lid (175 µm thick 4-pyrex, UniversityWafers) in H2O:H2O2:HCl (5:1:1) for 10 min at 80 °C, and in H2O:H2O2:NH3OH (5:1:1) for 10 min at 80 °C. (b) Pre-bonding the lid to the substrate by bringing surfaces together and manually applying pressure. (c) Fusion bonding of the lid to the substrate for 5 hours in N_2 atmosphere at 550 °C (5 °C/min ramp rate).

Dicing: Cutting nanofluidic chips of 25 µm x 25 µm size from the bonded wafer using a resin bonded diamond blade of 250 µm thickness (Dicing Blade Technology) at 35 krpm and 1 mm s^{-1} feed rate.

A.3 Preparation of DNA

Lambda phage DNA (dam-, dcm-, Thermo Fisher Scientific Inc, MA, USA) at 0.8 µM was stained with YOYO-1^{TM} (Life Technologies, Carlsbad, CA, USA) at a ratio of 1 dye molecule per 5 base pairs. TBE was added to final concentrations of between 0.02 and 5×TBE and beta mercapto

ethanol (BME) to a final concentration of 0.5 (BME is commonly used at concentrations up to 3 to suppress photodamage. Since we are interessted in observing photodamage we have reduced the amount of BME but retain 0.5 in order to have buffer conditions similar to those normally used for our experiments and to suppress some photo-bleaching).

A.4 Imaging system

All images were taken through an inverted Nikon Eclipse Ti microscope (Nikon Corporation, Tokyo, Japan) with a 100x oil immersion objective (CFI Apo TIRF, Nikon Corporation, Tokyo, Japan) and captured using an Andor Ixon back illuminated EM CCD camera (Andor Technology, Belfast, Northern Ireland). Films were acquired in epifluorescence using a Lumencor SOLA light engine$^{\text{TM}}$ (Lumencor Inc, OR, USA) and a FITC filtercube.

Appendix B Image segmentation

Here, we describe the procedure used to segment a gray-scale image into signal and background based on regional classification. In the classification challenge, we are given an image $\vec{I} = I(x, y)$ with x being the row and y the column. In addition to this information, we are in this study given the wavelength of the emitted light λ. The task is now to classify each pixel $I(x, y)$ for all x and y as either *signal* or *background*. To do this, we use the algorithm provided by Jens Krog which we now present for completeness. The algorithm is structured as followed:

1. Convolve $I(x, y)$ with a Laplacian-of-Gaussian filter, $LoG(x, y)$, to obtain $\tilde{I}(x, y) = I(x, y) * LoG(x, y)$ with
$$LoG(x, y) = \frac{x^2 + y^2 - 2\sigma^2}{\sigma^4} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{B.1}$$
where $\sigma = \sigma_{\text{PSF}}$ given by Eq (F.2).

2. Segment $\tilde{I}(x, y)$ according to
$$\tilde{I}_0(x, y) = \begin{cases} 1 & \text{if } \tilde{I}(x, y) \geq 0 \\ 0 & \text{if } \tilde{I}(x, y) < 0. \end{cases} \tag{B.2}$$

3. Trace all boundaries where $\tilde{I}_0(x, y) = 1$ borders to $\tilde{I}_0(x, y) = 0$ as to obtain closed contours.

4. For all closed contours compute a score h (for calculation of h see below).

5. Inspect the histogram of scores and set a threshold value to separate *false* contours from *real* contours.

With the above algorithm every closed contour now encloses a signal region, the objects are confined by their border pixels.

Let us now describe the calculation of the edge score value h. Given a set of boundary pixels belonging to a closed contour, the procedure goes as followed:

1. Compute the gradient direction for all pixels in $\tilde{I}(x, y)$ as [24]
$$\theta = \tan^{-1}\left(\frac{\tilde{I}_y}{\tilde{I}_x}\right), \tag{B.3}$$
with $\tilde{I}_x$ and $\tilde{I}_y$ being the partial derivatives of $\tilde{I}(x, y)$ with respect to x and y, respectively. In the computation of θ, keeping track of the quadrant we are located in is necessary to obtain the correct angle of direction.

2. Define a walking distance ω equal to the distance between the minimum value and the maximum value of a 1-d artificial binary edge convolved with a Gaussian and 1-d *LoG* filter using $\sigma = \sigma_{\text{PSF}}$ given by Eq. (F.2) in both convolutions. The artificial edge $E(x)$ is given by

$$E(x) = H(x) * G(x), \tag{B.4}$$

with $H(\cdot)$ being the Heaviside step function and $G(\cdot)$ the normalized Gaussian distribution. Compute the distance between the values of x that give the minimum and maximum value of

$$\Psi(x) = E(x) * LoG(x) \tag{B.5}$$

to obtain the walking distance as

$$\omega = |x_{\Psi_{\max}} - x_{\Psi_{\min}}|. \tag{B.6}$$

3. For each pixel in the closed contour walk a distance ω parallel to the gradient direction in both directions starting from the current pixel of interest. When walking in the negative gradient direction sum all visited pixel values of $\tilde{I}(x, y)$ to obtain $\tilde{h}_n$. When walking in the positive gradient direction sum all visited pixel values of $\tilde{I}(x, y)$ to obtain $\tilde{h}_p$. The score h_i for pixel i in the contour is given by $h_i = \tilde{h}_p - \tilde{h}_n$. The total score h for the contour is given by $h = \sum_i h_i$.

