Characterising DNA-protein interaction through the automated analysis of AFM images

Mingxue Du

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

The University of Sheffield

Faculty of Engineering

School of Chemical, Materials and Biological Engineering

December 2024

Abstract

Software for automatic image analysis has been widely applied to the characterisation of biomolecules as it allows for faster data processing and reduces bias. However, many tools developed for other bio-imaging fields are not directly compatible with atomic force microscopy (AFM), which is a high-resolution characterisation technique that enables the direct visualisation of biomolecules in a physiological environment. AFM presents unique challenges to the implementation of automatic image analysis tools because it stores data in special formats and is affected by artefacts that find no equivalence in other microscopy techniques.

While automatic analysis methods tailored for AFM images have been developed, most existing programs are project-specific and unsuitable for different sample types, while other programs see limited usage due to their cost or unsuitability for batch processing.

To address these issues and contribute towards the establishment of a versatile and accessible AFM bio-image analysis platform, this project improves existing automatic analysis methods and develops more methods to characterise the conformation of DNA and proteins. Since DNA is a flexible molecule that exhibits both intrinsic conformational variation and conformational changes induced by protein interaction, these characterisation methods will allow us to better understand how DNA functions and how its shape is altered by proteins.

Specifically, this thesis analyses the conformation of the DNA-manipulating topoisomerase TOP1, investigates the interaction between DNA and dumbbell-shaped autophagy receptor NDP52, and probes the effects of the nucleoid-associated protein HU on DNA bending angle.

The work in this thesis is built on open-source software and forms part of a general AFM image processing workflow. The methods developed can therefore be easily adapted for other projects that study the conformation of different biomolecules.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would first like to thank my primary supervisor Dr Alice L.B. Pyne, who has constantly offered support and guidance, even during trying times. I would also like to express my gratitude towards other members of my supervisory team, Dr Collin L. Freeman and Dr Robert Turner, who have provided invaluable feedback during different stages of my PhD.

I am grateful to have been part of a supportive research group; it has been a pleasure to work with every single member of the team, and I would like to offer special thanks to Daniel E. Rollins, who continuously provided experimental data, so that the characterisation methods I developed could be tested and applied.

Je dois beaucoup à mon amie Milena, qui m’a accompagnée et m’a aidée à continuer, surtout quand je manquais de motivation.

Non avrei potuto finire la tesi senza l’aiuto del mio amore Alberto. Ti ringrazio infinitamente. Mi hai dato il coraggio e il potere non solo di finire il mio lavoro ma anche di vedere un mondo più grande.

最后，我还想感谢四年间与我一直分处亚欧大陆两端的父母，给予我物质上的支持与精神上的鼓励。

Declaration

I, Mingxue Du, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means. This work has not been previously been presented for an award at this, or any other, university.

The following publications have arisen from the work in this thesis:

Á. dos Santos *et al.*, "Autophagy receptor NDP52 alters DNA conformation to modulate RNA polymerase II transcription," *Nat. Commun.*, vol. 14, no. 1, pp. 1–24, May 2023, doi: 10.1038/s41467-023-38572-9

E. W. Chan *et al.,* "The mechanism of action of the bacterial protein HU for DNA repair," in preparation.

Chapter 1 Introduction

1.1 Automatic image analysis as a tool in bio-imaging

1.1.1 Introduction to bio-imaging

Bio-imaging is a term that covers the visualisation of biological samples via microscopy, to determine their structure and link this to their function. Optical microscopy [1] was the first technique employed to observe biological materials in detail. Its resolution has continuously improved, allowing for the visualisation of finer structures through the invention of new imaging modalities, such as phase-contrast microscopy [2], which increases the contrast of transparent samples and makes them more visible.

Fluorescence microscopy is a sub-field of optical microscopy that utilises fluorescence emission to visualise tissues, cells and even single molecules. By enabling targeted staining, this technique provides information on specific features of interest, such as biomarkers of diseases. Super-resolution methodologies can be used to overcome the resolution limit (about 200 nm [3]) caused by the diffraction of optical waves, achieving a resolution of 20 nm [4]. The main drawback of this imaging method is the necessity of fluorescence labelling for samples that are not naturally fluorescent, as this introduces complexity in sample preparation and can negatively impact cellular functions [5].

Electron microscopy enables higher resolution due to the smaller wavelength of electrons as compared to visible light [6]. This technique can be broadly divided into two categories [7]: in transmission electron microscopy (TEM), samples are cut into thin slices that electrons can pass through, while in scanning electron microscopy (SEM), electrons interact with the surface of the sample. Both TEM and SEM operate in a vacuum, presenting challenges when characterising the structure of biological samples. Compared to metallic samples that are unaffected by the vacuum, biological materials require extra sample preparation steps such as dehydration and staining, which can alter the conformation of biomolecules.

This problem can be partially alleviated by cryo-electron microscopy, where samples are rapidly vitrified and maintained under cryogenic temperature [8]. By averaging over data from multiple images, near-atomic-resolution 3D structures can be reconstructed. Rare conformations or events may however be missed due to this averaging process.

Scanning probe microscopy (SPM) collects data through physical interaction between the sample and a probe; it eliminates the need for optical or electron waves, but may introduce a physical force to the sample [9]. One sub-category of SPM is scanning tunnelling microscopy [10], which measures the electric current as the probe scans the sample surface, and is only viable on conductive materials. Another sub-category is atomic force microscopy (AFM) [11], which does not rely on electric conductivity, and instead measures the force between the sample and the tip of the probe, building up a topographic map of the sample surface line by line, with its resolution in modern systems limited by the radius of the tip. AFM can characterise the structure of biomolecules with nanometre resolution, and it will be further discussed in Section 1.2.

From 17th century optical microscopes to modern electron microscopy, the field of bio-imaging has seen significant improvements in resolution, speed and the ability to capture live processes. These rapid advances in bio-imaging have led to new challenges: as biological structures are visualised at a greater speed with increasingly higher resolution, the large volume of bio-images generated has become impractical to analyse manually. In response, automatic image analysis tools have been developed and applied in all bio-imaging fields to perform structural determination across biological scales.

1.1.2 Current state of automatic bio-image analysis

Due to the large quantity of bio-images generated and the lack of objective standards on what should be measured, traditional methods of manually extracting usable data from these images are time-consuming and prone to selection bias; these issues can be addressed by the development of automatic image analysis programs, which enable a large amount of data to be processed within a short time; furthermore, these programs can produce quantitative data from qualitative information while minimising human error.

The automatic bio-image analysis workflow typically consists of three stages: the first stage, preliminary processing, prepares an image for further analysis principally through noise removal, though other modifications such as contrast enhancement [12] and image expansion [13] may also be performed depending on the nature of the dataset; the second stage recognises features of interest by segmenting the image into different regions [14], or by dividing the image into foreground and background [15]; the third stage involves the automatic measurement of parameters such as shape, size, colour and intensity, which allows features of interest to be quantitatively characterised.

The wide range of tasks carried out in the bio-image analysis workflow are synthesised through software packages such as Fiji [16], which is based on the platform ImageJ [17] and incorporates a variety of processing functions including image smoothening, binary masking and edge finding. These functions are achieved via established algorithms: for example, if a greyscale image is treated as a height map where each point represents a height value, the rolling ball algorithm [18] can be used to determine the threshold for background noise removal by calculating the position of the highest point on an imaginary ball that fits onto the image from below; the watershed algorithm [14] can be subsequently used to segment the image by simulating water filling from lower points on the height map and identifying the lines over which neighbouring bodies of water join each other, thus ensuring that different features of interest are separated and can be measured independently.

Recently, these methods have seen increased competition from the development and employment of more advanced algorithms and software based on deep learning and neural networks [19]–[22]. These advanced algorithms are powerful as they can be applied to new datasets and projects with minimal user input; however, the training of neural networks requires a large amount of data and computational resources; in addition, some algorithms rely on human supervision and manual labelling. Consequently, deep learning and neural networks are best suited for image analysis problems common to multiple modes of imaging, and cannot completely substitute conventional bio-image analysis methods.

1.1.3 Challenges in automatic bio-image analysis

Progress in automatic image analysis has brought about increasingly powerful tools to characterise biological samples, but a series of challenges need to be first addressed, in order that these tools can be adopted by more researchers and adapted for more projects. To start with, the development of advanced analysis methods, especially those based on neural networks, depends on a large and varied training dataset; this calls for further collaboration between biologists and software developers, so that data-sharing platforms can be created and best practises in data collection can be defined and implemented [23]–[25]. In response to this need, the criteria of FAIR (findability, accessibility, interoperability and reusability) data [26] have been proposed, which enable researchers and developers to find and use bio-imaging datasets, and allow software suited for the analysis of these datasets to be developed.

Another challenge lies in the accessibility of newly developed bio-image analysis methods. As an example, the widely used image processing package Fiji focuses on traditional algorithms, and has not yet incorporated deep-learning-based programs [27]. This means that researchers cannot always find suitable software to employ and modify, and that resources may be wasted on re-developing existing methods.

In cases where an established bio-image analysis program compatible with the needs of a project has been found, additional problems are posed by the selection and application of appropriate metrics to assess the validity of image processing and characterisation methods [28]. Existing software needs to be examined and adapted based on the properties of the sample and the image type when applied to a new project, so as to ensure accurate and objective results [25].

Many approaches to address challenges in automatic bio-image analysis are often tackled from the viewpoint of cells and optical microscopy [24]. This prevents the latest development from being easily accessed in all imaging fields; for example, established automatic image acquisition and analysis methods in fluorescent microscopy based on cells are not always applicable to the characterisation of single molecules due to differences in resolution and sources of errors [29], [30].

These challenges show that there is still vast room for improvement in the field of automatic bio-image analysis, especially in the form of constructing a general platform where high-quality data from different imaging techniques can be shared alongside a selection of well-maintained software tested on different sample types and imaging modes.

1.2 Automatic bio-image analysis in atomic force microscopy

1.2.1 AFM as a tool for bio-imaging

Atomic force microscopy (AFM) is an established technique for the characterisation of biomolecules that combines nanometre lateral resolution on individual molecules with operability in solution. Unlike optical or electron microscopy, the maximum resolution of AFM is not set by a diffraction limit, because AFM does not rely on waves; instead, it measures properties such as height using a tip attached to a cantilever that scans across the surface of the sample line by line, producing a topographic map based on the deflection of the cantilever caused by the tip-sample interaction at each point [11], [31]. At the time of invention, the resolution of AFM was less than 1 Å vertically and 30 Å laterally [31]; the latter has since been improved, with atomic resolution achieved on minerals [32] and sub-nanometre resolution reached for biological molecules [33].

As illustrated below in Figure 1.1, AFM imaging can be classified into different modes based on how the tip interacts with the sample [34], [35]. The first developed mode is contact mode, where the tip is in constant contact with the sample surface. While it is straightforward and suitable for solid materials such as metals, the lateral force between the tip and the sample can damage softer biological materials. To address this limit, tapping mode was subsequently invented, where the tip oscillates and is in intermittent contact with the sample surface, reducing the lateral force. This mode causes less damage and deformation to biological samples.

The AFM manufacturer Bruker developed PeakForce Tapping mode [35], [36], which has been employed for the data collection in this thesis (see Section 2.1 for details). Compared to the conventional tapping mode, PeakForce Tapping mode allows the tip to oscillate at a lower frequency with a larger amplitude. This results in slower imaging than tapping mode, but enables better control of the maximum force applied perpendicular to the sample surface (hence the name PeakForce). When operated in solution, PeakForce Tapping enables the conformation of biomolecules to be preserved under physiological conditions. As such, it is especially suited for the imaging of biomolecules with complex conformational variations and high levels of flexibility, which can be difficult to characterise through other techniques.

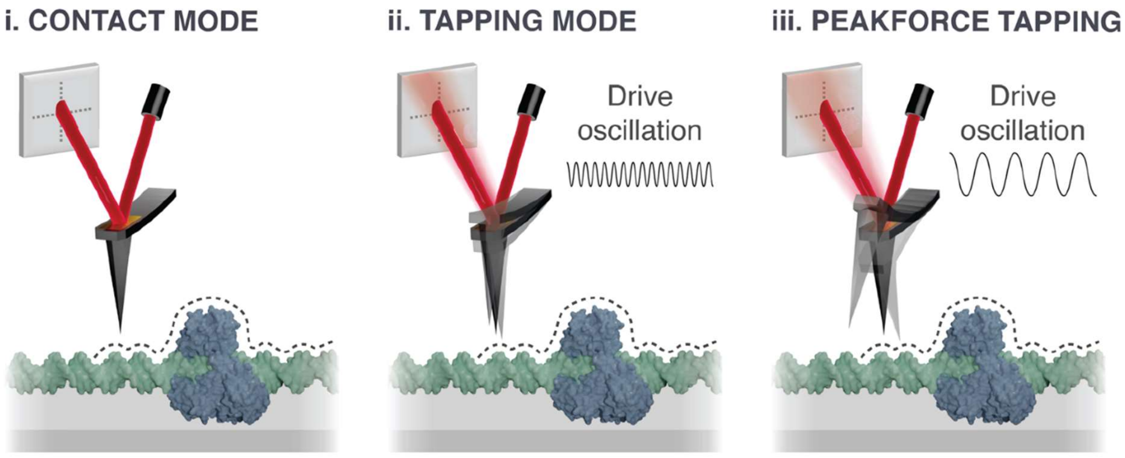


Figure 1.1 Schematic illustration comparing different AFM imaging modes. In contact mode (i), the tip is in constant contact with the sample; in tapping mode (ii), the tip is in intermittent contact with the sample and oscillates at a high frequency; in PeakForce Tapping mode (iii), the tip oscillates with a higher amplitude at a lower frequency, allowing the maximum normal force between the tip and the sample to be controlled, so that damage to soft biological samples can be reduced. Figure obtained from Main et al. [35]

In addition to the measurement of surface topography, AFM is also capable of characterising the mechanical properties of biomolecules. For example, PeakForce Quantitative Nanomechanics (PeakForce QNM) [36], a variation of PeakForce Tapping mode, constantly records and analyses force curves as the tip scans across the sample, thereby calculating the Young's Modulus and other mechanical properties at each point of the surface.

Although AFM provides unique information on the conformation of biomolecules, it also comes with its own drawbacks and limits. While unaffected by diffraction, the lateral resolution of AFM images is limited by the size of the tip [37]. If the radius of the tip is comparable to or larger than the features measured, the features will appear larger than their actual size [37], [38]. This is known as tip convolution and can be quantified using the models illustrated below in Figure 1.2. Its effect can be reduced by choosing a smaller tip, but commercially available AFM tips compatible with biological samples typically have a radius of several nanometres [39], which is not always small enough for features within single biomolecules. For example, the Bruker FastScan D [40], one of the AFM probes used by collaborators to collect experimental data for this thesis, advertises a nominal tip radius of 5 nm, which is insufficient to achieve double-helical resolution on DNA, since the helix diameter for B-form DNA is 2 nm [41]. It is possible to find better alternatives, such as the Bruker PeakForce-HIRS-F-B probe [42], which is specially designed for the characterisation of single biomolecules and has a nominal tip radius of 1 nm; however, nominally identical AFM probes vary in actual tip size, and the tip radius for PeakForce-HIRS-F-B can reach 2 nm, which still results in non-negligible tip convolution when imaging smaller features on biomolecules.

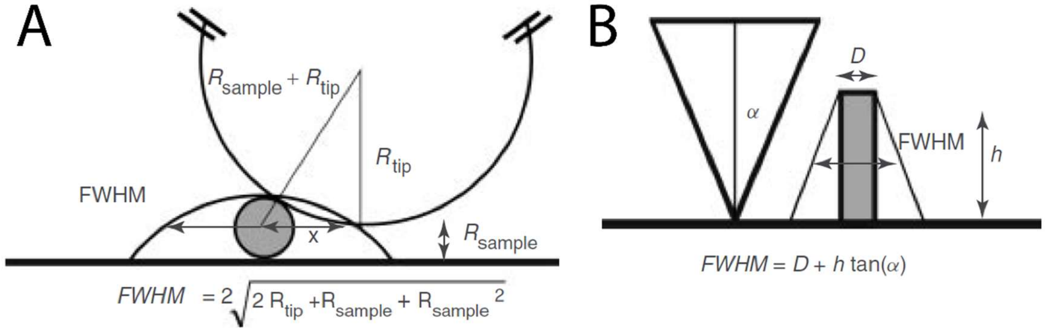

A
$R_{sample} + R_{tip}$
R_{tip}
FWHM
x
R_{sample}
$FWHM = 2\sqrt{2\,R_{tip} + R_{sample} + R_{sample}^{2}}$
B
α
D
FWHM
h
$FWHM = D + h\tan(\alpha)$

Figure 1.2 Quantifying the effect of tip convolution in AFM. The measured full width at half maximum (FWHM) is larger than the actual size of the feature if the tip is not small enough. The relation between the tip size, feature size and the measured FWHM can be approximated using the globular tip model (A) or the conical tip model (B). Figure obtained from Moreno-Herrero and Gomez-Herrero [37].

Other issues affecting AFM image quality include noise, tip contamination or damage, thermal drift, and unsuitable feedback parameters [38], [43]. The small size of the tip and cantilever means that they are particularly sensitive to environmental changes: the vibration caused by acoustic noise can interfere with the scanning of the sample, while temperature fluctuations can affect the mechanical properties of the cantilever [43]. AFM imaging also requires the user to manually select experimental parameters such as the gain applied to the feedback loop, which ensures that the tip closely follows the sample surface [44]. The value of the gain is constantly adjusted during each imaging session; if it is too high, the tip will move excessively and the image will contain too much noise; if it is too low, smaller features of interest will be ignored and the sample will not be accurately traced [43]. Errors are also introduced due to the line-by-line nature of AFM imaging [45]: the base height value can vary between neighbouring scan lines; in addition, defects referred to as scars may appear in the middle of single scan lines [46], [47], as shown below in Figure 1.3.

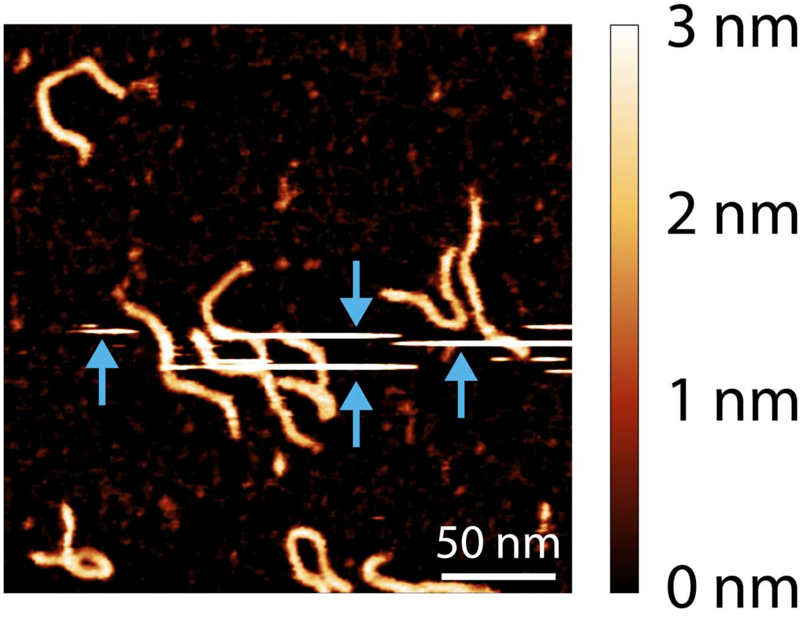


Figure 1.3 Example AFM image of short linear DNA molecules with scan line defects (also known as scars), indicated by blue arrows.

In brief, non-ideal environmental, instrumental, and user operational conditions can result in features on the image that do not represent the topography of the sample. These undesirable features, known as artefacts, can be identified and removed using specialised processing software [47]–[49]. This is essential to the further analysis of AFM data.

While artefacts can occur when imaging all types of samples, the unique nature of biological molecules in liquids presents additional challenges to AFM characterisation. For example, the solvent used can be subject to evaporation, thus changing the concentration of the solutes and potentially affecting the properties of the sample [37]. Since AFM relies on the interaction between the tip and the sample surface, the presence of liquid in between will also interfere with the oscillation of the tip, making imaging parameters difficult to control and feedback signals from the sample difficult to interpret [50]. Researchers have also pointed out that the slow scanning speed required for obtaining high-resolution data is not sufficient to capture certain biological processes that take place within the time scale of a few seconds [51], [52]. In summary, while AFM is a useful technique for the characterisation of conformational differences between different biomolecules and within the same molecule, many barriers need to be overcome to obtain accurate results for downstream data analysis. This can be achieved through the combination of optimised experimental design and automatic image processing.

1.2.2 Current state of automatic AFM image analysis

AFM enables us to visualise the structure of individual biomolecules and characterise their interactions. For this technique to reach its full potential, it is essential to employ automatic analysis methods, so that image artefacts can be corrected, and that DNA conformation can be measured quantitatively and efficiently.

Despite the constant emergence of new bio-image analysis tools, performing automatic analysis is not always straightforward or accessible in the field of AFM. This is because popular existing programs such as Fiji, while user-friendly and versatile, are not directly compatible with AFM images, due to a combination of several factors.

To start with, AFM data is stored in files of special formats that contain metadata on imaging parameters in addition to multiple channels, each corresponding to one set of surface measurements including height values and mechanical properties [53]. These are not standard image files, and are often unreadable by established image analysis software.

Moreover, current development and discussion of advanced automatic bio-image analysis methods primarily focus on optical microscopy and cells [17], [24], [25]. These methods are not optimised for AFM images at a single-molecular or higher resolution. In addition, optical microscopy images provide information on the colour of the sample, which can be used for tasks such as cell segmentation; by comparison, AFM images are greyscale, and features of interest can only be recognised based on size, height and shape. Although advanced analysis programs have been developed for other types of high-resolution and greyscale bio-imaging, such as transmission electron microscopy [20], they still fail to address AFM-specific needs for preliminary processing, including the removal of scan line differences and the handling of tip convolution.

The difficulty in employing deep-learning-based automatic image analysis methods for AFM images also lies in the lack of sufficient training data: hundreds and thousands of images can be produced in a single session of optical microscopy [25], but AFM imaging is much slower, and it is typical to generate only dozens of AFM images suitable for analysis following a week of experimental work.

Because of these barriers in adapting general bio-image analysis tools for AFM data, specialised programs for AFM image analysis have been developed, but they are often tailored for specific projects with limited wider applicability [54]–[57]. While there are a few software packages designed for general usage, they all have drawbacks: MountainsSPIP [48] is a costly commercial product; Gwyddion [58] is free but not suitable for batch processing; the recently developed NanoLocz [53] has limited functionality for linear molecules that follow a specific trajectory such as DNA.

To conclude, progress in automatic analysis for AFM data lags behind other bio-imaging fields. As part of a wider effort to bridge this gap and construct an accessible, versatile platform for automatic AFM image analysis, this thesis focuses on the development and application of automatic AFM image analysis methods, in the context of studying DNA-protein interaction. It is however worth noting that, while methods developed for general bio-image analysis are not always applicable to AFM data, general principles and standards [59] such as the collection of FAIR data are still highly relevant to the field of AFM.

1.2.3 Available methods for automatic AFM image analysis

Similar to other modes of bio-imaging, the procedure for automated image analysis in AFM varies between different projects and is dependent on the nature of the sample, but nearly all currently available methods can be grouped into three categories: preliminary processing, molecular identification, and measurements. In this section, automatic AFM data analysis algorithms from literature are described and discussed, since they can be combined to construct a single platform that contains functionalities suitable for various sample types.

AFM imaging can be performed on samples ranging from cell monolayers to single molecules, with varying parameters and methodologies needed at each length scale being probed. The focus of this thesis is the characterisation of DNA and protein using automated AFM analysis pipelines as initially developed by Beton *et al.*, which are shown in Figure 1.4 below. As such, this section mainly discusses analytical methods aimed at single biomolecules, although for future projects, the preliminary processing and molecular identification stages are readily adaptable to cover other sample types such as cells.

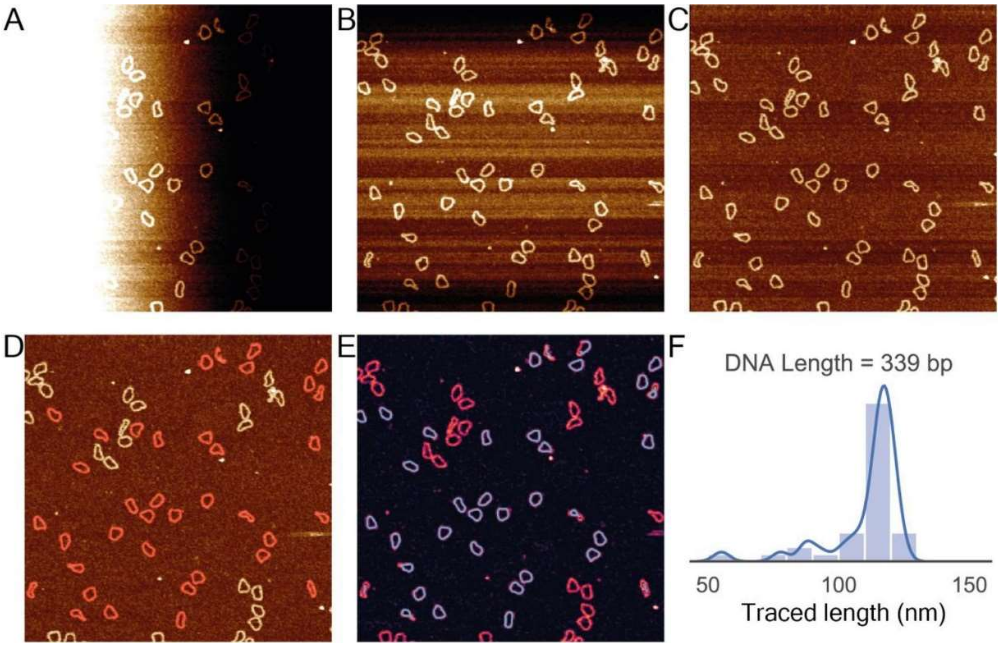


Figure 1.4 Automatic AFM image analysis pipeline developed by Beton et al. A: An example unprocessed AFM image of small circular DNA molecules. B: Sample tilt is removed during the preliminary processing stage. C: The height difference between scan lines is partially addressed. D: After fully eliminating the height difference between scan lines and correcting other image artefacts, DNA molecules are identified from the background and highlighted in red. E: To enable further measurements, identified molecules are transformed into smooth, continuous curves, which are shown in blue and overlaid on the AFM data. F: The contour length is measured for each identified molecule and the distribution is presented as a histogram. Figure obtained from Beton et al. [49]

1.2.3.1 Preliminary processing

Preliminary processing constitutes the first step of automatic AFM image analysis. It addresses issues common to AFM data of all sample types, and involves functions that are found in most AFM analysis software packages, notably the removal of line artefacts, plane artefacts, and noise.

As the surface of the sample is measured via line-by-line interaction with the tip, the image needs to be adjusted to eliminate the artificial height difference between scan lines (see Figure 1.4 (C) above, or Figure 2.1 (B) in Section 2.2.2.2). This process is achieved by aligning the median height of each scan line [46] or polynomial fitting [60]. Scars (see Figure 1.3) in the middle of a scan line can be identified from the image based on its extreme height and subsequently replaced with values interpolated from neighbouring pixels [61].

Plane artefacts such as image bowing are caused by thermal drift and hysteresis [38]; they can also be addressed with polynomial fitting [62], [63]. Another common source of 2D defect is image tilt, which occurs when the sample surface is not exactly perpendicular to the tip; plane fitting algorithms are used for this situation [60]. Figure 1.4 (B) and Figure 2.1 (C) provide examples of tilt removal.

Besides systematic artefacts, AFM images are also affected by noise, which can be caused by a combination of environmental and instrumental factors such as temperature fluctuations, mechanical vibrations and air currents [38]. Although the noise in AFM imaging is relatively low compared to other techniques [64], its complexity means that multiple types of noise filters are often required: low-pass filters (see Figure 1.9 (iii) in Section 1.4.1) deal with periodic noise and smooth sharp edges, while less common high-pass filters remove low frequency noise, which may be caused by an uneven support surface [55]. Many algorithms make use of Gaussian filtering (see Figure 2.1 (E)) to reduce general background noise [55], [56], [65]–[67], and median filtering to address local noise and irregularities [54], [55], [66], [68]. Fast Fourier transform filtering is used to target noise of certain frequencies [66], [69]. An advanced workflow combining multiple filtering methods has also been devised to target stripe noise, which is difficult to eliminate otherwise [70]. Since the noise distribution of AFM images varies according to the equipment, sample, and experimental condition, an automatic processing program designed for general use needs to include all major filter types to cover a wide range of situations.

In addition, the limit to lateral resolution caused by tip convolution can be addressed through algorithms referred to as localisation image reconstruction [13], where data from multiple images are merged through the multiplication of height value by probability at each peak height position, resulting in a final image with higher resolution. This method can be used to obtain detailed information on static samples, but it is unable to capture dynamic processes due to changes in peak positions.

Recent development in machine learning has led to the creation of automatic preliminary processing algorithms based on convolutional neural network [71]. This new approach is comparable in accuracy with conventional methods, and requires no prior knowledge of the artefacts present on the images. It can serve as an optional module on a comprehensive

AFM image analysis platform, but as is the case with all machine learning algorithms, its performance will be highly dependent on the similarity between training data and actual images; as such, it is not always preferred over traditional processing pipelines, and its suitability for images obtained under different experimental conditions needs to be further assessed.

1.2.3.2 Molecular identification

Following preliminary processing, the next step in the automated analysis for AFM bio-imaging is to recognise molecules of interest, so that they can be subsequently measured. This is principally achieved by selecting clusters of pixels whose heights are greater than a threshold value. In doing so, many molecular identification algorithms transform AFM data into a binary image consisting of only foreground and background [54]–[57], [65], [67], as shown in Figure 1.9 (iv) in Section 1.4.1, and this will result in the loss of height information that can be useful during further characterisation. A better alternative is to store the details of identified molecules within a separate mask [49], thereby keeping the original image intact.

Different approaches have been used to determine the height threshold that distinguishes molecules of interest from the rest of the image. The Otsu [15] method automatically separates features from the background assuming a bimodal height distribution; it has been successful on images with low noise level [54], but is not suitable for data containing features that belong to neither molecules of interest nor the background [55]. Pixels can also be identified as belonging to a molecule of interest based on their deviation in height from the mean value of the entire image; the minimum extent of this deviation is either determined automatically through iteration [55], [72], or designated via user input [58]. Alternatively, the user can directly provide an absolute height threshold [61], [66], and clusters of pixels higher than this value are automatically classified as features of interest, but this approach is unlikely to be viable for a large dataset with images obtained under different experimental conditions. As part of the work in this thesis, different height thresholding methods are compared in detail, as will be discussed in Section 3.1.1.2.

In addition to selecting molecules of interest via height thresholding, algorithms from other imaging fields have also been employed for the feature recognition on AFM images; for example, Hough transform, an established algorithm that recognises lines and circles from images [73], has been used to detect globular nucleosomes [54], though this is incompatible with the general identification of biomolecules exhibiting irregular shapes.

The features of interest identified based on height criteria are subject to further selection: molecules touching the image border are usually removed because they are likely incomplete and thus unsuitable for downstream measurements [55], [67]; other recognised molecules may be eliminated because they are too small or too large compared to the expected size for the molecule of interest [56], as demonstrated by the example of Figure 1.9 (v) in Section 1.4.1. A detailed description and comparison of area-based thresholding methods employed in this thesis can be found in Section 3.1.1.3.

While machine learning algorithms have also been applied to the field of automatic recognition on AFM images [74], [75], they are better suited for the classification of molecules into categories based on conformational differences, and are not currently optimised for the characterisation of conformational variation between individual biomolecules.

1.2.3.3 Measurements

Once molecules of interest have been identified, they can be quantitatively characterised via the measurement of various parameters. While the calculation process is straightforward for some of these parameters such as the area, other more advanced measurements can only be made following a further processing step referred to as tracing: to calculate the contour length of a linear molecule such as a DNA filament from one end to the other, the molecule cannot be treated by the automatic analysis program as an amorphous pixel cluster, but rather needs to be represented by a smooth, continuous, single-pixel-wide curve, as shown in Figure 1.4 (E) and Figure 1.9 (vii).

To facilitate this transformation, automatic measurement software generally employs a skeletonisation process where only the pixels required to maintain the general shape of the molecule are kept [49], [55], [56], [65], [76], as illustrated below in Figure 1.5.

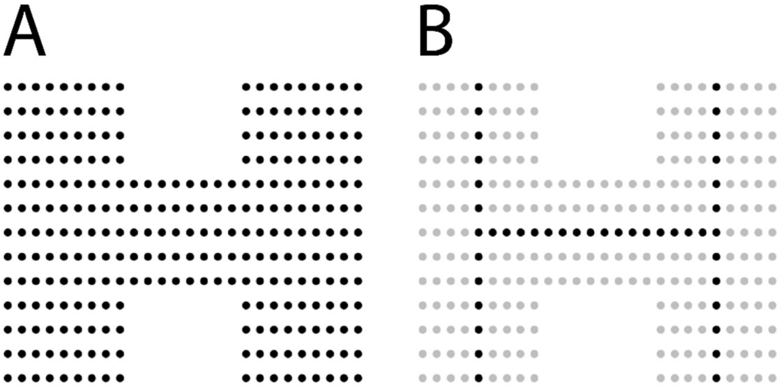


Figure 1.5 Schematic illustration of the skeletonisation process. The original shape (A) is reduced to a skeleton (B) with redundant pixels removed for the ease of future measurements. Only pixels that preserve the continuity and the general shape of the molecule are kept. Figure modified and redrawn from Zhang and Suen [77].

The skeletonisation of a molecule is followed by pruning [55]–[57], which refers to the removal of branches, so that only a single curve without crossings remains. This is necessary when the analysis is conducted on simple linear biomolecules, since the branches in this case are caused by impurities or unwanted overlapping; however, when the biomolecule forms legitimate crossings such as in the case of knots and catenanes [78], the pruning process becomes problematic. The tracing and characterisation of knots on AFM images is still difficult to achieve: while algorithms to handle complex paths have been developed for other types of bio-imaging [79], their application to AFM data is hindered by incompatible file formats and has not yet been realised. Similarly, quantitative analysis on topologically complex biomolecules can be performed based on 3D coordinates [80], but it does not easily integrate into the AFM image analysis workflow.

Currently available methods to perform measurements on the smooth, continuous curve focus on the contour length [55], [57], [68], [81], [82], curvature [56], [75], and bending angle [54], [76], [83] (see Figure 1.9 (viii) in Section 1.4.1 for an example). These algorithms are far from ideal, since apart from their inability to handle crossings, they also suffer from other drawbacks: their effectiveness and accuracy are often demonstrated on simulated data that do not completely represent the complexity and variation of real AFM images [55], [56], [83]; in addition, they may partially rely on manual input [81]–[83]. The workflow developed in certain projects are not applicable to other sample types; for example, many measurements of contour length and curvature rely on the identification of end points and therefore cannot be performed on circular or closed biomolecules [55], [56], [68]; likewise, some programs that calculate the bending angle need to first identify a protein binding site [54], [83], and have limited applicability if the sample does not involve interaction between two different types of biomolecules, or if the protein is too small to be visible.

Last but not least, similar to the two previous steps, machine learning methods can be employed to improve the accuracy of tracing and perform measurements such as curvature [75], although they cannot yet overcome the challenge of tracing knots and crossings.

1.2.4 Challenges in automatic AFM image analysis

As a rapidly growing field, the automated analysis of AFM images faces multiple challenges. Currently available automatic image analysis methods are best suited for samples with simple structures and repeated features, and it remains difficult to characterise complex conformations such as knots. For highly-detailed structures, while automation can be used during part of the image analysis workflow, the role of manual observation is still irreplaceable.

The limited ability to perform advanced measurements means that many properties of interest cannot always be measured directly, and proxy parameters are sometimes used instead. This does not fully utilise the advantage of AFM as a characterisation tool that directly visualises molecular structure in three dimensions. For example, the aspect ratio of circular DNA molecules has been used to represent the conformational changes induced by supercoiling to different extents, when the automatic identification and direct characterisation of structural changes and crossings would have been more suitable [84].

In addition, many automatic image analysis algorithms are developed by primarily testing on simulated data [65], [71], and can encounter difficulties when dealing with real experimental data: a generated image typically contains only the molecules of interest and noise, while multiple types of molecules of different sizes and shapes can feature on a real image.

The most serious limitation, however, lies in the lack of a comprehensive AFM image analysis platform to which methods developed as part of separate projects can be integrated. This inconvenience is worsened by the fact that commercial AFMs all have their own file formats; by contrast, data in other bio-imaging fields can be stored and presented in standard image files.

In the short to medium term, this situation can be alleviated by the development of a comprehensive AFM image analysis software package that serves as a basis for the general characterisation of biomolecules and can be easily customised to suit the need of individual projects. In the long term, the standardisation of AFM data format might be needed to facilitate the access to automatic image analysis algorithms developed by previous researchers.

1.3 DNA as a biomolecule of interest

1.3.1 The conformational variation of DNA

Deoxyribonucleic acid (DNA) is one of the most important biological molecules due to its role as the carrier of genetic information. To fully establish how such genetic and sequential information is stored, passed on and expressed, attention is drawn to changes in the structure of DNA. While the classic double helical model first characterised through X-ray diffraction by Franklin and Gosling [85] and proposed by Watson and Crick [86] has been extensively studied, the existence of conformational [41], [87], [88] and topological variation such as supercoiling [89] still presents challenges in the understanding of this molecule. These variations can affect and also be affected by the interaction between DNA and protein molecules, which occur as part of vital biological processes such as transcription.

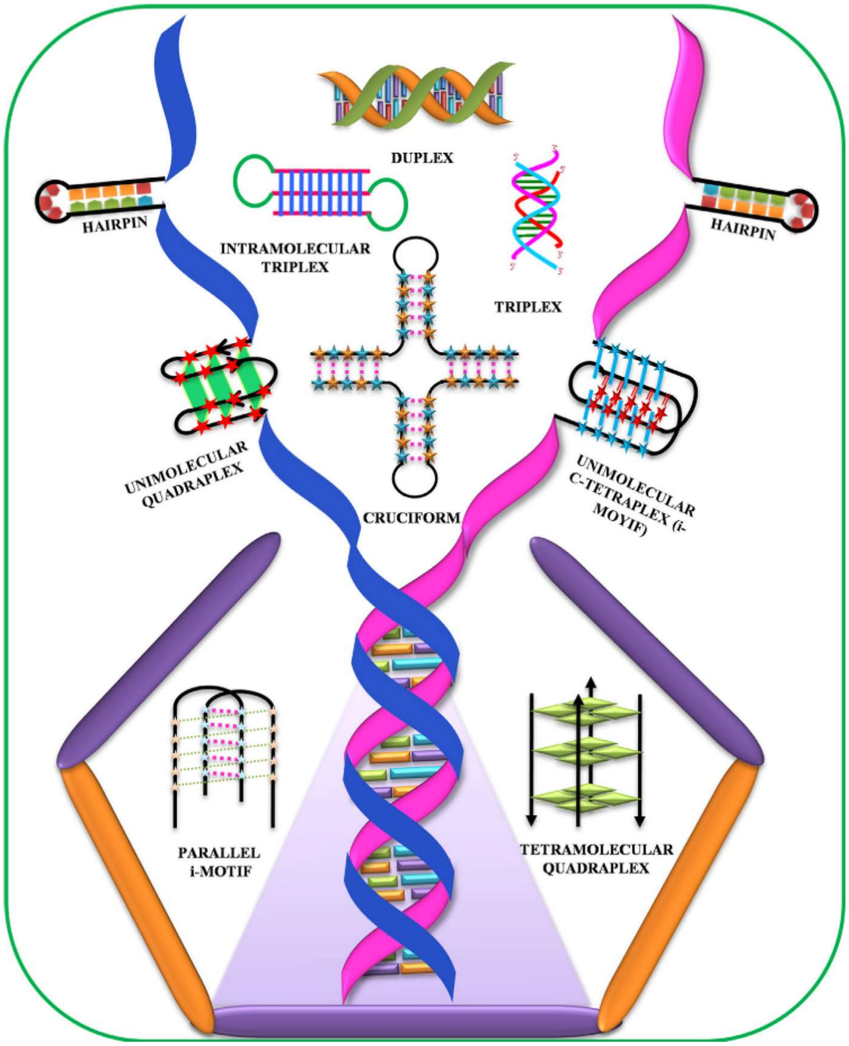


Figure 1.6 Schematic illustrations of alternative DNA conformations. Figure obtained from Kaushik et al. [87]

As illustrated in Figure 1.6 above, DNA can be found in a wide range of alternative conformations, such as triplexes [90], [91], quadruplexes [91], hairpins [92] and cruciform structures [93]. Cruciforms involve base-pairing within the same strand, and can be observed *in vitro* [94] when palindromes are present in the sequence. While their existence *in vivo* is contested [41], they are structurally similar to holiday junctions formed during meiosis. Triplexes are formed when a third strand, from either the same molecule or a different molecule, joins a double-stranded DNA through non-Watson-Crick base pairing. Triplex-forming oligonucleotides serving as the third strand are suggested to have therapeutic potential due to their ability to target specific sequences [95]. Structures with four strands of DNA known as quadruplexes are located within telomeres, which are specialised sequences at the end of chromosomes that play an import role in regulating aging and genome stability [96].

Variation also exists within the DNA duplex structure, which can be categorised into different forms, including B-DNA, A-DNA and Z-DNA. They differ from each other in diameter, handedness, shape of major and minor grooves, as well as number of base pairs per turn [41]. Moreover, the shape of a double helical DNA molecule can be affected by torsional stress, which arises from the overwinding and underwinding of the helix during biological processes such as replication and transcription. This type of topological change is referred to as supercoiling [89], [97].

Proteins play an important role in the regulation and modification of DNA conformation. For example, to regulate DNA supercoiling and release torsional stress, topoisomerases can break one or both DNA strands in front of the replication or transcription site transiently and re-join them. Apart from sequence-dependent intrinsic curvature, DNA *in vivo* is also bent by proteins that wrap them, such as histones in eukaryotes, which can be modified to change the accessibility of DNA and regulate gene expression [98]–[100]. The winding around histones induces conformational distortions, and causes DNA to be natively supercoiled. While histones and chromosomal structures do not exist in prokaryotic cells, the integration host factor (IHF) and HU proteins in bacteria play a similar role by compacting DNA, and they can induce bending at different angles [101]. Other proteins found to alter DNA structure in bacteria include the cyclic AMP receptor protein, which serves both as a transcription activator and repressor, and also leads to bending at specific angles [102]. Proteins that bind to DNA and regulate DNA conformation can also come from viruses: the bacteriophage 434 repressor protein is known to bend DNA at a specific site [103].

Due to the large variation of DNA structure and the frequent occurrence of bending, a quantitative characterisation is required to fully understand DNA conformation and the roles of different proteins in the spatial organisation, replication and transcription of DNA. Here we describe the use and limitations of common DNA characterisation methods, and discuss their suitability for investigating the mechanism behind DNA-protein interactions and understanding the function of DNA-manipulating proteins.

1.3.2 Common methods of DNA characterisation

Most of our understanding on DNA structure is obtained from X-ray diffraction (XRD) of fibres, which measures crystallographic parameters and has led to the discovery of the double helix [85], [86]. The main drawback of this method lies in the need to average over a large number of molecules and the incompatibility with non-periodic features [104]. While atomic-resolution XRD has been performed on single crystals to characterise sub-molecular structural variations of DNA [105], this technique relies on a homogenous sample and is unsuitable for the study of conformational changes without long-range order. Furthermore, XRD cannot always be used to characterise DNA-protein interaction, as some proteins are difficult to crystallise [106].

Another common method to study DNA and its interaction with proteins is nuclear magnetic resonance (NMR) spectroscopy, which characterises biomolecules in a solution-based environment, and provides structural information such as distances and angles between atoms. NMR has been employed to characterise DNA damage [107] and supercoiling [108], but this technique requires isotopic labelling and is not ideal when the molecule is larger than 100 nucleotides [109]; as such, it is best reserved for chemically synthesised small molecules.

Cryo-electron microscopy (CryoEM) is a powerful imaging technique capable of reaching angstrom resolution [8], [110], [111]. It can be used in conjunction with XRD and has been shown to provide useful information on double-stranded DNA breaks [112] and DNA-protein interactions [113]. While CryoEM preserves the natural structure of biomolecules, its accessibility is impeded by a complicated and costly specimen preparation process, where the sample is typically frozen rapidly and maintained under cryogenic temperature during the imaging process [114]. In addition, due to the low signal-to-noise ratio of this technique, the high-resolution molecular structure is obtained by averaging over multiple similar molecules, through the 3D reconstruction of 2D images [8], [111]. This means that CryoEM is less suited for the characterisation of flexible molecules due to their conformational heterogeneity.

Fluorescence techniques such as fluorescence anisotropy [115], [116] and fluorescence resonance energy transfer (FRET) [116]–[118] are regularly used in nucleic acid research, but their focus is limited to labelled regions within a molecule. Electrophoresis with agarose or polyacrylamide gel [67], [84], [108], [119] provides specific information such as the supercoiling level of DNA minicircles (small circular DNA molecules with a length of several hundred base pairs, extracted from bacteria plasmids [120]) without characterising the overall structure [84], [119]. These techniques do not allow for the visualisation of DNA as a whole, and are often employed to complement other methods.

The development of super computers has enabled researchers to predict the conformation of biomolecules through molecular dynamics (MD) simulations [84] [66]. These simulations are valuable for the understanding of DNA-protein interaction [121], but the amount of computing power required limits their usage to molecules of small size. In addition, the results of MD simulations are best interpreted in combination with experimental data.

In conclusion, as summarised in Table 1.1 below, while the wide range of techniques discussed above can all be employed to study DNA structure, none of them directly visualises the conformation variation of individual DNA and protein molecules in a physiological environment. This gap is filled by AFM, which can work in synergy with other techniques to provide results that can be verified in multiple ways.

Table 1.1 Comparison between different DNA characterisation techniques

Technique	**Type of data collected**	**Data analysis method**	**Highest resolution achievable on DNA**	**Compatible with single-molecule characterisation**	**Operable in a solution-based environment**
X-ray diffraction	Diffraction patterns	Structural parameters calculated from diffraction patterns	Atomic, for single crystals of short sequences [105]	No	No
Nuclear magnetic resonance (NMR) spectroscopy	Spectroscopic data measuring relaxation in a magnetic field	Structural parameters calculated from NMR spectra	Atomic, via isotopic labelling	No	Yes
Cryo-electron microscopy	2D projections obtained through electron scattering	3D structures reconstructed from multiple 2D projections	Near-atomic, around 4 Å [113], [122]	Yes, but at reduced resolution	In vitrified ice, no dynamics available
Fluorescence anisotropy and fluorescence resonance energy transfer (FRET)	Fluorescence intensity measured after applying polarised light or facilitating energy transfer between donor and acceptor fluorophores	Molecular orientations or distances calculated from changes in fluorescence intensity over time	Approaching 30 Å, for distance measurement with FRET [123]	Yes	Yes
Electrophoresis	Separation of DNA molecules by charge, length and conformation	Properties of DNA molecules measured through comparison with known reference molecules	Around 50 base pairs (bp) [124]	No	Yes
Molecular dynamics simulation	Atomic interactions or forces calculated based on changes in interatomic potentials	Molecular structures and conformations predicted through simulations	Atomic	Yes	Yes, for explicitly solvated simulations
Atomic force microscopy	Topographic images obtained through raster scans of sample surface	Features of interest identified from images and measured	Double-helical, 1-2 nm laterally and sub-nanometric vertically [84], [125]	Yes	Yes

1.3.3 The role of AFM in DNA characterisation

As discussed in Section 1.2.1, AFM is a bio-imaging tool that allows for the direct visualisation of biomolecules in a physiological environment. By scanning across the sample surface line by line with a cantilever, it produces a topographic map based on the interaction between the tip of the cantilever and the sample at each point, as illustrated below in Figure 1.7. Compared to other techniques, AFM does not require isotopic or fluorescence labelling, and is compatible with single-molecule characterisation. Since DNA conformation exhibits variation and is affected by interaction with proteins, AFM can be employed to characterise both intrinsic conformational heterogeneity, and conformational changes induced by protein interaction.

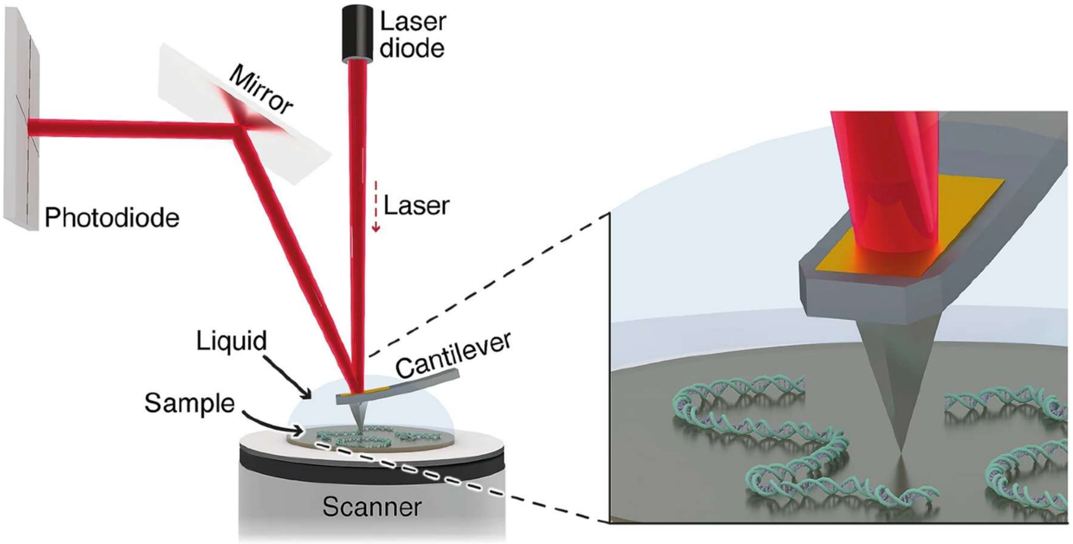


Figure 1.7 Schematic illustration showing the AFM imaging of DNA in solution. As the cantilever scans across the sample surface, interaction between DNA molecules and the tip causes the cantilever to bend and the laser beam to change its position on the photodiode; in this way, surface properties such as height values at each point are measured. Figure obtained from Haynes et al. [126].