Appendix C Diffusion to capture

DNA damage is the process by which free radicals interact with the backbone of the DNA. This process should reasonably depend on the amount of free radicals in the surrounding solution of the DNA. It is also reasonable to assume that the amount of nicks that happen along the DNA under a certain amount of time is proportional to the concentration of free radicals in the surrounding solution. To formalize this, a computation will be performed which links these loose assumptions into a mathematical model.

Let us now formalize the problem and state the assumptions. We begin by assuming that the DNA can be seen as a cylindrical object of radius a and length L. Outside the cylinder forming the DNA, we have another, larger, cylinder of radius b and length L enclosing the smaller one. In the volume between the cylinders a buffer with certain concentration $c(r)$ of free radicals exist. Furthermore, we assume that the concentration of free radicals only depends on the radial coordinate and is held constant at the surface of the outer cylinder. Yet another assumption is that all free radicals that arrive at the surface of the DNA will react, i.e., the DNA is a perfect absorber. These assumptions are in line with the calculations done in chapter 13.3 in [15].

With all the assumptions in place we can now move on to the mathematical formulation of the problem. The overall motion of the free radicals in the buffer can be modeled with the steady state diffusion equation, favorably expressed in cylindrical coordinates with only r dependence, as [25]

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(r\frac{\mathrm{d}\,c(r)}{\mathrm{d}r}\right) = 0. \tag{C.1}$$

The general solution to Eq. (C.1) is obtained as

$$c(r) = A + B\ln(r). \tag{C.2}$$

Let us at this point invoke two boundary conditions. The assumption about the DNA being a perfect absorber implies that $c(a) = 0$. Furthermore, the assumption that the concentration at the outer radius b is held constant at some value implies that $c(b) = c_0$. With these two boundary conditions and some manipulations Eq. (C.2) can be written as

$$c(r) = \frac{c_0}{\ln(b/a)}\ln(r/a). \tag{C.3}$$

With the final expression of the concentration in place we can use Fick's first law

$$j(r) = -D\frac{\mathrm{d}c(r)}{\mathrm{d}r} \tag{C.4}$$

to calculate the flux, $j(r)$. Utilizing our expression of the concentration in Eq. (C.3), we find the flux as

$$j(r) = -D\frac{c_0}{\ln(b/a)}\frac{1}{r}. \tag{C.5}$$

With the flux at hand we can now compute the number of free radicals that arrive at the surface of the DNA per time unit using the fact that

$$\frac{\mathrm{d}n}{\mathrm{d}t} = -j(r)A. \tag{C.6}$$

Here A represents the area exposed to the buffer, which in our case is the area of the inner cylinder given by $A = 2\pi aL$. Furthermore, we realize at this point that the quantity $\frac{\mathrm{d}n}{\mathrm{d}t}$ is quite a special one and represents, in physical terms, the total number of nicks per second, a quantity we denote as κ. With Eq. (C.6) evaluated at $r = a$ and the area of the DNA surface, we can write down the number of free radicals arriving at the surface per time unit, or equivalently the total nicking rate, as

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \kappa = D\frac{2\pi Lc_0}{\ln(b/a)}. \tag{C.7}$$

If numerical values of the included parameters are given, we can use Eq. (C.7) to estimate the total nicking rate of the DNA analytically. We can make the interesting observation that the nicking rate and concentration of free radicals seem to be linearly dependent on each other in steady state.

Appendix D The fraying distance

Here, we will deal with the fraying distance, ξ, and attempt to estimate it. To begin with, we need to look closer at structure of the DNA. In very simple terms, the DNA consists of two strands with numerous base-pairs connecting the strands, much like a ladder, as can be seen in Figure 10. The structure of the DNA stops two single-stranded nicks from forming a double-stranded cut if

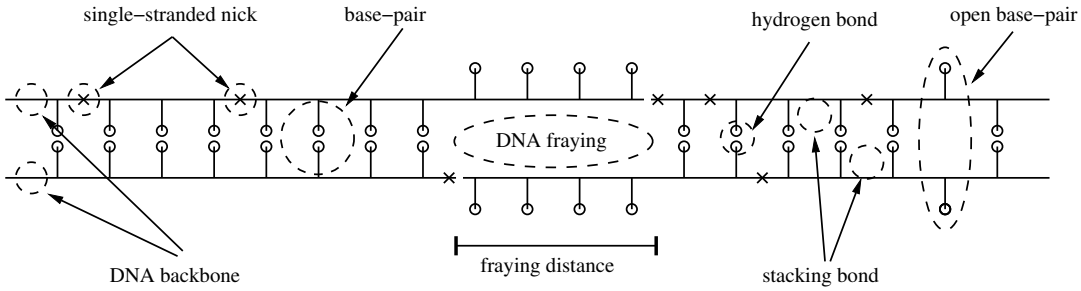


Figure 10: **A schematic illustration the dsDNA molecule and its constituent parts.**

they are not located between the same base-pairs, but only two a certain degree. Base-pairs can, through a stochastic process, *open up* which makes the structure deviate from the notion of a rigid ladder [26, 27, 28, 29]. The distance at which two single-stranded nicks can form a cut, $\xi/2$, is thus governed by the individual probabilities of base-pairs opening.

Let us start with the basics of base-pair opening. Here, we begin by assuming that the dye molecules (from the fluorecent staining) do not effect the base-pair opening process. We also assume that the energy needed to open a locked in base-pair is given by $\Delta G = 2\Delta G_{\mathrm{st}} + \Delta G_{\mathrm{hb}}$ [29]. Here, ΔG_{st} is the stacking energy and ΔG_{hb} the hydro-bond energy, respectively, which are marked in Figure 10. Since the DNA is nicked, we omit the ring factor as no internal bubble needs to be created. Assuming one base-pair is open, we only need to break a single stacking interaction and a hydrogen bond to open a second base-pair. This results in the total opening energy $\Delta\tilde{G} = \Delta G_{\mathrm{st}} + \Delta G_{\mathrm{hb}}$. According to [29] the probability to open up a base-pair associated with a certain energy cost, ΔE, is given by $P_{\mathrm{open}} = e^{\Delta E/RT}$, where ΔE in our case is either ΔG or $\Delta\tilde{G}$ and $R = 1.9872$ cal/(K mol) is the gas constant [30] and T the absolute temperature.

To estimate half the fraying distance we want to estimate the expectation value of the random number of consecutively open base-pairs. With P_{open} from above, we can write this down as

$$\langle m \rangle = \frac{e^{\frac{\Delta G_{\text{st}}}{RT}} \sum_{n=1}^{\infty} n e^{\frac{\Delta \tilde{G}}{RT} n}}{1 + e^{\frac{\Delta G_{\text{st}}}{RT}} \sum_{n=1}^{\infty} e^{\frac{\Delta \tilde{G}}{RT} n}}. \tag{D.1}$$

Let us set $k = e^{\frac{\Delta G_{\text{st}}}{RT}}$ and $r = e^{n \frac{\Delta \tilde{G}}{RT}}$, further assuming that $r, k < 1$. Some standard geometric series [21] turn the above expression into

$$\langle m \rangle = \frac{\frac{kr}{(1-r)^2}}{1 + \frac{kr}{1-r}} = \frac{kr}{(1-r)(1-r+kr)}. \tag{D.2}$$

We use this quantity to estimate the fraying distance as

$$\xi := 2 < m > +1 = \frac{2kr}{(1-r)(1-r+kr)} + 1. \tag{D.3}$$

Where we included the additional 1 to account for the spacing between one base-pair.

It is now time to calculate a numerical value for the quantities r and k, such that we can get a feeling for the magnitude of ξ. To do this we need values of ΔG_{st} and ΔG_{hb}. The temperature and salt dependent calculation for these quantities is done in Appendix E. For $T = 20$ °C and a sodium concentration of 0.1 molar we find that $r = 0.0298$ and $k = 0.0608$. Plugging these two values into (D.3) gives $\xi = 1.0038$ base-pairs. It is clear that the estimate of ξ is very close to one and that the effect of base-pair opening, under our assumptions, is small for experiments in room temperature with standard salt concentrations.

Appendix E Base-pair opening energy

Here, we will present the calculations of the energies needed to open up base-pairs, as a function of temperature and salt concentration, for both AT and GC base-pairs. The task of obtaining values for the stacking energy, ΔG^{st}, and the hydrogen bond energy, ΔG^{hb}, is not trivial. Measurements of the 16 different (10 unique) energies needed to open up an individual base-pair have been accomplished through a series of separated, non-chronological, experiments [31, 32, 33]. The challenge has been to separate the free energy required by breaking a bond between adjacent base-pairs and the complementary bases [33]. We will now present the calculations performed in [29], and in addition make an approximate for the total energy needed to open up any base-pair.

We begin by computing ΔG_{hb} for both AT and GC base-pairs as followed

$$\Delta G^{\text{hb}}_{\text{AT}} = \Delta G_{\text{AT}} - \frac{1}{4} \sum_{AT,TA,AA,TT} \Delta G^{\text{st}}_{KL}, \tag{E.1}$$

$$\Delta G^{\text{hb}}_{\text{GC}} = \Delta G_{\text{GC}} - \frac{1}{4} \sum_{GG,GC,CG,CC} \Delta G^{\text{st}}_{KL}. \tag{E.2}$$

Here, we have

$$\Delta G_{\text{KL}} = \Delta S (T^M_{\text{KL}} - T), \tag{E.3}$$

where KL is either AT or GC and $\Delta S = -24.85$ cal/mol K. T^M_{KL} is the melting temperature for the different bases and is given by

$$T^M_{\text{AT}} = 355.55 + 7.95 \ln\Big([\text{NA}^+]\Big) \tag{E.4}$$

for the AT base-pairs and

$$T^M_{\text{GC}} = 391.55 + 4.89 \ln\Big([\text{NA}^+]\Big) \tag{E.5}$$

for the GC base-pairs. The unit of T^M_{KL} is Kelvin. Furthermore, $[\text{NA}^+]$ is the concentration of sodium in the buffer.

Now, we continue by computing the stacking energies, $\Delta G^{\mathrm{st}}_{\mathrm{KL}}$. The stacking energies for $T = 37$ °C and [NA$^+$] mol is given in [29]. These stacking energies depend on temperature in the following way

$$G^{\mathrm{st}}_{KL,\mathrm{temp}} = G^{\mathrm{st}}_{KL} + 0.026(T - T_{\mathrm{ref}}) \tag{E.6}$$

with $T_{\mathrm{ref}} = 37\,^{\circ}\mathrm{C}$. They also depend on sodium concentration as

$$G^{\mathrm{st}}_{KL,\mathrm{salt}} = G^{\mathrm{st}}_{KL} - 0.2\ln\left(\frac{[\mathrm{NA}^+]}{C_{\mathrm{ref}}}\right) \tag{E.7}$$

with $C_{\mathrm{ref}} = 0.1$ molar.