The AFM imaging of DNA follows the same principles as the imaging of other biomolecules. It should however be noted that while sub-nanometre resolution has been achieved in the AFM imaging of biological molecules [33], this is performed on samples with highly periodic surface features such as the purple membrane [33], [50], which serves as a proton pump and exhibits a 2D hexagonal lattice structure [127]. Although the techniques and protocols used in these instances of successful high-resolution AFM imaging, such as the selection of an appropriate tip with a stiffness comparable to the sample [50], will be helpful for other biological materials in general, a similar level of resolution cannot be expected for less periodic samples, including single DNA molecules in solution. Nevertheless, due to the development of new imaging modes and sample preparation techniques, the image quality for non-periodic molecules has also been steadily improving over time, and double-helical resolution can be achieved in the case of DNA imaging [84], [125], as shown in Figure 1.8 below.

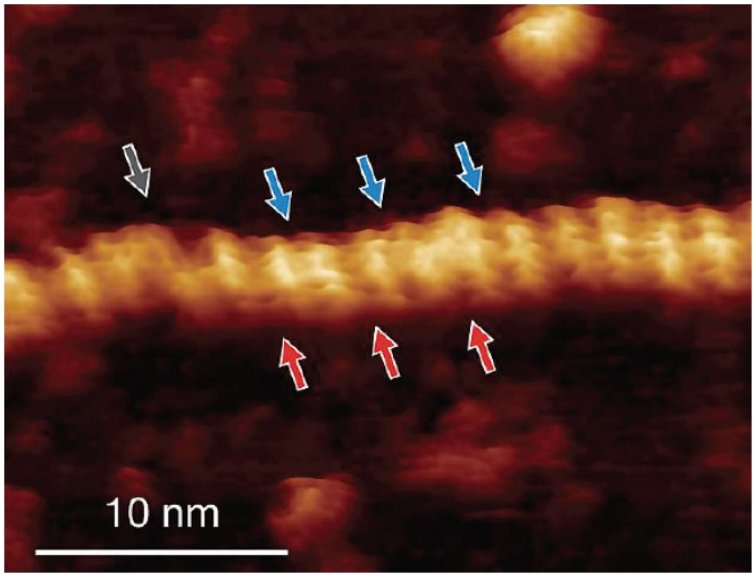


Figure 1.8 Example AFM image of DNA exhibiting double-helical resolution, with major and minor grooves indicated by red and blue arrows respectively, and local unwinding of the double helix indicated by a grey arrow. Figure obtained from Ido et al. [125]

It has been discussed that AFM imaging can be affected by unique artefacts, such as scan line differences and scars. In addition to general issues detailed in Section 1.2.1, another major challenge in the imaging of DNA lies in the use of immobilising agents. To allow the imaging to proceed, DNA and other biomolecules in the solution are first immobilised and adsorbed on a solid surface. Mica is typically chosen for this [68], [128], [129] because of its smoothness [130], but as both mica and DNA are negatively charged, the electrostatic repulsion needs to be bridged in order to immobilise the DNA [35]. Divalent metal ions such as Ni^{2+} and Mg^{2+} are commonly used for this purpose because of their positive charges [131]; however, they may affect the structure of DNA [35] and associated proteins such as topoisomerases [132]. Immobilisation can also be achieved via polymers such as poly-L-lysine (PLL) [133], [134] and poly-L-ornithine (PLO) [135]. Unlike metal ions that stay in the solution, they form a positively-charged layer between the mica surface and the imaged molecules, and the conformation of DNA may be altered from interaction with this layer [133], [136]. Care therefore needs to be taken when choosing the appropriate immobilising agent so as to minimise its side effects; this is particularly relevant to the characterisation of DNA-protein interaction, since conformational changes related to the immobilisation process can be confounded with those induced by the protein of interest.

To conclude, AFM allows for the high-resolution direct visualisation of both intermolecular conformational variation and conformational changes in individual DNA molecules. Similar to the case with other biomolecules, AFM imaging of DNA is not always straightforward, and challenges including the selection of immobilising agents need to be addressed. Automatic image analysis software can be employed to extract more information from AFM data, by dealing with image artefacts, recognising DNA and other biomolecules from the background, and making quantitative measurements of structure and conformation. Having described general AFM image analysis algorithms in Section 1.2.3, we now examine how they have been applied to characterise DNA, and how new contributions can be made to this field, so that more measurements can be made from biomolecules on AFM images, and that the effects of more DNA-manipulating proteins can be quantitatively investigated through automatic analysis.

1.4 Automatic AFM image analysis as a method of DNA characterisation

1.4.1 Applying AFM image analysis methods to DNA characterisation

Automated AFM image analysis methods have been applied to a wide variety of DNA structures, both on their own and in the presence of proteins. In the case of naked DNA, semi-automatic analysis performed on high-resolution AFM images has been used to measure parameters such as the periodicity and chiral angle of the double helix, and the results are in agreement with the predicted B-DNA structure [128]. Similarly, contour length measurement can be combined with the known number of base pairs to provide new information on the conversion between different DNA secondary structures [81]. The algorithms used in these studies require user input to specify the start and end of each linear DNA molecule. This allows for the accurate tracing and measurement of DNA trajectory, but the manual component limits the speed of characterisation.

Another semi-automated image analysis process has been used to quantitatively analyse G-wires [66], which are quadruplex DNA structures with applications in nanotechnology [137]. Following manual observation and classification of G-wires into multiple types, semi-automatic height profile measurement and analysis have been conducted to characterise the periodicity for each type. While this algorithm is designed for a particular sample and does not see much use in the general characterisation of DNA conformation, it demonstrates how automatic analysis on AFM images of DNA can be applied to various sub-molecular features.

Completely automatic analysis methods, by comparison, are typically employed for metrics that are simple to measure. Aspect ratio has been automatically measured to characterise the effect of supercoiling on 339 bp DNA minicircles, revealing that while supercoiling leads to more compacted shapes in general, it can also cause the formation of defects, which are able to partially release the superhelical stress and reverse the conformation compaction to a limited extent [84]. The successful automation in this case is in part due to the relatively simple, uniform structure and small size exhibited by minicircles; as such, it would be challenging to introduce similar algorithms to the study of more complex molecules.

The automatic measurement of DNA bending angle has been used to characterise the operation mechanism of DNA glycosylase, an enzyme that facilitates the repair of damage to DNA bases [76]. By comparing the bending angle distributions of damaged and undamaged DNA before and after treatment with DNA glycosylase, a likely mechanism for the protein to identify the damage site has been proposed. It is however worth noting that the automatic analysis in this study was conducted on ImageJ rather than AFM-specific software. While this approach enables the use of automatic analysis tools developed in other bio-imaging fields, it cannot be applied to AFM image analysis directly, because it disregards the multichannel nature of AFM data, and is not optimised to handle AFM-specific noise and artefacts.

In eukaryotes, DNA in the cell nucleus is wrapped around histones, forming complexes referred to as nucleosomes [138]. To characterise histone-DNA interaction, AFM imaging has been performed on nucleosomes reconstructed *in vitro* with short pieces of DNA and histones [54], [83]. This renders automatic recognition more challenging than for naked DNA, since the histone and the DNA are different in height yet form a single unit. Despite this difficulty, suitable molecular identification algorithms reaching an accuracy of 95% have been developed [83], and they have been applied to characterise the structure of nucleosomes as visualised by AFM. The combination of manual tracing and automatic bending angle measurement has been used to investigate the wrapping mechanism of different histone types. In addition, external factors such as salt concentration, have also been shown to affect the opening angles and influence gene accessibility [83].

In another similar study, the opening angle of the nucleosome has been automatically measured alongside the radius and relative position of the histone, as shown in Figure 1.9 below. This data has also been used to investigate and compare different types of histones, and it has been found that certain mutations in histones lead to larger opening angles, which indicates the genes in the surrounding DNA can be expressed more readily [54]. The main drawback of this study is that, the results of automatic measurement need to be manually inspected and occasionally corrected.

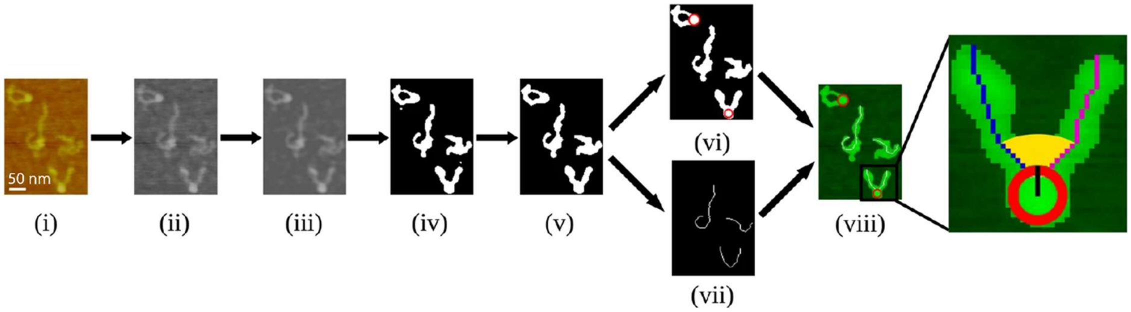


Figure 1.9 Workflow for the automatic measurement of the nucleosome opening angle. During the preliminary processing stage (i-iii), noise was removed from the original image (i) through initial filtering (ii) and additional low-pass filtering (iii). During the molecular identification stage (iv-v), pixels on the image were classified into foreground and background, with clusters of foreground pixels recognised as potential molecules of interest (iv); some identified molecules were subsequently removed as they were too small or too large, and were therefore likely to be noise or impurities (v). The final measurement stage (vi-viii) started with two parallel processes: bending point identification (vi) and tracing (vii); the results from these two processes were combined based on physical proximity to enable the automatic measurement of the opening angle (viii). Figure obtained from Würtz et al. [54]

Automated image analysis algorithms are also used to study dynamic processes such as the interaction of DNA with drugs. By taking time-lapse AFM images, daunorubicin, a replication inhibitor used for cancer treatment, is shown to affect the conformation of DNA [139]. The average height of DNA molecules is automatically measured as the concentration of daunorubicin is increased over time, and it is observed to decline at first before rising more quickly. The taller structure at a higher concentration suggests that the molecule is more compacted and twisted, causing replication to be more difficult to proceed.

In summary, DNA exhibits a wide range of structures that need to be studied for a better understanding of its function; as a suitable tool for the direct visualisation of biomolecules, AFM has been extensively employed for this purpose. The increased use of AFM is accompanied by the development of automated analysis software, which is essential for the quantification of AFM data and has been applied to various contexts. However, the capacity of currently available automatic analysis tools is limited, and the accurate measurement of more complex structures often requires manual input.

1.4.2 Automatic AFM image analysis with TopoStats

Given the limitations of existing automatic AFM image analysis software, to ensure that the work in this thesis makes a meaningful contribution, it is important to develop new methods by building on a software package that can be adapted for general usage in multiple situations. While a powerful program extensively employed by the AFM bio-imaging community does not yet exist, the open-source software package TopoStats [49] has the potential to fulfil this role.

TopoStats was originally created by Joseph G. Beton and Alice L. B. Pyne. Due to its initial dependency on modules from Gwyddion which were only compatible with outdated Python 2, the program was difficult to install and saw limited use by the wider community. To address these issues and increase the applicability of TopoStats, Neil Shephard *et al.* have been refactoring the code to produce a second version [61] that is independent from Gwyddion and more user-friendly. Ongoing improvements are still being made to this new version, including the incorporation of features developed as part of this thesis.

While the two versions differ slightly in functionality, as will be explained in Section 2.2, they share a similar workflow that can be summarised as follows: preliminary processing is first performed to handle image artefacts and reduce noise; pixel clusters satisfying certain height and area criteria are then automatically recognised from the background as molecules of interest; afterwards, statistics such as size, height and volume are measured for the identified molecules. This data can help us quantitatively understand the conformation of molecules of interest, and how such conformation changes as a result of intermolecular interaction. In the case of DNA and other linear molecules, the identified pixel clusters can be further simplified into single-pixel-wide curves, allowing for further characterisation such as the measurement of contour length. TopoStats also contains graphic production modules, which enable users to visualise the distribution of the measured variables.

While TopoStats is capable of automatically batch processing AFM datasets and obtaining statistical information on molecules of interest, it was initially designed for and used on the characterisation of DNA minicircles [49], [84], which are 115 nm long circular DNA molecules extracted from bacterial plasmids. This means that many modifications were required for the program to become more compatible with a larger variety of samples.

To start with, although the automatic identification algorithm was theoretically able to recognise molecules of all sizes, it was tailored for the recognition of DNA. The selection of other molecules with different shapes and sizes such as proteins could be facilitated through the adjustment of parameters, though before the refactoring, those parameters were hard-coded and sometimes difficult to locate within the code. This required improvements to be made to TopoStats, so that a parameterised pipeline can be established for the characterisation of different molecules including proteins, which need to be automatically recognised and measured as part of the research on how they affect DNA conformation.

Furthermore, the aforementioned process of turning pixel clusters into single-pixel-wide curves, known as tracing, was only successful for closed molecules forming a loop. This severely limited the applicability of TopoStats as tracing is essential for understanding how the conformation of DNA changes along its trajectory; in other words, the restriction of this feature to closed molecules meant that many projects using open, linear molecules with a start and an end could not rely on TopoStats. The code also lacked more advanced measurements such as curvature and bending angle, which are useful in providing a quantitative representation of DNA structure and locating abnormalities on the molecule. This will give us more information on the local conformational variation of DNA and the mechanism of DNA-protein interaction.

In response to these challenges, improvements have been made to TopoStats during the work in this thesis through the adjustment of existing functions and the development of new functions. The scope of the work involved is summarised in Table 1.2 below:

Table 1.2 Comparison between the functionalities of TopoStats before and after the work in this thesis

Function	**Availability in TopoStats before the work in this thesis**	**Availability in TopoStats after the work in this thesis**
Preliminary processing	Available	Available and unchanged
Molecular recognition	Available but only optimised for DNA minicircles	Available and customisable for DNA and protein molecules of different shapes
Measurements unrelated to tracing	Available	Available and unchanged
Tracing	Available only for circular molecules	Available for both circular and linear molecules
End-to-end distance measurement	Unavailable	Available
Contour length measurement	Available but only reliable for circular molecules	Available and reliable for both circular and linear molecules
Curvature measurement	Unavailable	Available
Bending angle measurement	Unavailable	Available

1.4.3 Potential applications for automatic AFM image analysis

1.4.3.1 Case of interest: TOP1

Due to the double helical structure of DNA, supercoiling stress is generated during biological processes such as replication and transcription. This can be addressed by topoisomerases, which are enzymes that catalyse topological changes in DNA. They relieve stress and straighten coiled DNA by transiently breaking one or both strands of DNA [140], [141].

Topoisomerases are classified into many different types based on their structure, function, and the location they are found in [140], [141]. One particular topoisomerase of interest is the human topoisomerase I (TOP1), the most studied enzyme in the category of type IB topoisomerases [140].

Composed of 765 amino acids and can be divided into 4 domains [140], TOP1 is involved in both DNA transcription and replication, and is therefore essential for genomic stability and cell proliferation [142]. This makes it a target for cancer chemotherapeutics [143]: TOP1 inhibitors, a group of drugs that prevent TOP1 from functioning, are employed in the treatment of cancer, but their clinical usage is limited by side effects and off-target toxicity [144], [145]. A better understanding of the mechanism of TOP1 operation is therefore required in order that TOP1 inhibitors can achieve their clinical potential.

AFM imaging is able to provide a direct visualisation of TOP1 structure, which is vital for investigating how TOP1 operates. Furthermore, through the use of automatic image analysis, statistics on TOP1 such as the distribution of its size can be obtained, which will allow us to study variation within a TOP1 population. As TOP1 is a globular protein with a relatively simple shape, it also provides a good starting point for testing the automatic recognition of different molecules.

1.4.3.2 Case of interest: NDP52

Some proteins are speculated to interact with DNA, but the mechanism of interaction and the effects on DNA conformation are unknown.

One such case of interest is NDP52, a relatively small protein with a 446 amino-acid sequence. It is known as an autophagy receptor, as it is associated with the process of removing undesired or damaged structures in the cell [146]. It is also involved in the organisation of actin filaments that connect neighbouring cells [147].

While the known functions of NDP52 do not include interaction with DNA, NDP52 was first discovered in the nucleus. In addition, as shown in Figure 1.10, its structure and sequence are highly similar to the protein CoCoA, which is known to be a secondary transcription coactivator in the nucleus [148]. Both NDP52 and CoCoA contain a kidney enriched inositol phosphatase carboxyl homology (SKICH) domain at the N terminal, zinc finger domains at the C terminal, and a coiled-coil domain in the middle connecting the two terminals [149]–[151].

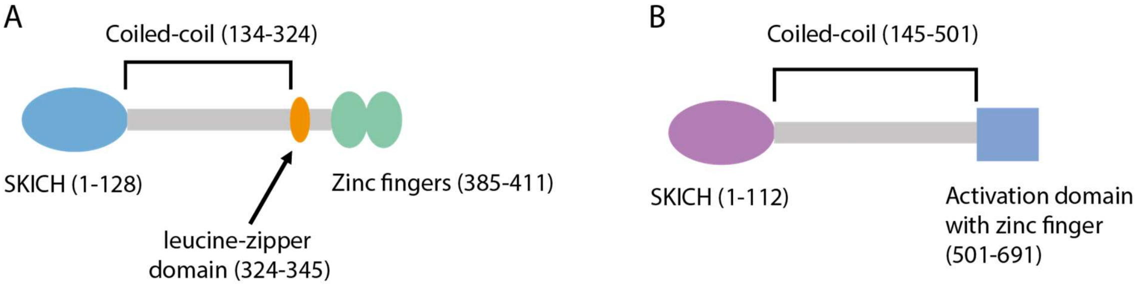


Figure 1.10 Schematic illustrations of NDP52 (A) and CoCoA (B). Image redrawn from dos Santos et al. [151]

As such, it has been speculated that NDP52 also has a role in the nucleus and is able to interact with DNA. To probe whether this is the case, our collaborators performed electrophoretic mobility shift assay on the protein [151], which showed that it has a high affinity to double stranded DNA.

To investigate what potential role NDP52 might have in the nucleus, immunofluorescence staining, electron microscopy and stochastic optical reconstruction microscopy (STORM) were performed to characterise the distribution of NDP52 within the nucleus. The results showed that they form clusters around transcription initiation sites, suggesting a potential involvement in transcription activation, similar to what is known about CoCoA. Furthermore, when transcription is suppressed, less clustering is observed and NDP52 showed more dynamic movements in the nucleus [151].

To characterise the nature of such clusters, size-exclusion chromatography with multi-angle light scattering (SEC-MALS) was performed to see whether NDP52 in solution operate alone. The results showed that NDP52 is prone to oligomerisation with the majority of the proteins in dimeric form, but higher oligomeric states are also found. This is in line with work using other techniques, including microscale thermophoresis and mass photometry.

While these data suggested that NDP52 could be involved in transcription activation, they did not provide direct information on the mechanism of interaction between DNA and NDP52 and the structure of potential complexes formed. A more thorough understanding of NDP52 structure is required to characterise its interaction with DNA and investigate which domain is more active in the process. However, the flexibility of the coiled-coil means that NDP52 can exhibit a variety of conformations, and as a result, no crystallography structure has been obtained to confirm its structure. Therefore, a direct visualisation of NDP52 and its interaction with DNA is needed to corroborate the previous understanding of its structure, and to probe whether DNA conformation is modified through such interaction. This information will allow us to determine what role NDP52 has in the nucleus and how it might affect the functioning of DNA.

1.4.3.3 Case of interest: HU

The nucleoid-associated proteins such as HU and IHF are known to compact bacterial DNA [152]. HU in particular is composed of an α-helix and multiple β-sheets that form two β-ribbon arms, which bind the minor groove of DNA [152]. Unlike IHF, it does not bind to a specific sequence, but is known to have a high affinity to distortions in DNA such as repair intermediates [153].

To further understand the operation mechanism of HU and the structure of the HU-DNA complex, multiple studies have been conducted on the bending of DNA as a result of HU interaction: a crystallography study suggests a bending angle of 105° to 140° [101], while another study using fluorescence resonance energy transfer measures a bending angle of 70° [116].

The discrepancy between these results calls for further investigation. Molecular dynamics simulations carried out by our collaborators suggest that the geometry of HU-DNA complex be classified into three modes, referred to as half wrapped, three-quarters wrapped and fully wrapped; they correspond to bending angle measurements of 100.2° ± 11.6°, 137.0° ± 7.6° and 161.1° ± 6.4° [154].

The existence of these distinct categories can be best verified through the direct visualisation of HU-induced bending via AFM imaging. Furthermore, automatic analysis methods that characterise the distribution of bending angles is ideal to complement the simulated data and provide more information on the role of the HU protein in bacteria gene expression.

1.5 The scope of this thesis

This thesis investigates how we could best utilise automatic AFM image analysis to characterise DNA conformation at the nanoscale including in the presence of DNA-manipulating proteins. The automatic approach would allow us to mass process images and obtain more quantitative information on DNA structure and interactions, which could be extrapolated to improve our understanding of the function of DNA.

As described in Section 1.2.3, the automatic analysis of AFM images can be divided into three stages: preliminary processing, molecular identification, and measurements. In this thesis, the development of image analysis methods builds on the existing software TopoStats [49] by improving the last two stages. This is achieved through adapting molecular identification for a wider range of molecules, and introducing new measurement methods.

Chapter 3 focuses on molecular identification and subsequent automatic measurements that can be carried out immediately without molecular tracing. While previous methods developed by Joseph G. Beton *et al.* and implemented via TopoStats were aimed at automatically recognising DNA from the background of AFM images, this chapter discusses how these methods could be customised to suit a larger variety of molecules of interests, using the examples of proteins with different shapes, one globular (TOP1) and one with sub-molecular units (NDP52). The successful characterisation of these proteins allows us to distinguish between different protein species based on their size and shape, and to characterise individual differences driven by flexibility or molecular heterogeneity between molecules of the same species. In the future, this knowledge can be used to characterise the profile of protein molecules that interact with DNA, so that we can understand whether they do so as single proteins or higher order oligomers, and whether they interact singularly or in multiple locations per DNA molecule. It will also be possible to determine whether one protein species preferentially interacts with DNA over another, as long as they are different in size or shape. Since AFM does not allow for the chemical identification of the type of the protein, or the characterisation of protein structure at an atomic level, the results of automatic AFM image analysis can be complemented by data obtained from other biochemical characterisation techniques to gain a better understanding of the molecular mechanisms driving these DNA-protein interactions.

Chapters 4 and 5 describe the development of new methods to quantify the structure of DNA, beyond the capacities of currently available functions in TopoStats, which are only able to analyse the structure of circular molecules that do not contain crossings, and cannot compute values such as the curvature at different parts of the DNA. Unlike proteins where polymer chains are folded to form complex higher order structures, DNA follows a clear trajectory that can be automatically extracted and quantified through tracing, as mentioned in Section 1.2.3.3. Successful tracing allows us to evaluate local changes to the DNA backbone, which can be induced by sequence, topology or by interaction with other molecules such as immobilising agents and DNA-manipulating proteins [35]. In Chapter 4, improvements to the tracing process are made, so that both linear and circular molecules can be traced, and that their conformations can be characterised through further measurements including the contour length and the end-to-end distance. These measurements can be applied to study conformational changes on DNA caused by interactions with proteins such as NDP52.

Building on the tracing and measurements developed in Chapter 4, we introduce more advanced DNA characterisation methods in Chapter 5. These are the measurements of curvature and bending angle, which are employed to investigate the effects of NDP52 and HU on DNA. They allow for the quantification of local conformational changes on specific points along the DNA trajectory. As proteins may bind to a specific DNA sequence or other special positions such as defect sites, by measuring the curvature along the DNA trajectory in the presence and absence of proteins, we can determine the binding preferences of these proteins and the effect of the binding on DNA structure. When DNA is considered as a single polymer of defined flexibility, its curvature and bending can also be used as a readout or approximation of its mechanics, as has been well described by polymer physicists for a long time [155].

It is worth remarking that the methods developed are not always successful at quantifying DNA conformation, due to limitations that are beyond the scope of this project, such as the failure of tracing for molecules exhibiting complex shapes or the low quality of the experimental data. These limitations can however be addressed or circumvented in future studies, where the methods developed in this thesis can be applied to characterise the effect of other proteins on DNA conformation.

1.6 Aims and objectives

The **aim of the thesis** is to develop and apply automatic AFM image analysis tools to quantify DNA and protein structures, and to describe changes in DNA conformation as a result of protein interactions.

The **objectives of the thesis** are listed as follows:

1. To determine the limits of existing manual and automatic AFM image analysis methods for the measurement of the size, conformation and structure of DNA and proteins;
2. To develop new automated algorithms, such as the measurement of curvature and bending angle, that can quantify changes in DNA conformation;
3. To investigate the effects of protein interaction on DNA conformation and corroborate results obtained from other techniques through the use of aforementioned automated algorithms.

To achieve these objectives, the thesis will focus on the automatic recognition and measurement of proteins on AFM images in Chapter 3. Subsequently, in Chapter 4, the thesis will discuss the characterisation of DNA and its conformational changes as a result of protein interaction. Finally, Chapter 5 will focus on the development of new tracing-based characterisation methods that enable automatic measurements of curvature and bending angle at defined points along the trajectory of DNA.

Chapter 2 Methods

2.1 Source of data

All analysis in this thesis was carried out on AFM data provided by collaborators. The methods they employed to perform AFM imaging are outlined below.

2.1.1 Source of TOP1 data

Analysis in Sections 3.1 to 3.3 was performed on AFM images of human type I topoisomerases (TOP1). These images were provided by Kavit H. S. Main from University College London, and the TOP1 protein molecules were obtained from the provider Inspiralis [156].

Sample preparation and AFM imaging were conducted following an established experimental protocol [126]: 20 µL of buffer solution containing 20 mM HEPES and 3 mM $NiCl_2$ was first pipetted onto the surface of freshly cleaved mica. The Ni^{2+} in the solution was used as the immobilising agent, which enabled negatively charged biomolecules to attach to the negatively charged mica. 4 µL of 21.5 ng/µL TOP1 solution was then added and gently mixed with the buffer. The sample was left alone for 30 minutes to allow time for adsorption. The TOP1 solution was washed away afterwards and replaced with the buffer through pipetting; this procedure was performed four times so as to completely remove protein molecules that were not bound to the mica surface. The imaging process was then carried out on Bruker Multimode 8 and FastScan Bio AFM models in PeakForce Tapping mode, using MSNL-10-E (nominal tip radius 2 nm) [157], PeakForce HIRS-F-B (nominal tip radius 1 nm) [42] and FastScan D (nominal tip radius 5 nm) [40] probes.

2.1.2 Source of NDP52 data

Daniel E. Rollins and Ália dos Santos (both based in the University of Sheffield) provided the data for the NDP52 project discussed in Sections 3.4 to 5.2. This included AFM images of full-length and truncated NDP52 for Chapter 3, and AFM images of DNA with and without NDP52 co-incubation for Chapters 4 and 5.

The NDP52 protein molecules were obtained from our collaborator Ália dos Santos. The full-length NDP52 (FL-NDP52) contained 446 amino acids, while the truncated protein, known as $_C$NDP52, contained only 82 amino acids (numbered 365 to 446) from the C-terminal domain. The DNA molecules were linearised from minicircles provided by Twister Biotech. They were produced by extracting a specific 339 bp sequence from bacteria plasmids.

The published experimental protocol as used for TOP1 and described in Section 2.1.1 was followed, with minor modifications depending on the immobilisation technique. FL-NDP52 and $_C$NDP52 were imaged with poly-l-lysine (PLL) as the immobilising agent. The cleaved mica surface was treated with 20 µL 0.01% PLL solution and incubated for 1 minute. The mica disc was then washed with Milli-Q ultrapure water so that PLL molecules unattached to the surface were removed. 3 to 8 ng of the protein dissolved in 20 µL of buffer solution containing 50 mM Tris and 150 mM NaCl was subsequently added. After 10 minutes of adsorption, the mica surface was washed four times with the buffer solution to remove unbound molecules.

Different sets of images for DNA were taken using Ni^{2+}, Mg^{2+}-Ni^{2+} exchange or poly-l-ornithine (PLO) as the immobilising agent. When Ni^{2+} was used, the mica surface was first treated with 20 µL buffer solution containing 20 mM HEPES, 10 mM NaCl and 3 mM $NiCl_2$. 7 ng of DNA with or without 40 ng of NDP52 was then added and the sample was left alone for 30 minutes. Prior to imaging, the mica surface was washed four times with the buffer solution.

For immobilisation with Mg^{2+}-Ni^{2+} exchange, the mica was initially treated with a different buffer solution containing 20 mM HEPES, 10 mM NaCl and 25 mM $MgCl_2$. The rest of the sample preparation process was the same as for immobilisation with Ni^{2+}, and Mg^{2+} was replaced by Ni^{2+} during the final washing stage, where the Ni^{2+}-based buffer solution was used.

PLO and PLL had been established to have similar structures and immobilising mechanisms [135], and the choice between them was based solely on their availability at the time of the imaging. The experimental procedure for PLO-immobilised DNA was therefore similar to that for PLL-immobilised NDP52, although the buffer solution was different and contained 20 mM HEPES with 50 mM NaCl. When applying the sample solution to the mica surface, 10 ng of DNA with or without 40 ng of NDP52 was used. Upon discovering that the DNA concentration was too high for image analysis, an additional set of experiments were conducted with the amount of DNA lowered to 7 ng.

All images were taken at room temperature with Bruker FastScan Dimension XR AFM in PeakForce Tapping mode using FastScan D probes (nominal tip radius 5 nm) [40].

2.1.3 Source of HU data

Analysis in Sections 5.3 and 5.4 was performed on AFM data from the HU project provided by Daniel E. Rollins and Elliot W. Chan. Three types of samples were imaged: undamaged DNA, damaged DNA without HU treatment, and damaged DNA with HU co-incubation. The DNA molecules were constructed by Elliot W. Chan and Jamieson Howard, both based in the University of York. Undamaged and damaged molecules were 303 bp and 305 bp long respectively, with the latter containing two additional flipped bases, two additional stacked bases and a base pair mismatch at the centre. The HU protein molecules were provided by Michelle Hawkins, also from the University of York.

For all three sample types, 20 µL 0.01% PLL solution was first pipetted to the mica surface for one minute and washed off with Milli-Q. The sample solution containing varying amount (5 ng for undamaged DNA; 2 or 5 ng for damaged DNA without HU treatment; 7.5 or 10 ng for HU-treated damaged DNA) of DNA was then applied. For HU-treated DNA, 1 or 2 ng of HU was also added, and the co-incubation was subsequently performed for one hour at either room temperature or 37.5°. Afterwards, the mica was washed four times with 10 mM Tris buffer solution. The images were obtained on Bruker Bioscope Resolve and Bruker MultiMode 8 AFM models using PeakForce Tapping mode with PeakForce HIRS-F-B probes (nominal tip radius 1 nm) [42].

2.2 Image analysis

2.2.1 Overview of image analysis

The image analysis methods developed and discussed in this thesis are built on the pre-existing Python program TopoStats [49]. Its original version (referred to as TopoStats 1) was created by Alice L. B. Pyne and Joseph G. Beton. It was based on the open-source software Gwyddion [58], which negatively impacted the longevity of TopoStats due to reliance on the outdated and unmaintained Python 2. To address this issue and improve the general quality of the code, Neil Shephard *et al.* refactored TopoStats into a new version [61] (referred to as TopoStats 2) that uses Python 3 and does not depend on Gwyddion; in addition, the workflow is organised into smaller modules that can be tested individually.

All versions and project-based variations of TopoStats are hosted on GitHub, a platform that enables collaborative working and version control for software. Methods development during the TOP1 and NDP52 projects (Sections 3.1 to 5.2) was achieved through making modifications to TopoStats 1, and the resulting code is available on the 'Curvature' branch of the GitHub repository (https://github.com/AFM-SPM/TopoStats/tree/Curvature). The development of bending angle measurement and its application to the HU project (Sections 5.3 and 5.4) were based on TopoStats 2, and the code is stored on the 'JeanDu/BendingAngle-Upgraded' branch (https://github.com/AFM-SPM/TopoStats/tree/JeanDu/BendingAngle-Upgraded).

Installation instructions can be found in the README.md file of the relevant branch. In this project, TopoStats 1 was edited and executed through the integrated development environment PyCharm. TopoStats 2 was edited in the same way but activated via the 'run_topostats' command in a virtual environment using Anaconda Powershell Prompt.

TopoStats 1 and TopoStats 2 operated under the same principles and general workflow. Both versions involved a configuration file (in .ini format for TopoStats 1 and .yaml format for TopoStats 2) that could be modified by the user for customisation purposes. AFM data in the form of .spm files was read by the program and stored as NumPy arrays [158], with each point on the image corresponding to a pair of X and Y coordinates and a height value. Note that .spm files contained several channels with measurements for different properties, but only the channel with height data was used for this project.

The automatic image analysis can be divided into three main steps: preliminary processing, molecular identification, and measurements. In addition, some measurements on DNA required another process known as tracing, where identified molecules were converted into smooth, continuous curves that followed the trajectory of the DNA. The work carried out in this project concerned molecular identification and measurements: Chapter 3 will discuss the adaptation of the identification process for different types of molecules, while Chapter 4 and Chapter 5 will focus on the improvement of the tracing algorithm and the development of new tracing-based measurement methods.

2.2.2 Preliminary processing

2.2.2.1 Preliminary processing in TopoStats 1

Preliminary processing for TopoStats 1 was performed in the script pygwytracing.py. To start with, the ‘traversedirectories’ function identified all .spm files in the relevant directory (specified via the configuration file) and its sub-directories. The software then iterated through each file and selected data from its ‘Height’ channel using the ‘choosechannels’ function. This data was uploaded to the Gwyddion data browser and later stored as a NumPy array.

While the data was in the data browser, the ‘editfile’ function adjusted and corrected the AFM images by calling several Gwyddion functions [46], [159]: ‘level’ removed the tilt of the sample while ‘align rows’ eliminated the differences between scan lines. Polynomial background noise was subtracted through ‘polylevel’, and line defects known as scars were addressed using ‘scars remove’. Further levelling was achieved through ‘flatten base’, which combined different levelling algorithms and sharpened the image. The mean height value was adjusted to zero through the function ‘zero mean’, which added a constant to the entire image. A Gaussian filter with σ = 1 was applied in the end to remove high frequency noise.

2.2.2.2 Preliminary processing in TopoStats 2

TopoStats 2 stored functions for different stages of the image analysis workflow in separate modules. Some of them could be activated and deactivated independently by entering ‘true’ or ‘false’ in the ‘run’ field under relevant sections of the configuration file.

Functions related to data input and output were located in the script io.py: ‘find_files’ identified .spm files in the designated directory, while the ‘load_spm’ method within the ‘LoadScans’ class was used to extract and load data from the ‘Height’ channel of each file. The preliminary processing of AFM images was subsequently carried out (if enabled by the user) via the ‘Filters’ class in the filters.py module: the ‘median_flatten’ method dealt with differences between scan lines; ‘remove_tilt’ addressed the issue of sample tilting; ‘remove_quadratic’ handled quadratic bowing; these three methods were sequentially applied twice to ensure the complete removal of artefacts. Afterwards, ‘average_background’ adjusted height values so that the background was set to zero, and ‘gaussian_filter’ applied a Gaussian filter to the image with a customisable σ value. Functions pertaining to the removal of scars were located in a different script scars.py, which could be disabled separately from other preliminary processing methods. If enabled, these functions were executed after both instances of ‘remove_quadratic’.

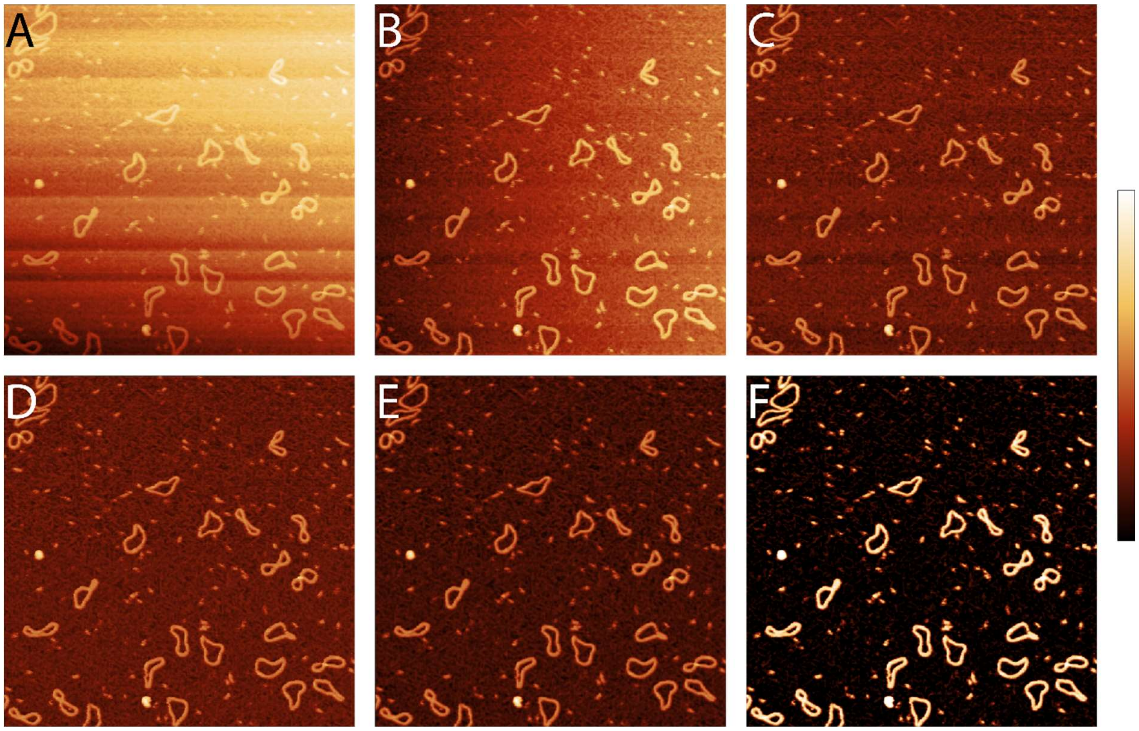


Figure 2.1 The preliminary processing workflow in TopoStats 2. A: An example unprocessed AFM image of DNA minicircles. Height scale 21 to 37 nm. B: The differences between scan lines are addressed through the 'median_flatten' method. Height scale -3 to 6 nm. C: Sample tilting and quadratic bowing are removed using the 'remove_tilt' and 'remove_quadratic' methods. Height scale -3.5 to 4.5 nm. D: The methods 'median_flatten', 'remove_tilt' and 'remove_quadratic' are reapplied to further remove image artefacts, and the background is adjusted to zero through 'average_background'. Height scale -1.5 to 6.5 nm. E: The 'gaussian_filter' method applies a Gaussian filter to the image. Height scale -1 to 6 nm. F: After preliminary processing, the display height scale of the resulting image can be configured by the user to highlight features of interest and maintain consistency across the entire dataset. In this thesis, the height scale of all AFM images is 0 to 3 nm unless otherwise specified.

2.2.3 Molecular identification

2.2.3.1 Molecular identification in TopoStats 1

TopoStats 1 performed molecular identification using the 'grainfinding' function in pygwytracing.py. This process was also referred to as masking because it produced masks of identified molecules overlaid on the original image. Pixel clusters were marked as potential molecules of interest or 'grains' if they deviated sufficiently from the mean height of the image (adjustable through the configuration file parameter 'theresholdingcriteria', the minimal deviation from the mean value in the Gaussian distribution, expressed in multiples of σ, the standard deviation) and were large enough in area (configuration file parameter 'minarea', in m^2).

Grains that were too large or too small (configuration file parameters 'maxdeviation' and 'mindeviation', in multiples of the median area among all grains on an image) and hence likely to be impurities or noise were removed through functions 'removelargeobjects' and 'removesmallobjects'. When there was no large impurity, the 'maxdeviation' was set to a very large value of 1000 to prevent legitimate molecules from being erroneously excluded. The configuration file parameters were adjusted through trial and error for different datasets and different features of interest from the same dataset.

To determine whether suitable parameters could be selected automatically rather than via manual observation, parameter sweeps were conducted on example images from the TOP1 dataset. For each image, this involved recording the number of grains identified as a function of the configuration file parameters, and consisted of two stages. In the first stage, the functions 'removelargeobjects' and 'removesmallobjects' were disabled, and the parameter 'thresholdingcriteria' was gradually increased from 0.1 to 2.0, while the parameter 'minarea' was adjusted from 10 × 10^{-18} m^2 to 100 × 10^{-18} m^2; the optimal values for these two parameters were then chosen based on results from the parameter sweep and manual observation. In the second stage, the 'thresholdingcriteria' and 'minarea' values were fixed, while the functions 'removelargeobjects' and 'removesmallobjects' were re-enabled, and the parameters 'maxdeviation' and 'mindeviation' were gradually adjusted; when selecting for all protein species on an image, 'maxdeviation' was increased from 1.0 to 10.0, while 'mindeviation' was changed from 0.0 to 1.0; when exclusively selecting for TOP1, 'maxdeviation' was changed from 1.0 to 5.0, and 'mindeviation' was adjusted from 0.1 to 1.0. During each stage, 20 equally spaced values within the specified range were used for each parameter, which corresponded to 400 parameter pairs being tested.

Since the parameter sweeps did not show distinct regions for different types of molecules (see Sections 3.2.2.1 and 3.2.2.2 for relevant results), the final parameters used were determined primarily through manual observation. The values of the parameters adopted in the TOP1 and NDP52 projects are listed below:

Table 2.1 Molecular identification parameters in TopoStats 1 used for different features of interest

Feature of interest	**thresholdingcriteria**	**minarea / m^2**	**maxdeviation**	**mindeviation**
TOP1	1.5	50e-18	1.5	0.5
All proteins on images of TOP1	0.2	30e-18	5.0	0.1
NDP52 monomers	0.5	50e-18	2.3	0.5
NDP52 oligomers	0.3	50e-18	1000	2.3
NDP52 terminals	1.5	50e-18	2.3	0.5
$_C$NDP52	1.5	50e-18	2.3	0.5
DNA single molecules	0.9	300e-18	1.3	0.9
DNA clusters	0.9	1350e-18	1000	0
All DNA molecules	0.9	300e-18	1000	0.9

The 'savefiles' function in pygwytracing.py subsequently exported processed AFM images before and after molecular identification as .tiff files, where the height value at each pixel was represented by a colour. Two main height colour scales known as 'gwyddion' (see Figure 3.1 (C) in Section 3.1.1.1) and 'afmhot' (see Figure 4.19 in Section 4.2.2.1) were used in this project. Pixels identified as belonging to a molecule of interest were highlighted with a user-specified colour and overlaid on the processed image (see Figure 3.1 (B) in Section 3.1.1.1).

2.2.3.2 Molecular identification in TopoStats 2

Molecular identification in TopoStats 2 was achieved through methods under the 'Grains' class in the module grains.py. Potential molecules of interest were first selected as grains via height thresholding based on standard deviation; two other height thresholding methods (Otsu and absolute height) were also available and will be discussed in Section 3.1.1.2. Following the initial identification, grains that touched the border were excluded using the 'tidy_border' method, since they were incomplete and unsuitable for subsequent measurements. Further removal of unwanted grains based on size was carried out through the 'remove_small_objects' and 'area_thresholding' methods, using absolute area thresholds specified in the configuration file.

Due to differences in experimental conditions, AFM images from different datasets in the HU project were processed with different molecular identification parameters. Their values are summarised below in Table 2.2:

Table 2.2 Molecular identification parameters in TopoStats 2 used for different datasets in the HU project

Dataset	**Standard deviation height threshold**	**Lower absolute area threshold / nm^2**	**Higher absolute area threshold / nm^2**
Undamaged DNA	2.0	300	2000
Damaged DNA without HU treatment	1.0	500	2000
Damaged DNA treated with HU	1.0	300	2000

In addition to processed images with and without overlaid masks of identified molecules, TopoStats 2 was also able to output images showing intermediate results of preliminary processing (see Figure 2.1 in Section 2.2.2.2 above) and molecular identification (see Figure 2.2 below). These were all produced via the 'Images' class in the plottingfuncs.py module.

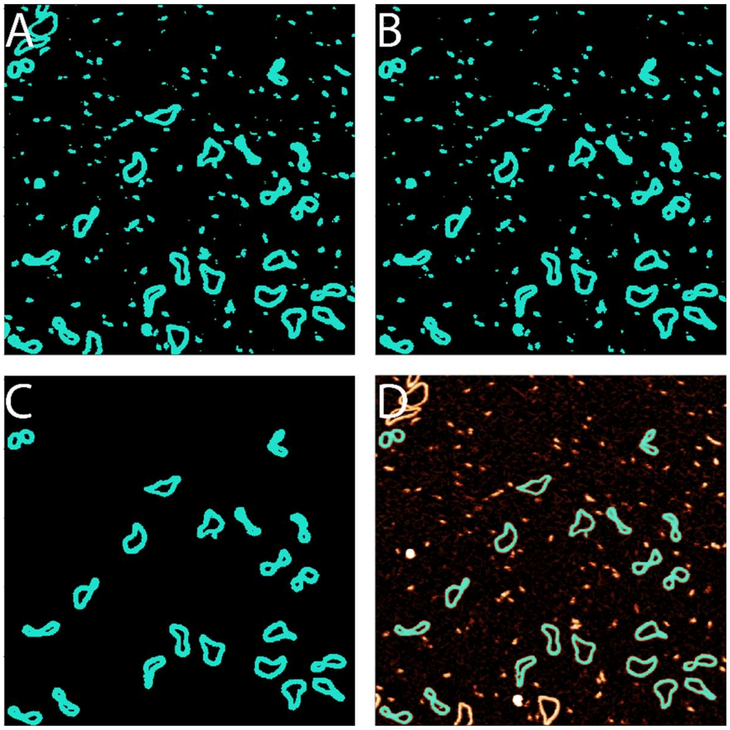


Figure 2.2 The molecular identification workflow in TopoStats 2. A: Clusters of pixels whose height values deviate sufficiently from the background are selected to be potential molecules of interest, as highlighted in cyan. B: Pixel clusters touching the border are removed with the 'tidy_border' method. C: Pixel clusters that are too small and likely to be noise are eliminated using 'remove_small_objects' and 'area_thresholding'. D: The remaining identified molecules are superimposed on the original AFM image as a mask.

2.2.4 Measurements

2.2.4.1 Basic measurements

In TopoStats 1, a list of parameters such as the area and maximum height of each identified molecule were directly measured with the 'grainanalysis' function using Gwyddion's grain value modules [160]. These were stored in the Pandas (a Python library) DataFrame 'grainstats' and then added to a list named 'appended_data', which was later exported as a .csv file using the 'savestats' function. In TopoStats 2, these measurements were performed via methods in the class 'GrainStats', located in the script grainstats.py, and the results were output into an all_statistics.csv file through the 'save_folder_grainstats' function in io.py.

Among all the basic parameters measured, the mean, median and maximum heights, the mean radius, the half lengths of the major and minor axes for the equivalent ellipse, and the volume were used to characterise the globular protein TOP1, as will be explored in Section 3.3. The maximum bounding distance (also known as maximum Feret diameter, see Figure 3.29 in Section 3.5.2.1.1 for an illustration) was employed to characterise the size of biomolecules with irregular shapes, such as NDP52 and DNA. This will be the subject of Sections 3.5.2 and 4.4.2 to 4.4.4.

Note that all measurements were based on pixels from the mask of a single identified molecule. For example, the mean height referred to the average height value for all pixels on a single identified molecule, rather than the average height of all molecules from an image, and the maximum height was the value taken from the highest pixel on the molecule. Similarly, the mean radius was calculated for each molecule as the average distance between its centre of mass and its boundary, and it did not refer to the average radius of all molecules from an image.

2.2.4.2 Tracing

The tracing process remained the same for both versions of TopoStats used in this project because the refactoring was not yet complete. Methods in the class 'dnaTrace' from the dnatracing.py script were used to transform an identified DNA molecule from a cluster of pixels to a smooth, continuous curve. Note that the naming convention for methods was different between the two versions of TopoStats, and here the format adopted in TopoStats 1 is used. The tracing workflow could be divided into five steps:

1. Skeletonisation: the pixels forming a molecule were reduced into a single-pixel-wide trace representing the shape of the DNA using the method 'getDisorderedTrace'; this trace was referred to as the skeleton of a molecule.
2. Categorisation: skeletons whose size were too small to be a legitimate DNA molecule were discarded, and the rest were categorised into linear and circular molecules by the method 'determineLinearOrCircular' based on the number of points that had only one neighbour; a circular molecule would contain no such point while a linear molecule would contain two.
3. Ordering: the pixels forming the skeleton were arranged into order following the DNA trajectory through the 'getOrderedTrace' method.
4. Fitting: 'getFittedTraces' adjusted the ordered trace to obtain a better representation of the DNA shape; this was done through taking the height of the pixels into account, instead of relying solely on the binary mask generated in the molecule identification process.
5. Splining: the fitted traces were transformed into smooth, continuous curves with the 'getSplinedTraces' method, which was particularly useful when the resolution of the image was low and the trace did not contain many pixels.

Since this project involved improvements to the tracing process, further discussions and illustrations of these steps will be reserved for Section 4.1.2.

Following the completion of tracing, methods within the 'dnaTrace' class were employed to output tracing-related data. Traces from different stages were overlaid on the original AFM image and exported as a .png file using the 'saveTraceFigures' method. Detailed information on the tracing results of a specific molecule, including a complete list of coordinates and their visualisation, could also be generated as .csv and .png files via the 'writeCoordinates' method.

2.2.4.3 Advanced measurements

The pre-existing ‘measureContourLength’ method was used to measure the contour length of each molecule by calculating the distance between each pair of neighbouring points and adding all distances together.

During this project, the ‘measureEndtoEndDistance’ method was developed to calculate the direct distance between two end points of a molecule. The ‘findCurvature’ method was created to calculate the curvature on each point of a molecule, as will be discussed in more detail in Section 5.1. The ‘saveCurvature’ and ‘plotCurvature’ methods were used to export curvature data respectively as part of a curvature_stats.csv file and as an image. The ‘analyseCurvature’ method was introduced to calculate the maximum curvature, mean curvature, and variance of curvature. In particular, the mean and variance were calculated in two different ways depending on how positive and negative curvature values were treated; this will be the focus of discussion in Section 5.1.4.2.

Bending angle measurement was developed in this project as the ‘measureBendingAngle’ method, which calculated the bending angle at a user-specified site on each molecule. A detailed description of the development process can be found in Section 5.3.

In TopoStats 1, the tracing-dependent advanced measurements were exported into a file named tracestats.csv. This was carried out via the ‘traceStats’ class in dnatracing.py. In TopoStats 2, these measurements were combined with the basic measurements into a single all_statistics.csv output file.

2.3 Data visualisation

The distribution of measured data was analysed and visualised through the plotting_old.py script in the ‘JeanDu/BendingAngle-Upgraded’ branch. The .csv files generated from the image analysis process were used to create histograms, kernel density estimation (KDE) plots and violin plots that featured the distribution of a specified parameter measured by TopoStats. The ‘computeStats’ function was used to calculate the peak and standard deviation of the distribution if required.

To produce a plot, the paths to the relevant .csv data files were first entered into the plotting_config.yml file. The requirements for each plot, including the parameter chosen (e.g. maximum height, contour length, maximum curvature), the plot type, the ranges of the X and Y axes, and the number of bins for histograms, were subsequently specified in the ‘plots’ section of the same file. The violin plot shown in Figure 3.27 (A) in Section 3.5.1.2.3 was produced by the function ‘plotviolin’, while the functions ‘plotdist’ and ‘plotdist2’ generated histograms and KDE plots overlaid on the same figure.

Chapter 3 Identifying and characterising molecules of interest

3.1 Introduction to automatic analysis of AFM images

3.1.1 Workflow and challenges of automatic molecule recognition

3.1.1.1 Overview of automatic molecule recognition workflow

The type of AFM images used in this project can be thought of as a 2D array of pixels each carrying a height value. These values are visually represented by colours, and as shown in Figure 3.1 (C) below, in the colour scheme used by this project, a brighter colour corresponds to a higher value. The automatic molecule recognition workflow builds on the preliminary processing stage, which deals with artefacts and produces a processed image, as exemplified by Figure 3.1 (A): the dark region is the background, while the brighter features are either molecules of interest or impurities.

These features can be distinguished from the background thanks to two factors: firstly, when collecting the data experimentally, the biomolecules are placed on a mica surface, which is atomically flat [130]; secondly, at the end of the preliminary processing stage, a height value is added to or subtracted from all pixels on the image, so that the background is set to 0 nm, which means that molecules can be identified based on their relative height from the background, ensuring consistency across different images.

As of result of these two factors, groups of pixels next to each other with their height deviating from zero can be automatically recognised as a single unit. This process of automatic molecule recognition is also referred to as masking, because a 'mask' is applied to cover all selected pixels, as demonstrated by Figure 3.1 (B), while the recognised pixel clusters are known as 'grains'.

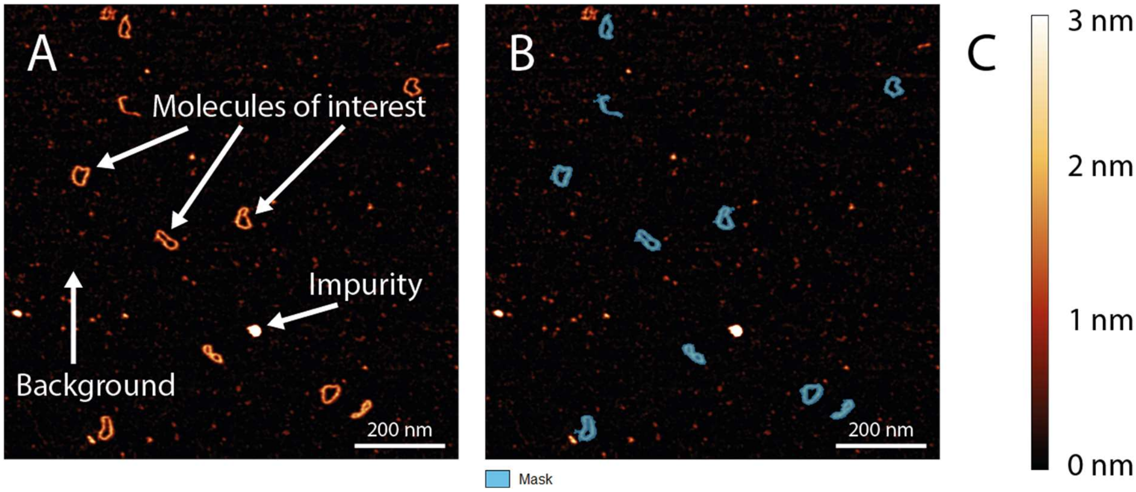


Figure 3.1 Automatic recognition of DNA minicircles from an AFM image. A: An example processed AFM image of DNA minicircles with different features annotated. B: The same image with molecules of interest (DNA minicircles) automatically recognised from the background and masked in blue. C: A

height scale demonstrating how colours on these processed images represent different height values. The same height scale is used for all AFM images in this chapter.

Since noise and impurities are also present on AFM images as clusters of pixels higher than the background, they are often mistakenly identified as molecules of interest and included in the mask. This means that following the initial selection of grains, the mask needs to be further refined to enable the exclusion of unwanted molecules. The initial selection process and the subsequent filtering process are based on height and area criteria, which will be discussed in more detail in the subsequent sections.

Compared to traditional methods of manually identifying molecules of interest, the automatic molecule recognition workflow is significantly faster while retaining a high accuracy. For the example AFM image in Figure 3.1, the masks produced from manual and automatic methods have a high level of overlap, which can be quantified by a Jaccard index of 0.82, as shown in the 'high quality' row of Figure 3.2 below. This index is the ratio of the number of pixels selected in both masks to the number of pixels selected in at least one mask, and a value of 1 represents complete overlap while 0 represents no overlap.

On the other hand, automatically identified masks exhibit less overlap with manually selected masks for images with lower resolution and more noise, such as the low-quality example image in Figure 3.2, where the Jaccard index for the two masks is 0.49. This is because selecting suitable height and area thresholding criteria is more difficult when the image quality is low, and as a result, established automatic molecule recognition parameters based on high-quality data may not be applicable to low-quality data.

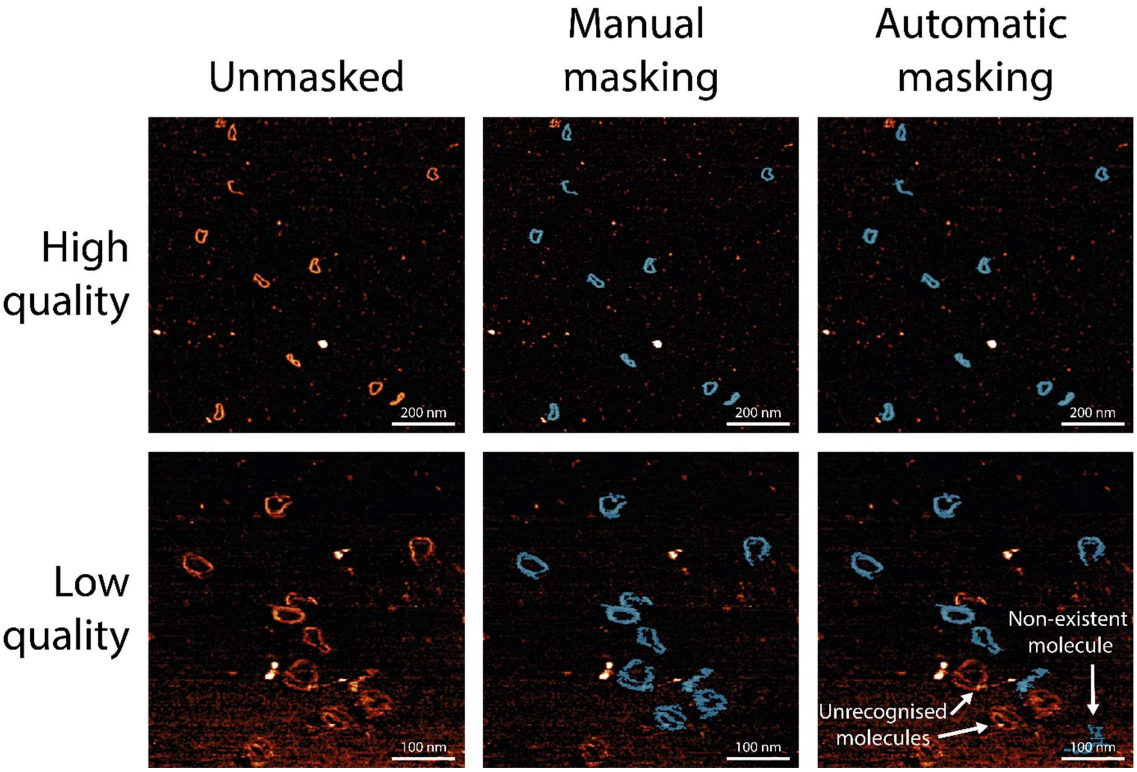


Figure 3.2 Comparison between manual and automatic masking for high-quality and low-quality AFM images of DNA minicircles. For the example high-quality image, similar masks are produced

from manual and automatic molecular identification, with a high level of overlap and a Jaccard index of 0.82. For the example low-quality image, automatic masking fails to recognise certain molecules and mistakenly identifies a non-existent molecule, resulting in less overlap with manual masking and a lower Jaccard index of 0.49.

In addition, while the functions to automatically select and filter molecules of interest are available from prior work, they have not been developed for or used on single protein molecules, which bring additional challenges through their small size. Because of this, different height and area thresholding methods are compared in this thesis to determine how selection criteria should be applied to identify molecules of interest while excluding other molecules or impurities. The methods for thresholding will need to be adapted based on the nature of the data to establish a framework under which automatic recognition algorithms can be systematically applied to molecules with different sizes and shapes.

3.1.1.2 Identifying molecules of interest by height

The first step of identifying molecules of interest is to select grains that exceed a height threshold. This is relatively straightforward for DNA molecules, which have an established height range and can be easily distinguished from the background and other features. The selection of proteins is however more difficult, because their height varies from species to species, and they are often imaged alongside other types of molecules. Therefore, when adapting automatic identification algorithms for proteins, it is worth examining several different ways of height thresholding, in order to determine which one is the most suitable.

3.1.1.2.1 Otsu

Otsu is a binary height thresholding method that automatically sorts all pixels into foreground and background based on their height value [15]. Due to the completely automatic nature of the thresholding, this method has the advantage of not requiring any user input or prior knowledge of the features of interest. It works well for images that contain only molecules of interest, and do not have much noise or other molecules. This is sometimes the case for images of DNA minicircles, but less likely to be applicable to images with more impurities. Moreover, when investigating DNA-protein interactions, protein molecules and can be imaged together with DNA, so they need to be distinguished not only from the background but also from DNA. In situations like this, the Otsu thresholding method is completely non-viable, as the height distribution is not binary.

3.1.1.2.2 Absolute height

Another candidate thresholding method is to select pixels that are higher than a user-provided absolute height value. This ensures that noise and impurities lower than the molecule of interest can be easily excluded. However, it does require prior knowledge about the molecule of interest, and the user often needs to experiment with different height thresholds, in order to identify the cut-off value that selects all molecules of interest while excluding other features as much as possible.

Another major limitation of this method is that, the ideal threshold can vary from image to image, even if they are for the same type of sample, since the absolute height values in AFM images could be affected by certain experimental factors, such as the condition of the tip or the immobilisation technique used. Therefore, using absolute values for height thresholding could be problematic when mass processing data taken over a long period of time and under different conditions.

3.1.1.2.3 Standard deviation

A third option of height thresholding is the standard deviation method, which selects pixels based on how far their height deviates from the mean height of the image. To achieve this, the mean height and standard deviation of all the pixels on the image are first calculated. The standard deviation is then multiplied by a user-provided multiplier, and the result determines the height threshold: pixels higher than the mean plus the result of the multiplication are selected, and clusters of selected pixels are identified as features of interest.

Similar to absolute height thresholding, the standard deviation method also requires user input, as well as some trial and error to find the best trade-off between including molecules of interest and excluding unwanted features. However, this method is less dependent on consistency in experimental conditions, and the multiplier obtained through testing on a small subset of images can be easily applied to a wider range of data, thus shortening the time required to determine the thresholding values.

3.1.1.2.4 Conclusions on height thresholding

After comparing different thresholding methods, it can be remarked that Otsu is problematic when there are multiple types of molecules present, while absolute height thresholding is not ideal when the condition varies between different images. Standard deviation should therefore be set as the default choice, as it covers the widest range of situations.

The other two methods can still be used under specific circumstances: Otsu is the best option when there is a clear binary distinction of the sample and the background; similarly, absolute height thresholding is useful when all data are imaged under similar conditions.

3.1.1.3 Filtering molecules of interest by area

As mentioned above in Section 3.1.1.1, it is often the case that not all identified grains are molecules of interest. For example, large aggregates or oligomers could be selected because they are higher than the height threshold. This means that further filtering based on area is needed to remove unwanted grains. Different area thresholding methods are also examined, so that their suitability can be compared.

3.1.1.3.1 Relative area

As using relative criteria proves to be the most successful for height thresholding, we first look into the selection of features of interest based on relative area. It is however worth noting that this thresholding is based on the area of grains, and not the area of pixels, as all pixels on the same AFM image have the same area. This distinction is important because relative thresholding relies on a sensible value for the mean area, and since an image contains many pixels but only a few grains, the mean area of grains can be more easily skewed by extreme values, which are usually caused by impurities or unwanted molecules.

This can be demonstrated by Figure 3.3 below, where the selection of proteins based on mean area is hindered by the presence of a DNA molecule when relative area thresholding is used:

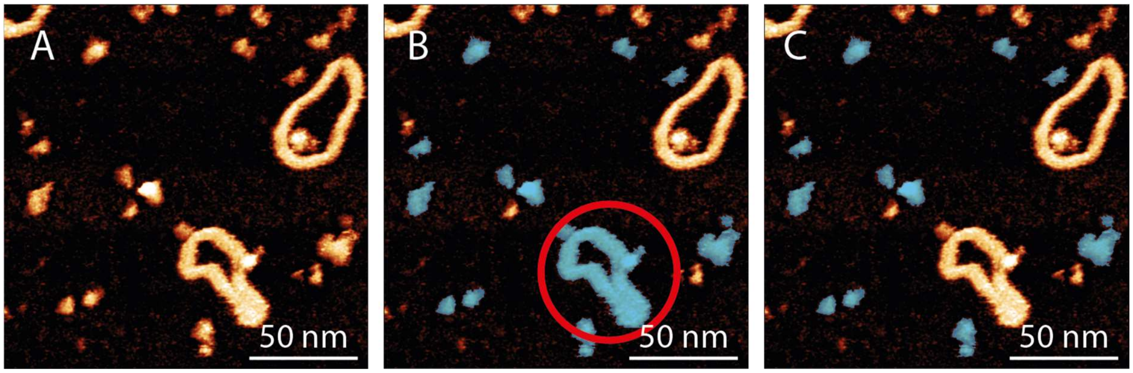


Figure 3.3 The automatic selection based on mean area can be affected by grains of extreme area values. A: Example AFM image containing TOP1 protein and DNA. B: The mean area of masked grains is skewed by the presence of the large DNA (circled in red). This makes it difficult to filter out the DNA using area criteria based on the mean and standard deviation. C: The desired result where proteins are masked and the DNA molecule is excluded.

Since the DNA molecule has a much larger area than the protein molecules in this image, the mean grain area is much higher than the area of a protein molecule. This means that in order to select for proteins based on mean area and standard deviation, the upper threshold needs to be below the mean. This criterion is not compatible with selecting proteins of interest in other images unaffected by grains of extreme sizes, where the upper threshold should be above the mean area. Using mean area as a reference point for filtering could therefore cause problems in mass processing.

A better alternative is to filter molecules based on the median area instead of the mean. For example, the user can initially decide that molecules between 0.5 and 1.5 times the median area are to be kept, then adjust the values based on the results of masking.

However, while the median area is less affected by extreme values, it can still be unreliable if the image contains three or fewer grains. In this case, multipliers suitable for other images could result in unwanted molecules being selected, or molecules of interest being excluded.

3.1.1.3.2 Absolute area

The second method is to filter grains based on absolute area, which means only molecules with an area falling into a user-defined absolute range are kept. For example, if the user is selecting for DNA molecules that are 2 nm wide and 150 nm long, they can use an absolute range of 200-400 nm^2. Similar to thresholding with absolute height, this method could fail when the data is taken under various different experimental conditions, as the same molecule could appear to have different areas when imaged under different resolutions or using different AFM tips.

On the other hand, compared to using the median value, selecting molecules of interest based on absolute area is unaffected by the presence of unwanted grains with different sizes, so this method is more suitable for noisy data where large contaminants, aggregates or oligomers are commonly present.

3.1.1.3.3 Conclusions on area thresholding

While Section 3.1.1.2.4 has concluded that relative height thresholding should be the default method for molecular identification based on height, due to its suitability for most situations, the same cannot be said of filtering based on relative area, because relative area is based on a small set of identified grains, whereas the relative height is based on the height distribution of every single pixel on the image. As a result, the mean or median grain area on a specific image does not always reflect the size of a molecule of interest.

The choice of the area thresholding method is therefore not clear cut, as it is dependent on the nature of the data. Absolute area should be used when the images contain many unwanted features of very different size from the molecules of interest, and relative area is more suitable when the imaging conditions, such as the tip used, differ significantly across images.

3.1.1.4 Future work

The aforementioned height thresholding and area filtering methods are sufficient in most cases. However, to further improve the accuracy of molecular identification and increase the applicability of automatic molecule recognition, more methods can be developed in the future to complement the existing ones.

3.1.1.4.1 Filtering molecules of interest by height

As previously discussed, clusters of pixels above a user-determined height threshold are identified as grains of interest. These grains are subsequently filtered based on area criteria. This workflow means that unwanted grains such as impurities that are much higher than molecules of interest but similar to them in area cannot be removed. To address this potential issue, it would be useful to introduce further height based filtering criteria, so that certain grains can be removed for being too high.

Since this step will take place after the grains are identified, it will be based on grains instead of pixels, and will therefore be more similar to the area selection process. Both absolute and relative thresholding methods should be considered when developing this feature in the future.

3.1.1.4.2 Filtering molecules of interest by shape

The current selection methods available to this project are all based on height or area. To further distinguish between molecules that are similar in both height and area, more advanced techniques such as machine learning can be employed to automatically classify and select grains based on shape.

There is ongoing work by another member of the research group on incorporating machine learning methods [161], though it is not within the scope of this project. It should be noted that these methods also require user input, either through manually labelling grains in the case of supervised classification, or by verifying the results and selecting the category of interest in the case of unsupervised classification.

While shape-based filtering methods should be added to the molecular identification pipeline in the future, they are best reserved as an option for particularly challenging cases, instead of a default step, because a balance needs to be struck between the accuracy of selection and the computational resources required: if molecules of interest can be satisfactorily identified based on height and area only, it is more efficient to use traditional selection methods, even when more advanced methods become available.

3.1.2 Workflow and challenges of automatic measurement

3.1.2.1 The benefits and limitations of automatic measurement

Once the molecules of interest have been recognised from the background, they can be automatically measured almost instantly in subsequent steps, when all images are batch processed by an automatic analysis program instead of examined individually. Not only does this approach save a significant amount of time compared to manual measurement, but it also reduces selection bias and human error. In addition, many parameters that are not straightforward to measure by hand, such as the mean height of all pixels belonging to an identified molecule, can be easily calculated as part of an automatic workflow. These parameters introduced through automatic measurement can give us more information on how the conformation of biomolecules varies within the population, allowing us to better quantify heterogeneity and make the most use out of AFM as a technique that visualises differences between samples of the same species. Furthermore, as automatic measurement enables the processing of a large amount of data, it allows for easy comparison between different groups of samples. This could be used to investigate how different experimental conditions or different treatments affect the conformation and structure of molecules of interest.

However, it should be noted that automatic measurement is not completely free from bias: while the results of the measurement are not affected by the preferential selection of molecules that fit better with proposed theories, as in the case of manual measurement, they can be affected by errors introduced during the previous step of masking, where molecules are identified from the background. This means that accurate masking is essential to the reliability of automatic measurement. To evaluate the masking quality, measurements made from different masks can be compared with each other, and also with data obtained from other sources, so that potential issues can be identified and addressed.

3.1.2.2 Parameters that can be measured

To quantitatively analyse the conformation of molecules of interest and to evaluate the accuracy of molecular identification, we first look into the range of parameters that can be automatically measured. Since AFM provides a height map where each individual molecule or oligomer adsorbed on a smooth mica surface is represented by an assortment of neighbouring pixels with varying heights, the variables chosen to measure features of interest are the height based in the direction perpendicular to the surface (the Z direction), the size based in the surface (the X-Y plane), and the volume based on a combination of the two.

These measurements allow us to obtain a better understanding of protein structure, and the extent of conformational variation between different molecules. While the structure of many proteins can be characterised through other techniques such as X-ray diffraction, the automatic analysis of AFM images will provide additional information on the variability in protein structure, including oligomerisation and aggregation, in a solution-based physiological environment, although the conformation of biomolecules may be affected by the process of adsorption to the mica surface.

3.1.2.2.1 Height

In a typical automatic characterisation program, there are three most common measurements that can be made in the Z direction for each grain: the mean height, the median height, and the maximum height.

For single globular proteins without visible internal structural variation such as TOP1, the measured mean and median heights on the same molecule should be similar, with both values lower than the maximum height. In addition, the distributions of these three height measurements across the population are expected to be similar. Deviation from this would indicate inaccuracy in the masking stage.

For more structurally complex proteins, information on their sub-molecular structures can be obtained by comparing the difference in distribution of the three height measurements. This will be further discussed in Section 3.5.1 using the example of the protein NDP52.

3.1.2.2.2 Size and shape

While characterising biomolecules in the Z direction is achieved through height measurements, characterising them in the X-Y plane is less straightforward, as there are more parameters that can be measured. The latter is also less accurate than the former, because data on the X-Y plane can be affected by tip convolution.

In the case of globular proteins, both the area and the radius are reflective of their size on the X-Y plane. It should be noted that proteins described as globular are not a perfect globe (for an example, see TOP1 molecules in Figure 3.5 below), and their projection on the X-Y plane could be closer to an ellipse than a circle. This introduces difficulties in finding the right parameters that describe the proteins. One possible solution is to measure the mean, median and maximum radius of each identified molecule. Another approach suitable for some molecules is to approximate them as ellipses, and measure the half length of their major and minor axes. The combination of the aforementioned measurements can provide information on both the size and the shape of globular molecules of interest.

These measurements can be applied to the majority of proteins that interact with DNA, though there are exceptions where the protein presents a more complex shape that cannot be approximated as a globe or an ellipsoid. An example of this is the dumbbell-shaped NDP52, and the parameters needed for its characterisation on the X-Y plane will be investigated in Section 3.5.2.

3.1.2.2.3 Volume

Since AFM images are 2D representations of 3D structures, molecules of similar sizes might appear to have different height and area values depending on how they are attached to the surface. The volume can therefore better represent the size of a globular protein. In automatic characterisation algorithms, it is obtained by summing the local volumes at each pixel, which are in turn calculated by multiplying the height of the pixel with the area represented by a pixel.

Similar to measurements of radius and area, while the volume can be used to characterise the size of globular proteins such as TOP1, it does not capture the 3D conformational variation of proteins with visible sub-molecular structures such as NDP52.

3.2 Identifying single globular proteins: TOP1

3.2.1 Background on TOP1

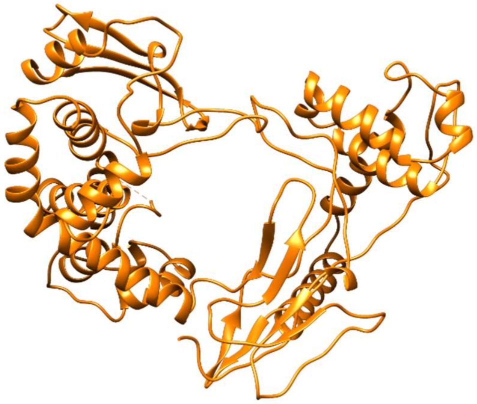


Figure 3.4 The crystal structure of human topoisomerase I (TOP1) [162], *visualised through the software Chimera* [163].

DNA conformation can be altered through protein interaction, and to understand the mechanism of such interaction, it is important to study the structure of proteins. One DNA-manipulating protein of interest is the human topoisomerase I (TOP1), which, like other topoisomerases, is an enzyme essential to biological processes such as transcription and DNA replication: the two strands of DNA are separated during these processes, and the double-helical structure of DNA means that such a separation causes increased twisting in the helix, thereby generating stress [89]. To preserve the structural stability of DNA and ensure the continuation of these biological processes, TOP1 releases the stress by catalysing the breaking and re-joining of one DNA strand [141].

The interaction between DNA and TOP1 is of particular interest to us because TOP1 can be used a target in cancer treatment [164]: a group of chemotherapeutics known as TOP1 inhibitors, such as camptothecin and its derivatives, are clinically employed to kill cancer cells by supressing TOP1 activity. However, these drugs also cause side effects due to their off-target toxicity, while their mechanism of inhibiting TOP1 is not fully understood. To explore the clinical potential of TOP1 inhibitors and increase its specificity in future drug designs, it is important to obtain a better understanding of TOP1 structure, which can be achieved through the automatic identification and measurement on AFM images.

In addition, since TOP1 exhibits a regular and globular shape, it is relatively straightforward to characterise, and the work on TOP1 can serve as the foundation for the automatic recognition of other molecules with more complex shapes.

3.2.2 Challenges of identifying and selecting for globular proteins

Characterising the structure of biomolecules gives us insight on their functions and how they interact with other molecules. For both DNA and protein, this characterisation starts by recognising them from an AFM image; however, the structures of DNA and proteins exhibit significant differences: DNA are flexible linear molecules that can form curves or loops, while proteins are typically spherical or composed of spherical units. This means that compared to DNA, it can be harder to distinguish proteins from noise and impurity, which also tend to be spherical. In addition, the size, shape and conformation of proteins are not only different from those of DNA, but also more diverse and variable, which makes it difficult to establish a fixed set of identification criteria like in the case of DNA. Another factor to take into account is that many proteins are also prone to oligomerisation, and to ensure the accuracy of automatic measurements, the oligomers need to be separated from monomers.

In summary, the main challenge of automatic selecting a globular protein such as TOP1 lies not in identifying the protein from the background, but in distinguishing it from other molecules. Here we discuss how we can apply the selection methods explained in Section 3.1.1 and adjust relevant parameters, so that TOP1 and TOP1-like proteins can be identified and unwanted features can be excluded.

3.2.2.1 Distinguishing TOP1 from other proteins such as BSA

One of the most common molecules that appear alongside TOP1 in the provided AFM data is bovine serum albumin (BSA), which is used for TOP1 stabilisation. It can be difficult to automatically distinguish TOP1 from BSA, since both are proteins with a globular structure. However, these two proteins still exhibit observable differences. As shown below in Figure 3.5, in a typical AFM image of TOP1, there are many different globular structures. Some of them are slightly larger, higher and more globular, and these can be identified as TOP1 (blue squares); most of them are smaller and more ellipsoidal, and these are identified as BSA (yellow circles).

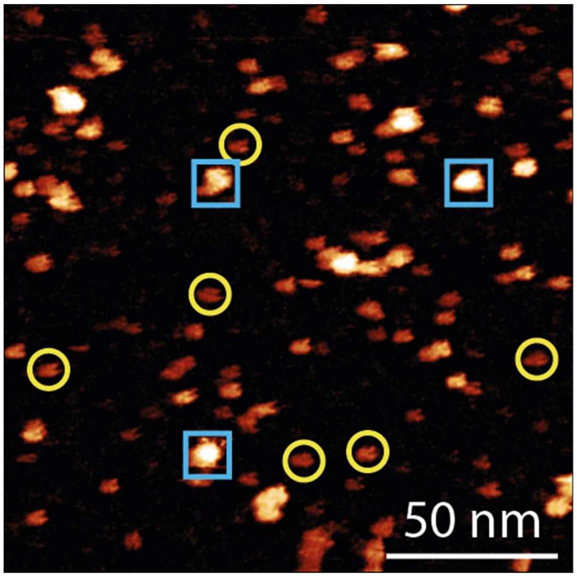


Figure 3.5 Typical AFM image of TOP1 that contains BSA. Examples of TOP1 molecules are marked with blue squares, while examples of BSA molecules are marked with yellow circles.

Based on these observable differences, automatic selective identification can be achieved through both increasing the height threshold during initial selection and narrowing the area thresholds in the subsequent filtering process, so that BSA can be excluded for being too small.

The ideal thresholding parameters need to be found out through trial and error, which involves repeatedly adjusting the parameters and verifying the quality of the resulting masks. To investigate whether this process can be automated to some extent, a two-stage parameter sweep was carried out on the example image above, comparing the numbers of grains identified from the image under different parameters, and the results are shown below in Figure 3.6.

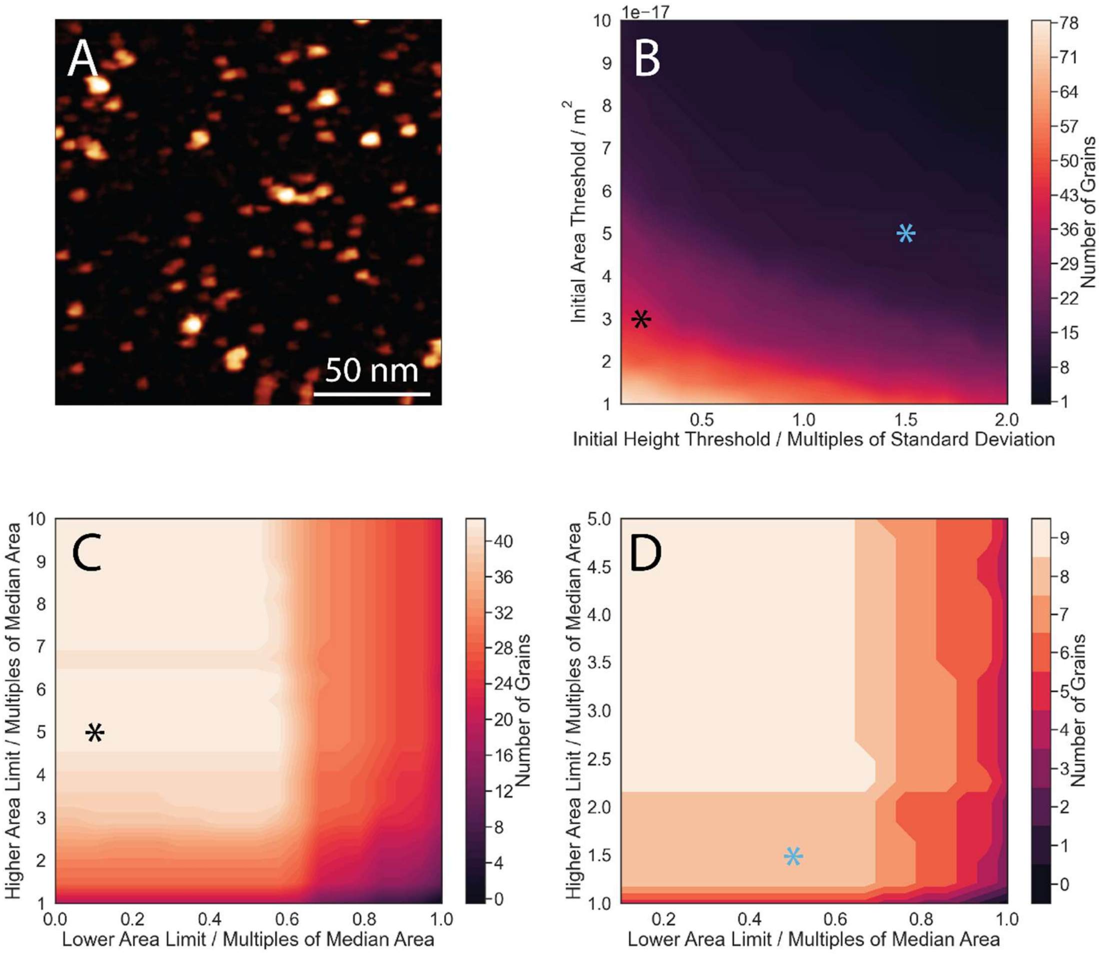


Figure 3.6 Parameter sweep on an example AFM image containing TOP1 and BSA. A: The unmasked example AFM image. B: During the first stage of the parameter sweep, the initial minimum height and area thresholds are gradually adjusted and the corresponding numbers of grains identified are shown; as there are no distinct regions, the optimal parameters to select for all proteins (initial height threshold = 0.1, initial area threshold = 3 × 10^{-17} m^2 or 30 nm^2, black asterisk) and to select for TOP1 only (initial height threshold = 1.5, initial area threshold = 5 × 10^{-17} m^2 or 50 nm^2, blue asterisk) are determined through manual evaluation. C: Applying the initial thresholding parameters to select for all proteins, the second stage of the parameter sweep is conducted by adjusting additional area-based filtering parameters; the optimal values (lower area limit = 0.1, higher area limit = 5.0, black asterisk) are selected based on results from the parameter sweep and manual verification. D: Applying the initial thresholding parameters to select for TOP1 only, the second stage of the parameter sweep is conducted by adjusting additional area-based filtering parameters; the optimal values (lower area limit = 0.5, higher area limit = 1.5, blue asterisk) are selected based on results from the parameter sweep and manual verification.

During the first stage, the initial thresholds for the minimum height and area requirements were gradually adjusted across a wide range. As shown in Figure 3.6 (B), the number of grains masked decreased as the initial height and area thresholds increased, but the parameter sweep did not result in distinct, clearly separable regions. Consequently, suitable parameters to mask all proteins (black asterisk) and to mask only TOP1 (blue asterisk) were chosen manually.

During the second stage (Figure 3.6 (C) for all proteins and (D) for TOP1 only), the thresholding parameters from the first stage were fixed at the previously chosen values, and a different set of area-based filtering parameters were gradually changed. While some distinct regions could be identified, especially when selecting for TOP1 only, these regions could not be accurately interpreted without manual observation and verification. This meant that the optimal area-based filtering thresholds to select for all proteins (black asterisk in Figure 3.6 (C)) and to exclusively select for TOP1 (blue asterisk in Figure 3.6 (D)) also needed to be chosen manually.

Both stages of the parameter sweep show that the selection of suitable thresholding parameters cannot be automated for the current example image, and that manual assessment of masking quality as demonstrated below in Figure 3.7 is necessary. This may be because the image contains noise and impurities of different sizes in addition to the two protein species.

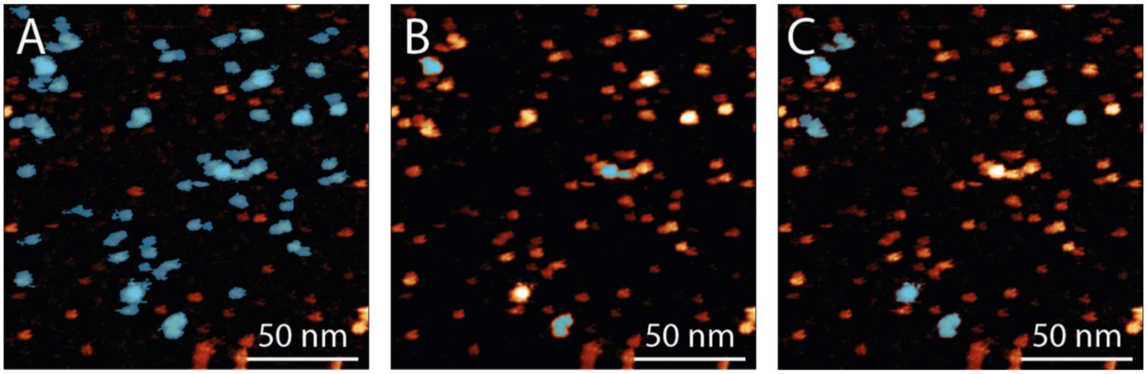


Figure 3.7 Adjusting parameters to select TOP1 and exclude BSA through trial and error. A: When the thresholds are too permissive, many unwanted molecules are selected. B: When the thresholds are too restrictive, too few molecules are selected and some TOP1 molecules are missed. C: The right thresholds are found to achieve optimal selection accuracy.

To verify whether the aforementioned example image is representative of the entire dataset, the two-stage parameter sweep process was applied to a different image with lower resolution, and the results are shown in Figure 3.8 below.

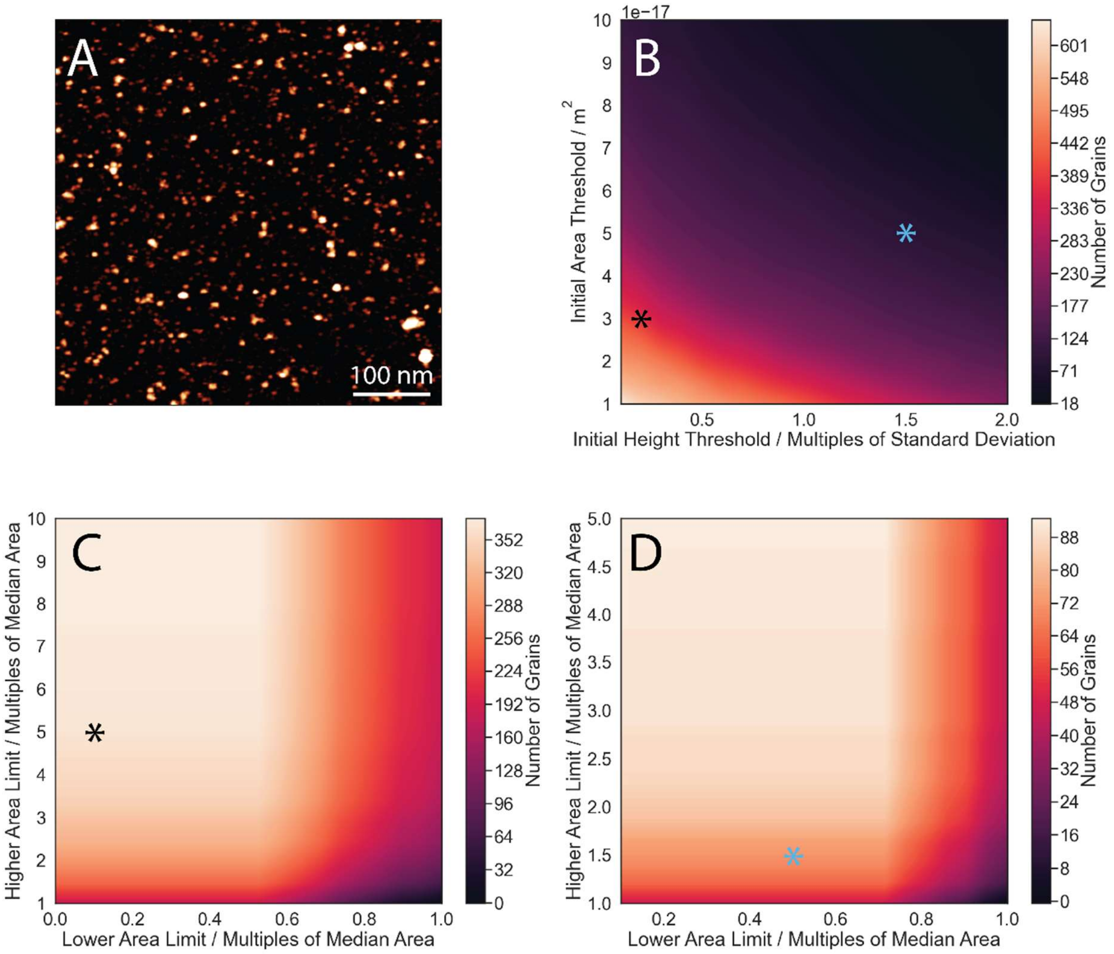


Figure 3.8 Parameter sweep on a lower-resolution example AFM image containing TOP1 and BSA. A: The unmasked example AFM image. B: During the first stage of the parameter sweep, the initial minimum height and area thresholds are gradually adjusted and the corresponding numbers of grains identified are shown; no distinct regions are identified, but manual verification confirms that the thresholds from Figure 3.6 to select for all proteins (initial height threshold = 0.1, initial area threshold = 3 × 10^{-17} m^2 or 30 nm^2, black asterisk) and to select for TOP1 only (initial height threshold = 1.5, initial area threshold = 5 × 10^{-17} m^2 or 50 nm^2, blue asterisk) are applicable to this image. C: Applying the initial thresholding parameters to select for all proteins, the second stage of the parameter sweep is conducted by adjusting additional area-based filtering parameters; while there are no distinct regions, manual verification confirms that the optimal values (lower area limit = 0.1, higher area limit = 5.0, black asterisk) from Figure 3.6 are applicable to this image. D: Applying the initial thresholding parameters to select for TOP1 only, the second stage of the parameter sweep is conducted by adjusting additional area-based filtering parameters; while there are no distinct regions, manual verification confirms that the optimal values (lower area limit = 0.5, higher area limit = 1.5, blue asterisk) from Figure 3.6 are applicable to this image.

In both the first stage involving initial height and area thresholds (Figure 3.8 (B)) and the second stage involving area-based filtering parameters (Figure 3.8 (C) for all proteins and (D) for TOP1 only), the parameter sweep plots for the lower-resolution image present a similar pattern to the plots for the higher-resolution image, with the main difference being higher numbers of identified grains during both stages. This is congruent with the observation that the image with lower resolution contains more proteins. Furthermore, while distinct, clearly separable regions are not observed from the parameter sweep plots, it can be confirmed through manual assessment that previously identified optimal thresholds to select for all proteins (black asterisks on Figure 3.8 (B) and (C)) and to select for TOP1 (blue asterisks on Figure 3.8 (B) and (D)) are still applicable to this image.

Overall, these results indicate that the automatic determination of suitable molecular identification thresholds through parameter sweeps is not viable for this dataset, and that the number of grains identified is not a reliable indicator of the masking quality. If combined with manual observation, the results from parameter sweeps can however serve as a starting point for the selection of suitable thresholds when analysing a new dataset.

3.2.2.2 Distinguishing TOP1 from DNA

To investigate the effects of TOP1 on DNA conformation, DNA is often imaged alongside TOP1. This presents an additional challenge in selecting for TOP1, though unlike BSA, the shape and size of DNA are both considerably different from those of TOP1, so the automatic exclusion of DNA is more straightforward in general.

Since DNA is longer and lower than TOP1, it can be filtered out by reducing the area threshold and raising the height threshold. Problems may arise when a TOP1 molecule is found next to a DNA molecule, and the two are identified as one single grain, resulting in the TOP1 not being identified. This can sometimes be dealt with by further raising the height threshold, so that the gap between the DNA and the TOP1 is no longer included in the mask. In some cases, no such gap exists because the TOP1 is attached to the DNA, as circled in blue in Figure 3.9 (A). This means that there is no easy way to separate the TOP1 from the DNA using the current methods available to this project, and the TOP1 molecules have to be excluded from the analysis, as shown in Figure 3.9 (B).

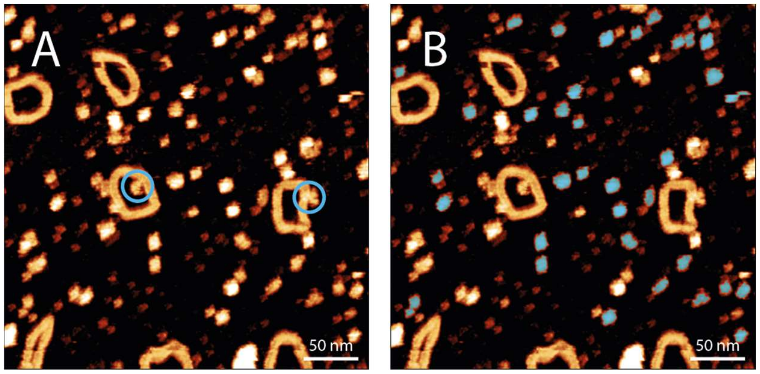


Figure 3.9 AFM images containing TOP1 and DNA minicircles. A: Some TOP1 molecules are attached to DNA, as circled in blue. B: Most TOP1 molecules are masked in blue, but the TOP1 molecules attached to DNA cannot be masked, because they would be identified as part of the DNA, rather than separate molecules; as a result, they are excluded from subsequent analysis.

To investigate whether the thresholds separating TOP1 from DNA can be determined via an automatic process, the two-stage parameter sweep was also carried out on example images containing DNA, as shown below in Figure 3.10.

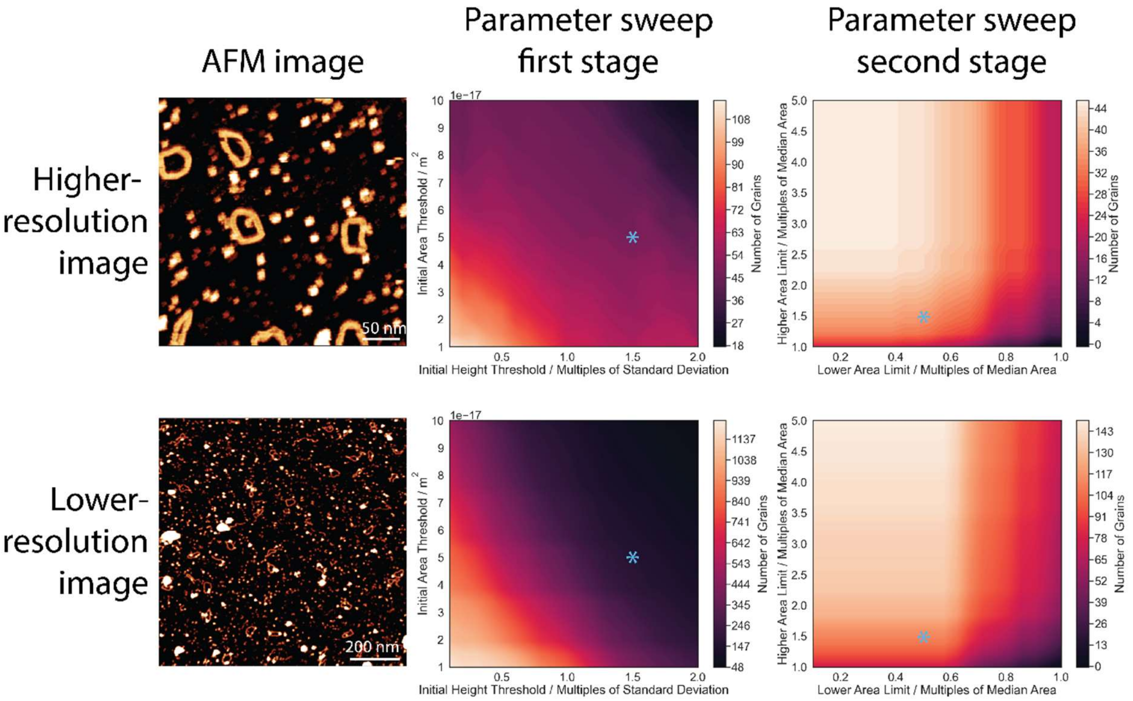


Figure 3.10 Parameter sweep on two example AFM images containing TOP1, BSA and DNA. During the first stage of the parameter sweep, the initial minimum height and area thresholds are adjusted and the corresponding numbers of grains identified are shown; during the second stage, the initial minimum height and area thresholds are fixed, while the additional area-based filtering thresholds are adjusted. Similar to the images without DNA shown in Figures 3.6 and 3.8, suitable parameters cannot be determined solely based on results from the parameter sweep due to a lack of distinct regions. It can be confirmed from manual verification that the parameters used for images without DNA to select for TOP1 during both the first stage (initial height threshold = 0.1, initial area threshold = 3 × 10^{-17} m^2 or 30 nm^2, blue asterisk) and the second stage (lower area limit = 0.5, higher area limit = 1.5, blue asterisk) are applicable to images containing DNA.

The results reveal that different regions of the parameter sweep plots are not clearly distinguishable from each other and cannot be directly mapped to different molecule species. Through manual assessment, it can however be confirmed that the previous thresholds (blue asterisks) to separate TOP1 from BSA are still applicable here, and can be used to mask TOP1 while avoiding DNA.

It is worth noting that the parameters identified in this project are specific to TOP1, and cannot be directly applied to other globular proteins in future projects. The methods employed here to determine suitable thresholds that exclusively select for TOP1, including both manual observation and automatic parameter sweeps, will however be useful in future projects, though their reliability will depend on the size and height of the protein in question, and the image quality of the dataset. On the other hand, if the protein happens to be similar to DNA in both area and height, the existing parameters will be insufficient, and new approaches such as machine-learning-based classification will be needed.

3.2.2.3 Distinguishing TOP1 from aggregates

Many proteins are known to form oligomers and aggregates, as shown in Figure 3.11. These features should be discarded when selecting for protein monomers, because their inclusion would interfere with subsequent measurements and statistical analysis.

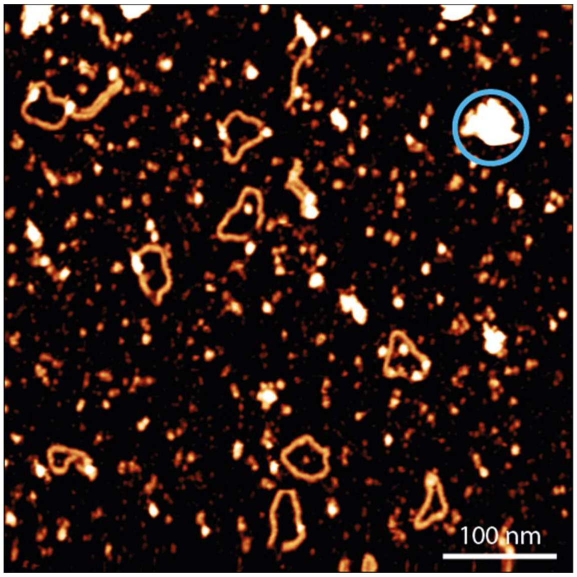


Figure 3.11 Example image of TOP1 and DNA that contains an aggregate circled in blue.

By lowering the threshold for the maximum allowed area, molecules that are too large can be removed from the masking, and aggregates can hence be filtered out. A potential threshold for maximum height as discussed in Section 3.1.1.4.1 will also be helpful, since it is possible that some oligomers or aggregates can be formed by proteins stacking in the Z direction perpendicular to the imaging surface, and they cannot be distinguished based on area alone.

3.2.2.4 Distinguishing TOP1 from miscellaneous impurities

Apart from the molecules discussed above, AFM images can contain a wide range of other impurities, which can be noise, dirt, or other known molecules. As they come from different sources, they can take different size, height and shape.

There is no one-size-fit-all solution to deal with all impurities, since the parameters and thresholds need to be changed depending on the nature of the impurities. When dealing with a large amount of data, the best approach is to test different parameters on a few selected images first.

This approach is not always successful, because the parameters identified through testing on a few representative images might not always work on every single image in the dataset, and the available filtering methods might not be sufficient to get rid of all the impurities. The limitation from the quality of the data and the methods available means that workarounds are sometimes needed, such as the exclusion of certain images that are hard to process. It is also useful to deal with impurities incorrectly identified as molecules of interest by excluding them in future measurement steps. For example, if there is a dirt particle that is higher than TOP1 but cannot be filtered out during the identification process due to the lack of a maximum height threshold, it can be excluded when performing measurements during later steps of the workflow.