At this point, we make the approximation that the only stacking energy we need is the average of all stacking energies

$$\Delta\tilde{G}_{\mathrm{st}} = \frac{1}{16}\sum_{KL}\Delta G^{\mathrm{st}}_{KL}. \tag{E.8}$$

With this, we can compute the total energy needed to open a single AT base-pair as

$$\Delta\tilde{G}^{\mathrm{1st}}_{\mathrm{AT}} = \Delta G^{\mathrm{hb}}_{\mathrm{AT}} + 2\Delta\tilde{G}_{\mathrm{st}} \tag{E.9}$$

and correspondingly for a GC base-pair we have

$$\Delta\tilde{G}^{\mathrm{1st}}_{\mathrm{GC}} = \Delta G^{\mathrm{hb}}_{\mathrm{GC}} + 2\Delta\tilde{G}_{\mathrm{st}}. \tag{E.10}$$

The energy to open a second base-pair, next to an already open one, we assume can be computed as

$$\Delta\tilde{G}^{\mathrm{2nd}}_{\mathrm{AT}} = \Delta G^{\mathrm{hb}}_{\mathrm{AT}} + \Delta\tilde{G}_{\mathrm{st}} \tag{E.11}$$

for a AT base-pair and

$$\Delta\tilde{G}^{\mathrm{2nd}}_{\mathrm{GC}} = \Delta G^{\mathrm{hb}}_{\mathrm{GC}} + \Delta\tilde{G}_{\mathrm{st}} \tag{E.12}$$

for a GC base-pair. Evidently the energies in Eq. (E.11) and Eq. (E.12) are also valid when we want to open the Nth base-pair, assuming the neighboring one is already open.

Appendix F Detailed simulation description – Model II

Here, we present the detailed procedure used to obtaining $r(t)$ for model II.

To simulate a single instance of $N_{\mathrm{cuts}}(t)$ for a specific set of parameters, we use the following method:

- Perform nicking simulation.
    1. Set $t_{\max}$ according to Eq. (F.1), $t = 0$ and initiate original molecule of length equal to the number of base-pairs.
    2. Generate a uniformly distributed random number $r \in [0, 1]$ and calculate the Gillespie time step given by $\tau = -\ln(r)/\sum_i \alpha_i$ with α_i being the nicking rate for each individual *nicking site* i which has not been nicked.
    3. If $t + \tau < t_{max}$ proceed to 4 otherwise go to step 10.
    4. Generate a weighted random integer F between 1 and the current number of fragments with weights according to the sum of the non nicked sites nicking rates on each fragment.
    5. Generate a random integer number N between 1 and the number of non nicked sites on the current fragment F weighted according to the nicking rate of each non nicked site.
    6. Place a nick on fragment F on the N:th non nicked position or split fragment in two if opposing nick exist within a distance $\xi/2$.
    7. If fragment been split update into two fragments accordingly.
    8. Save the current fragment constellation and time.
    9. Return to step 2.

10. Output all fragment lengths after all cut instances and the times of all cuts.

- Perform diffusion simulation.

1. Load output from the nicking simulation above.
2. Divide the time interval $[0, t_{\max}]$ into T intervals of length ΔT during which we want to create a *synthetic image*. Denote the time at the end of each interval, t_k, with $k \in [1, T]$, an image recording time.
3. Place the initial molecule at start positions = 0.
4. Compute the observed length of the molecule according to Eq. (G.1).
5. *Create a background image* four times longer than the initial molecule and 20 times higher than σ_{PSF}, also create a *signal image* of equal size as the background image with all pixel values being zero.
6. Set the stopping time equal to the time of the next cut: $t_{\text{stop}} = t_{\text{cut}}(count)$ if $count \leq count_{\max}$, otherwise go to step 16.
7. Place all non cutted fragments on same position as last simulation instance and replace the cutted fragment (unless on $count = 1$) with the two corresponding fragments next to each other conserving center of mass.
8. Compute the *hop-rates* for all fragments.
9. Generate a uniformly distributed random number $r \in [0, 1]$ and compute the waiting time $\tau = -\ln(r)/\mu$.
10. Sample one fragment to move with probabilities proportional to their corresponding *hop-rates*.
11. Sample with equal probability if fragment should be moved left or right.
12. Check if we can move the fragment due to spatial occupation of other fragments, if yes: save current positions as previous positions, move fragment and update time to $t = t + \tau$. If no: update time and save current position as previous position.
13. Check if we passed the current image recoding time.
    - If no: Go to step 14.
    - If yes:
      * *Add photon contribution* up until the image recording time to the signal image.
      * Perform *photon to readout procedure* on signal image.
      * Add background and signal image to obtain the final image, perform *segmention* on the final image after applying the *PSF filter*.
      * Count and save the number of detected molecules.
      * Do step 5 and add photon contribution from image recording time to current time to the new signal image.
      * Go to step 15.
14. *Add photon contribution* up until the current time.
15. Check if we passed a cut time, if yes: if $count < count_{\max}$ increment count by one and return to step 5, else go to step 16. if no: return to step 9.
16. Return number of detected fragments at all times an image was recorded.

To make the algorithm above complete we need to add some detailed information about certain steps:

- *nicking site*: one nicking site is the space between two base-pairs where a nick can happen. The total number of sites is given by $2(N - 1)$ where N is the number of base-pairs.

- $t_{\max}$: The time at which we stop the simulation. We calculate this time using Eq. (3.17) as

$$t_{\max} = \Delta T \cdot \text{floor}\left[\frac{\max\left(\left(\frac{\langle \bar{N}_{\text{cuts}} \rangle}{(\alpha^2 L \xi)}\right)^{1/2}, T_{\min}\right)}{\Delta T}\right] \tag{F.1}$$

  with $\langle \bar{N}_{\text{cuts}}) \rangle$ being the wanted number of cuts for the simulation and $T_{\min}$ the shortest required simulation time.