3.2.3 Results

After trial and error, suitable height and area parameters were determined, so that molecules of interest could be selected. The process of deciding appropriate parameters was aided by automatic parameter sweeps, but manual observation remained essential for the current dataset. As shown in Figure 3.12 below, for a given image of TOP1 that contained other molecules, all proteins could be selected from the background and masked; this mask could be further refined to select for only TOP1, excluding other proteins, aggregates and other impurities. The masking quality for all images in the dataset was reviewed manually, and four example images of different resolutions with and without DNA were further examined in detail. The results of the manual assessment confirmed that the same set of parameters could be applied to different images to achieve the same effect.

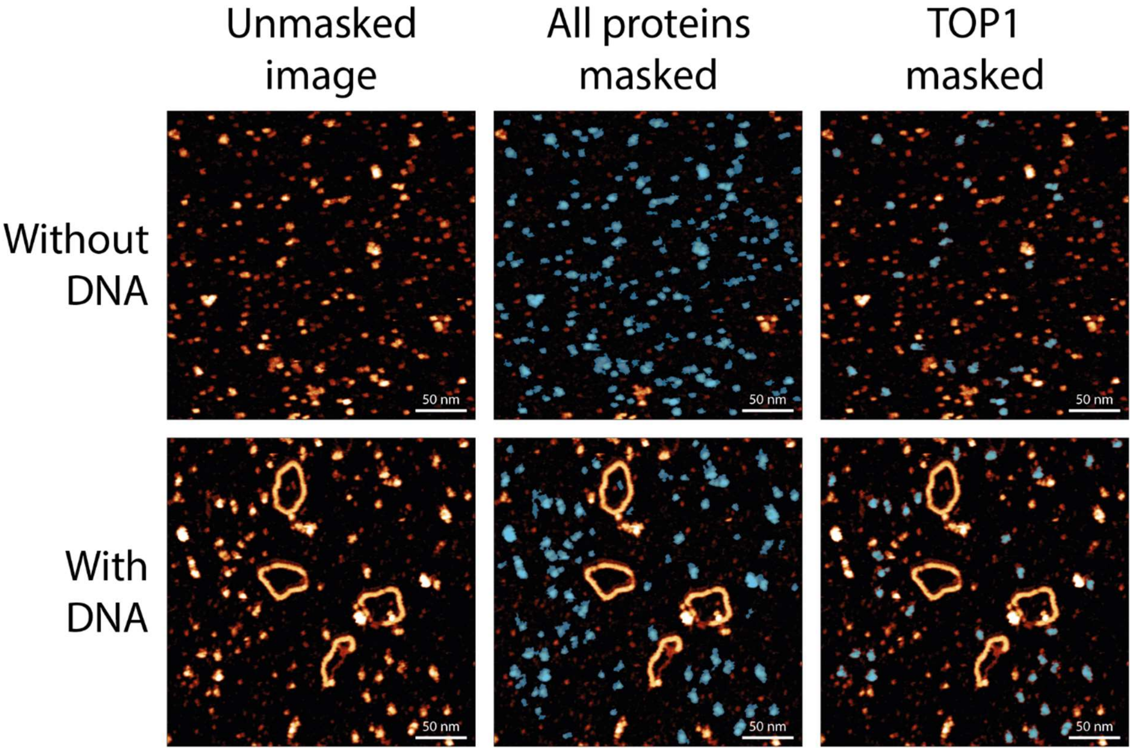


Figure 3.12 Masking AFM images of TOP1 with and without DNA. In both cases, a set of parameters can be used to identify all proteins from the background, while another set of parameters can be used to mask only TOP1.

Overall, 49315 molecules were automatically selected from 134 usable AFM images when the mask for all proteins was applied. When using the second mask to select for only TOP1, 15703 molecules were identified. The successful selection of TOP1 means that we are able to perform measurements and analyse the variation in TOP1 conformation.

3.3 Measuring single globular proteins: TOP1

3.3.1 Comparing different masks

3.3.1.1 The purpose of comparing different masks

As explained in Section 3.1.2.1, although automatic measurement reduces the bias from manual selection, it might introduce another form of bias due to inaccurate masking. To ensure that the automatic molecular identification achieves the desired result and that the subsequent measurements are reliable, the accuracy of the masking needs to be evaluated first. This can be done by comparing the measurements from different masks and assessing whether the difference in the distribution of these measurements matches with expectations. Since the main challenge to the successful identification of TOP1 is the presence of other protein molecules such as BSA, we hereby compare the measurements from the mask that covers all proteins with those from the mask that selects for only TOP1.

3.3.1.2 Comparing height measurements

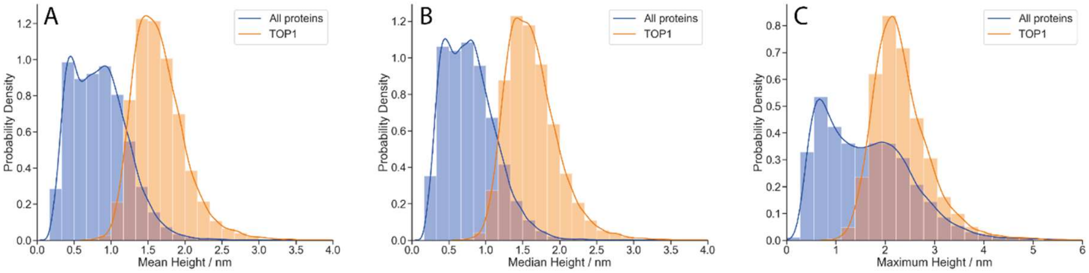


Figure 3.13 Kernel density estimation (KDE) plots and normalised histograms comparing distributions of measured mean (A), median (B) and maximum (C) heights between the masks selecting for all proteins (blue, N = 49315, number of images = 134) and only TOP1 (orange, N = 15703, number of images = 130). Note that as described in Section 2.2.4.1, measurements are made separately for each identified molecule, rather than based on all molecules from an image. For example, the mean height is averaged over all pixels from of a single identified molecule, and not over multiple identified molecules from an image. The same applies to all other measurements in this thesis.

Three different height measurements are made for each identified protein molecule, including the mean and median height values of all pixels forming the mask of the molecule, and the value of the highest pixel from the molecule. For all three height measurements, the mask selecting for TOP1 leads to peaks at higher values compared to the mask selecting for all proteins. In addition, the distribution curves for the TOP1 mask has only one peak, while the mask for all proteins produces curves with two peaks, which likely correspond to the two most common proteins, the lower BSA and the higher TOP1. This is supported by the observation that, in the distribution of the maximum height, the second peak for the all-protein mask matches the location of the peak for the TOP1 mask. While these two peaks do not align in the distributions of mean and median heights, with the peak for the TOP1 mask at a higher value than the second peak for the all-protein mask, this can be explained by the difference in height thresholds: as illustrated in Figure 3.14, when the all-protein mask is applied on a TOP1 molecule, the lower height threshold means that a larger area is covered, and part of the background is also included in the mask, resulting in lower mean and median height values; conversely, when the TOP1 mask is applied, the higher height threshold leads to the selection of only the core TOP1 region, which is higher than the periphery or the background, and consequently the measurements for mean and median heights are higher. The maximum height is not affected, as both masks cover the highest part of the TOP1 molecule in the centre.

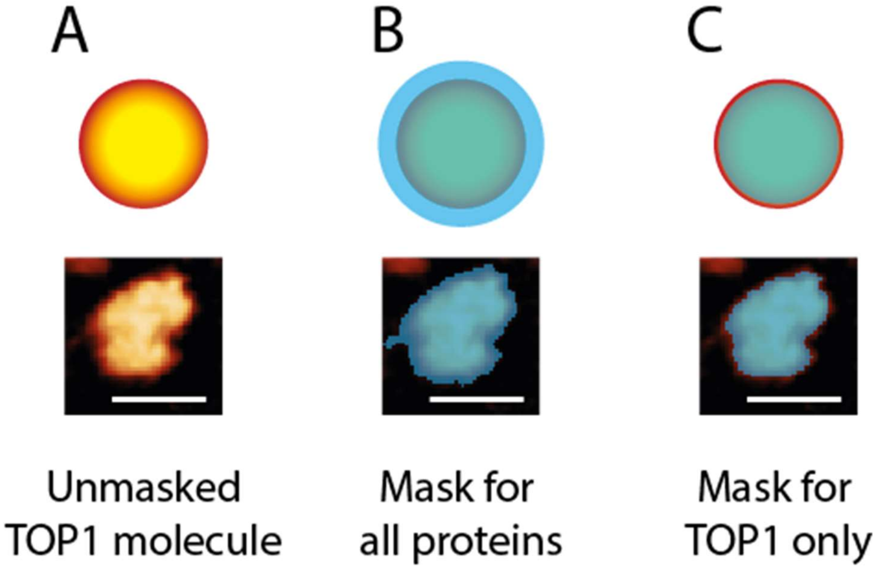


Figure 3.14 Schematic illustrations and AFM images demonstrating the different effects of the two masks (blue) on TOP1 (yellow and red). A: An unmasked TOP1 molecule is higher (yellow) in the centre and lower (red) at the periphery. B: The mask that selects for all proteins has a lower height threshold and therefore covers a larger area, which results in lower mean and median heights, but a longer radius and a larger area. C: The mask that selects for only TOP1 has a higher height threshold and covers TOP1 more precisely, which leads to higher mean and median heights, a shorter radius and a smaller area. The maximum height is not affected by this difference in masking. All scale bars are 10 nm.

These distributions confirm that the height measurements are valid and reflect the differences between the two masks. They show that compared to the all-protein mask, the TOP1 mask can not only exclude other proteins such as BSA, but also select TOP1 more precisely, distinguishing it better from the background. This suggests that the TOP1 mask is fit for purpose.

3.3.1.3 Comparing radius and area measurements

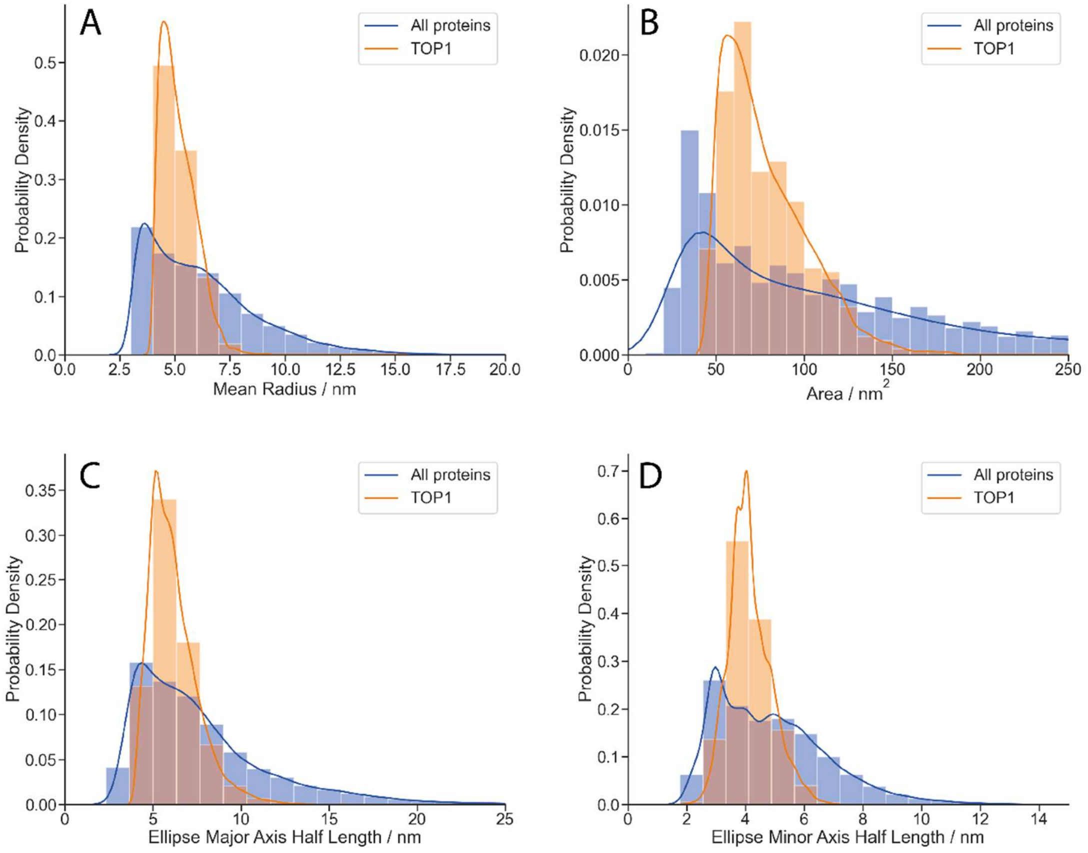


Figure 3.15 KDE plots and normalised histograms comparing distributions of measured mean radius (A), area (B), half lengths of major (C) and minor (D) axes for the equivalent ellipse between the masks selecting for all proteins (blue, N = 49315, number of images = 134) and only TOP1 (orange, N = 15703, number of images = 130). Note that as described in Section 2.2.4.1, the mean radius here is not averaged over multiple molecules; it is instead calculated separately for each molecule as the average distance between its centre of mass and its boundary.

The mean radius (average distance between the centre of mass and the boundary for each molecule), area and the ellipse axes lengths characterise the size and shape of the recognised molecules in the X-Y plane. Similar to the distribution for heights, these parameters based on the X-Y plane also exhibit peaks at lower values when the mask for all proteins is applied, reflecting the smaller size of BSA compared to TOP1. However, unlike in the case of height measurements, the distribution from the all-protein mask covers a wider range of values than from the TOP1 mask, especially at the higher end. This can also be attributed to the difference between the two masks illustrated in Figure 3.14 above: the all-protein mask uses a lower height threshold, and consequently covers a larger region surrounding a molecule of interest, which leads to higher values in mean radius, area, and ellipse axes lengths. This further demonstrates that the TOP1 mask is more precise than the all-protein mask.

3.3.1.4 Comparing volume measurements

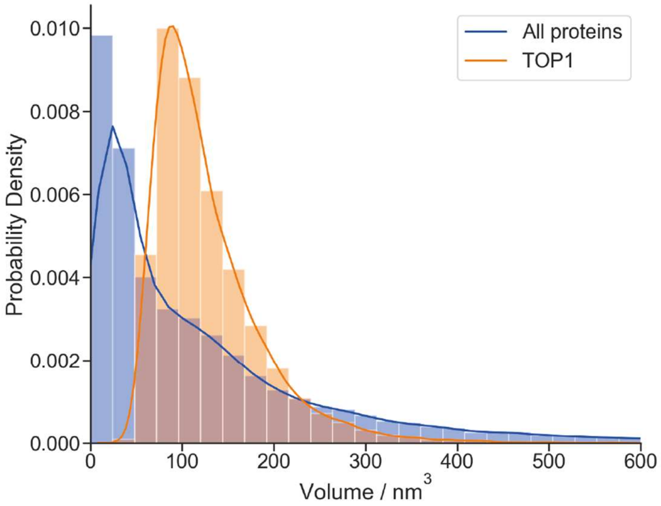


Figure 3.16 KDE plots and normalised histograms comparing distributions of measured volume between the masks selecting for all proteins (blue, N = 49315, number of images = 134) and only TOP1 (orange, N = 15703, number of images = 130).

The distributions of volume measurements are consistent with expectations based on height and area measurements: compared to the TOP1 mask, the all-protein mask results in a peak at a lower value, which corresponds to BSA molecules with lower volumes than TOP1, and matches with the distributions for both height and area. Due to the previously discussed difference in height threshold between the two masks, when identifying a TOP1 molecule from the background, the all-protein mask covers a larger region than the TOP1 mask, and the resulting volume measurement for the former is greater than for the latter. This difference is reflected in the graph comparing the two volume distributions, where the distribution for the all-protein mask reaches higher values.

3.3.1.5 Conclusions of comparison

By making semi-quantitative comparisons of the distributions for a wide range of parameters, we showed that the measurements corresponded well to the purpose of the two different masks. Distributions from the all-protein mask tend to be wider but have peaks at lower values, due to the inclusion of smaller BSA molecules, and the less precise identification of proteins, where part of the background is also covered by the mask, as shown in Figure 3.14. The narrower peaks at higher values for the TOP1 mask indicate that the automatic identification of TOP1 is successful and that the measurements from this mask are valid. To further characterise TOP1, we can compare these AFM measurements with data obtained from other characterisation techniques.

3.3.2 Characterising TOP1

3.3.2.1 Characterising TOP1 by height

Among the three established height measurements, the maximum height can be most easily compared to data from other sources. A previous AFM study measured the height of TOP1 manually [165] and obtained an average of 2.9 ± 0.4 nm. In addition, assuming a hemispherical structure [166], the maximum height of TOP1 can be approximated by its radius: dividing the molecular weight of 91 kDa [140] by an estimated density of 1.41 g/cm^3 [167], the expected molecular volume of TOP1 is calculated to be 1.07 × 10^{-19} cm^3 or 107 nm^3; using the volume formula for hemispheres, $V = 2/3\ \pi\ r^3$, an estimation of 3.7 nm can be obtained for the radius.

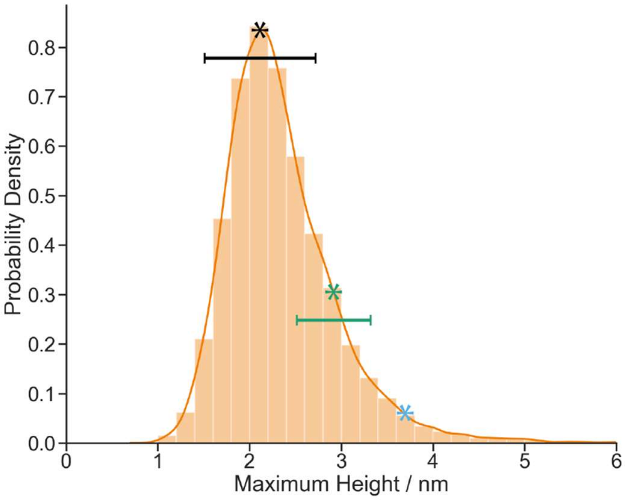


Figure 3.17 KDE plot and normalised histogram of automatically measured TOP1 maximum height. The peak value and standard deviation (black asterisk and error bar, 2.1 ± 0.6 nm, N = 15703, number of images = 130) are compared to measurements from a previous AFM study [165] *(green asterisk and error bar, 2.9 ± 0.4 nm) and the radius estimated (blue asterisk, 3.7 nm) based on molecular weight* [140] *and density* [167]*, assuming that the protein is hemispherical* [166].

As shown in Figure 3.17 above, in this project, the measured maximum height of 2.1 ± 0.6 nm is in line with the results from the previous AFM study, while both are smaller than the calculated radius. This difference could be explained by conformational changes when a 3D structure is restricted through attachment to a 2D surface.

3.3.2.2 Characterising TOP1 by radius and area

Automatic AFM image analysis algorithms facilitate the characterisation of TOP1 on the X-Y plane through a series of available measurements, including the mean radius, the area, and the axes lengths of the equivalent ellipse. Since it is not straightforward to compare the axes lengths with data obtained from other sources, this section focuses only on the mean radius and the area. As detailed in Section 3.3.2.1, if the shape of TOP1 is approximated as a hemisphere, based on the molecular weight and the estimated density, the radius is calculated to be 3.7 nm, which corresponds to an area of 43 nm^2 for the base plane of the hemisphere. These values are compared to the automatic measurements in Figure 3.18 below.

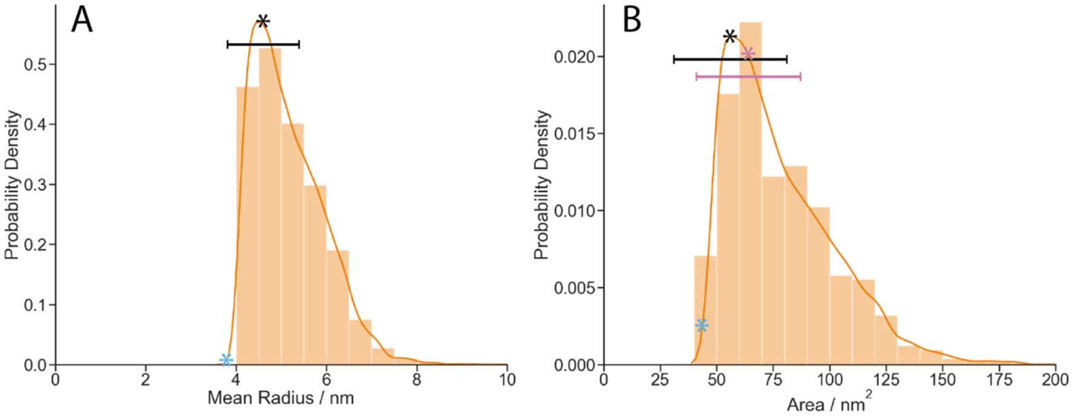


Figure 3.18 KDE plots and normalised histograms of automatically measured TOP1 mean radius (A) and area (B) (N = 15703, number of images = 130). The peak values and standard deviations for mean radius (black asterisk and error bar in A, 4.5 ± 0.8 nm) and area (black asterisk and error bar in B, 56 ± 25 nm^2) are larger than the radius (blue asterisk in A, 3.7 nm) and area (blue asterisk in B, 43 nm^2) estimated based on molecular weight [140] *and density* [167], *assuming that the protein is hemispherical* [166]. *The area calculated (purple asterisk and error bar in B, 64 ± 23 nm^2) from the mean radius is in agreement with the directly measured area.*

The measured peak values of the radius (4.5 ± 0.8 nm) and area (56 ± 25 nm^2) are both higher than the estimations based on the assumption that the protein exhibits a hemispherical structure. This further demonstrates that TOP1 has undergone conformational changes during adsorption, as the transition between moving freely in solution and being restricted by the 2D mica surface would affect the lowest-energy conformation. The differences between the estimated and the measured values in height, radius and area suggest that TOP1 becomes flatter when attached to a 2D surface, making the hemisphere approximation less viable.

In addition, as shown in Figure 3.18 (B), when applying the formula $A = \pi r^2$ to the measured mean radius, the area is calculated to be 64 ± 23 nm^2, which is in accordance with the directly measured area. The consistency between these two measurements confirms that, although TOP1 is not projected onto the X-Y plane as a perfect circle, the mean radius is a valid representation of its size, and therefore a suitable parameter for its characterisation.

3.3.2.3 Characterising TOP1 by volume

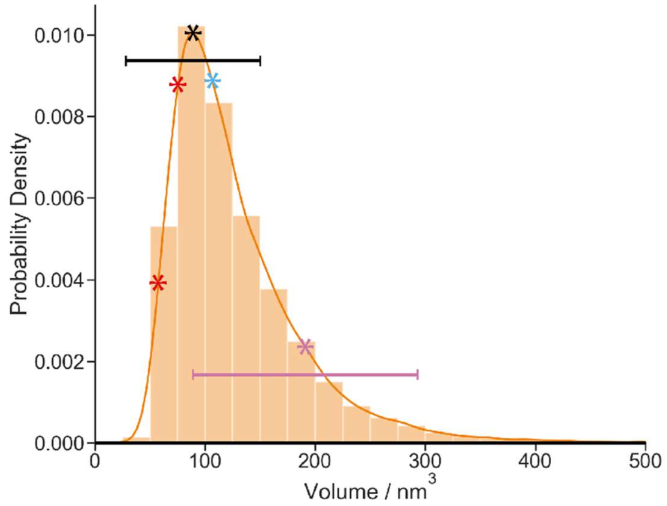


Figure 3.19 KDE plot and normalised histogram of automatically measured TOP1 volume. The peak value and standard deviation (black asterisk and error bar, 89 ± 61 nm^3, N = 15703, number of images = 130) are in line with the volume estimated (blue asterisk, 107 nm^3) based on molecular weight [140] *and density* [167], *and comparable to the volume calculated (purple asterisk and error bar, 191 ± 102 nm^3) from the automatically measured mean radius. The measured volume is also in agreement with volumes calculated via the software Chimera* [163] *(red asterisks, 57 nm^3 and 75 nm^3) using two different sets of crystallographic data obtained from X-ray diffraction* [162], [168].

The automatically measured volume, 89 ± 61 nm^3, is congruent with the estimation of 107 nm^3 based on the molecular weight of 91 kDa [140] and the density of 1.41 g/cm^3 [167], as described in Section 3.3.2.1. Unlike with the height, radius or area, a strong agreement is reached here because the volume estimation does not presuppose a hemispherical structure. Conversely, when the hemisphere assumption is used to calculate the volume from the automatically measured mean radius, the obtained value is 191 ± 102 nm^3, which is less consistent with the direct volume measurement. These results corroborate our findings from Section 3.3.2.1 and Section 3.3.2.2 that, the size of TOP1 can be accurately characterised through automatic measurements, and that the discrepancies between the data from different sources are mainly due to the inapplicability of the hemisphere model to AFM imaging where proteins are attached to a surface.

Human TOP1 has previously been characterised through X-ray diffraction (XRD) in complex with DNA, and the crystallographic data obtained varies from study to study. Using the software Chimera [163] which visualises the structure of biomolecules, the DNA fragment can be manually removed, as illustrated in Figure 3.20 below, and the volume of the remaining TOP1 molecule can be calculated from the crystallographic data. The resulting values based on different XRD studies, such as 57 nm^3 [162] and 75 nm^3 [168], are in agreement with AFM data, especially when taking into account the fact that cavities within the TOP1 molecule are excluded from Chimera calculations but included in AFM measurements.

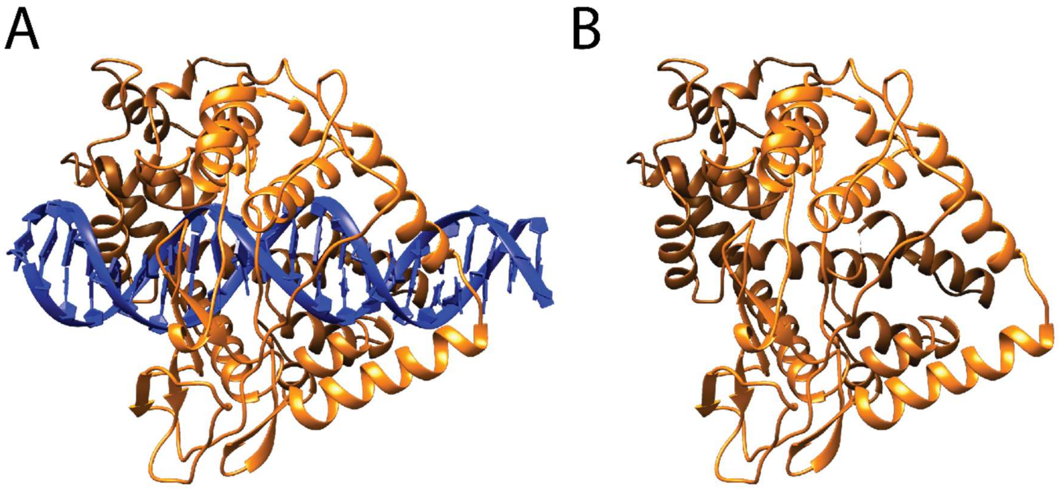


Figure 3.20 Removing DNA from TOP1 crystal structure using the software Chimera [163]*. A: The crystal structure of human TOP1 (orange) in complex with DNA (blue)* [162]*, as visualised through Chimera. B: The DNA is manually removed so that the volume of TOP1 can be calculated; this structure is also shown in Figure 3.4 from a different perspective.*

3.3.2.4 Conclusions and future work for the TOP1 project

In this project, suitable parameters for the characterisation of TOP1 in AFM images were identified. Automatic measurement of these parameters were performed on all recognised molecules of interest. Through comparison with data from other sources, including previous studies and theoretical estimations, the measurements of height, radius, area and volume obtained through automatic analysis algorithms were shown to be valid. In cases where values obtained through different methods were not in agreement with each other, the reasons for the mismatch were identified, such as the unsuitability of the hemispherical model for TOP1 molecules adsorbed on a 2D surface.

AFM imaging and automatic characterisation are confirmed as valuable tools to study globular proteins. In the future, these tools can be employed to measure and compare TOP1 imaged under different experimental conditions. They will also be useful for the quantification of conformational changes in DNA as a result of TOP1 interaction. Furthermore, the characterisation techniques used for TOP1 can be applied to future projects with other proteins, though different parameters may be needed to characterise the shape of proteins with more complex sub-molecular structures.

3.4 Identifying proteins with visible sub-molecular structure: NDP52

3.4.1 Identifying structures of interest on images of NDP52

3.4.1.1 Background on NDP52

It has been demonstrated that automatic AFM image analysis algorithms can be adapted for the recognition and measurement of single globular proteins such as TOP1. However, the role of such algorithms in the characterisation of molecules with more complex structures is yet to be established. Here we aim to establish this role by exploring the application of automatic AFM image analysis on a different protein, NDP52, that exhibits visible sub-molecular domains, and by identifying the adjustments that need to be made.

NDP52 was first discovered in the cell nucleus where DNA is stored [149]. It has been associated with biological processes such as autophagy [146], which refers to the removal of undesired or damaged structures in the cell, but takes place outside the nucleus. In order to probe the potential role of NDP52 inside the nucleus, which had not been identified in previous studies, our collaborators performed a series of biochemical analysis, and their results suggested that NDP52 could be involved in transcription initiation and interaction with DNA [151].

Since the function of a biomolecule is dependent on and can be predicted by its structure, more characterisation is needed to provide direct information on the structure of NDP52, so that its function during its complex interaction with DNA can be understood, and its different components can be related to the mechanism of its operation. For example, our collaborators' analysis suggested that when isolated, different domains of NDP52 exhibited varying affinity to DNA; a detailed investigation into the sub-molecular structure of DNA may allow us to further understand the mechanism behind such difference. However, the characterisation of NDP52 structure remains difficult, due to its small size (446 amino acids) and unique shape composed of two globular terminal domains connected by a flexible string-like domain in the middle known as the coiled-coil [149], as described in Section 1.4.3.2 and illustrated in Figure 3.22 (D). This structural complexity means that crystallographic data for complete NDP52 is not available from experiments, as the protein does not crystallise easily, although an AlphaFold [169], [170] prediction of NDP52 structure can be obtained, as shown in Figure 3.21 below.

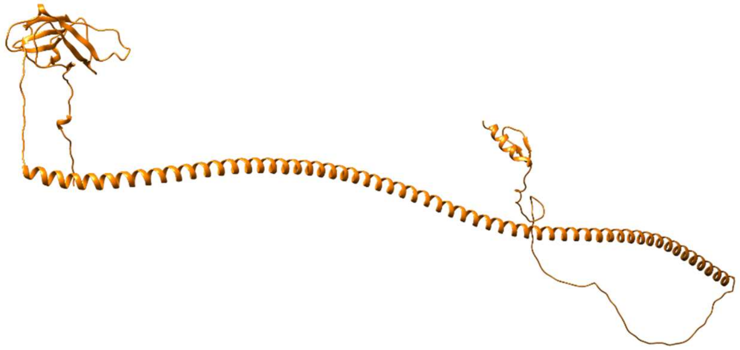


Figure 3.21 AlphaFold [169], [170] *prediction of NDP52 structure visualised through Chimera* [163]. *The terminal domains and the connecting coiled-coil in the middle can be distinguished.*

AFM can overcome these challenges because it allows for the direct visualisation of NDP52 in a solution-based environment, which can lead to a better understanding of NDP52 structure, and by extension the mechanism of NDP52-DNA interaction. Such increased understanding is facilitated in this project through the development of automatic analysis methods to identify and measure NDP52 imaged with AFM by Daniel E. Rollins and Ália dos Santos.

The rest of Section 3.4.1 will be dedicated to the automatic recognition performed on images of complete NDP52 molecules. This includes the identification of NDP52 monomers, oligomers, and terminal domains, because NDP52 is known both to exhibit sub-molecular structure and to form oligomers by combining with each other. Section 3.4.2 will discuss the recognition of truncated NDP52 containing mainly the C-terminal domain, which were imaged on their own because the experiments performed by our collaborators had suggested that this was the domain mainly responsible for interaction with DNA. Section 3.5 will focus on the measurement of NDP52 and its internal structure. Oligomers will be measured and studied in more detail in Section 3.6.

3.4.1.2 Selecting for the entire NDP52 molecule

3.4.1.2.1 The imaged structure of NDP52

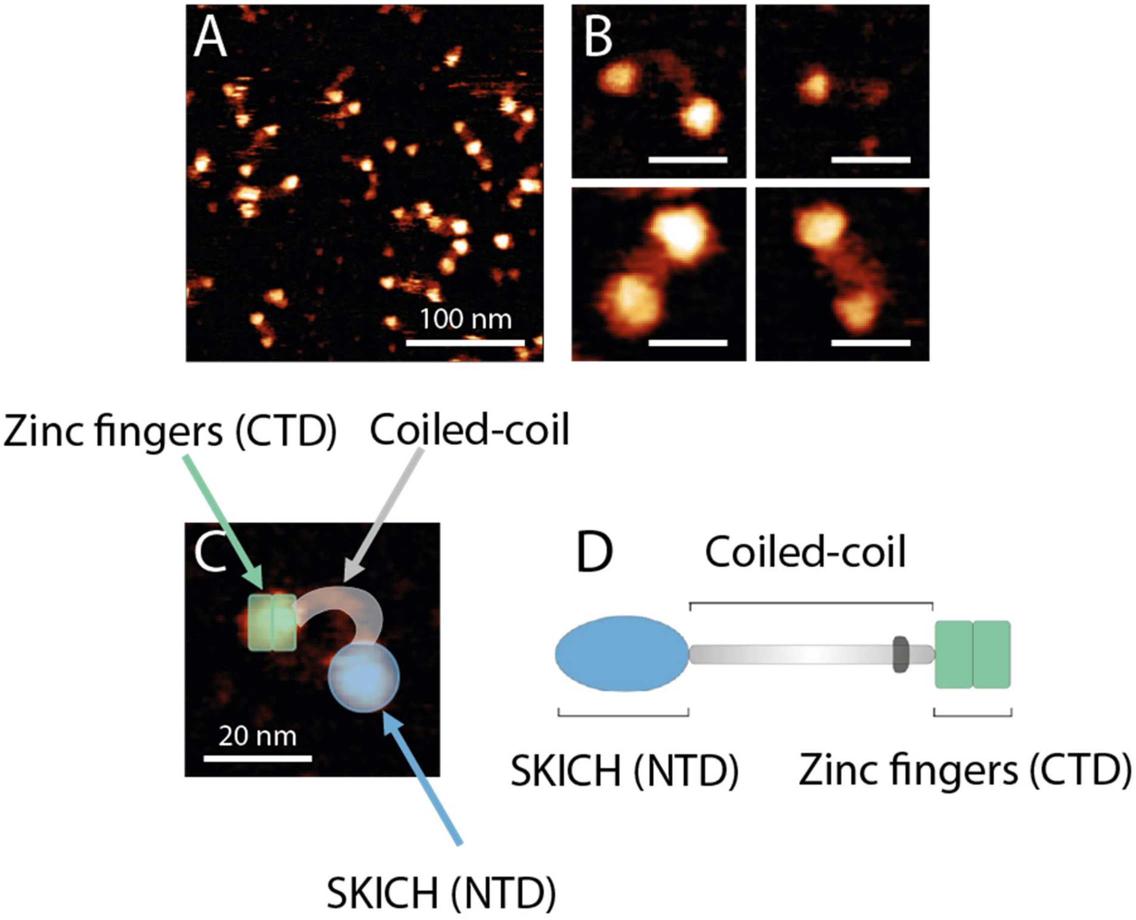


Figure 3.22 Comparison between the imaged and predicted structures of NDP52. A: An AFM image of dumbbell-shaped NDP52 proteins. B: Example close-up images of individual molecules; all scale bars are 20 nm. C: Each protein contains two terminal domains of different heights connected by a coiled-coil. D: The experimental observations correspond well to the predicted structure, according to which the SKICH at the N-terminal domain (NTD) is larger than the zinc fingers at the C-terminal domain (CTD).

Prior to performing automatic characterisation, we first inspect the processed images of NDP52 qualitatively, so as to obtain a better understanding of its structure and identify challenges in subsequent analysis.

As shown in Figure 3.22, AFM imaging confirms our previous understanding that NDP52 exhibits a dumbbell-shaped structure composed of two terminal domains linked by the coiled-coil. The terminal domains can be easily identified by eye, with one domain appearing to be higher than the other for most NDP52 molecules imaged. This is in agreement with the previous knowledge that the skeletal muscle and kidney enriched inositol phosphatase carboxyl homology (SKICH) domain at the N terminal is larger than the zinc fingers at the C terminal [149]. By comparison, the coiled-coil is lower than the terminal domains, which makes it more difficult to be distinguished from the background. In addition, the coiled-coil is observed to be very flexible, and as a result the protein does not have a single uniform shape, which could explain the difficulties in obtaining crystallographic data.

Having examined the appearance of NDP52 on AFM images, we now aim to develop quantitative analysis tools to probe the conformation of the protein, as NDP52 is so structurally heterogeneous that no clear single conformation can be generalised from qualitative observation. We start by recognising NDP52 from the background, which serves as the first step of the automatic image analysis workflow, and paves the way for further measurements and characterisation.

3.4.1.2.2 Challenges of selecting for two terminal domains connected by the coiled-coil

While the common problems faced when identifying molecules of interest from the background have been discussed in Section 3.2.2, the structure of NDP52 and the nature of the data available can lead to unique challenges, which are detailed below.

3.4.1.2.2.1 Selecting the coiled-coil

The coiled-coil is much lower than the rest of the molecule. This means that if the parameters for identifying TOP1 or other single globular proteins are used, only the terminal domains will be included in the mask, and not the coiled-coil. The height threshold can be lowered to rectify that, but this may lead to other problems, such as the inclusion of impurities and the drop in the accuracy of downstream measurements.

A delicate balance needs to be struck, in order to maximise the inclusion of the coiled-coil but minimise that of unwanted features. This can be done by taking the parameters for TOP1 as a starting point, and adjusting the height threshold through trial and error until the most suitable result is reached. It should be noted that, when performing such adjustments, the area threshold also needs to be modified to account for the larger size of NDP52 and to exclude impurities whose size is similar to that of TOP1.

3.4.1.2.2.2 Addressing impurities and noise

Whilst attempting to establish the height and area thresholds across the dataset provided, it was noticed that the AFM images were of varying quality. For some images, it was possible to find a set of parameters that allowed for the inclusion of the coiled-coil and the exclusion of most impurities; for others, this task proved to be difficult because of the amount of noise in the background.

Due to the impracticality of identifying parameters that worked for all images, a subset of images with relatively high quality were selected. The trial and error could then be performed on the selected images and the ideal parameters could be determined. We subsequently attempted to apply these parameters to the rest of the images, though as shown in Figure 3.23 in Section 3.4.1.2.3, this was not successful. However, we have established a method of testing on a smaller high-quality selection first before moving on to the entire dataset, which could be useful in future projects.

3.4.1.2.3 Results

By lowering the height threshold and increasing the area threshold from the values used for TOP1, the coiled-coil can be included in the mask and the entire NDP52 molecule can be identified from the background, although this only works for images within the high-quality selection.

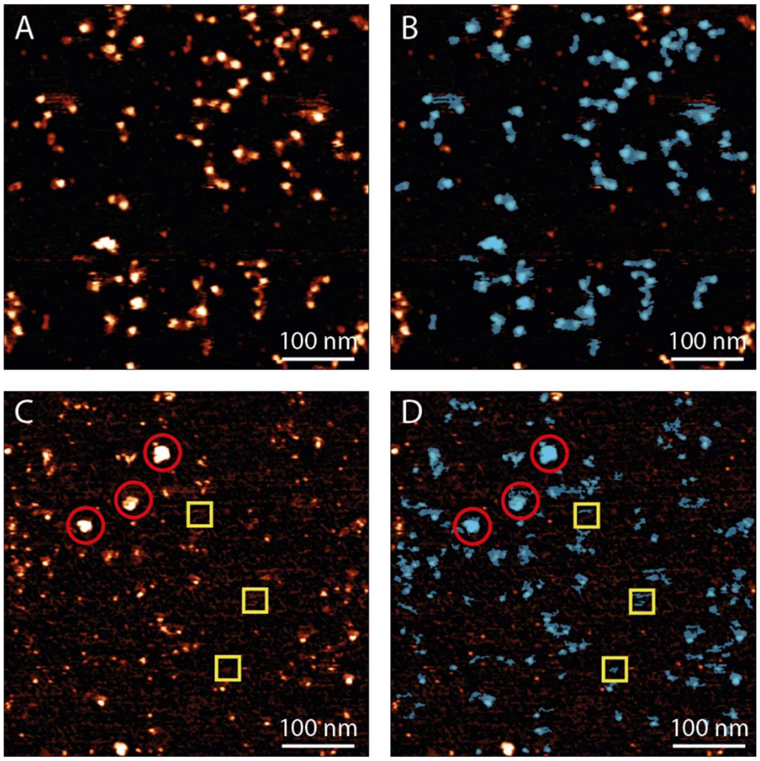


Figure 3.23 Selecting for the entire NDP52 molecules by adjusting parameters. The parameters established for high-quality data with a clear background (A: unmasked; B: masked) are successful in masking NDP52 monomers, but they do not work for low-quality data with a noisy background (C: unmasked; D: masked, with impurities marked in red circles and background noise falsely identified as molecules marked in yellow squares on both).

3.4.1.3 Selecting for the terminal domains

3.4.1.3.1 The purpose of selecting for the terminal domains

As illustrated in Figure 3.21, NDP52 is composed of the C-terminal domain (CTD), the N-terminal domain (NTD) and the linking coiled-coil. These three domains could play different roles in the functioning of NDP52, including its interaction with DNA. Our collaborators have probed the differences between the two terminal domains by isolating them and performing florescence spectroscopy assays. The results of their experiments suggest that the CTD has a higher affinity to DNA and is mainly responsible for NDP52-DNA interaction [151].

This discovery shows that it would be useful to characterise the terminal domains on their own, which could reveal the structural basis of their different functions. The first step to achieve this is to select for only the terminal domains on AFM images by separating them from the coiled-coil.

3.4.1.3.2 Challenges of selecting for the terminal domains

3.4.1.3.2.1 Excluding the coiled-coil

Section 3.4.1.2.2.1 discussed the adjustments of parameters required to identify the entire NDP52 structure, which includes both the terminal domains and the coiled-coil. However, when selecting for only the terminal domains, some of these adjustments need to be reverted, so that the coiled-coil can be removed from the mask. This means that compared to the parameters used for the recognition of the complete NDP52 molecule, the height threshold needs to be increased since the coiled-coil is lower than the rest of the molecule. The area threshold can also be lowered to account for the smaller size of the individual terminal domains and filter out large impurities, though the change in height threshold alone proved to be sufficient for the exclusion of the coiled-coil. The resulting thresholds should be similar to those used for TOP1, as terminal domains have a similar appearance to globular proteins.

3.4.1.3.2.2 Addressing impurities and noise

Some impurities or background noise are closer in size to the terminal domains than to the entire molecule. They can also be higher than NDP52, which presents an extra challenge to the recognition of terminal domains, as these impurities cannot be filtered out by adjusting height thresholds. To ensure that a suitable set of parameters can be found, only the high-quality images selected in Section 3.4.1.2.2.2 are used; in this way, the terminal domains can be more accurately identified with less interference from impurities and noise.

3.4.1.3.2.3 Distinguishing between the two terminal domains

The automatic distinction between the NTD and the CTD would enable separate measurement and characterisation of these two domains, which would allow us to quantify their structural differences.

The NTD is known to be larger than the CTD, and this difference can be observed on AFM images, as shown in Figure 3.22. This means that it is possible to tell the NTD apart from the CTD for a single given NDP52 molecule.

However, the automatic recognition of a specific domain across multiple molecules remains challenging, and attempts at finding a height threshold to mask only the CTD or only the NTD were unsuccessful. This suggests that there might be some overlap between the height distributions of the two terminals. Further analysis and comparison are needed to determine whether such automatic separation is theoretically possible.

3.4.1.3.3 Results

By adjusting height and area parameters, the terminal domains from the selected high-quality images can be masked, as shown in Figure 3.24 below.

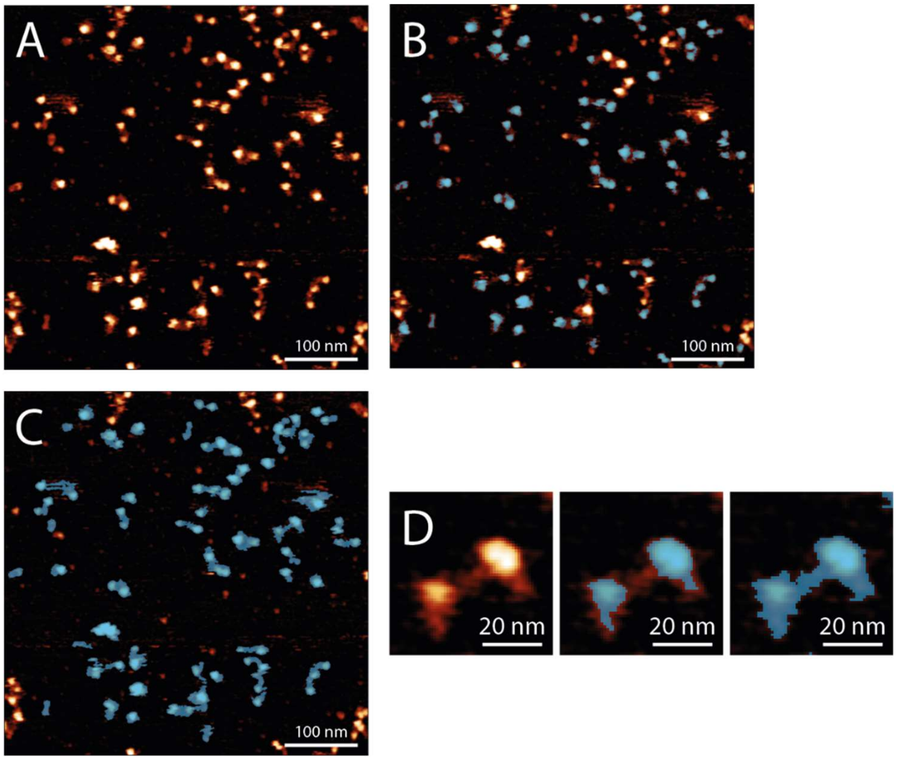


Figure 3.24 Selecting for the terminal domains. A: An unmasked AFM image of NDP52 molecules. B: The coiled-coil can be excluded by lowering the area threshold and increasing the height threshold, so that only the terminal domains are masked. C: The same image with the entire NDP52 molecule masked for comparison. D: A close-up comparison of an unmasked molecule, a molecule with only its terminals masked, and the entire structure being masked.

The NTD and the CTD cannot be automatically separated, since it was not possible to establish a mask that included only one of the terminal domains. However, the current mask of terminal domains where the NTD and the CTD are not distinguished can serve as a basis for further analysis and characterisation of the terminal domains.

3.4.1.4 Distinguishing between monomers and oligomers

3.4.1.4.1 The purpose of distinguishing between monomers and oligomers

Biochemical analysis using SEC-MALS and mass photometry has demonstrated that NDP52 molecules also interact among themselves, with the majority forming dimers and some forming higher oligomers [151]. Furthermore, as stochastic optical reconstruction microscopy (STORM) and cluster analysis performed by our collaborators have demonstrated that NDP52 clusters around transcription initiation sites [151], the oligomerisation behaviour could potentially play an important role when NDP52 is involved in DNA transcription. Therefore, to understand whether NDP52 operates alone or in concert and to obtain more structural information, it is necessary to identify and characterise the formation of dimers and higher order oligomers of NDP52, as observed in the AFM images. In addition, by identifying the parameters required to select for only oligomers, we can also establish a method to exclude them when analysing monomers, since a mixture of both would mean that accurate statistics on one single type of structure could not be obtained.

3.4.1.4.2 The imaged structure of oligomers

Before performing automatic analysis, we first assess their imaged structures manually in order to obtain relevant qualitative information. NDP52 dimers and trimers can be observed on many AFM images provided, especially those within the high-quality subset. These oligomers can be identified by eye based on the presence of more than two terminal domains closely linked together, as seen in Figure 3.25 below. However, the majority of molecules observed appear to be in monomeric form.

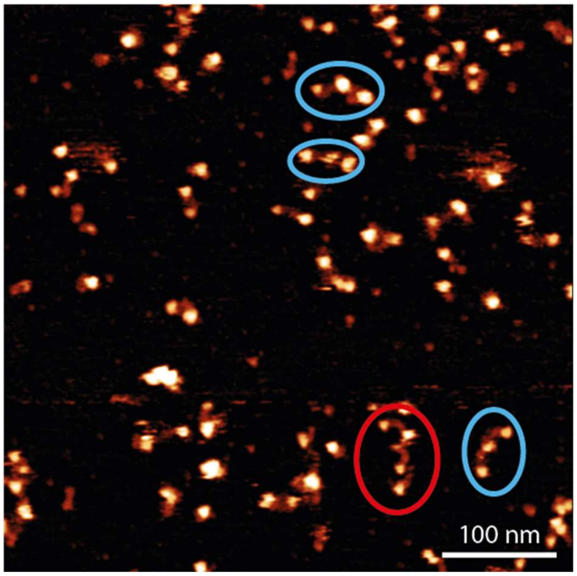


Figure 3.25 Example AFM image of NDP52 molecules containing oligomers, with dimers circled in blue and a trimer circled in red.

3.4.1.4.3 Challenges of distinguishing between monomers and oligomers

The main challenge to automatic identification regarding oligomers lies in the need to establish three different sets of related parameters. The first set should be used when the automatic molecule recognition is performed to characterise the conformation of NDP52: in this case, oligomers need to be excluded from the masking, as they would interfere with subsequent measurements and skew the distribution. The second set is needed to select for only oligomers, so that the operation mechanism of NDP52 oligomerisation can be studied. Thirdly, to calculate the proportion of oligomers among all NDP52 molecules, it is also necessary to identify parameters that select for both oligomers and monomers.

Since oligomers are distinguishable from monomers based on their size, their selective inclusion can be achieved through adjusting the area threshold. Three different area threshold ranges can be used to select for only monomers, only oligomers, and both types of molecules. Relative area was used for this instead of absolute area, as the dataset contained images obtained under different experimental conditions.

3.4.1.4.4 Results

By adjusting the area threshold, three different masks were established, allowing for the choice between monomers, oligomers, and both structures. To ensure consistency with other masks aiming at the automatic recognition of different features, only the high-quality images were used. As shown in Figure 3.26, the same example image that was used for the identification of terminal domains can now also be masked to select for monomers and oligomers. This further demonstrates that automatic AFM image analysis algorithms are capable of identifying and characterising different features from the same dataset through the modification of masking parameters.

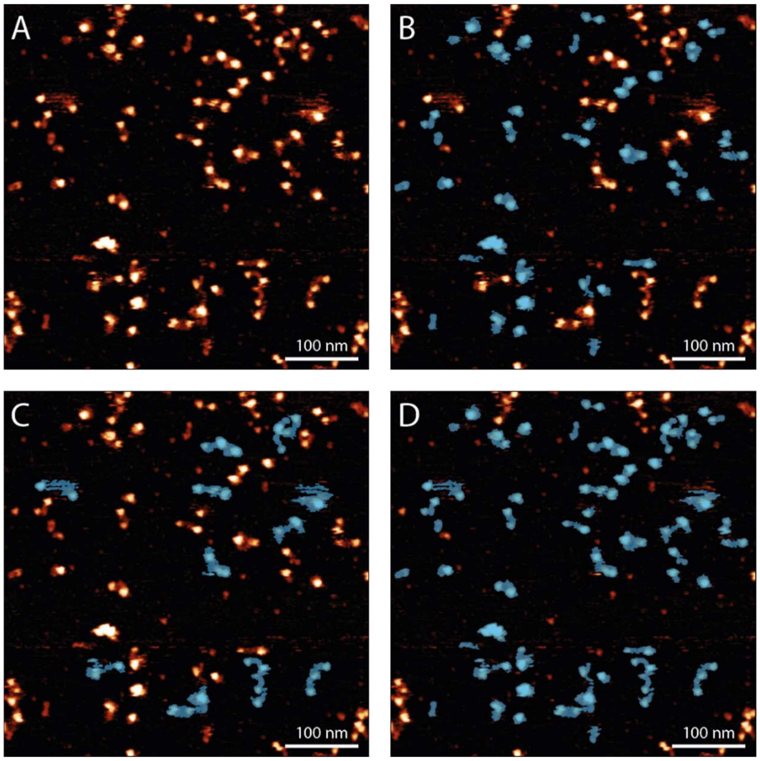


Figure 3.26 Distinguishing between oligomers and monomers through the adjustment of masking parameters. A: An unmasked AFM image of NDP52 molecules. B: Masking only monomers. C: Masking only oligomers. D: Masking both monomers and oligomers.

Upon successful automatic inclusion and exclusion of oligomers, further characterisation is needed to measure the proportion of oligomers within the imaged dataset, and to analyse the size of oligomers, so that the mechanism and role of NDP52 oligomerisation can be better understood. This will be discussed in Section 3.6.

3.4.2 Identifying the truncated C-terminal region ($_{C}$NDP52)

3.4.2.1 Background on $_{C}$NDP52

As described in Section 3.4.1.3.1, to probe the differences between the two terminal domains, our collaborators performed florescence spectroscopy assays on two types of truncated NDP52 molecules, the $_{C}$NDP52, composed of the CTD and part of the coiled-coil, and the $_{N}$NDP52, which contains the NTD alongside part of the coiled-coil. The results of their experiments showed that compared to $_{N}$NDP52, $_{C}$NDP52 has a higher affinity to DNA [151]. This signifies that the CTD could play a key role in the nuclear activities of NDP52. The characterisation of CTD structure on AFM images could therefore contribute to the understanding of NDP52-DNA interaction.

The automatic distinction between the two terminal domains on images of complete NDP52 molecules was unsuccessful, as we were unable to select one terminal without including the other in the masking. To circumvent this problem, the $_{C}$NDP52 was imaged by AFM on its own. This allows us to characterise the CTD without interference from the NTD, and to test automatic image analysis methods on a different type of sample.

3.4.2.2 The imaged structure of $_{C}$NDP52

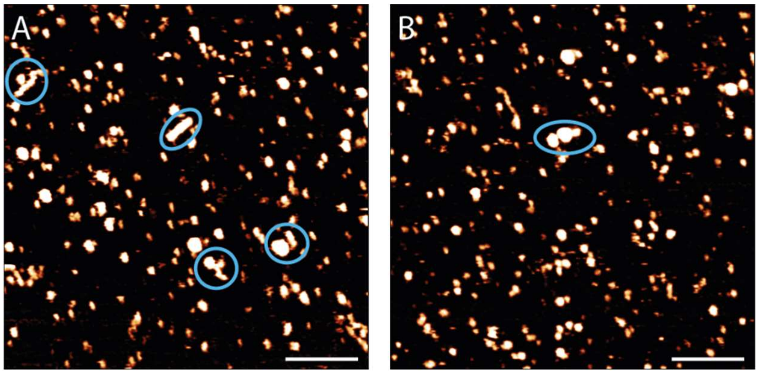


Figure 3.27 Example AFM images of $_{C}$NDP52, with oligomers circled in blue. Scale bars are 100 nm.

Similar to what was done for AFM images of full-length NDP52, before attempting automatic recognition and characterisation, we first manually observe the imaged structure of $_{C}$NDP52. As shown in Figure 3.27 above, these truncated molecules appear to be globular, which is in agreement with the CTD observed on images of complete NDP52. It can also be noticed that some molecules (circled in blue in the figure) are much larger and higher than the rest. These are identified as $_{C}$NDP52 oligomers, because they appear to be composed of several smaller globular proteins. The common presence of oligomers demonstrates that $_{C}$NDP52 readily interacts with each other, and suggests that the CTD could be responsible for the oligomerisation of full-length NDP52.

3.4.2.3 Challenges of masking $_{C}$NDP52

3.4.2.3.1 Addressing low-quality data

As $_{C}$NDP52 appears to be globular and does not have a visible sub-molecular structure, the parameter adjustment processed developed for TOP1 can be adopted. However, compared to TOP1, the automatic identification of $_{C}$NDP52 proved to be more challenging, because $_{C}$NDP52 does not exhibit a characteristic shape, and is harder to distinguish from impurities. Moreover, the data provided were taken under different imaging conditions, as shown in Figure 3.28 below. This hinders the identification of parameters that can be applied to all images.

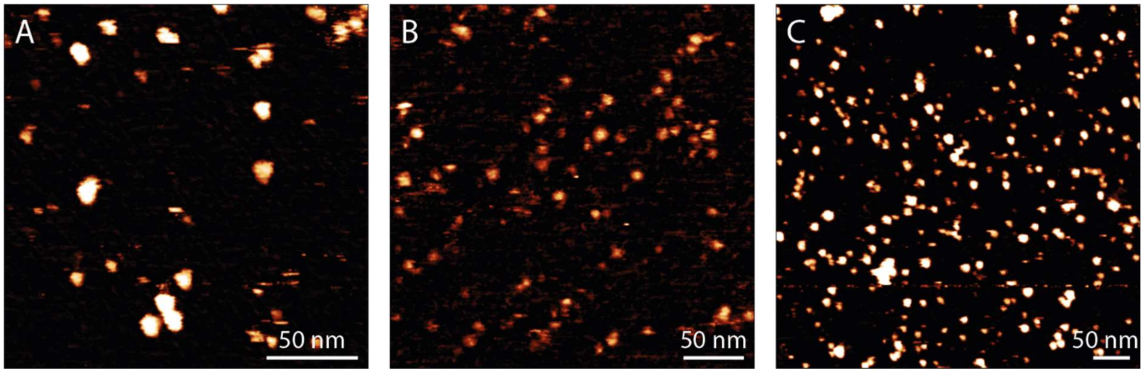


Figure 3.28 Example AFM images of $_{C}$NDP52. The proteins appear differently on each image due to varying imaging conditions.

3.4.2.3.2 Distinguishing between monomers and oligomers

Another factor that limits the automatic identification of $_{C}$NDP52 is the frequent presence of oligomers. These oligomers vary in size and height, and unlike the case of full-length NDP52, they are not always clearly distinguishable from monomers. While it would be useful to have three sets of parameters, for monomers, oligomers and both, this is not practical for the given dataset, because the variation between different images can be even larger than the difference between monomers and oligomers within the same image. Besides, as shown in Figure 3.29, some larger molecules do not have clear visible subunits, and it is difficult to ascertain whether they are oligomers or larger impurities. We therefore attempt to find a combination of height and area thresholds to only mask molecules that are likely to be monomers.

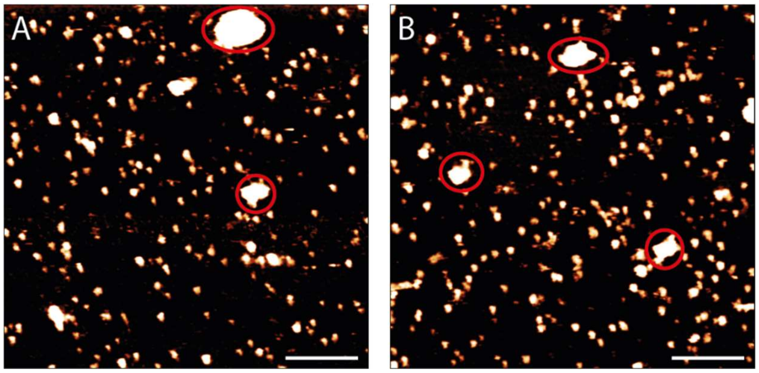


Figure 3.29 Example AFM images containing oligomers and large impurities. It can be difficult to identify whether the molecules circled in red are oligomers or larger impurities, since they are larger than monomers but do not have clearly distinguishable subunits. Scale bars are 100 nm.

3.4.2.4 Results

Due to low data quality and the tendency of $_{C}$NDP52 to oligomerise, it was particularly challenging to strike a balance between selecting all $_{C}$NDP52 monomers and excluding impurities adequately. The varying conditions across different images also meant that it was impractical to follow what was done for full-length NDP52 and select a subset of high-quality data.

However, we applied the masking parameters from the selection of terminal domains on images of full-length NDP52 (see Section 3.4.1.3), so that isolated $_{C}$NDP52 could be compared with the CTD as part of NDP52. As demonstrated in Figure 3.30 below, these parameters are not perfect, and they work to different extents on different images. On some images, certain legitimate $_{C}$NDP52 molecules are not included, while on other images, unwanted features such as oligomers or aggregates are mistakenly selected. Downstream comparison between the measurements made from these images and from images of full-length NDP52 can also be used to evaluate the validity of the automatic recognition.

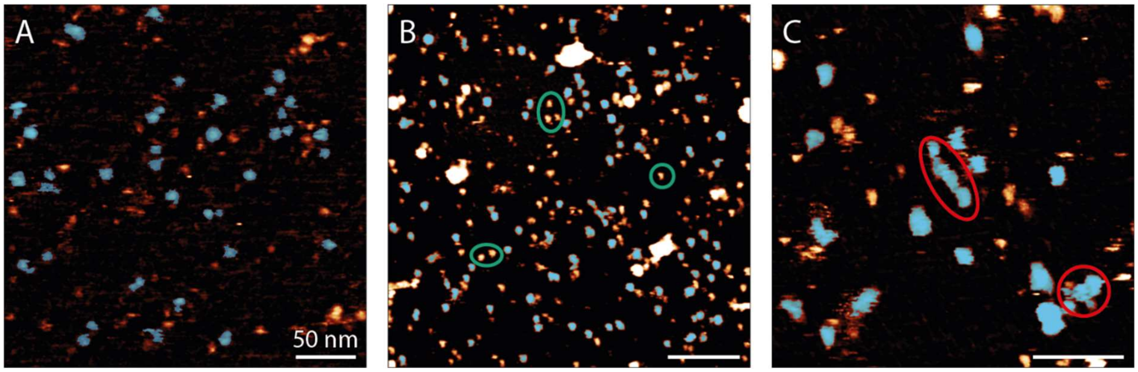


Figure 3.30 It is difficult to achieve accurate automatic recognition of $_{C}$NDP52 across data taken under different conditions with varying image quality. A: An image where $_{C}$NDP52 monomers are successfully selected. B: An image where some $_{C}$NDP52 monomers (examples circled in green) are left out when they should be selected. Scale bar is 100 nm. C: An image where some oligomers or aggregates (examples circled in red) are selected when they should be excluded. Scale bar is 50 nm.

The attempt at selecting $_{C}$NDP52 from the background further demonstrates that successful automatic recognition is highly dependent on the quality and consistency of the experimental data provided.

3.5 Measuring proteins and their visible sub-molecular structure: NDP52

3.5.1 Characterising full-length NDP52 and $_{C}$NDP52 by height

3.5.1.1 Characterising full-length NDP52 by height

To characterise the structure of NDP52 and obtain more quantitative information, automatic height measurements were performed on identified monomers.

Automatic AFM image analysis algorithms allow for the measurement of three different height values for each recognised molecule: the median, the mean and the maximum heights. As NDP52 exhibits a unique structure and its three components vary in height, the maximum height is reflective of the highest NTD, while the median and mean heights are determined by all three domains.

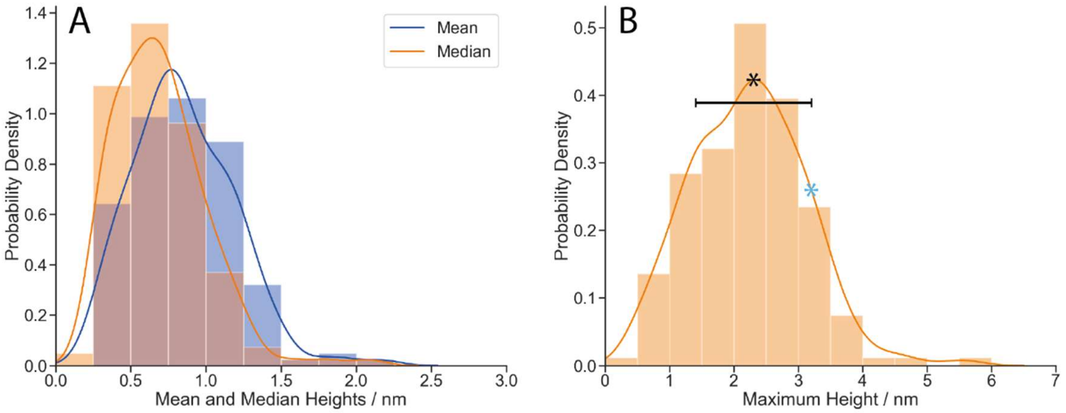


Figure 3.31 KDE plots and normalised histograms of automatic height measurements from selected high-quality images of full-length NDP52 (N = 162, number of images = 6). A: Comparison of mean (peak ± standard deviation 0.8 ± 0.3 nm) and median (0.6 ± 0.3 nm) heights. B: Comparison between the maximum height (peak and standard deviation marked in black asterisk and error bar, 2.3 ± 0.9 nm) and the diameter of the NTD (blue asterisk, 3.2 nm), calculated from molecular weight [149] *and density* [167].

As shown in Figure 3.31 (A) above, the mean value (0.8 ± 0.3 nm) is higher than the median (0.6 ± 0.3 nm), which is in agreement with the observation that the higher terminal domains are also larger than the lower coiled-coil: this means that the terminal domains on AFM images contain more pixels, and they are better represented than the coiled-coil when calculating the mean height value across the entire molecule.

The measured maximum height (2.3 ± 0.9 nm) is dependent on the height of the NTD, which is larger and higher than the other two domains. As shown in Figure 3.31 (B), to verify the accuracy of the automatic molecular identification and measurement, this value is compared to the calculated theoretical diameter of the NTD, which can be approximated as its height due to its globular shape. The molecular weight of this terminal domain, based on its 128-amino-acid sequence obtained from literature [149], is calculated to be 15.27 kDa. Using a density estimation of 1.45 g/cm^3 [167], its theoretical volume is calculated to be 17.49 nm^3, which corresponds to a diameter of 3.2 nm, assuming a globular shape.

The calculated value is higher than the automatic measurement, which can be explained by two reasons: firstly, the adsorption of a protein molecule to the mica surface could induce conformational changes and result in a lower and flatter shape, as discussed in Section 3.3.2 with the example of TOP1; secondly, when selecting for the entire NDP52 structure, the height threshold was lowered so that the coiled-coil could be included in the masking, and as a result, some unwanted lower structures such as broken NDP52 and impurities were also selected, as evidenced by the presence of a second, lower peak to the left of the main peak in Figure 3.31 (B). Examples of these unwanted structures are shown in Figure 3.32 below.

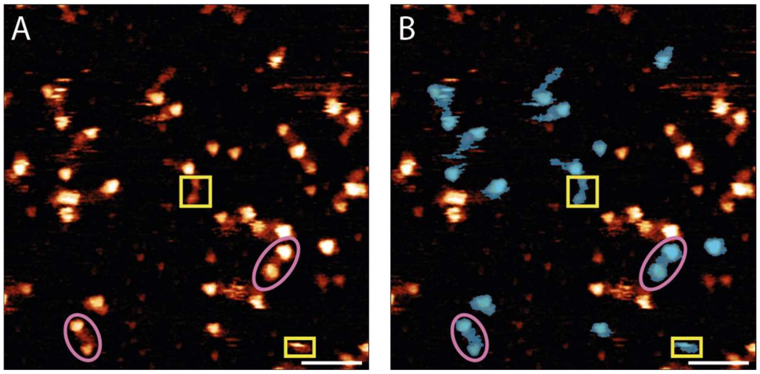


Figure 3.32 Unmasked (A) and masked (B) AFM images of NDP52, showing that some unwanted lower structures such as broken NDP52 and impurities (yellow rectangles) are included in the mask alongside legitimate NDP52 molecules (purple circles). Scale bars are 50 nm.

In conclusion, the height measurements of complete NDP52 molecules are in line with our expectations, but they also demonstrate that the unique shape of the NDP52 imposes limits on the accuracy of automatic molecular identification when noise and impurities are present.

3.5.1.2 Characterising terminal domains within full-length NDP52 by height

3.5.1.2.1 Manual measurement of NDP52 terminal domains

As discussed in Section 3.4.1.3, while it is possible to automatically recognise terminal domains of NDP52 from the background, we were unable to select only the CTD or only the NTD. In order to characterise the two terminal domains separately, Rollins manually measured the maximum heights of the CTD and the NTD using Gwyddion [58]: for a given NDP52 molecule on the selected high-quality images, the two terminal domains were distinguished by eye based on their height difference; their highest points were then identified and measured separately. 96 measurements were performed on a total of 48 molecules. The maximum height was found to be 3.0 ± 0.6 nm for the NTD and 2.1 ± 0.6 nm for the CTD (see Figure 3.34 in Section 3.5.1.2.3).

Here these measurements are used as a basis of comparison with automatic measurements, so that we can evaluate the accuracy of the latter, and determine whether it is theoretically possible to automatically select for only one terminal domain.

3.5.1.2.2 Automatic measurement of NDP52 terminal domains

To characterise the terminal domains faster and in greater quantity, the three established parameters of mean, median and maximum heights were automatically measured. As shown below in Figure 3.33, all three measurements exhibit two peaks, which correspond to the CTD and the NTD. However, the closeness between the two peaks suggests that the height distributions of these two domains overlap with each other to a great extent. Since these measurements are based on the automatic recognition of terminal domains from images of full-length NDP52, the overlap could be caused by inaccuracies during the masking process. To determine whether such is the case and to assess the reliability of the automatic recognition, the maximum height was compared with manual measurements, and this will be discussed in Section 3.5.1.2.3.

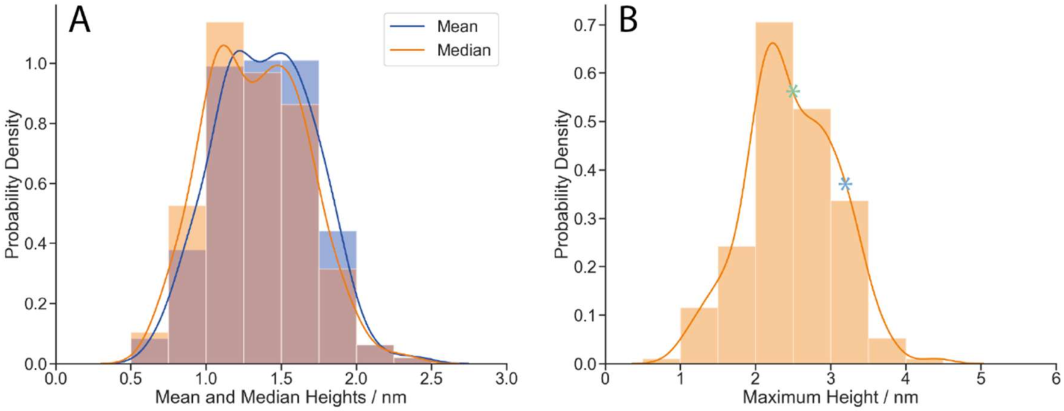


Figure 3.33 KDE plots and normalised histograms of automatic height measurements from identified terminal domains (N = 190, number of images = 6). A: Comparison of mean (two peaks at 1.2 nm and 1.5 nm) and median (two peaks at 1.1 nm and 1.5 nm) heights. B: Comparison between the maximum height (two peaks at 2.2 nm and 2.9 nm) and the diameter of the two terminal domains (CTD green asterisk, 2.5 nm; NTD blue asterisk, 3.2 nm), calculated from molecular weight [149] *and density* [167].

In addition, as mentioned in Section 3.5.1.1, the maximum height of a terminal domain can be approximated by its diameter, whose theoretical value can be calculated from the molecular weight and the density estimation. The 67-amino-acid long CTD has a molecular weight of 7.59 kDa [149] and an estimated density of 1.49 g/cm^3 [167], which results in a volume of 8.46 nm^3 and a diameter of 2.5 nm, while the values for the larger NTD with 128 amino acids are 15.27 kDa [149] and 1.45 g/cm^3 [167], corresponding to a volume of 17.49 nm^3 and a diameter of 3.2 nm. For both terminals, the theoretical diameter is longer than the measured maximum height. As previously explained, the difference can be attributed to conformational changes when a protein in solution is adsorbed onto a 2D surface.

It can also be noticed that when measured from terminal domains, the maximum height distribution exhibits peaks at higher values than when measured from NDP52 molecules recognised as a whole. This is in line with the understanding that during the masking of the entire NDP52 molecule, the challenges introduced by the coiled-coil resulted in the inclusion of broken molecules and impurities, which skewed the height distribution towards the lower end. The measurements made from terminal domains are therefore more accurate than those made from complete NDP52 molecules.

3.5.1.2.3 Comparing different measurements and assessing the masking accuracy of terminal domains

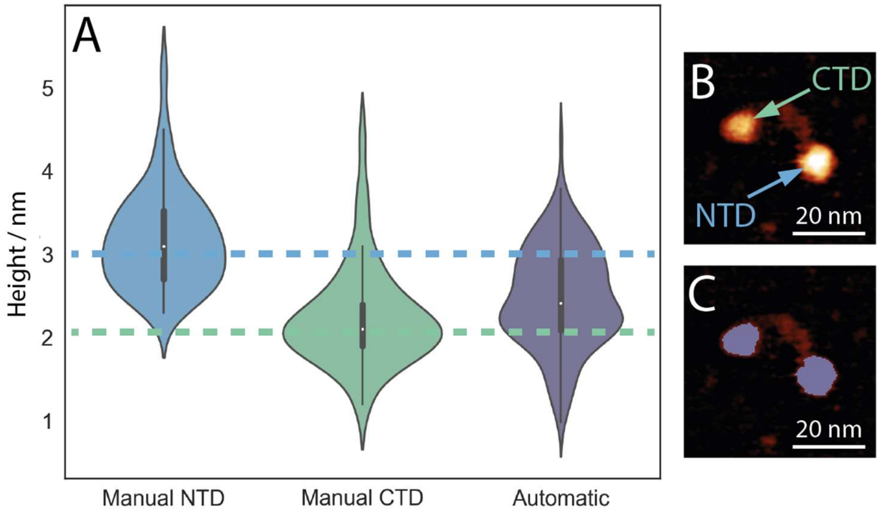


Figure 3.34 Manual and automatic height measurements of NDP52 terminal domains. A: Comparison between the maximum height distributions for the manual measurements (NTD in blue, peak ± standard deviation 3.0 ± 0.6 nm, CTD in green, 2.1 ± 0.6 nm, N = 48 for both, number of images = 6) and the automatic measurement (purple, peaks at 2.9 nm and 2.2 nm, N = 190, number of images = 6). The two peaks from the automatic measurement correspond well to the manual measurements on the two terminals. B: During manual measurement, the NTD and the CTD are measured separately, because they can be identified based on height difference. C: Both terminals are masked together during automatic molecule recognition.

To further assess the reliability of the automatic recognition and characterisation performed on terminal domains, the maximum heights measured by automatic and manual methods were compared side by side. As shown above in Figure 3.34, the peaks from automatic and manual measurements correspond well with each other. This indicates that the automatic measurements are accurate and representative of the structures of terminal domains.

The agreement between the two measurements confirms that the height distribution overlap between the NTD and the CTD is due to the structural variation of the protein and not introduced during data processing. It is therefore not possible to automatically identify one of the terminal domains using the current masking method.

A proposed alternative method would be to perform masking in two stages, instead of directly recognising all the terminals from the background: during the first stage, full NDP52 molecules are masked; during the second stage, the higher terminal and the lower terminal are identified within each masked molecule. This would allow for further characterisation of NDP52, such as measuring the difference in maximum height between the NTD and the CTD for each NDP52 molecule. The results could be subsequently compared across the entire population, to provide additional information on the conformational variation of NDP52 associated with the terminal domains.

This method would have limited success if applied to the current available data, because the coiled-coil is not always distinguishable from the background, which hinders the accurate selection of the full NDP52 molecule. However, it should be considered in future work involving proteins with visible sub-molecular components.

3.5.1.3 Characterising $_{C}$NDP52 by height

Compared to other parts of NDP52, the CTD was found to have a higher affinity towards DNA [151]; therefore, further characterisation of this terminal domain would be useful for the understanding of NDP52-DNA interaction. Since the CTD cannot be automatically selected from images of full-length NDP52, $_{C}$NDP52 (truncated NDP52 composed of the CTD and part of the coiled-coil) was imaged on its own. However, automatic molecule recognition on these images has proved to be challenging, due to image quality issues and the prevalence of oligomerisation, as discussed in Section 3.4.2.

To make the best use of available data and facilitate comparison between different types of molecules, we attempted to select for $_{C}$NDP52 monomers by applying the masking parameters from the identification of terminal domains within full-length NDP52 (FL-NDP52). We then performed automatic measurements of median, mean and maximum heights for each recognised $_{C}$NDP52 molecule, so that they could be compared with FL-NDP52. Here the measurements for FL-NDP52 terminals were used instead of the measurements for entire FL-NDP52 molecules, because previous analysis had shown that the automatic recognition is less accurate for complete FL-NDP52 molecules than for its terminals.

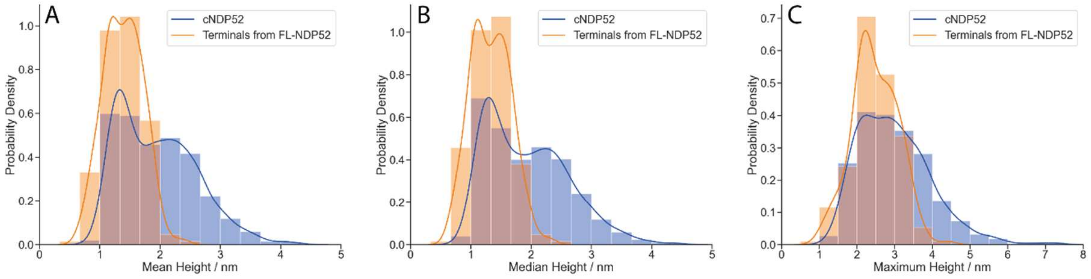


Figure 3.35 KDE plots and normalised histograms comparing between $_{C}$NDP52 (blue, N = 1323, number of images = 17) and terminals recognised from full-length NDP52 (FL-NDP52, orange, N = 190, number of images = 6), through automatic measurements of mean (A, peaks for $_{C}$NDP52 at 1.3 nm and 2.2 nm, peaks for FL-NDP52 terminals at 1.2nm and 1.5 nm), median (B, peaks for $_{C}$NDP52 at 1.3 nm and 2.2 nm, peaks for FL-NDP52 terminals at 1.1 nm and 1.5 nm) and maximum (C, peaks for $_{C}$NDP52 at 2.3 nm and 2.8 nm, peaks for FL-NDP52 terminals at 2.2 nm and 2.9 nm) heights.

As the CTD is smaller than the NTD, the heights obtained from $_{C}$NDP52 were expected to be lower than those from FL-NDP52 terminals. However, the results of the automatic measurement are different from speculated: for all three height parameters, the $_{\mathrm{C}}$NDP52 exhibits two peaks; while the first peak corresponds well with the first peak from FL-NDP52 terminals, the second peak for mean and median heights are at much higher values compared to the second peak from FL-NDP52 terminals.

This deviation from expectation is likely caused by $_{\mathrm{C}}$NDP52 oligomers, which could not be eliminated from the automatic image recognition process, as demonstrated by Figure 3.30 (C) in Section 3.4.2.4. These oligomers have a wider height distribution, which is in agreement with the observation that they have more diverse conformation than monomers of terminal domains. Their mean and median heights are also greater than those of both the CTD and the NTD from FL-NDP52.

Overall, the comparison between the heights measured from different samples confirms that the CTD has a maximum height of about 2.2 nm when attached to a surface. This is slightly smaller the calculated theoretical diameter of 2.5 nm because of adsorption-related conformational changes (see Section 3.5.1.2.2 above). The comparison also provides further proof that when isolated from the rest of the molecule, the CTD is prone to interaction with each other and oligomerisation. It can thus be proposed that the CTD plays a crucial role in the oligomerisation of FL-NDP52, though further AFM study on isolated NTD would be needed to ascertain whether such is the case.

3.5.2 Characterising full-length NDP52 and $_{\mathrm{C}}$NDP52 by size

3.5.2.1 Measuring the size of full-length NDP52

3.5.2.1.1 Parameters to measure

To fully characterise a protein, its conformation on the X-Y plane also needs to be studied. In the case of TOP1, this was achieved through the automatic measurement of area and mean radius. However, while these parameters are well suited for globular proteins, they cannot be used to characterise the structure or size of FL-NDP52, which does not have a regular shape. It is therefore necessary to devise different parameters that are not only capable of representing NDP52 conformation variation, but can also be measured automatically without requiring user input.

The maximum and minimum bounding distances are suitable for this purpose. These are the longest and shortest possible distances between two parallel lines touching the edge of a molecule, as illustrated in Figure 3.36 below, and they can be thought of as representing the length and width of an irregularly shaped molecule.

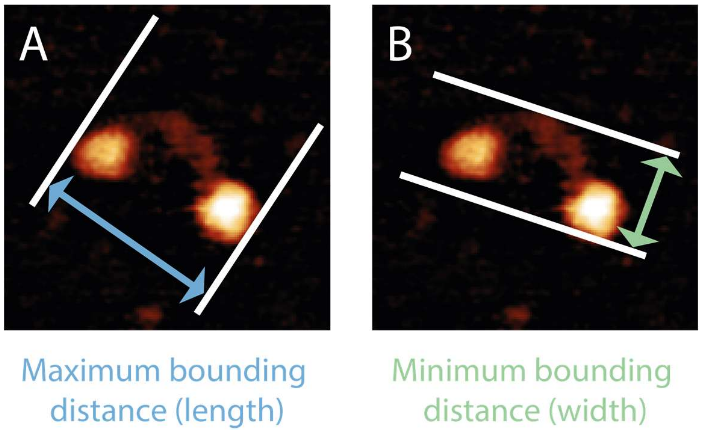

A
B
Maximum bounding distance (length)
Minimum bounding distance (width)

Figure 3.36 Illustration of maximum (A) and minimum (B) bounding distances, using an NDP52 molecule imaged by AFM as an example. These distances can be measured with the aid of two parallel lines touching the edge of the molecule: as the lines rotate, the distance between them changes; the longest and shortest possible distances are the maximum and minimum bounding distances. They are also referred to as the length and width, and can be used to characterise the size and shape of irregular molecules such as NDP52. For a similar illustration made on a DNA molecule, see Figure 4.36.

3.5.2.1.2 Results

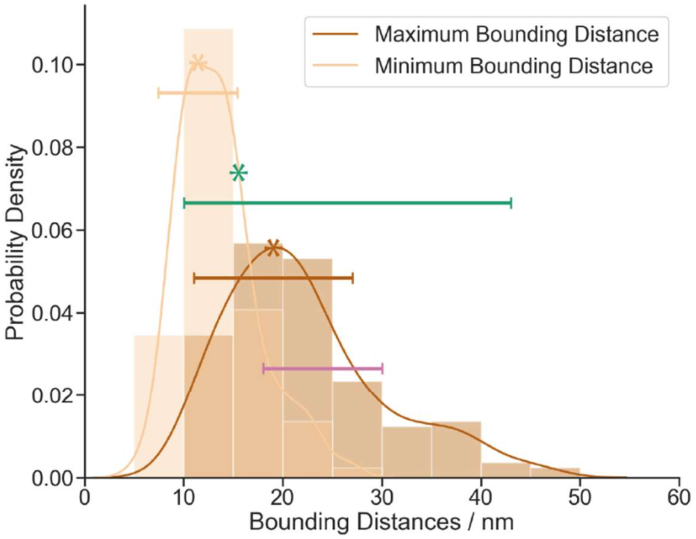


Figure 3.37 Distributions of maximum (dark orange, peak and standard deviation 19 ± 8 nm) and minimum (light orange, 11 ± 4 nm) bounding distances for FL-NDP52 (N = 162, number of images = 6), compared to the diameter measured from dynamic light-scattering (green asterisk and error bar, 10-43 nm peaking at 15.5 nm) [151] *and the end-to-end distance measured from small-angle X-ray scattering (purple error bar, 18-30 nm)* [151].

As shown in Figure 3.37 above, the maximum bounding distance of FL-NDP52 was measured to be 19 ± 8 nm, while the values for the minimum bounding distance were 11 ± 4 nm. The distribution is wider for the maximum bounding distance than for the minimum, because the former is more affected by the flexibility of the coiled-coil, while the latter is theorised to be determined by the diameter of the terminal domains.

To assess the validity of the measurements, we compared them to data from our collaborators obtained through different techniques: dynamic light-scattering showed that the diameter of NDP52 ranged from 10 to 43 nm, peaking at 15.5 nm [151]; small-angle X-ray scattering was used to estimate the end-to-end distance and obtained a range of 18-30 nm [151]. The results from both techniques corresponded well to the maximum bounding distance measured on AFM images, and this suggests that despite challenges with the masking process, the automatic measurement of bounding distances could be used as a reliable method to characterise the size and structure of NDP52.

3.5.2.2 Measuring the size of $_C$NDP52

$_C$NDP52 is globular and does not exhibit an unusual shape. However, in order to facilitate comparison between the size and structure of FL-NDP52 and $_C$NDP52, the maximum and minimum bounding distances were still chosen as parameters to measure for the characterisation of $_C$NDP52 on the X-Y plane.

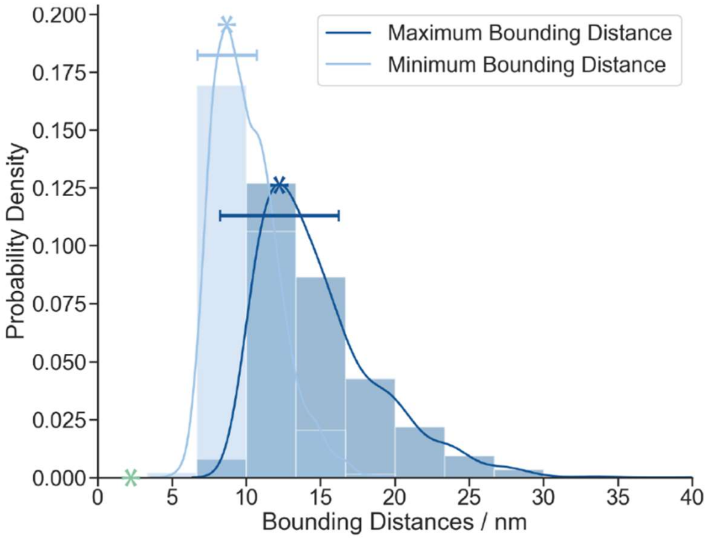


Figure 3.38 Distributions of maximum (dark blue, peak and standard deviation 12 ± 4 nm) and minimum (light blue, 9 ± 2 nm) bounding distances for $_C$NDP52 (N = 1323, number of images = 17). By comparison, the measured maximum height of the CTD (green asterisk, 2.2 nm, see Section 3.5.1.2) is much smaller.

Here the maximum and minimum bounding distances for $_C$NDP52 are 12 ± 4 nm and 9 ± 2 nm respectively. Both distributions exhibit much higher values than the measured maximum height of CTD from FL-NDP52 (2.2 nm, see Section 3.5.1.2), the measured maximum height of $_C$NDP52 monomers (2.3 nm, see Section 3.5.1.3), and the theoretical diameter (2.5 nm) calculated from molecular weight [149] and estimated density [167]. This suggests that the spherical model is not accurate for globular proteins imaged by AFM, since the attachment of the molecule to the surface can induce conformational changes, causing the protein to be flatter and lower, while increasing the area projected onto the X-Y plane.

It can be noticed that the distribution for both parameters contains secondary peaks at higher values. These are likely caused by oligomers, which cannot be excluded from automatic recognition and are also the source of additional peaks for height measurements.

3.5.2.3 Comparing the size distributions of full-length NDP52 and $_{C}$NDP52

3.5.2.3.1 Comparing measurements of the same parameters

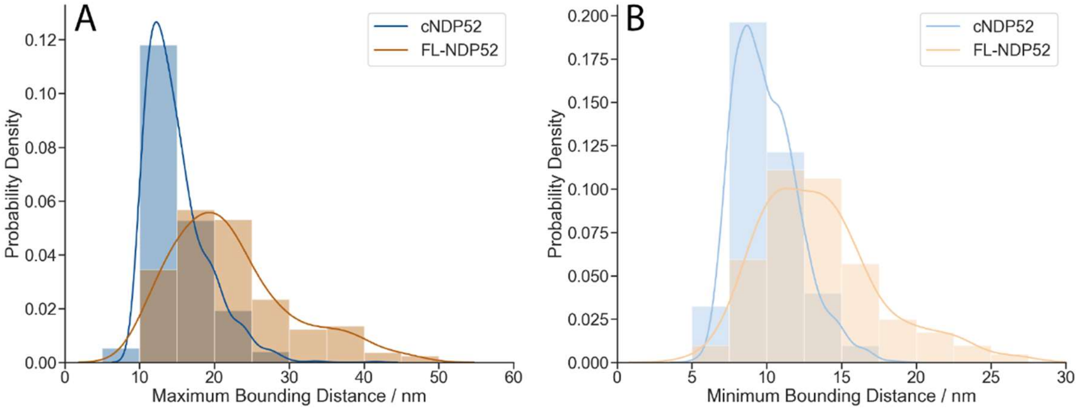


Figure 3.39 Comparisons of maximum (A, peaks and standard deviations 12 ± 4 nm for $_{C}$NDP52, 19 ± 8 nm for FL-NDP52) and minimum (B, 9 ± 2 nm for $_{C}$NDP52, 11 ± 4 nm for FL-NDP52) bounding distances between $_{C}$NDP52 (blue, N = 1323, number of images = 17) and FL-NDP52 (orange, N = 162, number of images = 6).

To ensure that the conclusions derived from FL-NDP52 and $_{C}$NDP52 bounding distances are accurate, measurements of the same parameters are compared between different species. As shown in Figure 3.39 above, for both maximum and minimum bounding distances, when compared to $_{C}$NDP52, FL-NDP52 exhibits a wider distribution and a peak at a higher value. This matches our expectation that, the FL-NDP52 is not only larger than the $_{C}$NDP52, but also has greater conformational variation, because the coiled-coil is seen to be very flexible, ranging from being coiled to being fully extended. The agreement between the measured data and the manual observation of NDP52 structure confirms that bounding distances are suitable parameters for the automatic characterisation of irregularly shaped molecules.

3.5.2.3.2 Comparing the minimum bounding size of full-length NDP52 with the maximum bounding size of $_{C}$NDP52

It was mentioned in Section 3.5.2.1.2 that, since FL-NDP52 exhibits a dumbbell shape, its minimum bounding distance could be dependent on the diameter of the terminal domains. To investigate whether this is the case, the minimum bounding size of FL-NDP52 is compared with the maximum bounding size of $_{C}$NDP52, and the result is shown below in Figure 3.40.

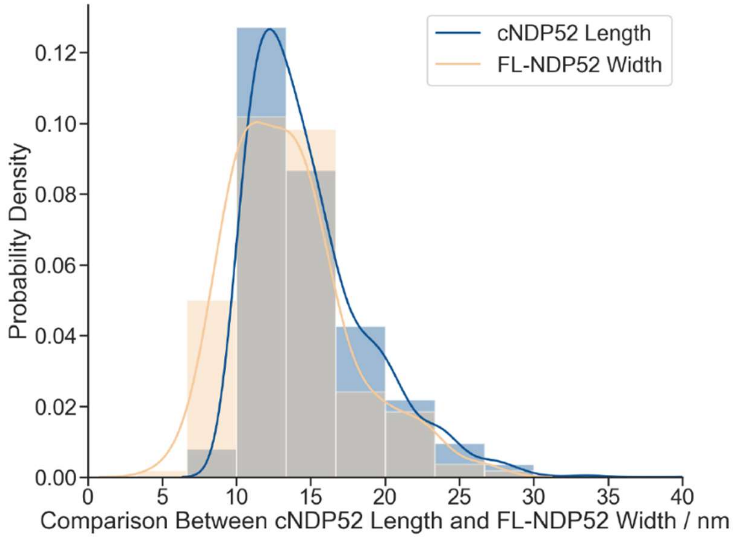


Figure 3.40 Distributions of the maximum bounding distance (length) for $_C$NDP52 (light orange, peak and standard deviation 12 ± 4 nm, N = 1323, number of images = 17) and the minimum bounding distance (width) for FL-NDP52 (dark blue, 11 ± 4 nm, N = 162, number of images = 6). It can be noticed that the two peaks coincide with each other.

When the two distributions are superimposed on each other, it becomes evident that their peaks largely coincide. This confirms the validity of our hypothesis that the width of the FL-NDP52 is in part determined by the diameter of the terminal domains.

Another noticeable similarity between the two distributions is that both are asymmetric, with more data to the right of the peak than to the left, although the cause is different for the two species. In the case of FL-NDP52, its minimum bounding distance is often longer than the diameter of the terminal domains: this happens when the flexible coiled-coil is not fully extended, as exemplified by the molecule in Figure 3.36 (B). The asymmetric distribution of $_C$NDP52 maximum bounding distance, on the other hand, is mainly due to the presence of oligomers.

3.6 Measuring protein oligomers: NDP52

3.6.1 Measuring the proportion of oligomers

3.6.1.1 Comparing automatic and manual measurements

Through the adjustment of masking parameters, we were able to exclusively select for oligomers from AFM images of FL-NDP52. This will allow us to further investigate the structure of oligomers through the automatic measurement of their height and size. However, before performing these characterisations, the reliability of the automatic recognition needs to be assessed. This can be done by comparing the proportion of molecules automatically identified as oligomers with the results of manual identification.

Rollins manually counted that 15% of NDP52 monomers have joined each other to form oligomers, whereas the automation software recognised 40 oligomers and 162 monomers, showing that oligomers make up 20% of all NDP52 molecules. Note that these two percentages do not refer to the same metric: during the manual measurement, an oligomer is counted based on the number of monomers it contains, so the combination of a dimer and three monomer would be counted as an oligomerisation rate of 40%; by contrast, the same group of molecules would lead to a calculated proportion of 25% with automatic recognition, where oligomers cannot be separated into components and are always identified as a whole. Therefore, to allow for direct comparison between the two values, the percentage calculated from the automatic recognition needs to be converted: if all 40 oligomers are assumed to be dimers, there are 80 NDP52 monomers oligomerising out of a total of 282, which is equivalent to a proportion of 33%. This is over twice the value of 15% obtained from manual counting. In addition, if trimers and higher-level oligomers are taken into account, the converted percentage will be even higher.

To ensure that the discrepancy is not caused by errors during the automatic identification process, the molecules recognised as oligomers are re-examined, and it was found out that while a few of them are in fact monomers, their number is small and does not account for the 18% difference in percentage.

The lack of agreement between the automatic and manual measurements could thus be explained by the fact that the former was performed on the subset of high-quality images, while the latter also used lower-quality images that are unsuitable for automatic processing. The oligomers on those lower-quality images are not always clearly identifiable, and consequently, the number of oligomers is likely underestimated during the manual counting.

3.6.1.2 Comparing AFM and SEC-MALS measurements

To further assess the suitability of AFM imaging and automatic analysis as a method for the characterisation of oligomers, the calculated rate of oligomerisation from AFM images is compared with results from another technique, size exclusion chromatography and multi-angle light scattering (SEC-MALS), which was used by our collaborators to measure molecular weight.

SEC-MALS measurements present a main peak at about twice the molecular weight of NDP52, and a smaller peak at a higher value, suggesting that the majority of NDP52 molecules in solution are dimerised, while the rest are mostly in higher oligomeric forms [151]. Although the proportion of oligomers measured through SEC-MALS is much higher than observed on AFM images, this can be explained by the difference in protein concentration: 1 mg/ml NDP52 solution was used for SEC-MALS, while during AFM imaging, 3-8 ng of the protein was added to a 20 µl buffer solution, which is equivalent to 150-400 ng/ml; this means that the concentration used for AFM is lower than for SEC-MALS by several orders of magnitude.

It would be impractical to increase the concentration used in AFM imaging to allow for better comparison with SEC-MALS data, because NDP52 molecules need to be attached to the mica surface, and a high concentration would lead to too much overlapping, which would obstruct both manual observation and automatic identification.

In conclusion, due to the difference in NDP52 concentration required by techniques of different nature, we were unable to establish a direct comparison between the rates of oligomerisation. However, the combination of AFM and SEC-MALS data demonstrated that NDP52 is prone to oligomerisation at varying levels of concentration.

3.6.2 Measuring the height of oligomers

Having probed the rate of oligomerisation, we now attempt to obtain more information on the structure of NDP52 oligomers through automatic characterisation in the Z direction, which can be achieved through the measurement of mean, median and maximum heights.

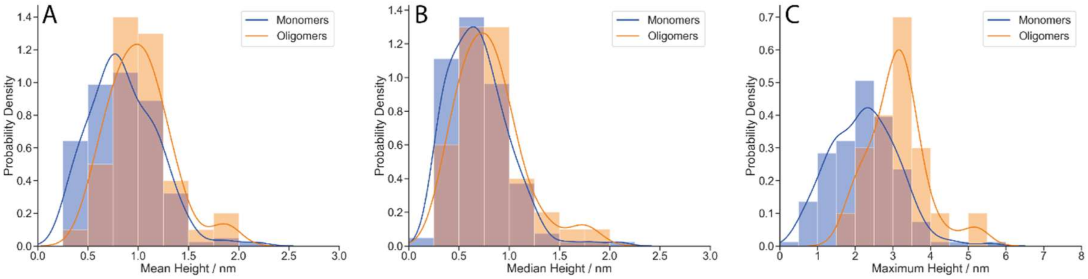


Figure 3.41 KDE plots and normalised histograms comparing the mean (A, peak and standard deviation 0.8 ± 0.3 nm for monomers, 1.0 ± 0.3 nm for oligomers), median (B, 0.6 ± 0.3 nm for monomers, 0.7 ± 0.3 nm for oligomers) and maximum (C, 2.3 ± 0.9 nm for monomers, 3.2 ± 0.7 nm for oligomers) height distributions of monomers (blue, N = 162, number of images = 6) and oligomers (orange, N = 40, number of images = 6).

As shown in Figure 3.41 above, compared to monomers, the distributions for all three height measurements of oligomers have shifted to the right, with the maximum height exhibiting the greatest shift. There is however a large overlap between the distributions of the two populations, even in the case of maximum height. This suggests that some NDP52 molecules oligomerised by stacking on top of each other vertically, causing an increase in height measurements, while other oligomers were formed by monomers joining each other on the X-Y plane, which did not lead to changes in the Z direction. Moreover, it can also be speculated that the extent of vertical stacking varies: most molecules overlap with each other over one terminal domain, and a few molecules experience more complete stacking, resulting in the small peak towards the right end of the height distributions. These fully stacked molecules can be identified on AFM images through manual observation, as they appear to be much higher than other oligomers, as shown in Figure 3.42 below.

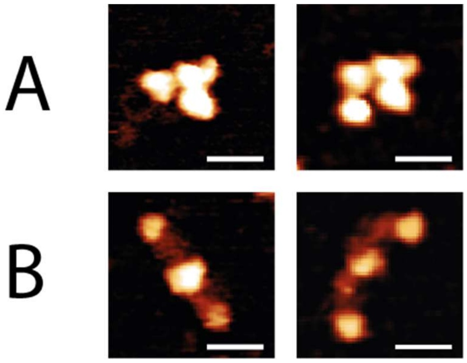


Figure 3.42 Oligomers that are completely stacked (A) appear much higher compared to other oligomers where there is only partial overlapping (B). All scale bars are 20 nm.

In summary, the height characterisation shows that NDP52 can interact with each other in different orientations when forming oligomers.

3.6.3 Measuring the size of oligomers

Following the established characterisation method on the X-Y plane for NDP52 monomers and $_{C}$NDP52, we measured the maximum and minimum bounding distances of NDP52 oligomers, as these parameters have been shown to represent the size and structure of molecules with an irregular shape.

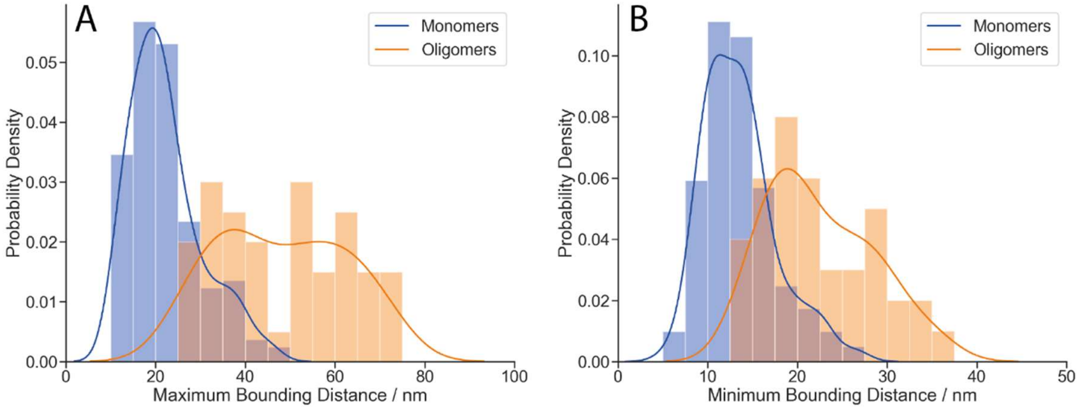


Figure 3.43 KDE plots and normalised histograms comparing the maximum (A, peak and standard deviation 19 ± 8 nm for monomers, two peaks at 38 nm and 57 nm for oligomers) and minimum (B, 11 ± 4 nm for monomers, 19 nm and 28 nm for oligomers) bounding distances of monomers (blue, N = 162, number of images = 6) and oligomers (orange, N = 40, number of images = 6).