- *synthetic image*: a synthetic image created to mimic a real image of DNA in nanochannels used to count number of detectable fragments in.

- *create background image*: Generate a complete background image with image counts according to the procedure in Appendix M using $\lambda = \lambda_{\text{bg}}$.

- σ_{PSF}: The standard deviation of the theoretical point spread function [34]. We calculate it as

$$\sigma_{\text{PSF}} = \frac{1.22\lambda_{\text{YOYO1}}}{2L_p \text{NA}} \tag{F.2}$$

  with λ_{YOYO1} being the wavelength of the incoming light, L_p the pixel size and NA the numerical apparatus assuming that it can be approximated as Gaussian with the width equal to the Rayleigh resolution limit [35].

- *hop-rates*: The hop-rate for each fragment is given by [17]

$$\nu = \frac{D}{h^2} \tag{F.3}$$

  with D given by Eq. (H.3) and h being the pixel size.

- μ: The sum of all hop-rates, in both left and right direction is denoted as $\mu = 2\sum_i \nu_i$.

- *add photon contribution*: We here generate for each signal pixel a Poisson number with $\lambda = \lambda_{\text{sig}}(dt/\Delta T)$ with λ_{sig} being the mean number of photons that arrive in the acquisition time ΔT and dt being the time from the previous hop time to, either the current time, or the image time.

- *photon to readout procedure*: Here we apply the procedure described in Appendix M skipping the step of generating number of incoming photons according to the Poisson distribution and instead use the number of collected photons in the signal pixels as input value N_{ph} in Eq. (M.3).

- *segmentation*: Division of an image into signal and background regions, respectively. For details see Appendix B.

- *PSF filter*: Convolution of the image with the theoretical Gaussian point spread function [34] as filter.

The algorithm outputs the observed number of fragments as a function of time which obeys the following relation

$$N_{\text{frag}}(t_k) = N_{\text{cut}}(t_k) + 1, \tag{F.4}$$

from which we can obtain number of cuts at all times, $N_{\text{cut}}(t_k)$. Here, we convert $N_{\text{cut}}(t_k) = \max\left(N_{\text{cut}}(t_j)\right)$, with $j \in [1, k]$, to obtain the cumulative max number of fragments in time.

After simulating M instances of $N_{\text{cut}}(t_k)$, $N_{\text{cut}}^{(n)}(t_k)$, with $n \in [1, M]$, we obtain

$$\langle N_{\text{cut}} \rangle(t_k) = \frac{1}{M}\sum_{i=1}^{M} N_{\text{cut}}^{(i)}(t_k), \tag{F.5}$$

which is the model II version of Eq. (3.17).

Appendix G Extension of DNA in nanochannels

Here, we theoretically estimate the extension of DNA in nanochannels. A lot of effort has been devoted to this matter [36]. In the theoretical estimates, the observed length of a linear dsDNA molecule in a nanochannel is formulated as a function of the Kuhn length, l_K, the depletion width, δ, the channel dimensions and the effective DNA width w [37, 38]. The way this estimate is formed depends on the type of regime working with. Commonly, one distinguish between 4 regions: *classical Odijk*, *de Gennes*, *extended de Gennes* and *bulk phase* listed in decreasing level of confinement [39].

In this study we are primarily interested in estimating the DNA extension for the extended de Gennes regime due to the size of the nanochannels used in the experiments. For DNA molecules in nanochannels abiding the conditions of the extended de Gennes regime, an exact theory of the mean extension and variance exists [40]. Based on that theory, it is possible to write down the mean extension μ and the variance σ^2 with bounds on the errors as following [41]

$$\mu = 0.9338(84)\left(\frac{l_K w}{D_{\mathrm{w}} D_{\mathrm{h}}}\right)^{1/3} L, \tag{G.1}$$

$$\sigma = 0.364(17)\left(L l_K\right)^{1/2}. \tag{G.2}$$

Where the Kuhn length, l_K, is equal to twice the persistent length, D_{w} and D_{h} the width and height of the almost squared channel and L the contour length of the polymer. For numerical values of w and l_K for buffers with various ionic strengths we refer to [37]. In this study the experiments were performed using a buffer strength of 0.5×TBE. Furthermore, $D_{\mathrm{w}} = 100$ nm and $D_{\mathrm{h}} = 150$ nm.

The contour length of the DNA is not only dependent on the number of base-pairs in the polymer, but also on the amount of staining with fluorescent dyes. In this study we will make the same assumption as [37] and assume that each dye contribute with 0.44 nm to the contour length. Usual levels of staining ranges from 5 to 10 base-pairs per dye molecule. We denote this ratio of dyes per base-pair as ρ. Given a molecule with N_{bp} base-pairs, a length per base-pair of 0.34 nm and a staining ratio ρ, the contour length becomes

$$L_{\mathrm{cont}} = 0.34 N_{\mathrm{bp}} + 0.44 \rho N_{\mathrm{bp}} \text{ nm}. \tag{G.3}$$

We note that the affect of staining on l_K have not reached a consistent answer in the literature so far [42, 43].

For λ-phage DNA used in this study ($N_{\mathrm{bp}} = 48490$ bp and $\rho = 1/5$), we use Eq. (G.3) to estimate $L_{\mathrm{cont}} = 20754$ nm or ≈ 130 pixels (1 pixel = 160 nm). The mean extension μ, given by Eq. (G.1), can from this be compute to $\mu = 8269$ nm or ≈ 52 pixels. For the later computation we used, in addition to $L_{\mathrm{cont}} = 20754$ nm, the numerical values for a 0.5×TBE buffer from [37] for w and l_K as well as $D_w = 100$ nm and $D_h = 150$ nm.

Appendix H Diffusion of DNA in nanochannels

Diffusion is an extensively researched field [44] and is applicable to many physical processes in nature. Let us now turn to some basic formulations of diffusion. The simplest form of diffusion is Brownian motion in one dimension. The probability density function, $f(x,t)$, for a one dimensional particle can be written as [44]

$$\frac{\mathrm{d}f(x,t)}{\mathrm{d}t} = D\frac{\mathrm{d}^2 f(x,t)}{\mathrm{d}x^2}. \tag{H.1}$$

D is the diffusion coefficient for the particle and has the unit m^2/s. The solution to Eq. (H.1) for a particle initially located at x_0 is given by

$$f(x,t) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{(x-x_0)^2}{4Dt}}. \tag{H.2}$$

We see that the solution is a normal distribution with mean x_0 and variance $2Dt$. Equation (H.2) describes the time evolution of the probability density function for the Brownian particle.

We now move on to diffusion of DNA in nanochannels. In this study we will make the assumption that each base-pair contributes to the total drag according to the Rouse model [45]. This assumption together with Einsteins relation [44], $D = k_B T/\zeta$, with ζ being the friction coefficient, k_B Boltzmanns constant and T the absolute temperature, results in $D \propto L^{-1}$, with L being the contour length of the DNA. With $D \propto L^{-1}$, we can estimate the diffusion constant D as a function of an arbitrary contour length L according to

$$D = D_0 \frac{L_0}{L}, \tag{H.3}$$

where D_0 is some predetermined reference diffusion constant for the corresponding contour length L_0.

Here, we note that *blob theory*, in difference to the Rouse model, predicts that $D \propto \mu^{-1}$, with μ being the observed extension of the DNA [7]. In addition, [7] suggested that a model which combines Rouse diffusivity and blob theory fits their data better than any of them, individually. This is of interest since their experimental conditions are similar to those of this study. In any case, Eq. (H.3) lets us predict the diffusion constant of a DNA molecule with arbitrary contour length, if we know it, given a predetermined reference diffusion constant for blob theory as well as Rouse diffusivity. This holds true since Eq. (G.1) predicts that $\mu \propto L$.

Appendix I Comparison – synthetic and real images

Here, we show four images from two synthetic movies generated using the procedure described in section 3.3.3. We also show, for comparison, four images of two individual molecules from the experimental movie in Figure 2. We begin by presenting the synthetic images generated using the numerical values found in Table 1 as input for the simulations and $\alpha = 0.002$ 1/s$\cdot$bp, in Figure 11.

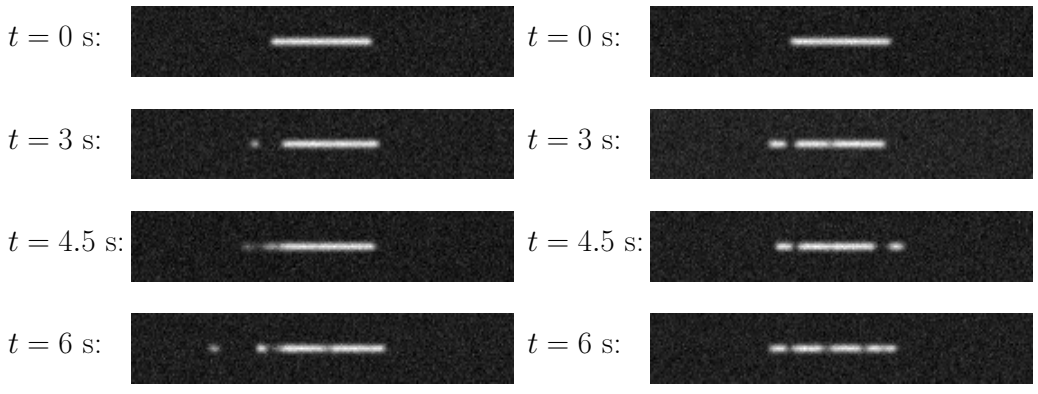


Figure 11: **Example images of two different molecules from two synthetic movies.** Here, we present two sets of four images taken at different times from two synthetic movies generated using the procedure in section 3.3.3. The numerical values used in the simulations of both image sets in Figure 11a and 11b, respectively, can be found in Table 1 with the addition of $\alpha = 0.002$ 1/s $\cdot$ bp. We can observe that the DNA molecule (seen as the bright region in contrast to the darker background regions) for $t = 0$ s (in both Figure 11a and 11b) is horizontally aligned and intact. As time progress, we observe that visible cuts start to appear and the molecules fragment into smaller fragments which diffuse in different directions.

We continue by presenting the real images in Figure 12.

Let us here comment on the images presented in Figure 11 and 12. To begin with, there are clear similarities between the synthetic and real images indicating that the synthetic movies have potential to mimic the experimental movies satisfactorily. Both the synthetic and real images contain an observable blurring of the edges around the fragments, which adds to an uncertainty in the number of actual fragments in each image. Furthermore, we can observe that the rate of