In agreement with what was expected, oligomers exhibit higher values for maximum and minimum bounding distances when compared to monomers. Both distributions for oligomers contain two peaks; as discussed in Section 3.6.2, this could be caused by NDP52 molecules combining with each other in different orientations, leading to oligomers with different conformations: those that were formed mainly through interaction on the X-Y plane are expected to have larger bounding distances than others that involved a higher extent of vertical stacking.

In addition, it can be remarked that the distribution of minimum bounding distance for oligomers has the same main peak position as the maximum bounding distance for monomers. However, unlike the comparison between $_{\mathrm{C}}$NDP52 and FL-NDP52 explored in Section 3.5.2.3, the smallest dimension for oligomers is not theoretically determined by the largest dimension of monomers. This similarity in distribution could be coincidental, and may change if the sample size of oligomers is larger. More data would be needed in future studies to draw a more definitive conclusion.

3.7 Conclusions

This chapter has demonstrated that automatic AFM image analysis algorithms can be adapted for different biomolecules, including proteins of regular and irregular shapes. By establishing a set of customisable parameters, molecules of interest can be selected from the background based on height and area criteria, while unwanted molecules are excluded.

Accurate selection depends on identifying an appropriate set of parameters and is often limited by image quality. For example, noise and impurities can prevent molecules of interest from being distinguished from their surroundings, while excessive clustering can prevent neighbouring molecules from being separated during automatic recognition.

Overall, the thresholding methods work best when the molecules imaged are homogeneous in both height and area. When this is the case, there is a high level of overlap between the results of automatic and manual molecular identification, which can be quantified by a high Jaccard index above 0.8. Conversely, when the molecules exhibit a large degree of conformational variation, or when the images contain excessive noise and impurities, the Jaccard index is lower, at around 0.5. The difference between high and low quality images is also manifest when performing parameter sweeps, where higher quality leads to regions that are more distinct. It should however be noted that regardless of image quality, manual observation is currently required to determine the most suitable thresholding parameters for a new sample type, because the automatic parameter sweep only provides information on the number of molecules identified from each set of parameters, and a stable number does not necessarily represent accurate molecular recognition, especially when there are multiple types of molecules in the image.

One possible approach to address datasets with varying image quality is to select a subset of high-quality data and adjust parameters based on them. Those parameters can then be applied to the wider dataset, although this may not be successful if the image quality for the wider dataset is too poor. A better alternative for future automatic analysis is to work alongside experimentalists and ensure that the AFM images they obtained are of sufficient resolution and unaffected by contamination or clustering.

While a completely automatic pipeline is theoretically possible via more advanced algorithms such as machine learning, it will require a higher amount of computational resources and thus render the automatic image analysis software less accessible. This is particularly challenging when the image quality is low, or when the dataset contains multiple types of molecules, due to the increased difficulty in identifying patterns without manual input. As a result, machine-learning-based methods are less justifiable for poor quality data, when taking both the financial cost and the environmental impact into account.

When best practices are followed during both imaging and processing, molecules of interest can be measured automatically and instantly, and their structures can be characterised on the X-Y plane and in the Z direction. By comparing this data with manual measurements and results from other techniques, the quality of the automatic image analysis can be evaluated, and it has been shown that where the data quality permits, the automatic measurement provides a good representation of the molecules of interest.

The combination of AFM imaging and automatic characterisation was applied to study TOP1 and NDP52. TOP1 was shown to be a globular protein, and it could be distinguished from other smaller globular proteins such as BSA; NDP52 was observed to exhibit a unique dumbbell shape with a large conformational variation, which can be attributed to the flexibility of the coiled-coil. Automatic analysis of AFM images also corroborated findings from other techniques that NDP52 readily oligomerises.

Automatic measurements of TOP1 and NDP52 were carried out both in the Z direction and on the X-Y plane. While the values from the former are in theory more reliable, as they are less affected by tip convolution, this data does not show any advantage in characterising proteins by height over area; in fact, when compared with estimations calculated from molecular weight and density, AFM height measurements are lower than expected, while AFM area measurements are larger than expected, suggesting that TOP1 and NDP52 undergo conformational changes during the adsorption to the mica surface, resulting in a flatter and lower structure. This highlights the limitations in characterisation biomolecules with AFM imposed by the necessity of surface immobilisation.

Chapter 4 Tracing and splining DNA, a flexible curved molecule

4.1 Introduction to tracing DNA molecules within AFM images

4.1.1 The role of tracing DNA

This thesis aims to quantitatively characterise DNA conformation and its changes as a result of protein interaction, through developing tools for the automatic analysis of AFM images. While Chapter 3 discussed the recognition and measurement of DNA-manipulating proteins, this chapter will mainly focus on characterisation methods unique to DNA.

Proteins can be directly measured once they are recognised from the background. DNA molecules, on the other hand, are curved lines that follow a trajectory; therefore, to locate specific features such as sharp bends in relation to this trajectory, another step is needed after molecular identification, so that the DNA can be vectorised and represented as a smooth, continuous curve. This additional step is known as tracing, and it is of great value to the automatic characterisation of DNA, because it allows for subsequent measurements such as contour length and end-to-end distance, which provide vital information on the length and shape of DNA. For example, through measuring the contour length, we can find out whether a piece of DNA has been cut short as a result of an interaction; through measuring the end-to-end distance, we can find out whether a piece of DNA has been made more compact.

Building on existing software that contains tracing functions, this chapter improves a process known as splining (see Section 4.1.2.6 and Section 4.2) so that successful completion of tracing can be achieved for DNA molecules that do not form a loop. The improvement in tracing also paves the way for the development of new measurement algorithms including curvature and bending angle, which will be discussed in Chapter 5. In addition, this chapter identifies other areas for future development within the tracing process.

4.1.2 The workflow and challenges of tracing DNA

4.1.2.1 Overview of the tracing workflow

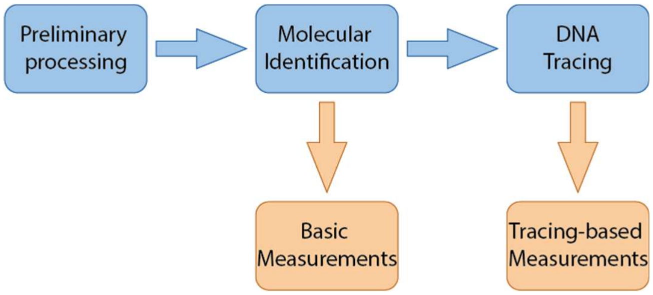


Figure 4.1 The workflow for the automatic characterisation of DNA on AFM images. Improvements made on molecular identification and basic measurements were explored in Chapter 3, while Chapter 4 discusses the improvement of DNA tracing and the application of tracing-based measurements.

In automatic AFM image analysis, tracing takes place after molecules of interest have been identified from the background, as illustrated in Figure 4.1 above. This process starts from all pixels recognised as part of a DNA molecule, and transforms them into a single-pixel-wide, smooth and continuous curve that follows the trajectory of the DNA. The successful completion of the tracing process is challenging due to multiple factors: since DNA is a flexible polymer chain, it does not have a regular shape; therefore, it cannot be simplified into a straight line or approximated by a well-defined curve such as a parabola; moreover, as DNA is small in size, when attempting to determine the direction of the trajectory at certain points such as sharp bends or crossings, there might not be enough pixels to be used as reference points.

To address these challenges, a tracing pipeline needs to be established and tested on a variety of DNA molecules. In this project, the development of such a pipeline is built on an existing workflow [49] which contains five main steps: skeletonisation, shape categorisation, ordering, fitting and splining, as illustrated in Figure 4.2 below:

1. During skeletonisation, the mask for a DNA molecule is simplified into a backbone known as the ‘skeleton’. This is done by gradually removing peripheral pixels, so that only a single-pixel-wide skeleton representing the shape of the molecule is retained.
2. During shape categorisation, a molecule is classified into either circular or linear, with the former involving a closed loop and the latter exhibiting an open curve with two ends. This is necessary because subsequent steps are performed differently based on the shape of the molecule.
3. At the ordering stage, the pixels of the skeleton are arranged into order so that the trajectory of the molecule can be followed.
4. At the fitting stage, the height data from the entire mask is taken into consideration, so that the location of some pixels on the backbone are slightly adjusted to coincide with the highest points. This ensures that the skeleton represents the shape of the DNA and not just the geometric centre of the mask.
5. Finally, during the splining process, the pixels from the adjusted skeleton are joined together to produce a smooth, continuous curve known as a ‘spline’, which will allow us to measure the extent of bending alongside the trajectory of DNA and locate features of interest.

These steps will be discussed in detail in Sections 4.1.2.2 to 4.1.2.6.

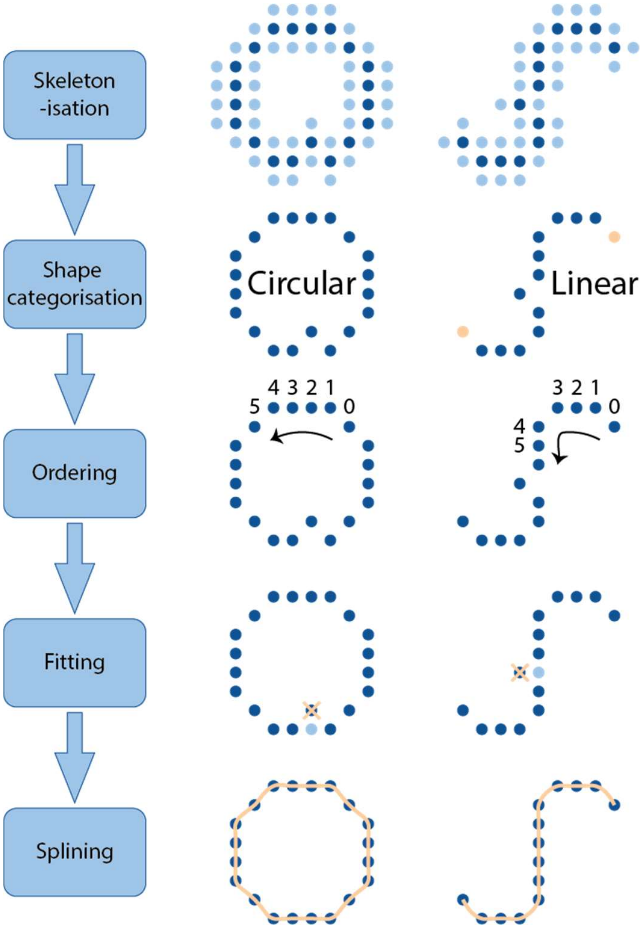


Figure 4.2 This project is built on an existing workflow for tracing DNA which consists of five main steps: during skeletonisation, pixels forming the DNA molecule are simplified so that only a 'skeleton' (dark blue dots) remains; during shape categorisation, each molecule is classified into linear or circular based on the presence of end points (highlighted in light orange); at the ordering stage, the pixels forming the skeleton are arranged into order, so that the trajectory of the DNA can be followed; at the fitting stage, the position of some pixels are slightly adjusted based on the height values of the pixels; at the splining stage, the discreet pixels are transformed into a smooth, continuous curve.

During this project, this workflow was tested on and applied to many different AFM images containing DNA, and it was discovered that DNA molecules cannot always be traced successfully. The most restricting issue identified was that, the splining process of the existing workflow was not successful for DNA molecules classified as linear; this was investigated and rectified, which will be discussed in Section 4.2. Furthermore, DNA can present as complex shapes with crossings, which introduces an additional difficulty in tracing, because it is hard to automatically determine the correct branch after a crossing, especially when the number of pixels is limited, as illustrated by Figure 4.3 below. While there is ongoing work in another project aimed at addressing this challenge, it is not yet fully developed or integrated into the AFM image analysis pipeline. Consequently, the applicability of tracing-based measurements is currently limited when the DNA molecules exhibit complex shapes. This will be discussed in more detail in Sections 4.4.1 and 4.5.1.

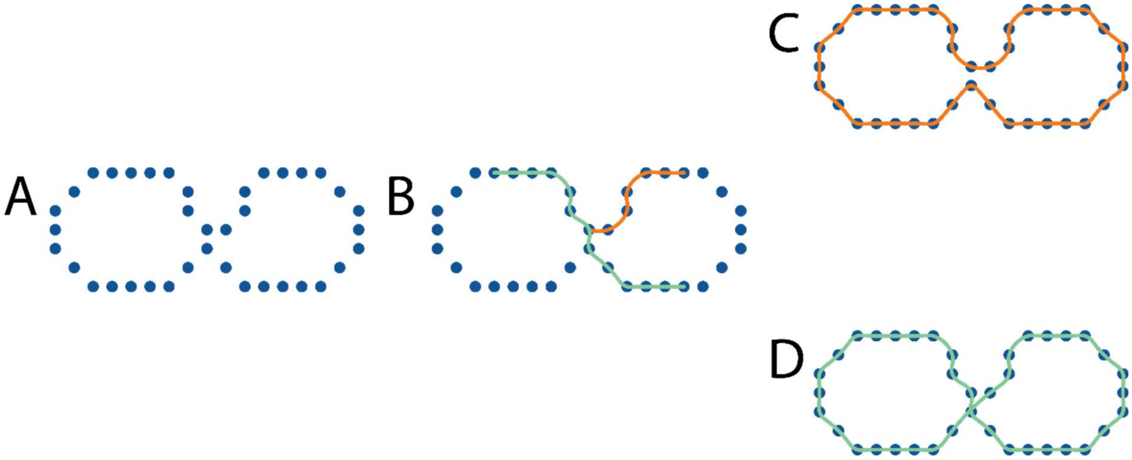


Figure 4.3 Tracing is difficult when crossings are involved due to the presence of branches. A: Schematic illustration of an example molecule containing a crossing that introduces challenges to tracing. B: There are two possible branches at the crossing, with the correct one following the trajectory of the molecule shown in green, and the incorrect one shown in orange. C: When the incorrect branch at a crossing is selected, the tracing of the entire molecule is erroneous. D: When the correct branch is selected, the trajectory of the molecule is followed accurately.

4.1.2.2 Skeletonisation

4.1.2.2.1 Workflow of the skeletonisation process

During the molecular identification process, a mask has been created that covers all pixels of a DNA molecule. However, these pixels form a solid shape without directionality, and they do not allow for measurements such as curvature to be made at specific points of the molecule. Therefore, to enable further characterisation of DNA, the mask must be simplified into a single-pixel-wide backbone, which can then be gradually transformed into a vectorised curve representing the trajectory of DNA during subsequent steps, as introduced in Section 4.1.2.1 above.

The process of extracting the backbone from the mask is referred to as skeletonisation, since the resulting backbone is the skeleton of the molecule. In this project, skeletonisation is achieved by combining an established method [77] with an additional pruning step to remove branches, as illustrated in Figure 4.4 below.

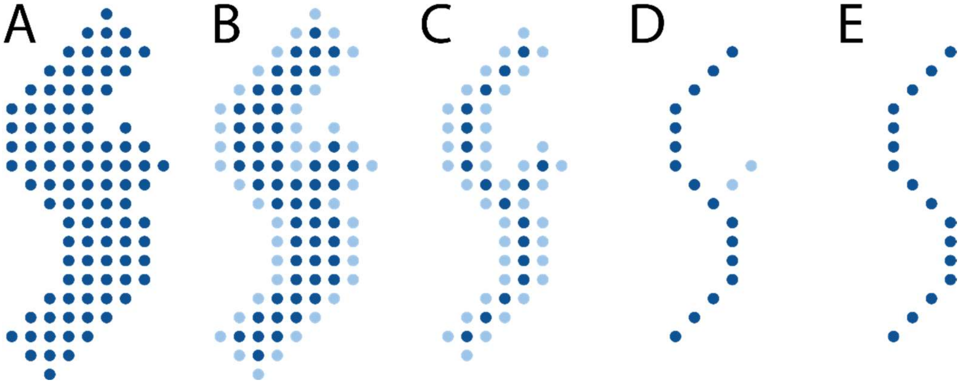


Figure 4.4 The process of skeletonisation. A: Schematic illustration of a mask composed of pixels. This is the starting point of skeletonisation. B & C: Initial skeletonisation following an established method

[77]*, where peripheral pixels are gradually removed. D: The additional pruning step where branches caused by impurities or attached proteins are removed. E: The end result of skeletonisation.*

The backbone produced by the skeletonisation process is also referred to as a disordered trace, since the tracing algorithm has not yet arranged the pixels into the correct order that follows the trajectory of the molecule.

4.1.2.2.2 Limitations of the skeletonisation process

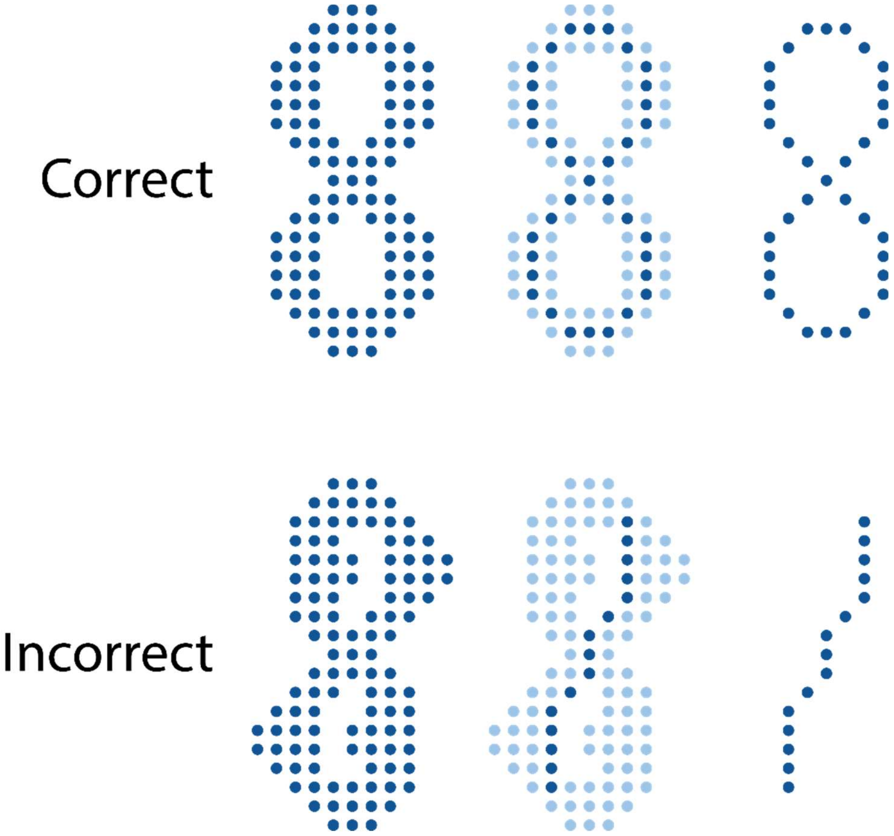


Figure 4.5 Comparison between correct and incorrect skeletonisations of a molecule shaped like the number 8. The incorrect skeletonisation occurs when the entire molecule is regarded as a solid shape and the internal gaps are not taken into consideration; consequently, pixels are only removed from the outside, and the molecule is skeletonised into a single line.

The aforementioned skeletonisation process works well for simple curves, but its reliability is limited for complex shapes where crossings are involved. For example, as shown in Figure 4.5 above, a molecule with the shape of the number 8 may be skeletonised into a single line if the two internal gaps are too small and disregarded, causing the molecule to be treated as a solid shape. This is particularly prone to happen when the number of pixels is small, and the spacing between neighbouring pixels is large, as shown in Figure 4.6 below. It is therefore better to take higher-resolution images of complex molecules to minimise the chance of incorrect skeletonisation.

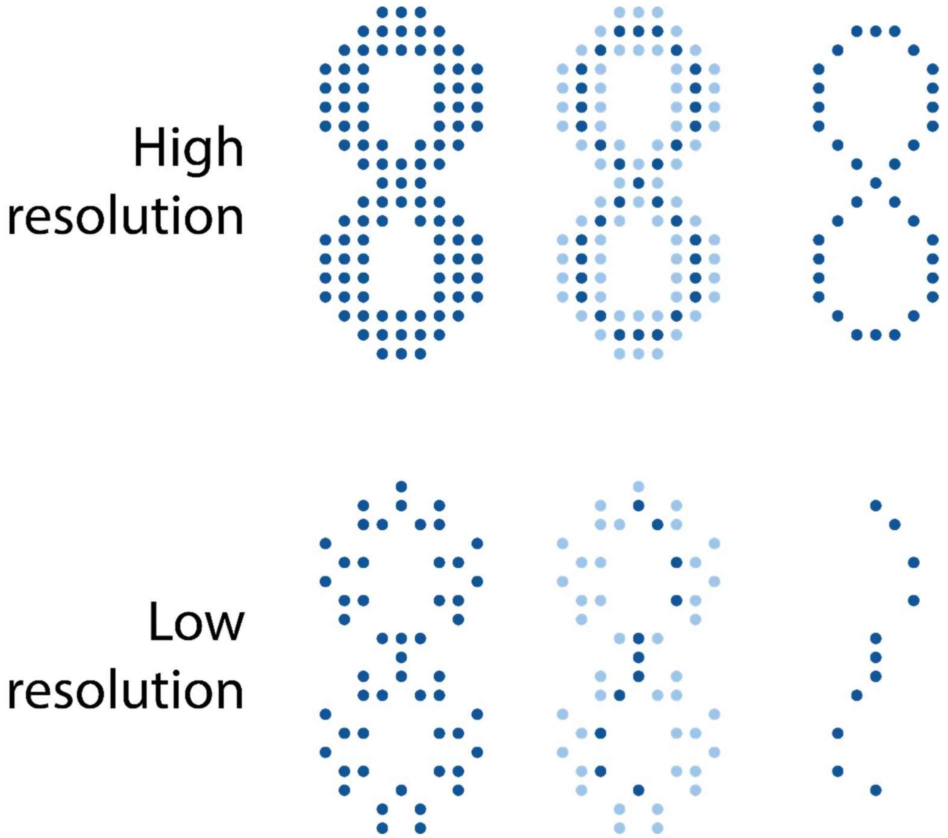


Figure 4.6 Comparison between correct and incorrect skeletonisations of the same molecule imaged with different resolutions. When the image resolution is low, the number of pixels is low and the average spacing between pixels is large; this increases the likelihood for the internal gaps of the 8 shape to be ignored, and for the molecule to be erroneously skeletonised into a single line.

Besides the difficulties with crossings, another limitation of the skeletonisation process identified during this project is that, since the algorithm removes pixels from the border, the ends of molecules are cut short, as illustrated in Figure 4.7 below. This is problematic because it will introduce inaccuracies when measuring contour length and end-to-end distance, or when locating features of interest in relation to the length of the DNA.

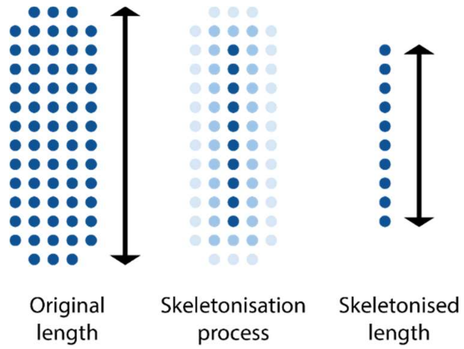


Figure 4.7 The skeletonisation process gradually removes pixels from the border of the molecule, which results in the molecule being shortened. This will affect downstream measurements of contour length and end-to-end distance.

At the moment, no method to circumvent this issue has been implemented. However, in the future, this can potentially be addressed by artificially re-adding the ends of the molecule after skeletonisation.

A third limitation is that the current skeletonisation algorithm does not remove pixels simultaneously from all directions [77]. This leads to inconsistencies where the same molecule is skeletonised differently when placed in different orientations.

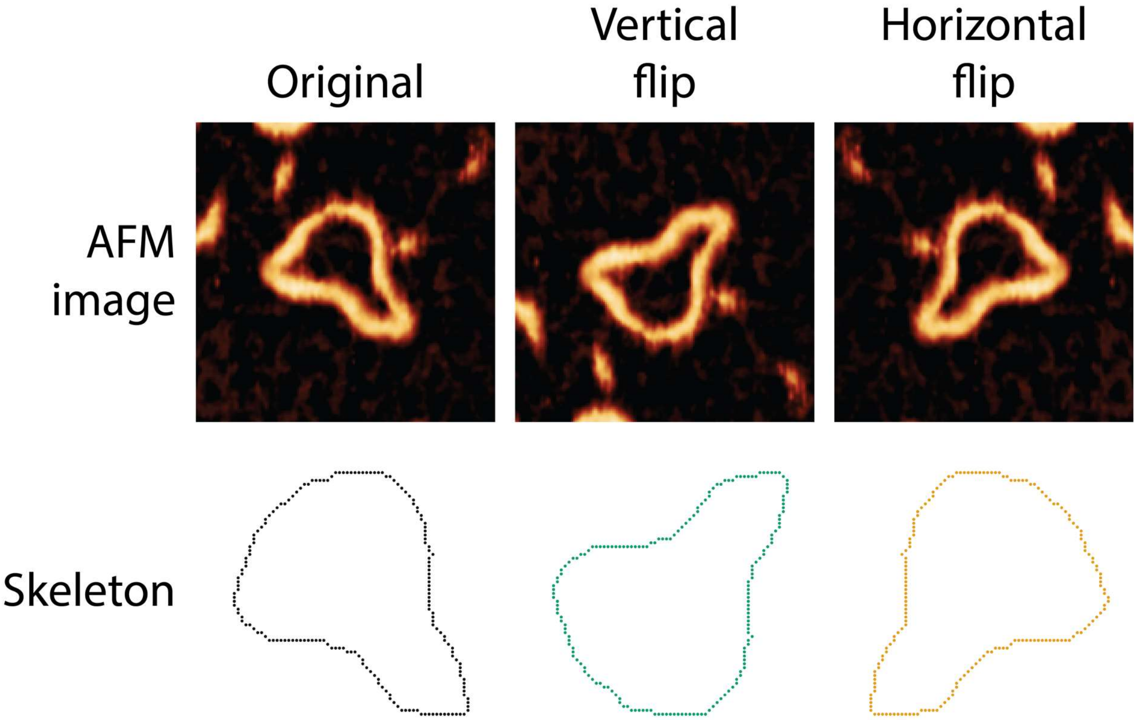


Figure 4.8 Skeletons produced from horizontally or vertically flipped AFM images retain the general shape of the DNA molecule.

For example, as compared in Figure 4.8 above, when an AFM image is flipped horizontally or vertically, the resulting skeleton exhibits the same general shape as the DNA molecule, and is similar to the skeleton directly produced from the original image. However, closer observation reveals that the exact positions of pixels from different skeletons do not match with each other. As presented in Figure 4.9 (A), when skeletons from the original image and the vertically flipped image are overlaid on each other, their extent of overlap can be quantified by a relatively low Jaccard index of 0.53, where 1 represents complete overlap and 0 represents no overlap; similarly, Figure 4.9 (B) shows that the extent of overlap between the skeletons from the original and the horizontally flipped images is only slightly higher, with a Jaccard index of 0.60. These inconsistencies in the skeletonisation process can be observed on all DNA molecules regardless of their shape complexity, although their effect on the overall tracing accuracy is inconsequential when the shape of the molecule is simple.

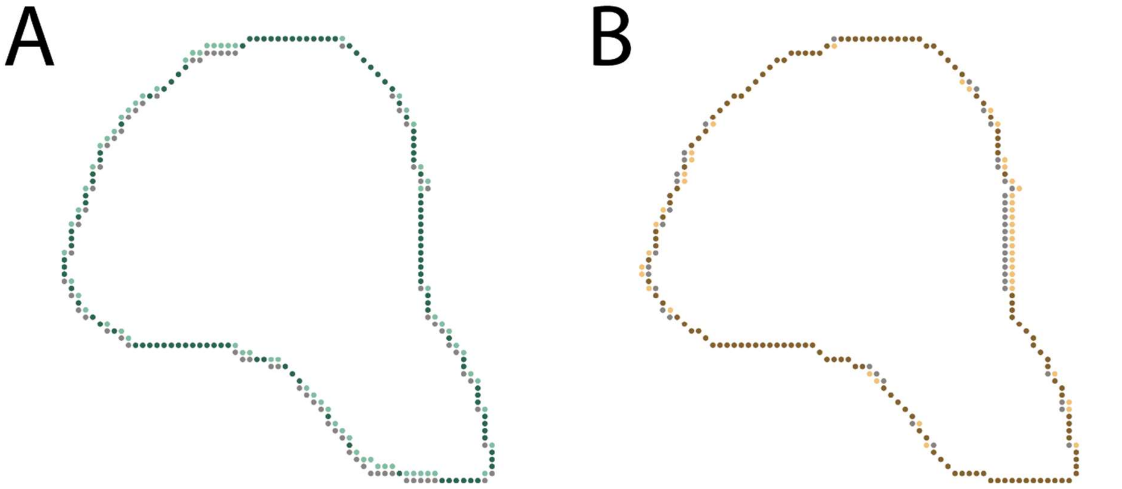


Figure 4.9 Skeletons produced from flipped images do not completely overlap with the skeleton produced from the original image. A: The skeleton produced after vertically flipping the molecule (light green) is overlaid on the original skeleton (grey), and only partial overlap (dark green) is observed; the Jaccard index for these two skeletons is 0.53, indicating a limited extent of overlap. B: The skeleton produced after horizontally flipping the molecule (light orange) is overlaid on the original skeleton (grey), and only partial overlap (dark orange) is observed, with a slightly higher Jaccard index of 0.60.

It is also of interest to investigate whether similar issues will occur when an AFM image is rotated by 90°. This is challenging as AFM images are taken line-by-line in a raster scan pattern, and rotating by 90° can cause issues for the automatic image processing workflow. For example, the current algorithm to address scan line artefacts is not effective when the scan lines have been rotated. A potential solution to this in the future is to improve the upstream preliminary processing stage so that the user can manually select the orientation of scan lines.

4.1.2.3 Categorising into linear or circular molecules

4.1.2.3.1 Workflow of the categorisation process

While the skeleton produced by the previous step represents the shape of the molecule, it does not yet allow us to locate features of interest on the DNA, because the pixels it contains have not been arranged into the correct order. To put them into order, it is necessary to first identify a pixel as the starting point. This process would be different between ‘open’ or ‘linear’ molecules that naturally have two ends, and ‘closed’ or ‘circular’ molecules for which the ordering can start at any point. Therefore, to ensure that the pixels forming both types of molecules can be ordered, the molecules need to be categorised first, so that the suitable ordering method can be used for each molecule according to its shape.

The skeleton is a single-pixel-wide curve, which means that two pixels at each end of a typical linear molecule have one neighbour, while pixels in the middle of the same molecule have two neighbours. On the other hand, all pixels on a circular molecule have two neighbours. Because of this difference, a molecule can be automatically classified into linear or circular based on whether it contains pixels with only one neighbour.

To achieve this, the number of neighbours for each point on the skeleton is first counted. As shown in Figure 4.10 below, since the pixels are located on a grid, each point has eight surrounding spaces, and other pixels found within these eight spaces are counted as neighbours.

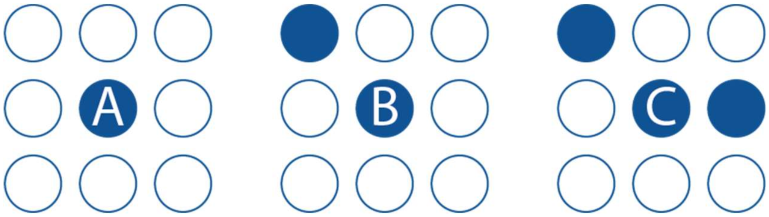


Figure 4.10 In order to classify molecules, the number of neighbours for each pixel is counted based on the occupancy of its eight surrounding spaces. A: A pixel with no neighbours. B: A pixel with one neighbour. C: A pixel with two neighbours. For a typical circular molecule, all pixels on the skeleton have two neighbours, while for a typical linear molecule, two of its pixels have only one neighbour.

In the following step, the algorithm searches the skeleton for pixels with only one neighbour. If no such pixel is found, the molecule is classified as circular (Figure 4.11 (A)). If at least one such pixel is found, the molecule is classified as linear. This is because while most linear molecules have two ends and therefore two pixels with one neighbour (Figure 4.11 (B)), some linear molecules form a partial loop (Figure 4.11 (C)) and only contain one pixel with one neighbour. These molecules are classified as linear because they have a clear starting point for the subsequent ordering process; in fact, they can be considered as having two ends with one end located next to the middle of the molecule.

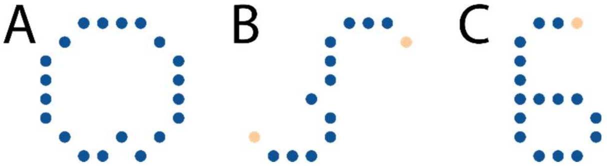


Figure 4.11 Schematic illustrations of different skeletons with points having only one neighbour coloured in light orange. If a molecule exhibits no such points, it is classified as circular (A); if it exhibits two such points, it is classified as linear (B) since it has two open ends; if the molecule is a partial loop and exhibits only one such point, it is also classified as linear (C), because it contains a clear starting point for the ordering of pixels.

4.1.2.3.2 Limitations of the categorisation process

Since categorisation is based on the result of skeletonisation, the reliability of the former is dependent on the successful completion of the latter. Consequently, molecules that exhibit complex shapes cannot always be categorised accurately due to difficulties with their skeletonisation. For example, a molecule shaped like the number 8 may be miscategorised as linear, because it has been erroneously skeletonised into a single line with two ends, as shown in Figure 4.5 and Figure 4.6 from Section 4.1.2.2.2.

Moreover, in certain cases, even molecules with relatively simple shapes can be placed in the wrong category: as shown below in Figure 4.12, when two ends of a linear molecule are too close to each other, they could be joined together during masking, resulting in a skeleton that resembles a loop. This causes the molecule to be mistakenly classified as circular.

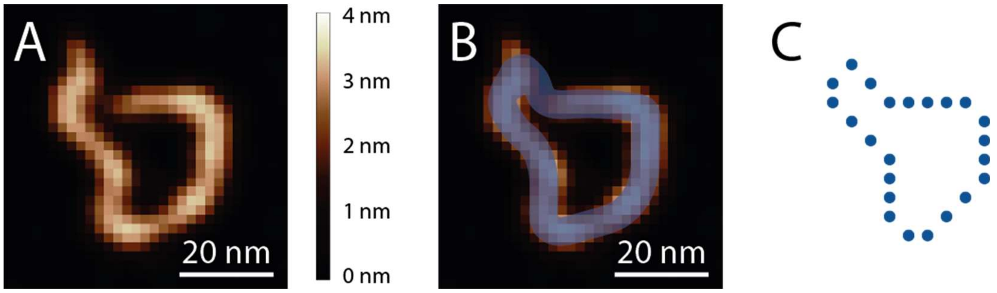


Figure 4.12 An example linear molecule that is misclassified as circular because its two ends are too close to each other. A: AFM image of the molecule accompanied by its height scale. B: Schematic illustration of the mask, where the two ends are joined. C: Schematic illustration of the skeleton, which is a closed structure and does not contain points with only one neighbour; as a result, the molecule is wrongly classified as circular.

4.1.2.4 Ordering

4.1.2.4.1 Workflow of the ordering process

The automatic DNA tracing program stores each pixel on the skeleton as a pair of coordinates. These coordinates need to be rearranged into an order that follows the trajectory of the molecule, so that features of interest on the DNA can be located and measurements such as contour length can be performed.

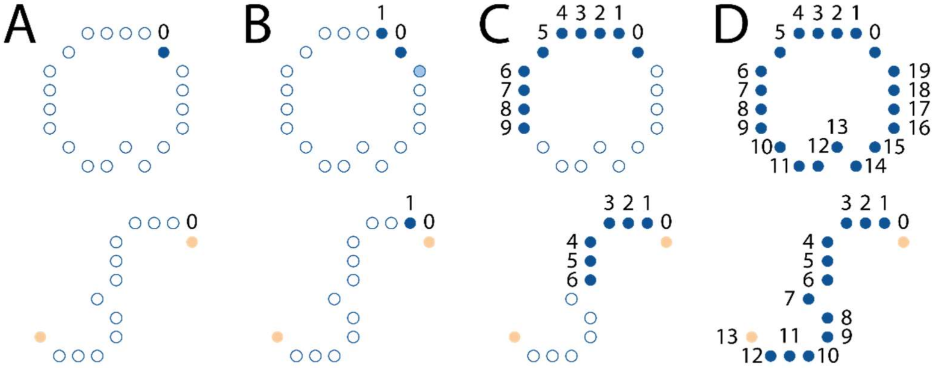


Figure 4.13 Pixels forming circular (top row) and linear (bottom row) molecules are added to an ordered list one by one, so that they follow the trajectory of the DNA. A: The starting pixel (number 0) on the ordered list is chosen randomly for circular molecules, but needs to be a point with only one neighbour (both points with one neighbour are marked in orange) for linear molecules. B: The next pixel (number 1) on the list is the closest neighbour to the starting pixel; there are two valid choices for circular molecules (the alternative option is marked in light blue) but only one for linear molecules. C: The ordering process continues in the same way for circular and linear molecules. D: The ordering process is completed when the starting pixel (for circular molecules) or the other end (for linear molecules) is reached.

To achieve this, an ordered list is created, and the coordinates are added to the list one by one, as shown in Figure 4.13 above: firstly, the coordinates of the starting point are added to the list; as mentioned in Section 4.1.2.3.1, for circular molecules, this can be any point on the skeleton, while for linear molecules, this needs to be a point that has only one neighbour.

In the next step, the point closest to the starting point is identified and added as the second item on the list; for linear molecules, there is only one such point, while for circular molecules, there are two potential candidates, and either of them can be chosen.

The rest of the ordering process is the same for both types of molecules: among the remaining points that have not been ordered, the closest one to the second point is added as the third item on the list, and the closest point to the third is added as the fourth item. This repeats and continues until the starting point of a circular molecule or the other end of a linear molecule has been reached. If multiple closest points have been identified at the same time, the point that leads to the smallest change in angle is chosen as the next point on the list, as illustrated in Figure 4.14 below.

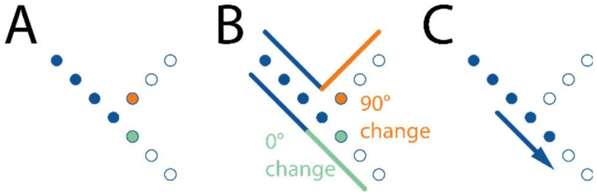


Figure 4.14 Determining the correct path at a crossing based on the change in angle. A: Two points (in orange and green) are identified as the next closest point at the same time. B: The change in angle induced by each point is calculated, with the orange point leading to a 90° change and the green point leading to no change. C: The green point is therefore chosen as the next point on the ordered list, and the ordering continues in the direction indicated by the arrow.

The result of the ordering process is referred to as an 'ordered trace', which can then be adjusted and converted into a smooth, continuous curve in subsequent steps.

4.1.2.4.2 Limitations of the ordering process

It has been discussed that skeletonisation and categorisation do not always work for molecules with complex shapes that involve crossings; however, even in cases where skeletonisation and categorisation have been successful, crossings can pose new challenges for the ordering process: as shown in Figure 4.15 below, the next closest pixel at a crossing might not follow the trajectory of the molecule, and as a result, the ordering continues in the wrong direction, and part of the molecule is discarded. In the future, this can potentially be addressed by prioritising maintaining the direction of the curve over identifying the closest pixel.

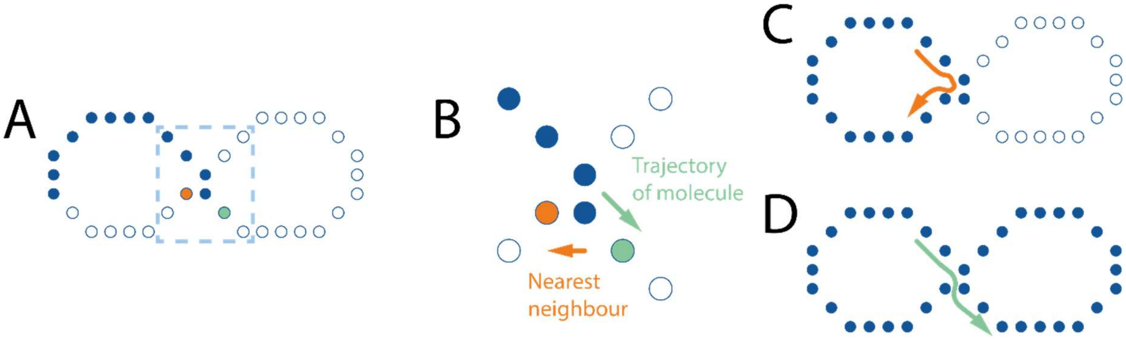


Figure 4.15 An example molecule with a crossing that causes the ordering process to be problematic due to the nearest neighbour not being on the trajectory of the molecule. A: When arranging the pixels into order, a crossing is encountered, as highlighted in the dashed box. B: Close-up view of the highlighted crossing, where the nearest neighbour (orange) is not the correct pixel (green) that allows the trajectory of the molecule to be followed. C: If the nearest neighbour is chosen at the crossing as the next item on the ordered list, some pixels will be removed from the ordered trace; because of this, part of a molecule can be discarded during ordering. D: If the pixel that follows the trajectory of the molecule is chosen instead, all pixels will be ordered correctly and the entirety of the molecule will be preserved.

4.1.2.5 Fitting

4.1.2.5.1 Workflow of the fitting process

The tracing process until this point has been based on the mask of the DNA, which is a collection of all pixels that have been identified as part of the molecule. This approach is not ideal, because the mask makes no distinction between pixels on the backbone of the DNA and pixels on the periphery. Since the skeleton is the geometric centre of the mask, it might not accurately reflect the shape of the DNA if the periphery included in the mask is asymmetrical.

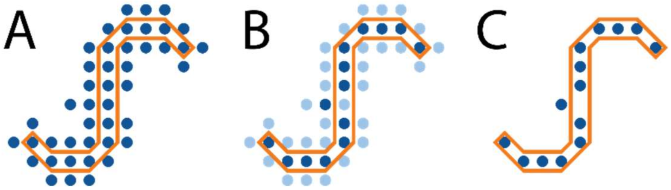


Figure 4.16 The skeleton is the geometric centre of the mask, but does not always correspond to the backbone of the DNA. A: Schematic illustration of all pixels forming a mask with the backbone of the molecule outlined in orange. B: During the skeletonisation process, some pixels (light blue dots) are removed and only the geometric centre of the mask is retained. C: The resulting skeleton does not coincide exactly with the backbone of the molecule.

The fitting stage has been devised to address this issue by adjusting the ordered trace, so that the height difference among points forming the mask can be taken into consideration, and that the position of the pixels on the skeleton can be more representative of the DNA backbone.

This process is carried out for each pixel of the ordered trace as illustrated in Figure 4.17 below: to start with, the direction perpendicular to the trajectory of the DNA at that point is calculated; subsequently, all pixels from the mask located along this perpendicular direction are identified; the heights of these pixels are then compared; finally, the highest pixel among them replaces the original pixel of reference on the ordered trace if they are different. When the fitting process has been completed for all pixels, a new skeleton is produced, and it is referred to as the 'fitted trace'.

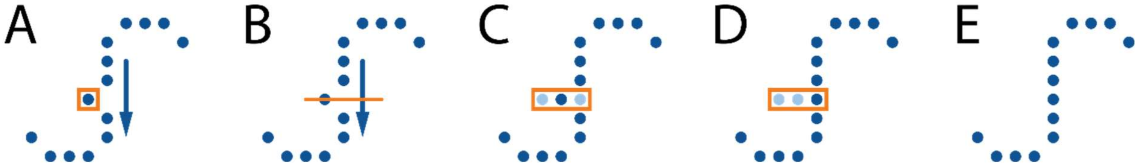


Figure 4.17 Schematic illustrations of the fitting process. A: For each pixel on the ordered trace, the direction of the DNA trajectory around that pixel is calculated (blue arrow). The pixel of interest in this instance is highlighted with an orange box. B: The direction perpendicular to the trajectory is subsequently identified (orange line). C: All pixels from the mask located along this perpendicular direction are identified. D: The highest among these pixel is chosen to replace the original pixel of interest, so that the skeleton represents the backbone of the molecule more accurately. E: The resulting new skeleton.

4.1.2.5.2 Limitations of the fitting process

When proteins or high impurities are attached to the DNA, the fitting process could produce a point that is too far from the rest of the backbone, as shown in Figure 4.18 below. This does not reflect the trajectory of the DNA, and can result in erratic lines instead of smooth curves during the following splining stage, since a higher pixel does not always indicate proximity to the DNA backbone, and may instead be caused by unwanted high features. While this can be addressed by increasing the smoothness parameter of the splining function (see Section 4.2.2.2), a more effective approach would be to take the height difference between pixels into account while skeletonising the molecule, so that the fitting stage would not be needed. However, this new approach is beyond the scope of this project, as it would require a drastic modification of the skeletonisation algorithm.

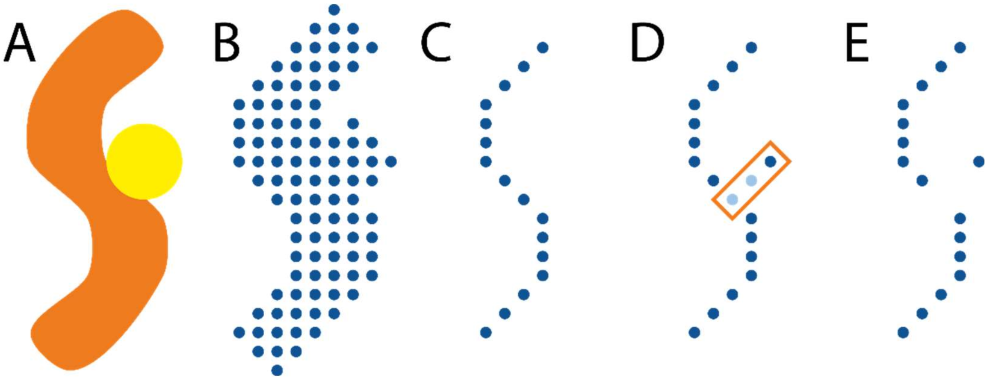


Figure 4.18 The fitting process can introduce inaccuracies when a higher protein is attached to the DNA. A: Schematic illustration of a DNA molecule (orange) with an attached higher protein (yellow) as imaged by AFM. B: The mask of the molecule, where no distinction is made between the DNA and the protein. C: The ordered trace of the molecule follows the trajectory of the DNA, because the protein has been removed during skeletonisation. D: During the fitting process, the heights of all pixels from the mask are taken into consideration, and a pixel that does not fall on the backbone of the DNA is selected because the protein is higher. E: The resulting fitted trace does not reflect the trajectory of the molecule, and can cause further issues for the following splining process.

An alternative method to prevent deviation from the DNA backbone would be to compare different candidate traces after splining with the mask, instead of automatically selecting the trace with the highest pixels. This would allow traces that cause erratic lines in the splining process to be replaced based on their lack of overlap with the mask. Compared to redesigning the skeletonisation algorithm, this method would be more compatible with the current tracing workflow, and could be explored in the near future.

4.1.2.6 Splining

4.1.2.6.1 Workflow of the splining process

After successful fitting, the skeleton is adjusted to more accurately represent the backbone of the DNA, but it is still composed of unconnected and sparsely distributed pixels. This does not allow us to measure the contour length along the trajectory of the molecule or to locate features of interest such as sharp bends. Therefore, in order to enable further characterisation, the fitted trace needs to be transformed into a smooth, continuous curve.

This transformation, known as splining, is achieved through the Python function 'scipy.interpolate.splprep' [171], which fits a polynomial curve between each neighbouring pair of data points in N-dimensional space based on adjustable parameters, as will be further discussed in Section 4.2. To ensure the smoothness of the resulting curve and prevent single misplaced pixels from heavily affecting the splining process, the fitted trace is first divided into several sub-traces, as shown in Figure 4.19 below. The 'scipy.interpolate.splprep' function is then applied to each sub-trace to generate multiple curves. Finally, these curves from different sub-traces are averaged to produce one single smooth, continuous curve, which is referred to as the splined trace. It serves as the starting point for further measurements and characterisations based on the trajectory of the DNA.

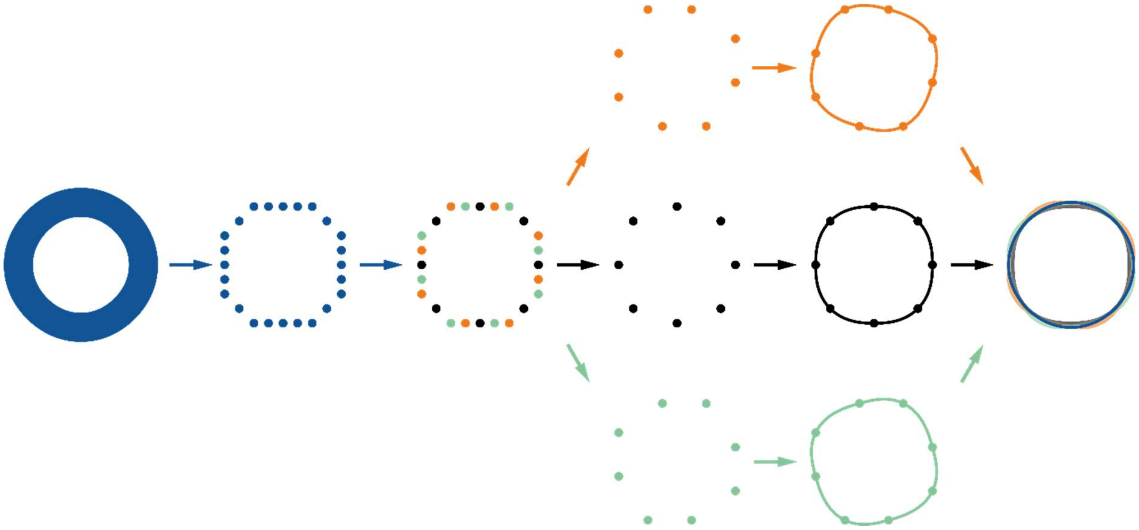


Figure 4.19 Schematic illustration of the splining process. The fitted trace (blue dots) obtained from the DNA molecule (thick blue ring) is divided into three sub-traces (orange, black and green dots). Each sub-trace is transformed into a continuous curve separately, and these curves are averaged in the end to ensure smoothness.