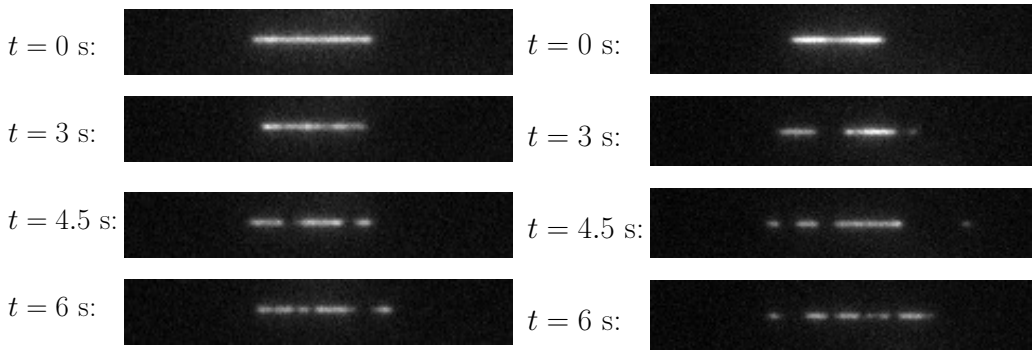
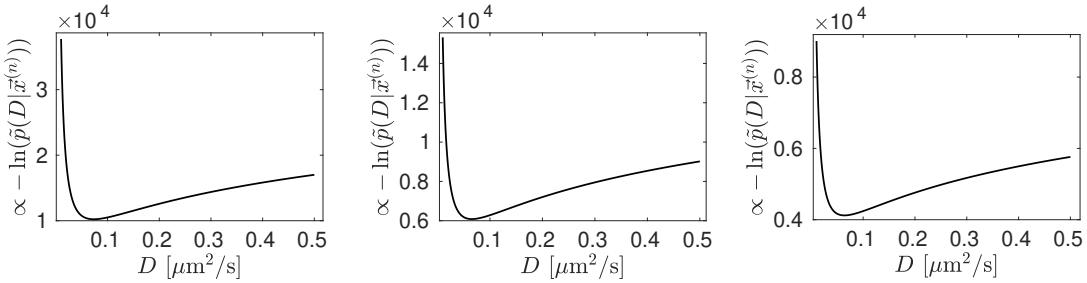


Figure 12: **Example images of two different molecules from one experimental movie.** Here, we present two sets of four images taken at different times of two different molecules from the movie in Figure 2. We can, as for the case in Figure 11, observe that the DNA molecule (seen as the bright region in contrast to the darker background regions) for $t = 0\,\mathrm{s}$ (in both Figure 12b and 12a) is horizontally aligned and intact. On the other hand, we can now observe a difference in length between the two initial molecules. Furthermore, we can also observe that the intensity varies along each of the molecules. As for the synthetic images, we can here observe, as time progress, that visible cuts start to appear and the molecules fragment into smaller fragments.

diffusion for the synthetic and real images seems to be on the same order of magnitude comparing the spread of fragments at the same time instances. On the other hand, there are also clear differences between the synthetic and real images. Firstly, we can observe that the DNA molecules in the real images differ in length at the first time instance, which is not the case in the synthetic images. Secondly, we can see that the intensity along the fragments in the real images, in difference to the synthetic ones, varies substantially. Lastly, we can also observe that the total intensity of the fragments in the real images seems to decrease with time, which we do not observe in the synthetic images.

Appendix J Additional material for estimation of diffusion constant

Here, we present three negative log-likelihood plots from the diffusion constant estimations obtained using three different values of dt. The plots can be seen in Figure 13.



(a) Here, $\Delta t = 2\,\mathrm{s}$ and $D = 0.073 \pm 0.002\,\mu\mathrm{m}^2/\mathrm{s}$.

(b) Here, $\Delta t = 5\,\mathrm{s}$ and $D = 0.065 \pm 0.003\,\mu\mathrm{m}^2/\mathrm{s}$

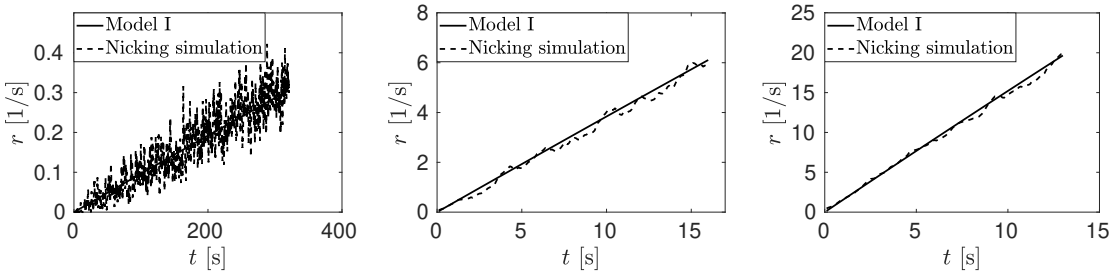
(c) Here, $\Delta t = 9\,\mathrm{s}$ and $D = 0.064 \pm 0.003\,\mu\mathrm{m}^2/\mathrm{s}$

Figure 13: **Three examples of negative log-likelihood plots from the diffusion constant estimation.** Here, we have plotted the negative log of the right-hand side of Eq. (3.24), which is proportional to $-\ln\big(\tilde{p}(D\,|\,\vec{x}^{(n)})\big)$, for three different values of Δt. We can in each of the plots observe that there exists a well behaved minimum and the variance around these minimum values are relatively small.

We can in Figure 13 observe that the negative log-likelihood functions are well behaved for all three cases of Δt.

Appendix K Comparison between $r(t)$ from model I and the nicking simulation

Here, we present how $r(t)$ from model I compares to the actual cutting rate obtained from the nicking simulation for three different nicking rates. To simulate $r(t)$ from the nicking simulation we used the numerical values of the needed parameters, in addition to the nicking rate, from Table 1. The results of the comparisons can be seen in Figure 14.



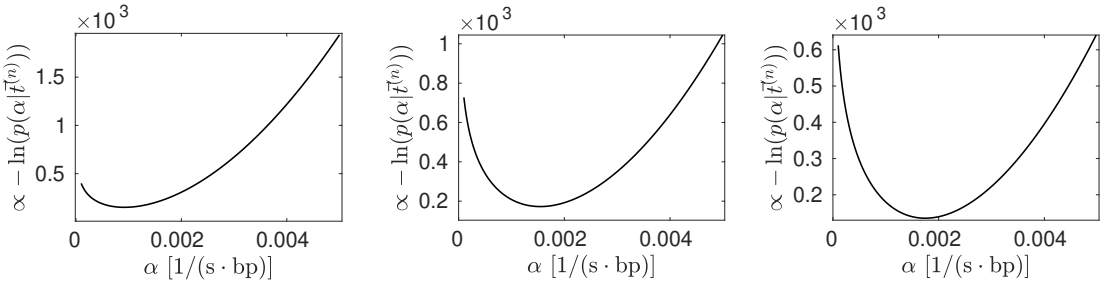
(a) $\alpha = 0.0001$ 1/s · bp. (b) $\alpha = 0.002$ 1/s · bp (c) $\alpha = 0.004$ 1/s · bp

Figure 14: **Comparisons between $r(t)$ from model I and the nicking simulation.** Here, we plot the analytical $r(t)$ from model I together with the simulated $r(t)$ obtained through the nicking simulation for three different values of the nicking rate. We observe that the two different models for $r(t)$ agree well for all three nicking rates. Furthermore, we see how the simulated version of $r(t)$ tends to fluctuate more for lower values of the nicking rate, which can be seen when comparing the result in Figure 14a with those in Figure 14b and 14c. We conclude that $r(t)$ from model I can accurately describe the cutting rate obtained from the single-stranded nicking simulation.

In Figure 14, we can see that $r(t)$ from model I and $r(t)$ obtained from the nicking simulation agree well for all three nicking rates. This indicates that the approximations we made when deriving $r(t)$ in model I are good for the parameters used in this study. Furthermore, we observe that the fluctuations in $r(t)$ obtained from the nicking simulations increase for decreasing values of the nicking rate.

Appendix L Nicking rate estimation on experimental data using model I

Here, we present estimations of the nicking rate for the three sets of blurred time series of cuts obtained from the experimental movies of different illumination strength used in section 4.4, using $r(t)$ from model I. The results can be seen in Figure 15.



(a) 25 % illumination strength. $\alpha_{est} = 0.0009 \pm 0.00005$ 1/s · bp

(b) 50 % illumination strength. $\alpha_{est} = 0.0015 \pm 0.00006$ 1/s · bp

(c) 75 % illumination strength. $\alpha_{est} = 0.0018 \pm 0.00009$ 1/s · bp

Figure 15: **Estimations of nicking rate on experimental movies using model I.** Here, we have plotted the negative log of the right-hand side of Eq. 3.23 using $r(t)$ from model I for experimentally obtained blurred time series of cuts. The blurred time series of cuts have been obtained with oxygen as driving gas for three different illumination strengths, 25, 50 and 75 % respectively. We observe that the increase in estimated nicking rate is not linearly proportional to the illumination strength.

We observe in Figure 15 that the estimated nicking rates increase for increasing illumination strength. Furthermore, we can conclude that the increase in the estimated nicking rate is not linearly proportional to the increase in illumination strength.

Appendix M Generating an artificial image

Here, we describe the process of generating a synthetic image according to the background distribution for a EMCCD camera. We follow the procedure in [46]. Note that, we describe the procedure used to generate one image count given that N_{ph} photons hit the pixel, a procedure that should be repeated for all individual pixels in the synthetic image.

To begin with, we assume that the number of *incoming photons* is

$$N_{\text{ph}} \sim P(\lambda), \tag{M.1}$$

i.e., N_{ph} is a random number with Poisson distribution of mean and variance λ (we use $\sim$ to denote that we draw a random number from the distribution that follows)

$$P(\lambda) = \frac{\lambda^i e^{-\lambda}}{i!}. \tag{M.2}$$

Where, i belongs to the natural numbers and $\lambda > 0$. Let us now assume that we are given λ and have generated one number N_{ph}. From this, the number of *incoming electrons* is given by

$$N_{\text{ie}} \sim P(\tilde{\lambda}) \tag{M.3}$$

where $\tilde{\lambda} = N_{\text{ph}}\text{QE} + c$, with QE being the quantum efficiency and c the clock-induced charge. The number of *outgoing electrons*, N_{oe}, obtained after the electron multiplication process is

$$N_{\text{oe}} \sim \Gamma(N_{\text{ie}}, EM_{\text{gain}}) + G(0, \sigma_R). \tag{M.4}$$

Here, $\Gamma(x, y)$ is the Gamma distribution, $G(0, y)$ is the normalized Gaussian distribution centered around zero, EM_{gain} the electron multiplying gain factor and σ_R the read noise of the EMCCD camera. The final output value in terms of pixel counts is

$$N_{\text{out}} = \min\left(\text{floor}\left(\frac{N_{\text{oe}}}{e_{\text{ADU}}}\right) + N_0, 65535\right), \tag{M.5}$$

where e_{ADU} is the electrons per analog-to-digital unit, N_0 the baseline offset and 65535 the largest possible image count.

In order to create the complete synthetic image we now need to generated equally many instances of N_{out} as there are pixels in the image. Furthermore, in this study we used the numerical values in Table 2 to generate the synthetic images.

Table 2: Numerical values used in the production of artificial EMCCD image counts.

Parameter	Value
QE	1
c	0.002
EM_{gain}	245
σ_R	74
e_{ADU}	45
N_0	100