Since the splining process is vital to many characterisation methods developed in this project, it was examined carefully for potential improvements. In particular, it has been verified that splitting the fitted trace into sub-traces is necessary, because if this approach is not used, the shape of the resulting curve will be limited by the position of the pixels and thus less reflective of the DNA, as shown in Figure 4.20 below.

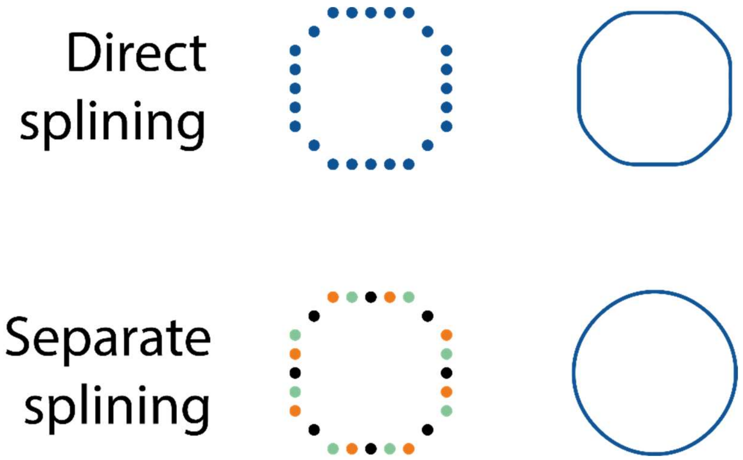


Figure 4.20 Comparison between splining directly and separating the trace into sub-traces first. When the splining takes place directly on the fitted trace, the shape of the resulting curve is affected by the position of the pixels, which can only be located at integer intervals of the smallest length unit. In contrast, when the separation and averaging method is applied, the resulting curve is smooth and more representative of the shape of the DNA.

4.1.2.6.2 Limitations of the splining process

The splining process from the existing software was only functional for DNA molecules classified as circular, and did not work on those classified as linear. This presented a significant obstacle to the characterisation of DNA, because linear molecules are widely used in projects that aim to investigate the effects of proteins on DNA conformation. Without the splined trace, measurements including end-to-end distance, contour length, curvature, and bending angle cannot be accurately calculated. Therefore, prior to developing more tracing-based characterisation methods, the splining process was examined and improved first, so that it could be applied to all molecules. This will be discussed in detail in the following Section 4.2.

4.2 Splining for linear molecules

4.2.1 Identifying the problems of existing code on splining linear molecules

To enable splining for linear molecules, it was necessary to first identify the reason why the existing software was unable to produce splines from them. Upon examination of the code, it was discovered that the splining algorithm employed in the tracing workflow differed between circular and linear molecules. As mentioned in Section 4.1.2.6.1, for circular molecules, the Python function ‘scipy.interpolate.splprep’ was used to perform interpolation. In comparison, the attempted splining for linear molecules used the ‘scipy.interpolate.interp1d’ function, which had been designed for interpolating 1D functions. This is not suitable for purpose, because linear molecules do not always correspond to a 1D function; instead, it represents a 2D curve, where multiple points could share the same X coordinate but have different Y coordinates.

Based on these findings, it was concluded that the failure of splining for linear molecules was caused by an incompatible interpolating function; to resolve this, a different function was needed. The function that worked successfully for circular molecules, ‘scipy.interpolate.splprep’, was tested on linear molecules.

This function has several adjustable parameters, among which the most important ones are degree of the spline, periodicity and smoothness. For circular molecules, the degree of the spline was set to the default value of 3, which meant that curves were approximated as third-order polynomials. This was kept for linear molecules to ensure consistency. However, the periodicity and smoothness, which were both set to 2 for circular molecules, needed to be adjusted to for linear molecules to prevent the interpolated curve from forming a closed loop or behaving erratically. These adjustments are described in Section 4.2.2.

4.2.2 Implementing new splining algorithm for linear molecules

4.2.2.1 Changing periodicity

The periodicity for circular molecules is set to 2, which ensures that the data is considered as periodic and that a closed loop can be produced. When the same value was applied to linear molecules, the resulting curves were also closed loops, as illustrated below in Figure 4.21.

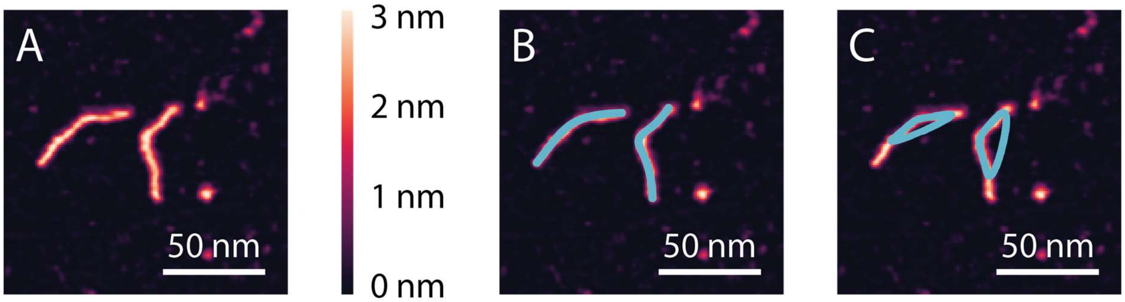


Figure 4.21 Incorrect splining of linear molecules. A: An AFM image of linear DNA molecules; the height scale attached is used for Figures 4.21, 4.22, 4.23, 4.26, 4.28, 4.34, 4.36, 4.45, 5.13, 5.18 and 5.19. B: The ideal result of splining is drawn as blue curves. C: When the periodicity is set to 2 (the value used for circular molecules), the splining function incorrectly generates closed loops from linear molecules.

To address this, we first attempted to lower the periodicity to 1 for a test image, but this did not lead to any noticeable change in the shape of the curves produced. When the value was further lowered to 0, the linear molecules were splined correctly. After applying the new parameter to a wide range of images and verifying the results, it could be confirmed that 0 is the ideal periodicity value for linear molecules.

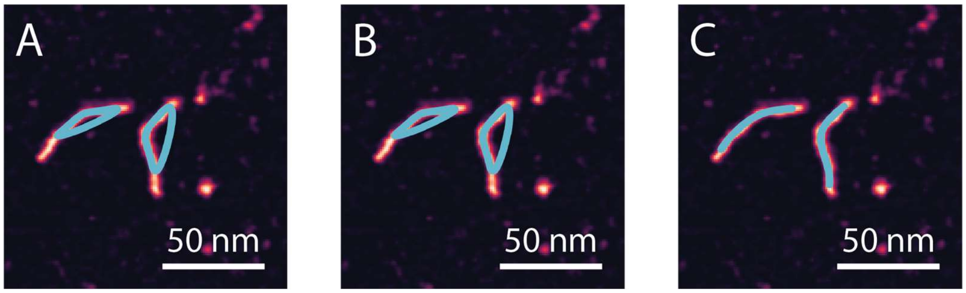


Figure 4.22 Adjusting the periodicity parameter for the splining of linear molecules. When the periodicity is set to 2 (A) or 1 (B), the splining function generates closed loops. When the periodicity is reduced to 0 (C), curves that accurately follow the trajectory of the molecules can be produced.

4.2.2.2 *Fixing erratic splines*

4.2.2.2.1 Investigation of the cause

Whilst testing the new splining algorithm for linear molecules, another issue was identified: on rare occasions, erratic curves deviating substantially from the trajectory of the DNA could be produced, as shown in Figure 4.23 below:

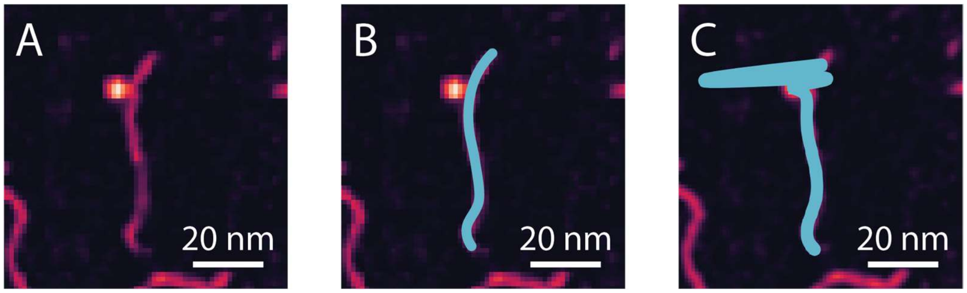


Figure 4.23 Erratic splining can occur, especially when a protein molecule is next to the DNA. A: An AFM image of a linear DNA molecule with an attached protein. B: The ideal result of splining is drawn as a blue curve. C: The splining function produces an erratic curve that deviates substantially from the trajectory of the DNA.

To investigate the cause of this, we examined the intermediate results produced after each step of the tracing workflow: it was discovered that, as discussed in Section 4.1.2.5.2, the protein caused the fitting process to introduce pixels that did not lie on the DNA backbone. Consequently, when the splining algorithm attempted to generate a cubic curve involving these pixels, a spline that did not follow the trajectory of the molecule was produced.

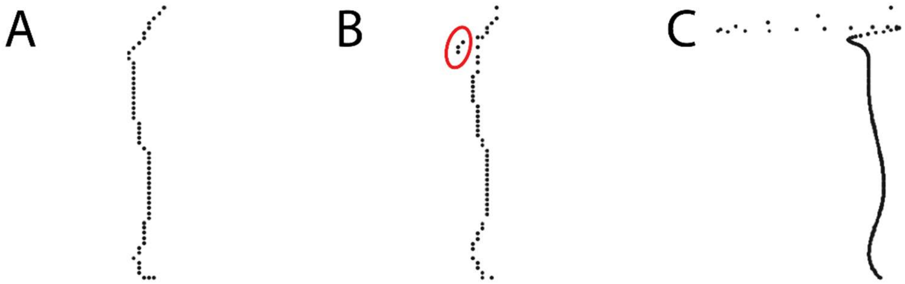


Figure 4.24 Comparison of results from different stages of the tracing workflow. A: The trace produced after the ordering stage is correct and accurately represents the trajectory of the DNA. B:

The fitting stage incorrectly introduces pixels that deviate from the backbone of the molecule, as circled in red. C: Due to the erroreous pixels from the fitting stage, the final splining stage generates more points that deviate even further from the molecule, forming erratic splines.

In addition, since the spline was obtained by taking the average of three curves produced separately from three sub-traces, each of these three curves was inspected. As demonstrated by the example in Figure 4.25 below, in most cases the erratic spline was solely caused by irregularities from one particular sub-trace (Figure 4.25 (C)). This indicates that it may be possible to automatically evaluate the reliability of the final spline by calculating the similarity between the sub-splines.

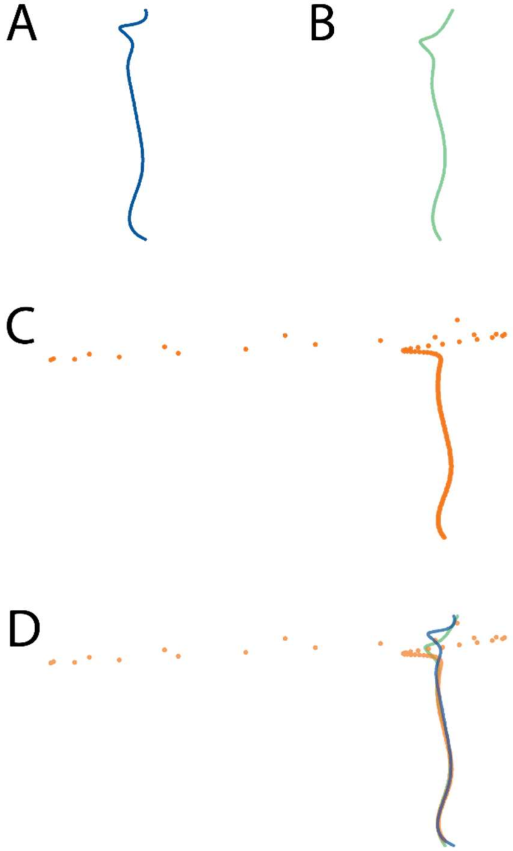


Figure 4.25 Comparison between three different sub-splines from the same DNA molecule. A & B: The first two sub-splines follow the general trajectory of the molecule. C: The third sub-spline contains many pixels that deviate substantially from the DNA trajectory; this leads to an erratic curve when the three sub-splines are subsequently combined and averaged. D: The three sub-splines are overlaid on each other for comparison.

4.2.2.2.2 Solutions

Once the source of erratic splining was identified, multiple solutions were proposed and tested to fix this issue or minimise its effect. The first solution considered was to perform splining directly after the ordering step for DNA molecules with proteins attached. Such DNA molecules could be automatically identified by calculating the similarity between sub-splines: if the sub-splines differed too much from each other, the splining would be performed for a second time using the ordered trace rather than the fitted trace. However, this approach cannot completely prevent the reoccurrence of erratic splining, because the ordered trace might also contain points that are far from the rest of the molecule.

The second solution considered was to increase the smoothness parameter of the 'scipy.interpolate.splprep' function. This parameter is a trade-off between the closeness and the smoothness of the polynomial fit, and a larger value introduces more smoothing at the potential expense of accuracy. To test the effectiveness of this solution and to determine the ideal value for linear molecules, a parameter sweep was carried out on both regular molecules that did not introduce tracing errors and irregular molecules that led to erratic splining. The contour length of test molecules was automatically measured while the periodicity was adjusted from 0 to 2, and the smoothness was increased from 0 to 8. Table 4.1 below lists the results for the example irregular molecule from Figures 4.23-4.27, and for an example regular molecule from Figure 4.28.

Table 4.1 Contour length measurements of example molecules when periodicity and smoothness values are adjusted

Periodicity	Smoothness	Contour length of an example irregular molecule / nm	Contour length of an example regular molecule / nm
0	0	114.2	99.4
	1	163.0	99.1
	2	255.0	98.9
	3	112.8	98.7
	4	112.1	98.6
	5	111.6	98.6
	6	111.1	98.6
	7	110.8	98.4
	8	110.8	98.3
1	0	222.3	150.4
	1	222.5	149.7
	2	221.3	148.5
	3	219.1	148.3
	4	214.9	147.6
	5	213.5	146.9
	6	212.0	146.6
	7	208.8	146.1
	8	206.7	145.9
2	0	222.3	150.4
	1	222.5	149.7
	2	221.3	148.5
	3	219.1	148.3
	4	214.9	147.6
	5	213.5	146.9
	6	212.0	146.6
	7	208.8	146.1
	8	206.7	145.9

The parameter sweep showed that as the smoothness value increased, the measured contour length decreased for both regular and irregular molecules; in particular, the contour length for the example irregular molecule stabilised and fell in line with expectations at a periodicity of 0 and a smoothness of 3 and higher.

Since the measured contour length was only an approximation of tracing quality, these results needed to be interpreted in conjunction with manual observation, which showed that when the smoothness was high, the occurrence of erratic splining was eliminated, but sharp bends were less likely to be traced accurately. A balance needs to be struck between preserving information and preventing erratic splining, and the appropriate smoothness value may differ depending on the nature of the dataset, as well as the number of data points present: if the sample contains a high amount of impurities, the smoothness needs to be higher so that other molecules do not affect the tracing of DNA; if the sample is clean, the smoothness should be lower to best preserve the trajectory of the DNA. For the data used in this project involving DNA and NDP52, the smoothness value is chosen to be 5, which was found by trial and error to eliminate most of the erratic splining without compromising the accuracy of the tracing, as demonstrated by the test molecule in Figure 4.26 below.

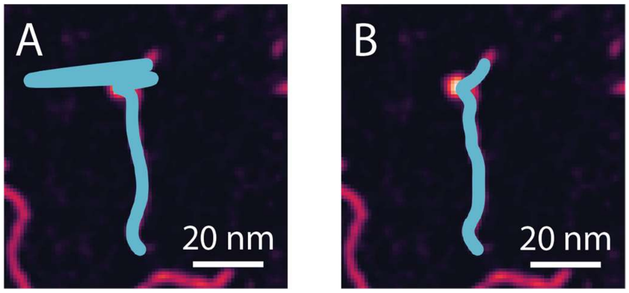


Figure 4.26 The result of the splining changes when the smoothness is increased from 2 (A) to 5 (B). Erratic splining is eliminated while the trajectory of the DNA is preserved.

However, this solution is still not ideal as it requires trial and error to identify the most suitable parameters for each dataset. In addition, since the splining is dependent on the 'scipy.interpolate.splprep' function, which is part of the regularly updated SciPy package, different smoothness values might be needed for different SciPy versions. In the future, a more robust long-term solution can be devised by taking the height of each pixel into account within the skeletonisation process; this will remove the need for the fitting stage and significantly reduce the likelihood of pixels being produced outside the DNA backbone during the automatic tracing workflow.

A third potential solution was to calculate the median position of pixels on the three sub-splines instead of the mean position. As shown in Figure 4.27 below, when tested on the example molecule, the median spline successfully prevented erratic curves from occurring but led to discontinuities. Since a smooth, continuous curve was essential for downstream analysis and measurements, such as the calculation of contour length, this solution was not adopted.

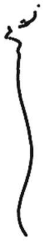


Figure 4.27 The median spline is no longer erratic but contains discontinuities. It cannot be used in subsequent measurements of contour length or curvature.

4.2.3 Results

The underlying issue that prevented the successful splining of linear molecules has been identified and addressed. Instead of using the unsuitable 'scipy.interpolate.interp1d' function, the more appropriate 'scipy.interpolate.splprep' function was chosen to achieve the interpolation of discrete pixels. By combining qualitative observation with quantitative contour length measurements from parameter sweep tests, the ideal values of periodicity and smoothness were identified to be 0 and 5 for linear molecules, as opposed to 2 for both in the case of circular molecules. As a result, all molecules with simple shapes can now be transformed into a smooth, continuous curve. This means that features of interest can be located in relation to the end points of a DNA molecule, and that more advanced measurements such as curvature can be developed and applied to real data.

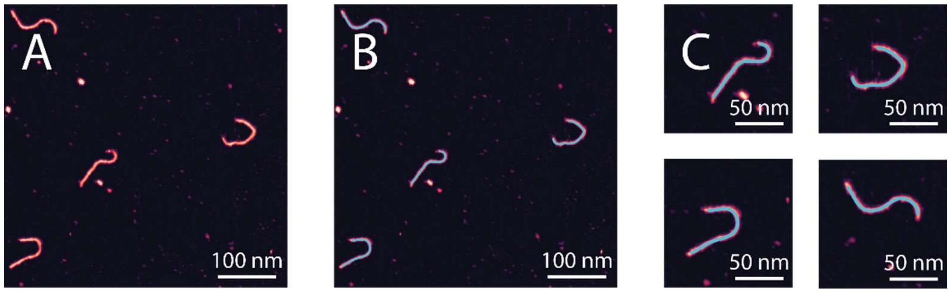


Figure 4.28 Successful splining of linear molecules. A: An AFM image of linear DNA molecules. B: After the improvement of the splining algorithm, these linear molecules can now be represented by a smooth, continuous, single-pixel-wide curve, known as a spline. C: Closed-up views of the splines.

4.3 Characterising the effects of NDP52 on DNA: overview

4.3.1 Introduction to the characterisation of NDP52-DNA interaction

The improvement of the splining algorithm means that we are now able to characterise linear DNA molecules with new tools based on tracing. One application of these tools is to probe possible conformational changes of DNA induced by NDP52 interaction.

As discussed in Section 1.4.3.2 and Section 3.4.1.1, NDP52 was first discovered in the cell nucleus, but previous work associated it with roles in the cytoplasm. Using other techniques such as stochastic optical reconstruction microscopy (STORM), our collaborators demonstrated that NDP52 has a potential nuclear role related to transcription initiation [151].

While the imaged structure of NDP52 has been characterised in Chapter 3, in order to obtain further information on how NDP52 might operate in the nucleus, it is necessary to also perform AFM imaging and automatic analysis on DNA. With the aid of tracing-based tools, we can explore different methods and determine the best way to characterise and compare the structure of DNA with and without NDP52 interaction, thereby identifying the role of NDP52.

4.3.2 Untraced AFM images of DNA with and without NDP52

To investigate the effects of NDP52 on DNA structure, the experimentalists Daniel E. Rollins andÁlia dos Santos performed AFM imaging of DNA with and without NDP52 co-incubation, using three different immobilising agents (see Section 2.1.2): Ni^{2+} only, Mg^{2+}-Ni^{2+} exchange, and poly-L-ornithine (PLO). Example images are shown in Figure 4.29 below:

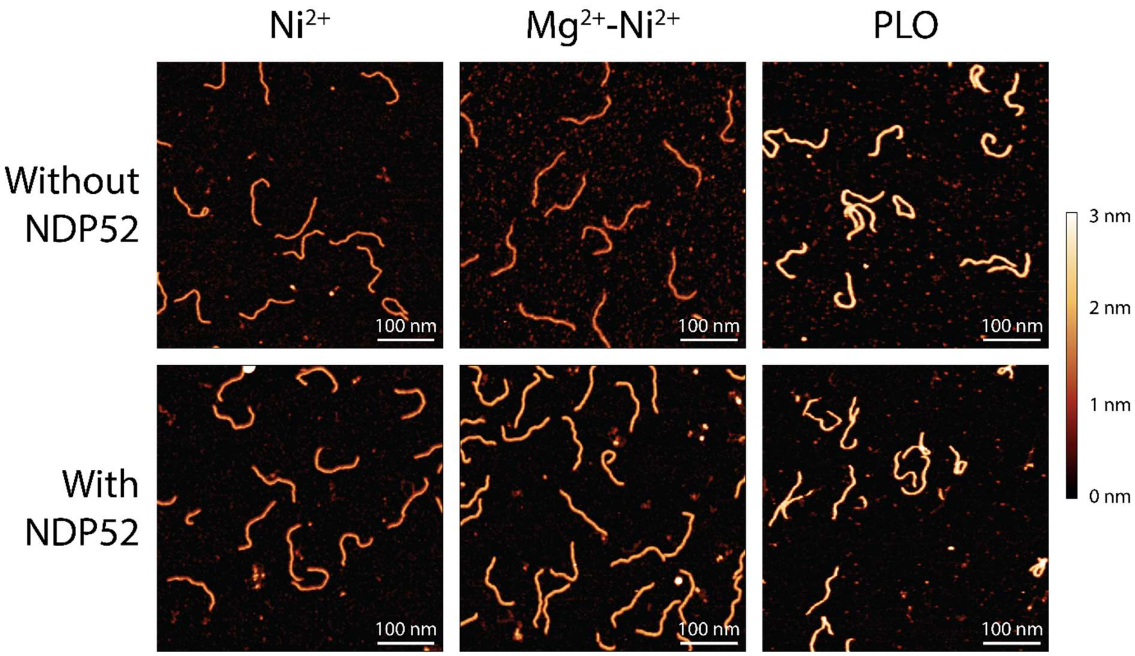


Figure 4.29 Comparison of DNA with and without NDP52 interaction, imaged using three different immobilising techniques: Ni^{2+} only, Mg^{2+}-Ni^{2+} exchange, and poly-L-ornithine (PLO). The height scale is used for Figures 4.8, 4.29, 4.30, 4.32, 4.39 and 4.42.

Before delving into tracing and quantitative analysis, we first qualitatively observed the images of DNA, since this could allow us to identify the nature of potential conformational changes and select the most suitable methods of automatic characterisation. It was noticed that when metal ions were used as the immobilising agent, the effects of NDP52 on DNA were not apparent; in contrast, when PLO was used, DNA with NDP52 co-incubation exhibited three types of characteristic structures, as shown in Figure 4.30 below: firstly, some DNA molecules contained a localised sharp bend; secondly, two parts of the same DNA molecule could be joined together, forming a loop; thirdly, two or more different DNA molecules could be joined together, forming a bridge. The latter two appeared to be formed under the same mechanism, which means that the looping can be understood as intramolecular bridging.

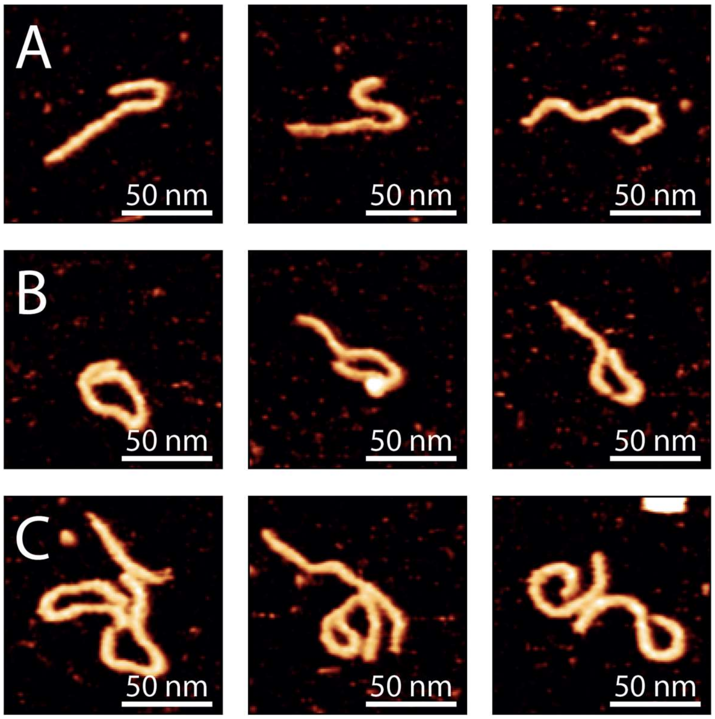


Figure 4.30 Examples of bending (A), looping (B) and bridging (C) observed on NDP52-treated DNA molecules when immobilised with PLO.

To quantitatively confirm these observations, the proportions of DNA involved in bending, looping and bridging were manually counted on five randomly selected AFM images from each experimental condition, as shown below in Figure 4.31. Molecules were excluded from the count if they were obstructed by large impurities or touching the image border, because the incompleteness rendered it impossible to determine their exact conformation.

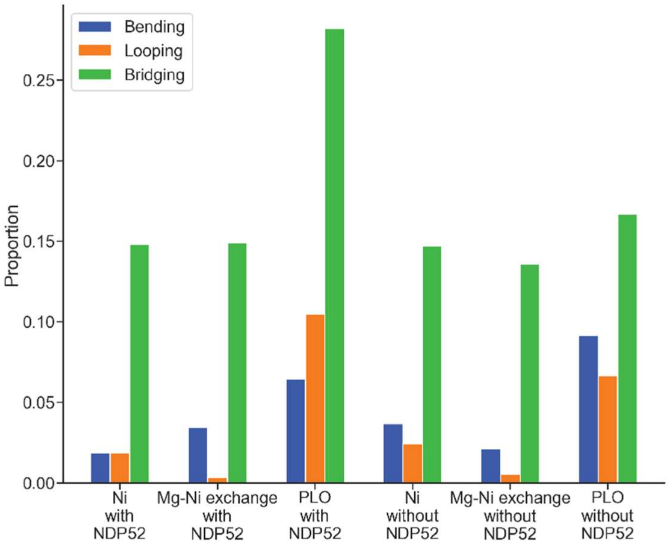


Figure 4.31 Bar chart comparing the proportions of DNA molecules involved in bending, looping and bridging under different experimental conditions. The number of DNA molecules is manually counted for five randomly selected example AFM images from each of the six categories: Ni^{2+} with NDP52

(proportions of molecules involved in bending, looping and bridging are 0.019, 0.019, and 0.148 respectively; N = 108), Mg^{2+}-Ni^{2+} exchange with NDP52 (proportions are 0.035, 0.003, and 0.149; N = 289), PLO with NDP52 (proportions are 0.065, 0.105, and 0.282; N = 124), Ni^{2+} without NDP52 (proportions are 0.037, 0.025, and 0.147; N = 163), Mg^{2+}-Ni^{2+} exchange without NDP52 (proportions are 0.021, 0.005, and 0.136; N = 567), and PLO without NDP52 (proportions are 0.092, 0.067, and 0.167; N = 120).

The results from manual counting were consistent with qualitative observation, and showed that NDP52 treatment led to a higher proportion of DNA molecules with conformational changes only when PLO was used as the immobilising agent. This suggests that compared to Ni^{2+} and Mg^{2+}-Ni^{2+} exchange, PLO is better at retaining the effects of NDP52-DNA interaction. Furthermore, this shows that conformational changes in DNA could be caused by both the immobilising agent and NDP52; in order to establish the role of the latter, the effects of the former also need to be characterised. This will be further discussed in Sections 4.3.3.1 and 4.4.3.

However, the reliability of manual counting is limited due to its subjectivity. For example, it is not always clear whether a molecule can be classified as exhibiting a localised sharp bend, since the threshold to qualify a bend as sharp is chosen arbitrarily, and different people may classify the same shape differently. In addition, the proportion of molecules involved in bridging is likely overestimated, because molecules physically overlapping on each other without being chemically joined can be misclassified as bridging. These issues highlight the need for more objective methods to characterise DNA conformation through automatic molecular identification and measurements, where the same criteria are consistently applied across all images.

It is also worth noting that, while the bending, looping and bridging demonstrated that interaction between DNA and NDP52 had taken place, the protein was rarely found alongside those structures. This indicates that the formation of the NDP52-DNA complex was transient: after the interaction, NDP52 was detached from DNA and subsequently washed away during the imaging process. It is therefore essential to devise a method that quantitatively characterises the observable differences between the structure of DNA with and without NDP52 interaction, since the NDP52-DNA complex cannot be characterised directly.

4.3.3 Challenges of characterising the effects of NDP52 on DNA

4.3.3.1 Addressing the effects of immobilising agents

Immobilising agents are indispensable for the AFM imaging of DNA. This is because the imaging of biomolecules in solution is performed on mica, which has an atomically flat surface but also repels DNA as both are negatively charged under neutral pH conditions [35]. To address this, positively charged divalent metal ions or polymers are added to the solution, so that DNA molecules can be attached to the mica surface.

However, as described in Section 4.3.2, it could be noticed through qualitative observation that the conformation of DNA differed depending on the immobilising agent used. As shown in Figure 4.29, even when there was no NDP52 co-incubation, DNA molecules imaged with PLO appeared to be more compact than those imaged with divalent metal ions. This could be explained by a combination of several factors: firstly, the layer of polymers on mica can interact with DNA and prevent it from relaxing; secondly, compared to metal ions, PLO is better at preserving the conformational changes made by NDP52; thirdly, when polymers are used as an immobilising agent, the imaged structure is similar to a 2D projection of a 3D structure [135], while if metal ions are used, DNA molecules equilibrate into the lowest energy conformation in 2D [172].

This difference between metal ions and PLO suggests that data obtained using different immobilising agents cannot be directly compared. Prior to characterising the effects of NDP52 on DNA conformation, it is necessary to first determine the most suitable immobilising agent: on the one hand, since Ni^{2+} and Mg^{2+} are less able to retain the conformational changes of DNA caused by NDP52, PLO appears to be more appropriate for this project; on the other hand, PLO can directly interfere with the observed conformation of DNA, introducing a confounding factor.

To resolve this dilemma, once a suitable method for the characterisation of DNA conformation is established, it should be first used to quantify the effects of different immobilising agents on DNA without NDP52 interaction. This will allow us to choose the most suitable immobilising agent and to establish a baseline conformation, so that immobilisation-related alternation to DNA structure can be distinguished from possible changes induced by NDP52.

4.3.3.2 Addressing the low signal-to-noise ratio

During the co-incubation of NDP52 and DNA, not all DNA molecules interacted with the protein. Moreover, it is not straightforward to determine from the AFM image whether a DNA molecule has interacted with NDP52, because the interaction would be transient and the protein would not remain attached to the DNA. The situation is further complicated by the varying image quality of the data provided, with some polluted by impurities and aggregates. The combination of the factors listed above means that any potential effects of NDP52 can only be observed on a small subset of the DNA molecules that have gone through NDP52 co-incubation. As a result, changes to DNA conformation caused by NDP52 are difficult to quantitatively characterise, since the automatic analysis of AFM images takes all identified DNA molecules into account.

4.3.3.3 Distinguishing between single molecules and joined molecules

Among the three observed structures related to NDP52-DNA interaction, bending and looping affect single DNA molecules, while bridging leads to joined molecules. This difference means that the quantitative characterisation of bridging needs to be separate from that of bending and looping, because single molecules and joined molecules should not be grouped together when measuring DNA and characterising the distribution of parameters such as the end-to-end distance.

It is therefore necessary to distinguish between single and joined molecules at the masking stage, which can be achieved by adjusting the area threshold. When investigating bending and looping, only molecules smaller than the threshold value should be masked, since this ensures the exclusive selection of single molecules; in comparison, for the characterisation of bridging, molecules larger than this value also need to be masked so that joined molecules can be recognised and measured.

4.3.3.4 Choosing the right parameter to quantitatively represent DNA conformation

As explained in Sections 4.3.3.1 and 4.3.3.2, the effects of NDP52 interaction are only present on a subset of DNA molecules, and these effects can also be confounded by interference from immobilising agents. To address these challenges, when characterising the conformation of DNA, it is crucial to select suitable parameters that can be reliably measured and are capable of quantitatively reflecting the changes brought onto DNA by NDP52: bending, looping and bridging.

Several potential parameters can be considered: the identification of DNA from the background allows us to measure the bounding distance, as was done for NDP52; the improved splining for linear molecules means that more advanced methods of automatic characterisation are also available, including the measurement of contour length and end-to-end distance. These parameters are all representative of DNA conformation in some way, and they need to be examined one by one, so that the viability of each of them can be determined.

4.4 Characterising the effects of NDP52 on DNA bending and looping

4.4.1 Assessing the viability of characterising DNA through end-to-end distance

4.4.1.1 Introduction to end-to-end distance measurement

Since all DNA molecules used in this project are of the same length, the end-to-end distance of each molecule is reflective of its compactness and can be used to quantify the extent of its bending and looping. With the implementation of splining for linear molecules, this parameter can now be measured by automatically calculating the distance between the two ends of the splined trace. To verify whether the results produced are representative of the shape of the molecule, it is necessary to first assess their accuracy before utilising the end-to-end distance for the characterisation of DNA conformation. This can be done by comparing the values from automatic measurement with those measured manually.

4.4.1.2 Assessing the accuracy of end-to-end distance

4.4.1.2.1 Comparing manual and automatic measurements

As it is impractical to conduct manual measurement over the entire dataset, the comparison between manual and automatic measurements was achieved on a few example images. Rollins selected three images with different immobilising agents and manually measured the end-to-end distance of all single DNA molecules found within them that did not touch the image border.

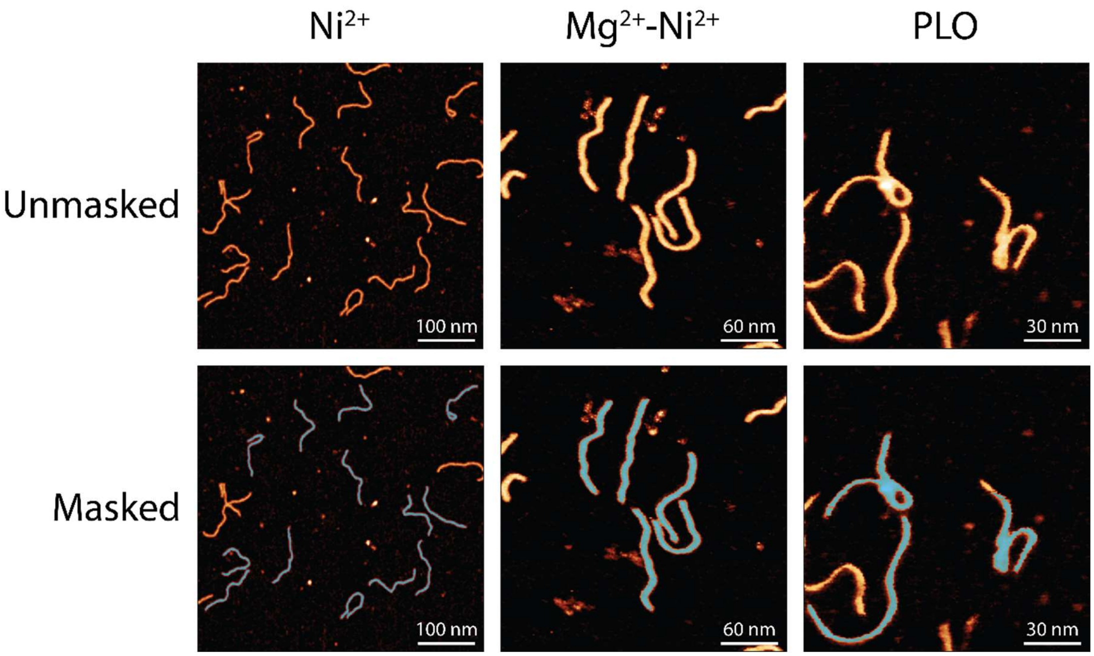


Figure 4.32 The example images chosen for comparison between manual and automatic measurements. The unmasked images are used for manual measurement, while the masked images are the basis of automatic measurement.

21 molecules in total (13 from the Ni^{2+} image, 5 from the Mg^{2+}-Ni^{2+} image, and 3 from the PLO image) were measured in this way, and the results were subsequently compared with values obtained from the automatic analysis algorithm. However, as shown in Figure 4.33 below, the two sets of measurements did not agree with each other. For most molecules, the automatic algorithm produced a lower end-to-end distance than the manual measurement; on average, the automatic measurement is lower than the manually obtained result by 7.6 nm or 12.9%.

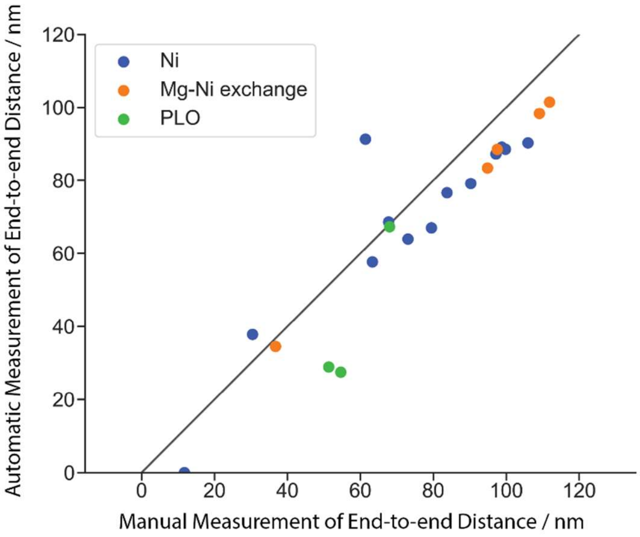


Figure 4.33 Comparison between manual and automatic measurements of end-to-end distance for DNA molecules imaged with three different immobilising agents (N = 21, number of images = 3). For

each molecule, the automatically calculated end-to-end distance is plotted against the manually obtained value, and the corresponding data point falls on the central X = Y line when the two measurements agree. However, it can be observed that most points are located below the line, closer to the X axis; this indicates that the values from automatic measurement are lower than those from manual measurement by an average of 7.6 nm or 12.9%.

4.4.1.2.2 Identifying the cause of discrepancies

Automatic measurement of end-to-end distance resulted in systematically lower values than manual measurement. This could be caused by inaccuracies introduced during the automatic tracing process. To verify whether such is the case, we observed the splined traces generated from the example images used in the comparison.

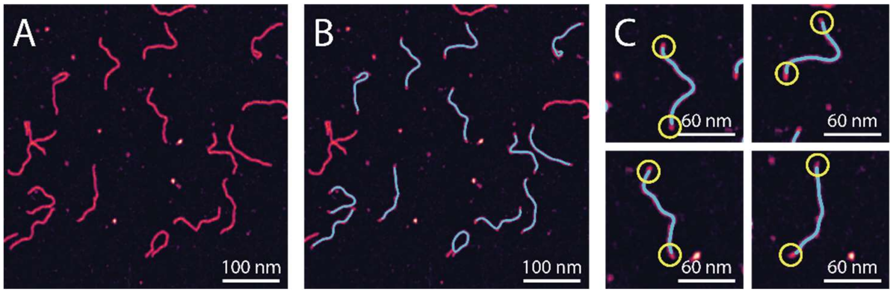


Figure 4.34 The splined traces generated from an example image used for the comparison of end-to-end distance measurements. A: The untraced image of DNA molecules with Ni^{2+} as the immobilising agent. B: Splined traces are shown in blue. C: Close-up views of the splined traces; as highlighted in yellow circles, it can be observed that the traces do not completely cover the ends of the molecules; because of this, the values from the automatic end-to-end distance measurement are lower than expected.

As shown in Figure 4.34 above, the splined traces were slightly shorter than the trajectory of the molecule, with both ends being cut off. The source of this error has been subsequently identified and detailed in Section 4.1.2.2.2: during the skeletonisation process, the ends of a molecule are considered to be part of the periphery and therefore removed.

While the skeletonisation causes molecules to be shortened, this issue *per se* does not render the automatic tracing completely unviable, because all molecules are equally affected, and valid comparison between different molecules can still be made for certain tracing-based parameters such as contour length. However, in the specific case of the end-to-end distance, although the lengths of the removed ends are similar across all molecules, the issue affects the measured value to different extents depending on the shape of the molecule, as illustrated by Figure 4.35 below. As a result, the accuracy of the end-to-end distance measurement varies from molecule to molecule, and this parameter is not a reliable representation of DNA conformation at this stage.

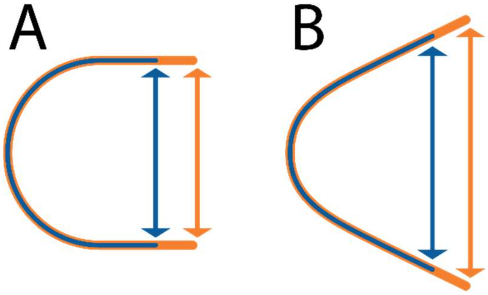


Figure 4.35 The shortening of molecules during skeletonisation affects the end-to-end distance measurement differently, depending on the shape of the molecule. A: For some molecules, the end-to-end distance of the actual molecule (orange) is equal to the measurement from the shortened trace (blue). B: For other molecules, the shortened trace can lead to a shortened end-to-end distance. It is therefore unviable to characterise DNA conformation through the automatic measurement of end-to-end distance.

4.4.1.3 Conclusions of assessment

Due to the shortening of molecules during skeletonisation, the end-to-end distance cannot be accurately measured for all molecules, and it can lead to inconsistencies for molecules of different shapes. This parameter in its current form is therefore not suitable for characterising DNA bending and looping induced by NDP52. In the future, the skeletonisation can be improved to take the height at each pixel into account, so that the ends of the molecules can be distinguished from the sides. This will prevent the ends from being cut off and improve the accuracy of the automatic end-to-end distance measurement, allowing it to become a viable parameter. For this project, other parameters to characterise DNA bending and looping will be investigated.

4.4.2 Assessing the viability of characterising DNA through bounding distance

4.4.2.1 Introduction to bounding distance measurement

To quantitatively describe the change in compactness as a result of NDP52 interaction, a different parameter needs to be selected for the automatic characterisation of DNA conformation. The maximum and minimum bounding distances are potential candidates. As explained in Section 3.5.2.1.1, these parameters are especially suited for irregular shapes, and they refer to the maximum and minimum distances between two parallel lines that touch the boundary of a molecule.

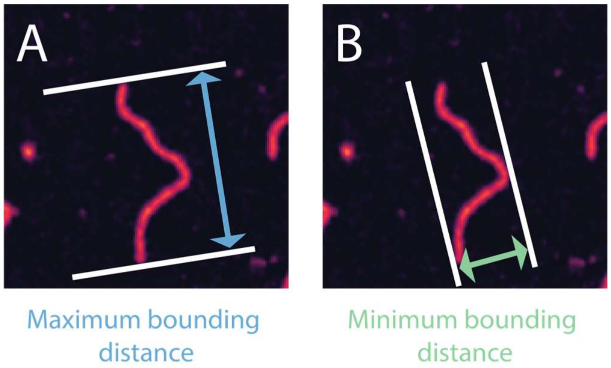


Figure 4.36 Illustrations of the maximum (A) and minimum (B) bounding distances on an example DNA molecule as imaged by AFM. These parameters refer to the longest and shortest possible distances between two parallel lines that touch the edge of a molecule. For a similar illustration made on an NDP52 molecule, see Figure 3.36.

The automatic measurement of bounding distances does not require tracing, and is therefore unaffected by problems with skeletonisation. The maximum bounding distance in particular is considered for the characterisation of DNA bending and looping, because it reflects the compactness of a molecule better than the minimum bounding distance, as shown in Figure 4.37 below.

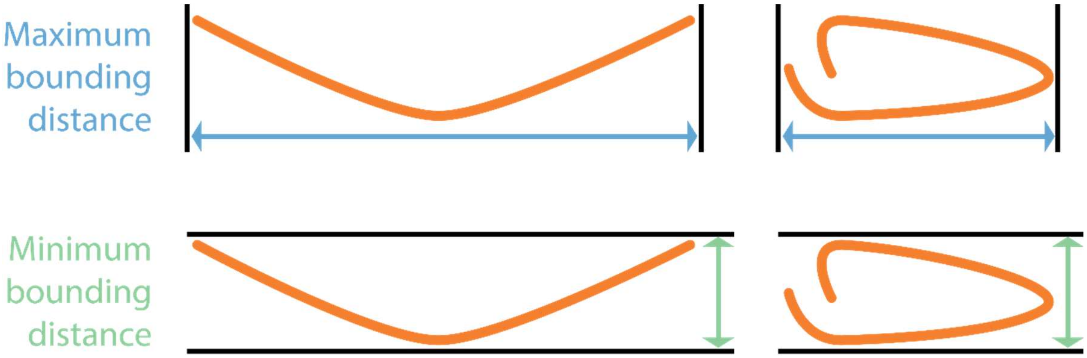


Figure 4.37 The maximum bounding distance is a more suitable parameter than the minimum bounding distance for the characterisation of DNA bending and looping. This is because the former is a better measure of compactness: a relaxed molecule (left) and a compact molecule (right) are always different in maximum bounding distance (top), but can be similar in minimum bounding distance (bottom).

Similar to the case of the end-to-end distance, before using the maximum bounding distance to characterise all DNA molecules in the dataset, the reliability of the automatic measurement needs to be first assessed, and this can also be achieved through comparison with manual measurement on example images.

4.4.2.2 Comparing manual and automatic measurements

Rollins performed manual measurement of maximum bounding distance on DNA molecules from the same three images as used in Section 4.4.1.2, and the results were compared with those from automatic measurement, as shown in Figure 4.38 below.

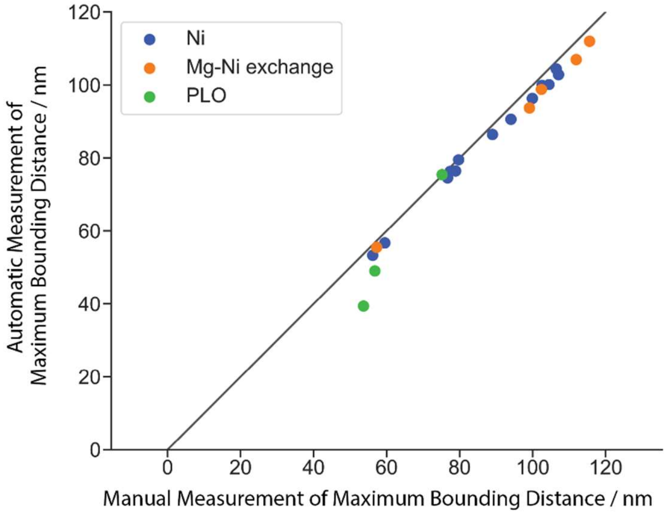


Figure 4.38 Comparison between manual and automatic measurements of maximum bounding distance for DNA molecules imaged with three different immobilising agents (N = 21, number of images = 3). For each molecule, the automatically calculated maximum bounding distance is plotted against the manually obtained value. Most data points fall on or next to the central X = Y line, which indicates that the two measurements are in good agreement, with the automatic measurement being slightly lower by an average of 3.6 nm or 4.7%.

The two maximum bounding distance measurements showed good agreement, with the automatically calculated value being slightly lower by 3.6 nm or 4.7% on average. This suggests that unlike the case of end-to-end distance where the automatic measurement was lower by 12.9%, the automatic measurement for the maximum bounding distance can be considered accurate.

4.4.2.3 Conclusions of assessment

The assessment has shown that the automatic measurement of maximum bounding distance is accurate and reliable. Compared to the end-to-end distance, it has the advantage of being able to quantify the compactness of DNA without relying on tracing, which means that it is not affected by the shortening of molecules during the skeletonisation process. The maximum bounding distance is therefore selected as the most suitable parameter for the characterisation of conformational changes on DNA related to bending and looping.

4.4.3 Investigating the effects of immobilising agents through bounding distance

4.4.3.1 Lowering the concentration of DNA immobilised with PLO

As discussed in Section 4.3.3.1, once a parameter has been established as a valid representation of DNA conformation, it needs to be first used to characterise the effects of immobilising agents. Accordingly, automatic DNA identification and maximum bounding distance measurement were performed on images taken with Ni^{2+}, Mg^{2+}-Ni^{2+} exchange, and PLO, so that comparisons could be made between the three datasets. However, a challenge was encountered during the automatic identification process: since two or more DNA molecules can appear to be in contact with each other on AFM images, either due to the formation of chemical bonds or through physical overlapping, the resulting clusters were sometimes incorrectly recognised as molecules of interest. This needs to be addressed because the comparison of maximum bounding distances is only valid when the values are all measured on single DNA molecules of the same length.

The removal of DNA clusters from the list of identified molecules can be achieved by adjusting the upper area threshold, which is calculated in reference to the median area of all potential molecules of interest on a given image, as described in Section 3.1.1.3.1. When Ni^{2+} or Mg^{2+}-Ni^{2+} exchange is used for immobilisation, DNA clusters are infrequent, and the median area is equivalent to the area of a single DNA molecule; clusters can therefore be removed through discarding molecules whose area deviate too far from the median value. By contrast, when PLO is involved, this process is less straightforward, because the proportion of joined or overlapping DNA molecules is higher; consequently, the median area can be larger than the size of a single molecule for some images, which leads to difficulties in the determination of a suitable area threshold that consistently excludes DNA clusters across all images.

To resolve this issue and enable the automatic selection of single molecules for data immobilised with PLO, the cause of the frequent clustering was analysed: immobilising agents are not known to create chemical bonds between DNA molecules; however, since PLO-based immobilisation can be described as projecting a 3D structure onto a 2D surface [135], the clusters are likely formed by unconnected DNA molecules superimposing on each other during adsorption onto the mica. If this is the case, the overlapping can be prevented by reducing the amount of DNA in the solution.

To verify this hypothesis experimentally, Rollins lowered the concentration of DNA and performed additional AFM imaging with PLO as the immobilising agent. The results were compared with previous PLO-based images. It is observed that when DNA molecules are less concentrated, they no longer form clusters unless treated with NDP52, as shown in Figure 4.39 below:

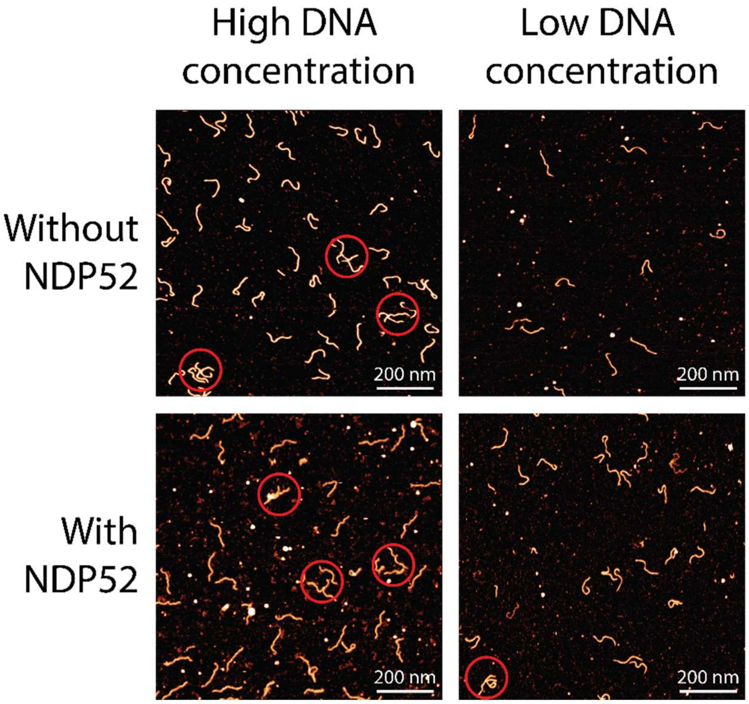


Figure 4.39 Comparison between DNA with and without NDP52 treatment at two different concentrations. DNA clusters (circled in red) occur frequently at high concentration, both with and without NDP52 treatment; by comparison, when the concentration is lowered, DNA molecules only cluster when they have undergone NDP52 interaction.

This is congruent with our understanding that PLO does not chemically combine DNA molecules, but it can result in overlapping during immobilisation when the concentration of DNA is too high. Since the newly obtained low-concentration data is more suited for the automatic identification of single molecules and the removal of clusters, the high-concentration PLO dataset is excluded from subsequent analysis, so that the accuracy of further measurements and characterisation can be ensured.

4.4.3.2 Comparing the effects of different immobilising agents on untreated DNA

The relation between DNA conformation and the immobilising agent used can now be characterised through the automatic identification and measurement of single molecules. The comparison of maximum bounding distance distributions was first performed on DNA molecules that were not co-incubated with NDP52, so as to avoid confounding factors and isolate the effects of Ni^{2+}, Mg^{2+}-Ni^{2+} exchange, or PLO.

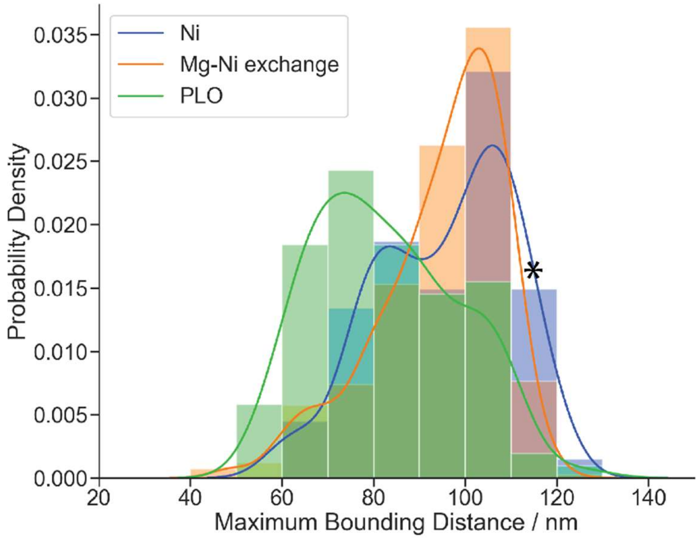


Figure 4.40 KDE plots and normalised histograms comparing the maximum bounding distance distributions of untreated DNA imaged with Ni^{2+} (blue, main peak and standard deviation 106 ± 15 nm, second peak at 83 nm, N = 134, number of images = 7), Mg^{2+}-Ni^{2+} exchange (orange, 103 ± 14 nm, N = 419, number of images = 5) and PLO (green, 74 ± 16 nm, second peak at 104 nm, N = 103, number of images = 19) as immobilising agents. For reference, all DNA molecules used for the imaging have 339 base pairs, which is equivalent to 115 nm in contour length, as marked by the black asterisk.

The maximum bounding distances for DNA immobilised with Ni^{2+} and Mg^{2+}-Ni^{2+} exchange are measured to be 106 ± 15 nm and 103 ± 14 nm respectively. They are both slightly smaller than 115 nm, the contour length of the DNA used in the experiment. By contrast, the maximum bounding distance for the PLO group is measured to be 74 ± 16 nm, indicating that the DNA molecules are more compact and less relaxed. In addition, the distributions for the Ni^{2+} and PLO groups exhibit notable secondary peaks; they are likely caused by smaller sample sizes, which amplify the effect of impurities and uncommon structures.

In accordance with the conclusions from qualitative observation, the results from automatic characterisation demonstrate that the choice of immobilising agent influences the imaged conformation of DNA: compared to those immobilised with metal ions, DNA molecules immobilised with PLO exhibit shorter maximum bounding distances and more compact shapes.

4.4.3.3 Comparing the effects of different immobilising agents on DNA treated with NDP52

It has been theorised that immobilising agents can impact the retention of conformational changes on DNA induced by NDP52 co-incubation, as discussed in Section 4.3.3.1. Therefore, to fully establish the effects of immobilising agents on the characterisation of NDP52-DNA interaction, we also compared the maximum bounding distance distributions of NDP52-treated DNA imaged with Ni^{2+}, Mg^{2+}-Ni^{2+} exchange, and PLO.

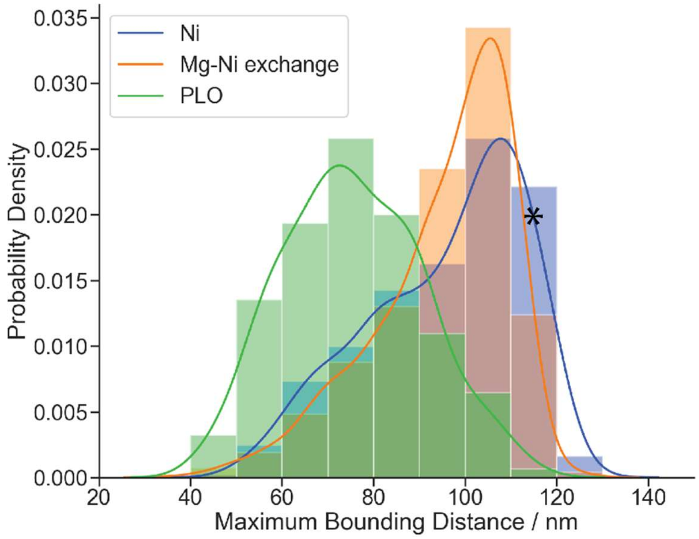


Figure 4.41 KDE plots and normalised histograms comparing the maximum bounding distance distributions of NDP52-treated DNA imaged with Ni^{2+} (blue, main peak and standard deviation 108 ± 17 nm, N = 492, number of images = 25), Mg^{2+}-Ni^{2+} exchange (orange, 105 ± 15 nm, N = 1203, number of images = 29) and PLO (green, 73 ± 15 nm, N = 155, number of images = 16) as immobilising agents. The black asterisk marks the contour length (115 nm) of the DNA molecules.

As shown in Figure 4.41 above, the maximum bounding distance of DNA molecules that have undergone NDP52 treatment is measured to be 108 ± 17 nm for Ni^{2+} based immobilisation, 105 ± 15 nm in the case of Mg^{2+}-Ni^{2+} exchange, and 73 ± 15 nm when PLO is used. By comparing these values with the results obtained in Section 4.4.3.2, it can be observed that for all three immobilising agents, the peak and standard deviation of the distribution do not differ much between DNA with and without NDP52 treatment. This is because among DNA molecules co-incubated with NDP52, only a subset have experienced actual interaction with the protein. Furthermore, this subset is under-represented since NDP52 can bridge neighbouring DNA molecules and result in clusters, while the measurement of maximum bounding distance only applies to single molecules. Due to these factors, it is challenging to characterise the effects of NDP52 through statistical approaches.

However, when the immobilising agent is PLO, a difference in shape can be noticed between the distributions for treated and untreated DNA: when NDP52 is involved, a higher proportion of DNA molecules are found on the lower end of the distribution. To obtain more information on the nature of NDP52-DNA interaction and its potential effects, further investigation of this shift is required, which will be discussed in Section 4.4.4. Since no comparable changes in maximum bounding distance distribution are observed for DNA immobilised with metal ions, it can be concluded that PLO is the most suitable immobilising agent to capture possible conformational changes of DNA caused by NDP52. The remaining analysis in this chapter is therefore all conducted on the low-concentration PLO-based dataset.

4.4.4 Characterising the effects of NDP52 on DNA through bounding distance

To directly visualise the shift in maximum bounding distance distribution between DNA with and without NDP52 co-incubation, the two datasets are plotted on the same figure.

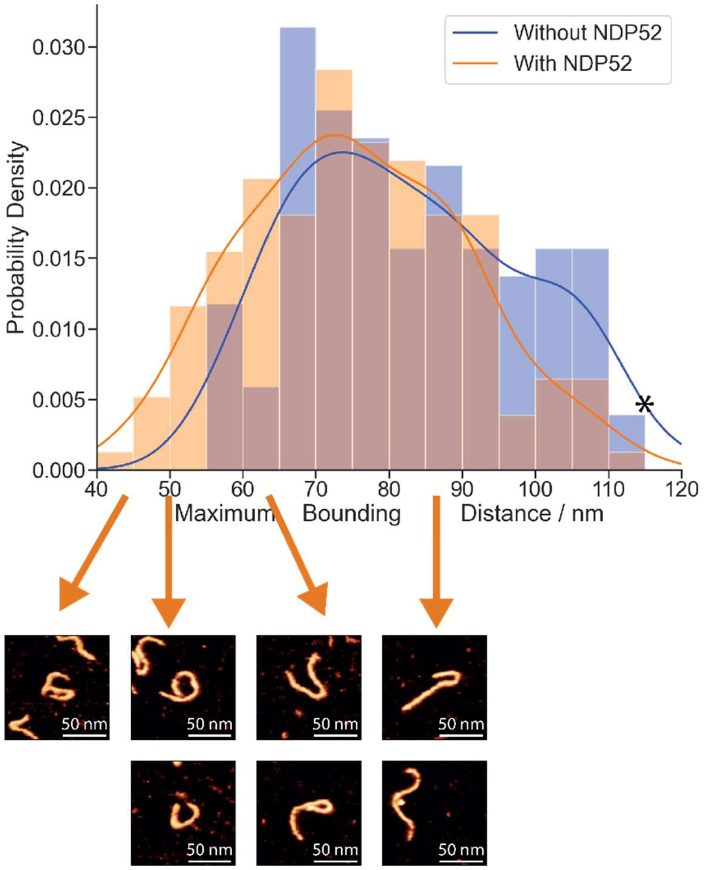


Figure 4.42 KDE plots and normalised histograms comparing the maximum bounding distance distributions of DNA with (orange, peak and standard deviation 73 ± 15 nm, N = 155, number of images = 16) and without (blue, 74 ± 16 nm, second peak at 104 nm, N = 103, number of images = 19) NDP52 treatment, both imaged with PLO as the immobilising agent. The DNA contour length (115 nm) is marked with a black asterisk. Examples of DNA molecules affected by NDP52 taken from different parts of the distribution are attached.

As the linearised 339 bp minicircles used in this experiment are approximately 115 nm in contour length, the comparison shows that under PLO immobilisation, DNA molecules in both groups exhibit compaction to some extent, but NDP52 treatment increases the proportion of molecules whose maximum bounding distance falls below 65 nm, and decreases the proportion of molecules with a maximum bounding distance greater than 95 nm. In addition, as demonstrated by the first three columns of the histogram, only DNA molecules that have been co-incubated with NDP52 can be found to have a maximum bounding distance lower than 55 nm.

To quantitatively determine the extent of correlation between a lower maximum bounding distance and NDP52 treatment, statistical analysis is performed to calculate the probability of NDP52 being present when the maximum bounding distance of a DNA molecule is lower than a given threshold, using the Bayesian theorem

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

where A refers to the event that a DNA molecule has undergone NDP52 treatment, and B refers to the event that the measured maximum bounding distance for a DNA molecule is lower than the given threshold. P(A) is calculated to be 155 / (155 + 103) = 0.601, while P(B) is the number of molecules (both with and without NDP52 treatment) below the maximum bounding distance threshold, divided by the total number of molecules, 258. P(B|A) refers to the probability of a molecule falling under the given maximum bounding distance threshold, provided that it has been treated with NDP52; it is the number of molecules under the threshold that have gone through NDP52 treatment, divided by the total number of molecules under NDP52 treatment, 155. P(A|B) refers to the probability of the molecule having been treated by NDP52, provided that it falls under the given maximum bounding distance threshold. The results of the calculation for different thresholds are shown below in Table 4.2:

Table 4.2 Probabilities of NDP52 being present for DNA molecules under different maximum bounding distance thresholds

Maximum bounding distance range	Number of molecules without NDP52 (N_{wo})	Number of molecules with NDP52 (N_w)	P(A) (= 155 / (155 + 103))	P(B) (= (N_{wo} + N_w) / (155 + 103))	P(B|A) (= N_w / 155)	P(A|B)
< 55 nm	0	14	0.601	0.054	0.090	1.000
< 60 nm	6	26		0.124	0.168	0.813
< 65 nm	9	42		0.198	0.271	0.824
< 70 nm	25	56		0.314	0.361	0.691

The Bayesian analysis shows that for a DNA molecule with a maximum bounding distance below 55 nm, 60 nm, 65 nm or 70 nm, the probabilities that it belongs to the NDP52-treated dataset are 1.000, 0.813, 0.824 and 0.691 respectively. This suggests that the presence of a maximum bounding distance lower than 70 nm is more likely caused by NDP52 treatment than by PLO or natural occurrence.

These results provide quantitative evidence that NDP52 is capable of altering the shape of DNA, and demonstrate that the bending and looping observed on NDP52-treated DNA cannot be solely attributed to the effects of PLO. By comparing example DNA molecules from different parts of the maximum bounding distance distribution, it can be further concluded that the effects of NDP52 interaction on DNA conformation vary, ranging from the formation of loops on some molecules to localised bending on others.

4.5 Characterising the effects of NDP52 on DNA bridging

4.5.1 Assessing the viability of characterising DNA through contour length

4.5.1.1 Introduction to contour length measurement

Having characterised bending and looping, we now attempt to investigate the third observable effect of NDP52 on DNA conformation known as bridging, where two or more DNA molecules are chemically joined together. This can potentially be quantified by measuring the contour length: since all DNA molecules used in the experiment are of the same size and sequence, a substantial increase in contour length indicates that multiple molecules have been combined.

While the automatic measurement of contour length was available prior to this project, it was only applicable to circular molecules. This is because the algorithm functions by calculating the distance between every pair of neighbouring points on a smooth, continuous trace and subsequently adding all the calculated distances together; the generation of the smooth, continuous trace requires splining, which was only possible for circular molecules. The implementation of splining for linear molecules during this project means that the algorithm can now be applied to a wider range of shapes.

It is worth noting that since the contour length measurement is based on tracing, its accuracy is limited by the shortening of molecules during the skeletonisation process, as described in Sections 4.4.1.2.2 and 4.1.2.2.2. However, unlike in the case of end-to-end distance, this issue affects all molecules equally, and the contour length can still be a viable parameter for the characterisation of DNA bridging if the measurement in relation to other molecules is used instead of the absolute value. To further explore this possibility, we need to examine the reliability and suitability of automatic contour length measurement as a tool for quantifying conformational changes on DNA.

4.5.1.2 Methods of assessment

4.5.1.2.1 Comparison with the expected value

The automatic measurement needs to be assessed for its accuracy, so as to ensure that it is representative of DNA conformation. While the contour length cannot be easily measured manually, the assessment can be achieved by comparing the results of automatic measurement with the expected value calculated from the number of base pairs.

It has previously been established that the molecules used in this experiment contain 339 base pairs and that their theoretical contour length is 115 nm. Therefore, if the tracing were completely accurate, the distribution curve would exhibit a peak at this value. As the ends of molecules have been erroneously cut off, a narrow peak slightly lower than 115 nm can serve as an indication that the contour length measurement is otherwise reliable.

4.5.1.2.2 Manual observation

To characterise the bridging of DNA induced by NDP52, the automatic measurement needs to be accurate for both single and joined DNA molecules, but the latter do not have a fixed reference contour length, since bridging can occur at any part of a molecule. Because of this, the reliability of the measurement for bridged molecules cannot be assessed through comparison with expected values.

Potential errors may be introduced into the contour length measurement during either the tracing or the calculation stage. While the bridging of molecules does not affect the reliability of the calculation process, it can render the tracing more challenging. Therefore, the key to assessing the accuracy of measurement on joined molecules is the evaluation of tracing quality. This can be accomplished by manually observing the splined traces and comparing them to the trajectory of corresponding molecules.

4.5.1.3 Assessment

4.5.1.3.1 Measuring the contour length of untreated DNA

The automatic contour length measurement was first applied to DNA without NDP52 treatment, so that the results can be compared with the theoretical value of 115 nm.

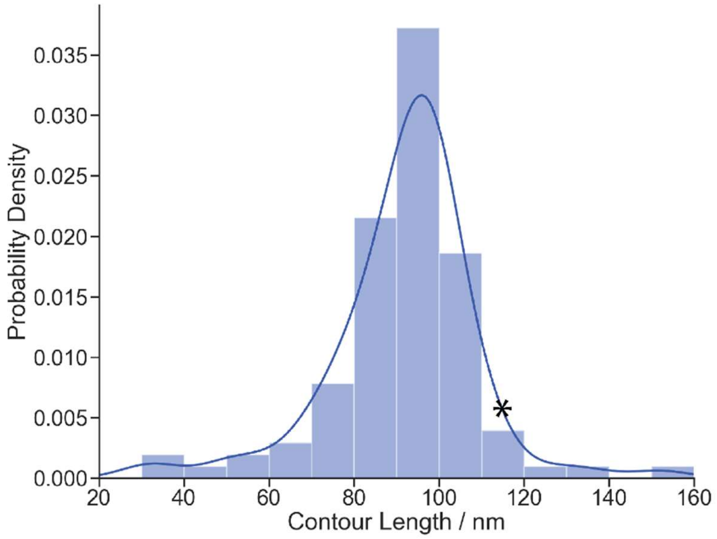


Figure 4.43 KDE plot and normalised histogram of automatically measured contour length (peak and standard deviation 96 ± 17 nm, N = 102, number of images = 19) for DNA untreated by NDP52 and imaged with PLO as the immobilising agent. The theoretical contour length of 115 nm is marked with a black asterisk for reference. Note that the sample size is slightly smaller compared to the distribution of maximum bounding distance from the same dataset (N = 103, number of images = 19, see the PLO part of Figure 4.40 in Section 4.4.3.2); this is because the tracing algorithm used in the NDP52 project is dependent on an external program [58] *that fails to recognise molecules too close to the image border, and the maximum bounding distance measurement is unaffected since it does not require tracing.*

As shown in Figure 4.43 above, the distribution of the measured contour length exhibits a peak at 96 ± 17 nm, which is 17% shorter than the reference value of 115 nm. To establish whether this is entirely caused by the known error with skeletonisation as discussed in Section 4.1.2.2.2, the traces produced were manually observed, and instances of inaccuracies could be identified, especially on molecules with more compact shapes. This suggests that in addition to the shortening of molecules from the ends during skeletonisation, the reliability of the contour length measurement may also be limited by inaccurate tracing for complex shapes. Another possible source of error is the low resolution, as a few missing pixels can lead to a shortening in measured contour length of several nanometres.

Molecules with a contour length longer than 115 nm were also identified and manually assessed. In one instance, because of damage to the sample prior to AFM imaging, the two strands of a DNA molecule had been separated and concatenated; in this case, the increased contour length measurement accurately represents the combination of the two single strands. Other molecules exhibit shapes that are challenging for the tracing algorithm: for example, as described in Section 4.1.2.3.2, when the two ends of a linear molecule are close to each other, the DNA can be misclassified as circular and its ends can be erroneously joined, resulting in an artificially high contour length measurement.

4.5.1.3.2 Measuring the contour length of DNA treated with NDP52

To further assess the suitability of the contour length as a parameter for the characterisation of joined molecules, we performed automatic measurement on DNA treated with NDP52. The area threshold was adjusted so that both single molecules and joined molecules can be included.

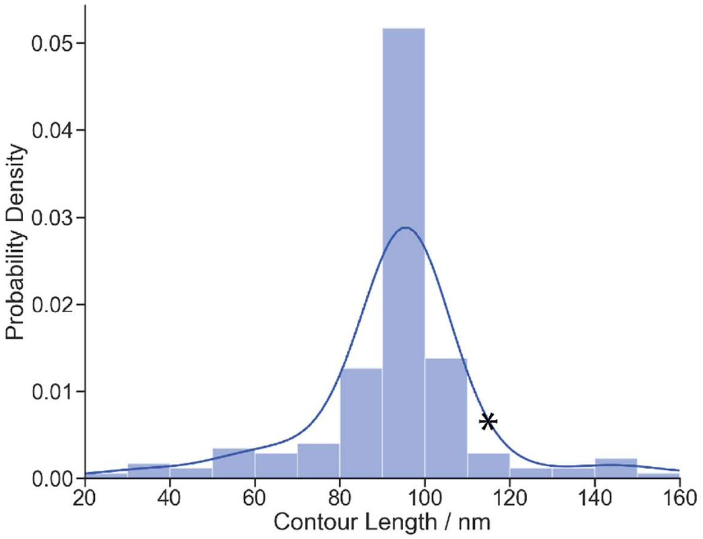


Figure 4.44 KDE plot and normalised histogram of automatically measured contour length (peak and standard deviation 96 ± 25 nm, N = 179, number of images = 18) for NDP52-treated DNA imaged with PLO as the immobilising agent. The theoretical contour length of 115 nm is marked with a black asterisk for reference. Note that the number of images is higher compared to the distribution of maximum bounding distance measured from the same dataset (N = 155, number of images = 16, see the PLO part of Figure 4.41 in Section 4.4.3.3); this is because two images only contain joined DNA molecules, and are therefore not counted when exclusively selecting for single DNA molecules to measure the maximum bounding distance.

Compared to the measurements in Section 4.5.1.3.1, the peak of the contour length distribution for NDP52-treated DNA falls on the same location of 96 nm, but is wider with a standard deviation of 25 nm. In addition, a slight increase in the proportion of molecules with a higher contour length than 115 nm can be observed, which corresponds to bridged DNA molecules, as confirmed through manual verification. These results are in line with previous findings that NDP52 is known to induce a range of different changes on DNA, leading to greater conformational variety, but as drastic alternations only occur on a small number of molecules, they are hard to be characterised through statistical methods and are not always captured by the peak value.

However, manual observation revealed that the trace is less likely to follow the trajectory of the molecule when crossings are introduced, as illustrated in Figure 4.45 below. This demonstrates that the measurement of contour length is not yet consistent enough across molecules of different shapes, and it is therefore not viable for the characterisation of DNA bridging.

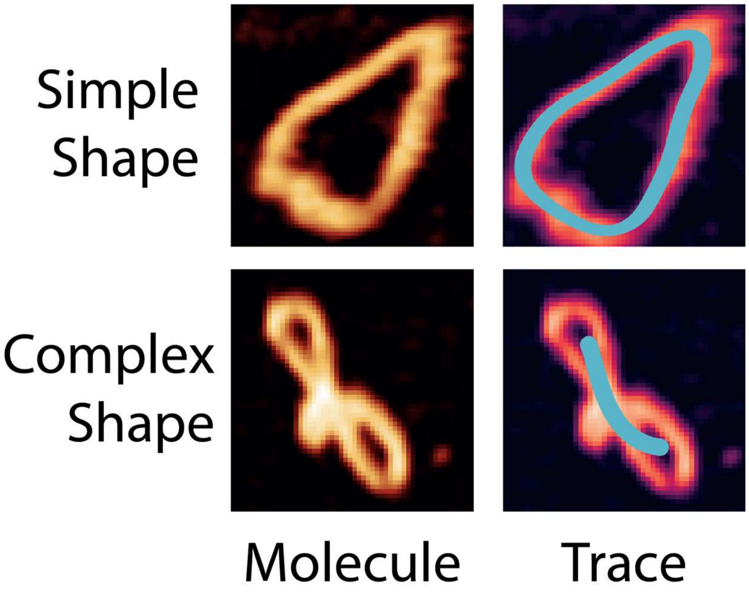


Figure 4.45 Comparison between tracing molecules with simple shapes and complex shapes involving crossings. Simple shapes can be traced accurately while complex shapes often result in failure of tracing.

4.5.1.4 Conclusions of assessment

The measurement of contour length suffers from two main drawbacks: firstly, the removal of the ends causes the measured value to be shorter than the real value, although this affects all molecules equally and is still compatible with the characterisation of changes in conformation; secondly, the reliability of the measurement is dependent on the accuracy of tracing, and the latter is insufficient when complex shapes and crossings are involved. The second factor severely limits the viability of contour length measurement in this project, as NDP52 is known to induce crossings by joining neighbouring molecules and different parts of the same molecule. It can thus be concluded that the contour length is not a suitable parameter for the quantification of DNA bridging until the tracing algorithm is improved and becomes compatible with more complex shapes. At the current stage, this method of characterisation can find its use in other projects where molecules do not exhibit crossings.

4.5.2 Measuring the proportion of joined molecules

While tracing-based measurements are not yet viable, the extent of bridging can be quantified by calculating the proportion of joined molecules. This requires the exclusion of single molecules during masking through the application of area-based filtering criteria.

As discussed in Section 4.4.3.1, relative area thresholding was used for the selection of single molecules when characterising bending and looping; since the data involved was obtained under different experimental conditions and with different immobilising agents, it would be impractical to establish absolute area thresholds that were compatible with every image. However, for the measurement of joined molecules, only the low-concentration dataset immobilised with PLO is used, where all DNA molecules have been imaged under similar conditions. This allows the adoption of absolute area thresholding, which enables more accurate and consistent selection of features of interest.

A threshold of 1350 nm^2 was established through trial and error, and molecules with a larger area than this value were considered to have undergone bridging. By applying this selection criterion to images of NDP52-treated DNA, we identified 7 joined molecules; in comparison, for DNA without NDP52 co-incubation, no molecule was found to be larger than 1350 nm^2. Combing these findings with the results from Section 4.4.4, where 155 single molecules could be identified from images of NDP52-treated DNA, we can calculate the extent of bridging: if each joined molecule is composed of 2 single molecules, there were 169 (7 × 2 + 155) molecules before NDP52 interaction, of which 14 or 8.3% were involved in bridging. As expected, this is lower than the estimated proportion (0.282 or 28.2%, see the 'PLO with NDP52' part of Figure 4.31 in Section 4.3.2) of molecules involved in bridging for the dataset with a higher DNA concentration, because the higher concentration both leads to more frequent bridging and causes more overlapping molecules to be misidentified as bridging.

As the sample size is relatively small, this measurement can also be carried out manually, but the methods of characterisation established through analysing this dataset can be applied to future projects where manual counting is no longer feasible.

4.6 Conclusions

This chapter has shown that, by representing linear DNA molecules as smooth, continuous curves known as splines, rather than as discrete pixels, the automatic tracing workflow has been improved from its previous state where it was only compatible with circular DNA molecules. Tracing-related methods to characterise conformational changes in DNA, including the automatic measurement of end-to-end distance and contour length, have been developed and tested on real data, although their applicability is currently limited. This is caused by a number of issues that affect the tracing accuracy, such as the removal of the ends of linear molecules during skeletonisation.

Further changes to the tracing process are required so that related automatic measurements can become more usable. For example, to address the aforementioned problem with the ends of linear molecules, the skeletonisation stage needs to be adjusted to take the height of pixels into account. This would also allow molecules with crossings or highly compact structures to be more reliably traced.

A major issue for this chapter was the difficulty of making accurate tracing-based measurements on the dataset available, due to the compaction of DNA in the presence of the protein NDP52 and the low image quality across the dataset. However, the effects of NDP52 on DNA conformation were visualised and successfully characterised through automatic AFM image analysis using a non-tracing-based method, which calculated the maximum bounding distance, or the largest measurement that can be taken across the DNA molecule, rather than along its traced backbone. Through our AFM data we showed that NDP52 is capable of interacting with DNA and that its effects vary from molecule to molecule. These effects were broadly categorised into three modes: bending, looping and bridging.

Our collaborators have discovered through STORM and coimmunoprecipitation assays that the spatial distribution of NDP52 in the nucleus underwent significant changes when transcription is inhibited, causing it to cluster less around transcription initiation sites and interact more with proteins related to DNA replication and damage response [151]. Combining this information with our results from AFM analysis that NDP52 is able to bind and compact DNA, we propose that in addition to involvement in the transcription process, NDP52 has a second nuclear function related to DNA shape regulation.

Furthermore, this project also highlights the important role of immobilising agents during the characterisation of conformational changes on DNA. Distinct differences in DNA conformation were observed when different immobilisation methods were used for the same DNA substrate. In future projects of a similar nature, prior to investigating the DNA-protein interaction, it will be crucial to first select a suitable immobilising agent and determine its possible effects on DNA conformation. In the next chapter, the different effects of immobilising agents on DNA conformation will be used to evaluate the performance of curvature measurement as a new tracing method that could potentially provide more information on the effects of NDP52.

Chapter 5 Measuring curvature and bending angle to study the effects of NDP52 and HU on DNA structure

5.1 Developing curvature measurement methods to characterise changes in DNA conformation

5.1.1 Introduction to curvature measurement

While the extent of DNA bending and compaction can be quantified through the automatic measurement of end-to-end distance and maximum bounding distance, these parameters only provide information on the overall shape of DNA. A different method of characterisation is required to measure the bending directly and to obtain more details such as the location of bending sites. In this project, we attempt to achieve this through the automatic measurement of curvature along the trajectory of the molecule: the value calculated at each point can be used to quantify the local bending, while the mean curvature of the entire molecule represents its general extent of compactness.

Due to limitations with tracing identified earlier in the project, including the shortening of molecules and the inaccuracies with shapes containing crossings, the curvature measurement may not yet be viable for the characterisation of DNA conformational changes induced by NDP52, as this involves the frequent presence of molecules with complex shapes. However, the automatic curvature measurement developed based on imperfect traces can still serve as a proof of concept. Furthermore, this new method of characterisation can be employed in other projects where molecules exhibit simple shapes; for example, if a protein is known to bend DNA without inducing crossings, the point of maximum curvature can be identified to locate and characterise the site of the DNA-protein interaction. In the future, when the tracing-related problems have been addressed, the automatic curvature measurement could be applied to more situations, such as the characterisation of DNA conformational changes related to supercoiling.

5.1.2 Calculating curvature by approximating an infinitesimal arc

5.1.2.1 Workflow of calculating curvature

The first attempt at automatically calculating the curvature was made by applying the formula

$$Curvature = \frac{\theta}{L}$$

where θ refers to the change in tangential angle over an infinitesimal arc, while L represents the length of said arc. In practice, these values can be approximated using the neighbours of the point of interest, as illustrated in Figure 5.1 below: for a given point of interest A, its left and right neighbours from the splined trace are referred to as B and C; the change in tangential angle (θ) is approximated by the angle between BA and AC, which is computed based on the coordinates of the three points; the length of the arc (L) is approximated by the distance between M1 (midpoint of BA) and M2 (midpoint of AC).

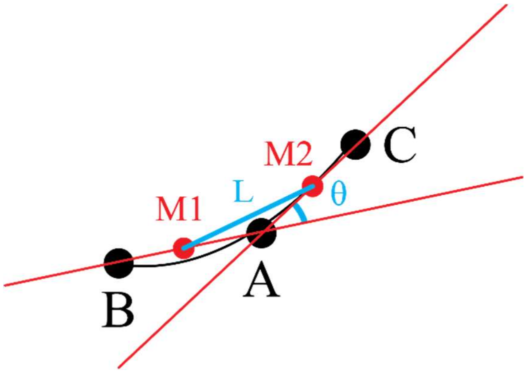


Figure 5.1 Schematic illustration of the curvature calculation process. The curvature at a point of interest A can be calculated using the formula Curvature = ϑ/L, where ϑ is the change in tangential angle over an infinitesimal arc, and L is the length of the arc. If the two points neighbouring A are marked as B and C, and the midpoints of BA and AC are marked as M1 and M2, then ϑ can be approximated by the angle between BA and AC, while L can be approximated by the distance between M1 and M2.

5.1.2.2 Assessing the accuracy of curvature measurement

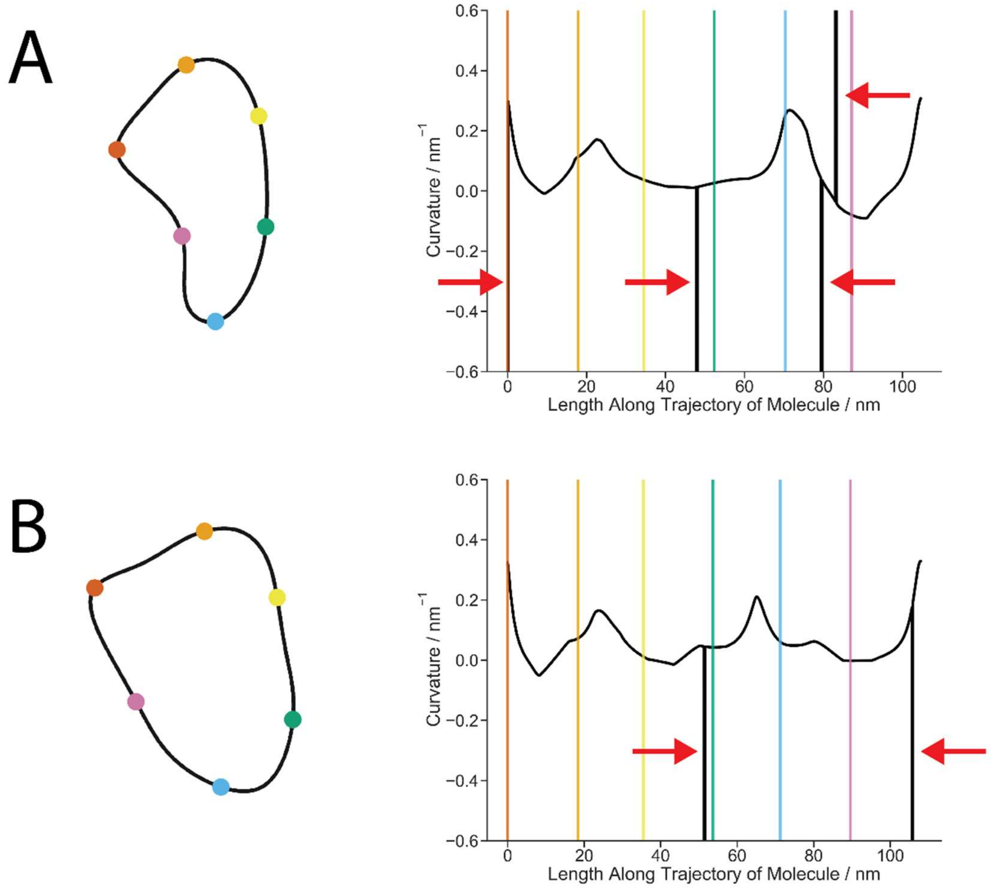


Figure 5.2 Comparison between the trajectory (left) and the curvature measurement (right) for two circular molecules, A and B. Reference points marked on the trajectory of the molecules correspond to vertical lines of the same colour on the curvature plots. Bends on the molecules are well represented in the curvature plots by peaks; for example, the peak next to the blue reference line on the curvature plot of molecule A matches with the sharp bend around the blue reference point. However, erroneous measurements can be observed in the form of abnormally high curvature values, as indicated by red arrows on the curvature plots.

To evaluate the reliability of this calculation method, an initial assessment of the automatic curvature measurement was performed, as shown in Figure 5.2 above: a few example DNA molecules with simple shapes were selected; their curvature values were obtained and plotted against the location of the corresponding points on the molecule; the result was then compared with the trajectory of the DNA. Six reference points were marked with different colours, allowing us to map features of interest on a molecule to its curvature plot.

This comparison showed that bends on the molecules were well represented by peaks on the curvature measurement plots. However, there were several instances of abnormally high curvature values for each molecule, which suggested that the measurement algorithm was flawed. Closer inspection revealed that these irregularities occurred at localities where the point of interest and its neighbours formed a vertical line. Since the calculation was based on dividing an angle over a distance, and the angle was determined through the inverse tangent function, it was likely the case that vertical lines resulted in angles with abnormally large values, which subsequently led to erroneous curvature measurements.

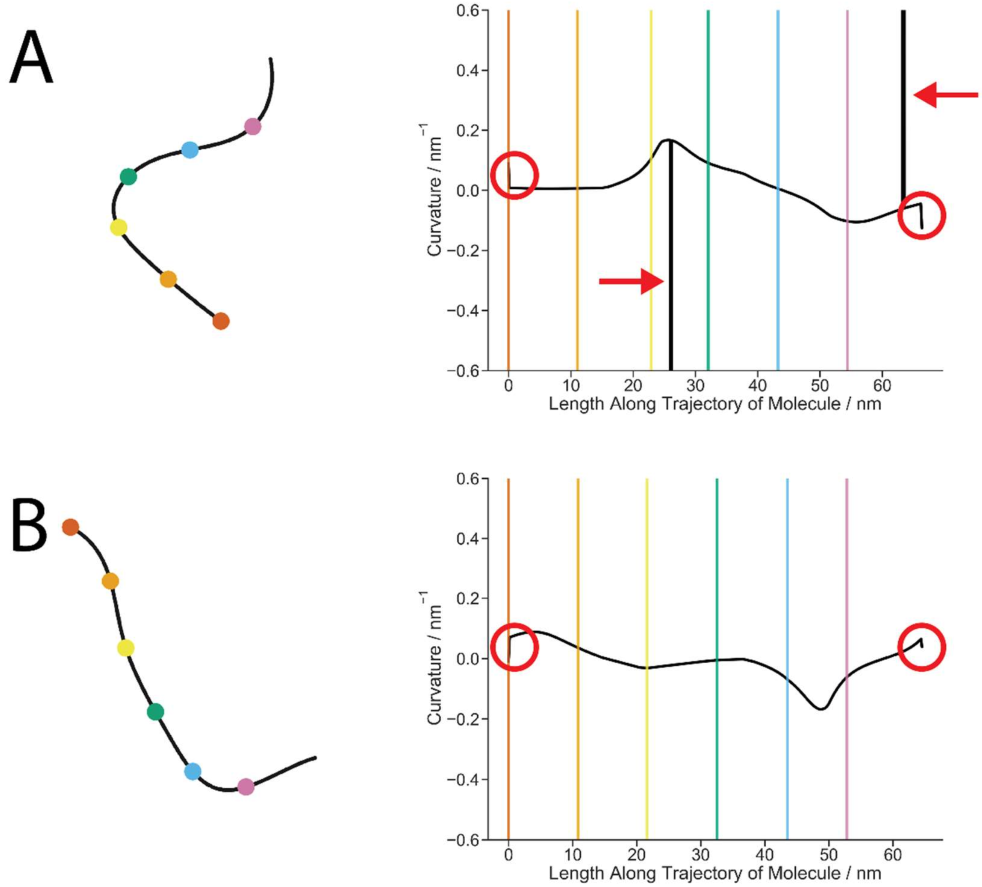


Figure 5.3 Comparison between the trajectory (left) and the curvature measurement (right) for two linear molecules, A and B. Reference points marked on the trajectory of the molecules correspond to vertical lines of the same colour on the curvature plots. In addition to aforementioned abnormally high values indicated by red arrows, errors can also be observed at the starting and ending points of the measurement, as circled in red.

In addition, errors were observed at the start and end of linear molecules, as circled in red in Figure 5.3 above. This was because the end points of linear molecules only had one neighbour, which was incompatible with the calculation method that required coordinates from neighbours on both sides.

5.1.2.3 Reducing irregularities by using more neighbouring points

To address the issues identified during the assessment, we attempted to improve the curvature measurement by including more points in the calculation process, as illustrated in Figure 5.4 below: instead of using only the immediate neighbours, we averaged over multiple neighbours on each side of the point of interest, as this was expected to smooth the curvature plot and reduce irregularities.

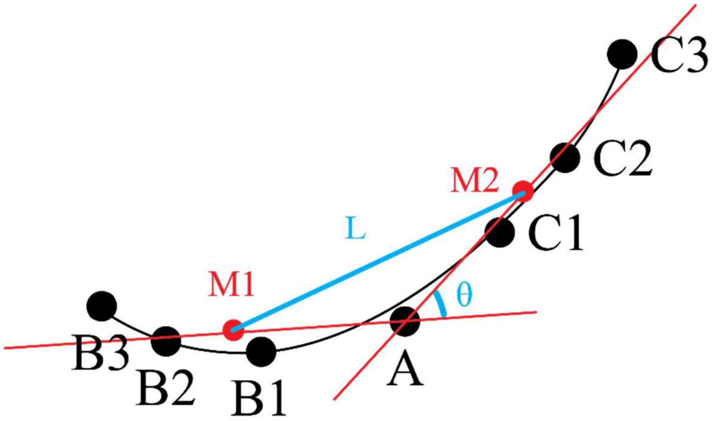
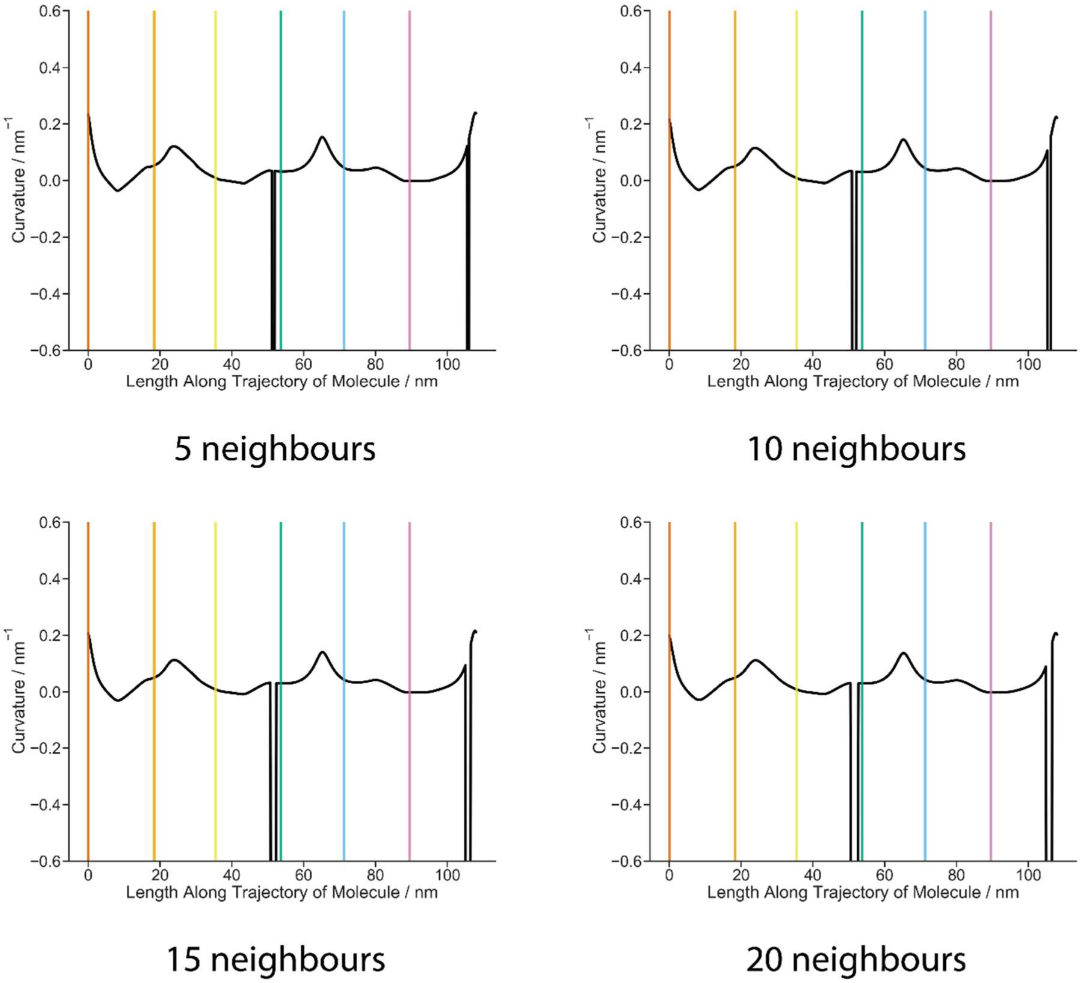


Figure 5.4 The number of neighbouring points involved in the curvature calculation is increased. In this example, three neighbours on each side of the point of interest (A) are used; the left neighbours are marked as B1, B2 and B3, while the right neighbours are marked as C1, C2, and C3. The curvature at point A is still calculated using the formula Curvature = ϑ/L, with ϑ being the change in tangential angle over an infinitesimal arc, and L referring to the length of the arc. L is approximated by the distance between M1 and M2, but instead of using the midpoint of a line segment, M1 is calculated by averaging over B3, B2, B1 and A, while M2 is the average point of A, C1, C2 and C3. The change in tangential angle ϑ is approximated as the angle between M1A and AM2.

To establish the effect of this improvement, the curvature measurements for the same example molecule were plotted repeatedly while the number of neighbours used on each side of a given point of interest was increased. As shown below in Figure 5.5, the curvature plot became smoother when more points were used, but the abnormally high values persisted, although their corresponding peaks became wider. This means that adjusting the number of neighbours involved in the calculation was not sufficient to resolve the issue of irregularities, and that vertical parts of a molecule caused persistent discontinuities in the angle calculation. In addition, there was still no good way to handle end points of linear molecules. Both of these factors suggested that a different method of curvature measurement should be devised to provide more accurate results.



5 neighbours

10 neighbours

15 neighbours

20 neighbours

Figure 5.5 Curvature plots for molecule B from Figure 5.2 obtained by averaging over 5, 10, 15 or 20 neighbours on either side of the point of interest during the calculation process. When the number of neighbours involved is increased, the curvature plot becomes smoother; the irregularities change from sharp lines into wider peaks but they are not eliminated.

5.1.3 Calculating curvature using the ‘numpy.gradient’ function

5.1.3.1 Workflow of calculating curvature

Since the calculation of curvature based on approximating the length of an infinitesimal arc proved to be unsuccessful, we adopted a different approach using the formula

$$Curvature = \frac{d^2x \cdot dy - \ d^2y \cdot dx}{(dx^2 + dy^2)^{1.5}}$$

where dx and d^2x refer to the first and second order gradient of the x coordinate, while dy and d^2y represent the first and second order gradient of the y coordinate. These values are all calculated for each point on the splined trace through an existing Python function ‘numpy.gradient’ [173], which operates on a list of coordinates and handles end points differently, thus preventing the generation of irregular values at the ends of linear molecules. However, this introduces another challenge for circular molecules: since they do not have end points, additional measures are required so that the start and end of their coordinate lists are not treated as such. To achieve this, two copies are appended before and after the coordinate list, forming a longer new list; when the ‘numpy.gradient’ function is subsequently applied to calculate the first and second order gradients, no point from the original list is considered an end point, since the original list has become the middle portion of the new list. Once the gradient calculation is completed, the two copies at the start and end are no longer needed and therefore removed.

A $(x_1, y_1)\ (x_2, y_2)\ldots\ (x_n, y_n)$

B $(x_1, y_1)\ (x_2, y_2)\ldots\ (x_n, y_n)\ (x_1, y_1)\ (x_2, y_2)\ldots\ (x_n, y_n)\ (x_1, y_1)\ (x_2, y_2)\ldots\ (x_n, y_n)$

C $(dx_1, dy_1)\ (dx_2, dy_2)\ldots\ (dx_n, dy_n)\ (dx_1, dy_1)\ (dx_2, dy_2)\ldots\ (dx_n, dy_n)\ (dx_1, dy_1)\ (dx_2, dy_2)\ldots\ (dx_n, dy_n)$

D $(d^2x_1, d^2y_1)\ (d^2x_2, d^2y_2)\ldots\ (d^2x_n, d^2y_n)\ (d^2x_1, d^2y_1)\ (d^2x_2, d^2y_2)\ldots\ (d^2x_n, d^2y_n)\ (d^2x_1, d^2y_1)\ (d^2x_2, d^2y_2)\ldots\ (d^2x_n, d^2y_n)$

E $(dx_1, dy_1)\ (dx_2, dy_2)\ldots\ (dx_n, dy_n)$ $(d^2x_1, d^2y_1)\ (d^2x_2, d^2y_2)\ldots\ (d^2x_n, d^2y_n)$

Figure 5.6 Calculating curvature for circular molecules using the ‘numpy.gradient’ function. A: Points on a DNA molecule are stored in the automatic image analysis program as a list of coordinates. B: For a circular molecule, two additional copies (red) of the coordinate list are produced, and they are combined with the original list; this ensures continuity by preventing the start and end of the original list from being treated as end points. C: The ‘numpy.gradient’ function is applied to the longer new list, generating a list of first order gradients. D: The ‘numpy.gradient’ function is applied to the list of first order gradients, generating a list of second order gradients. E: The two additional copies are removed; the first and second order gradients of the original list are kept and used to calculate the curvature at each point of the molecule.

5.1.3.2 Assessing the accuracy of curvature measurement

5.1.3.2.1 Qualitative observations

The example molecules from Section 5.1.2.2 were used to assess the accuracy of this new curvature calculation method. As shown in Figure 5.7 below, compared to the previous attempt (see Figure 5.2), the irregularly high values no longer existed, and the peaks on the curvature plots were still in accordance with the bends on the DNA.

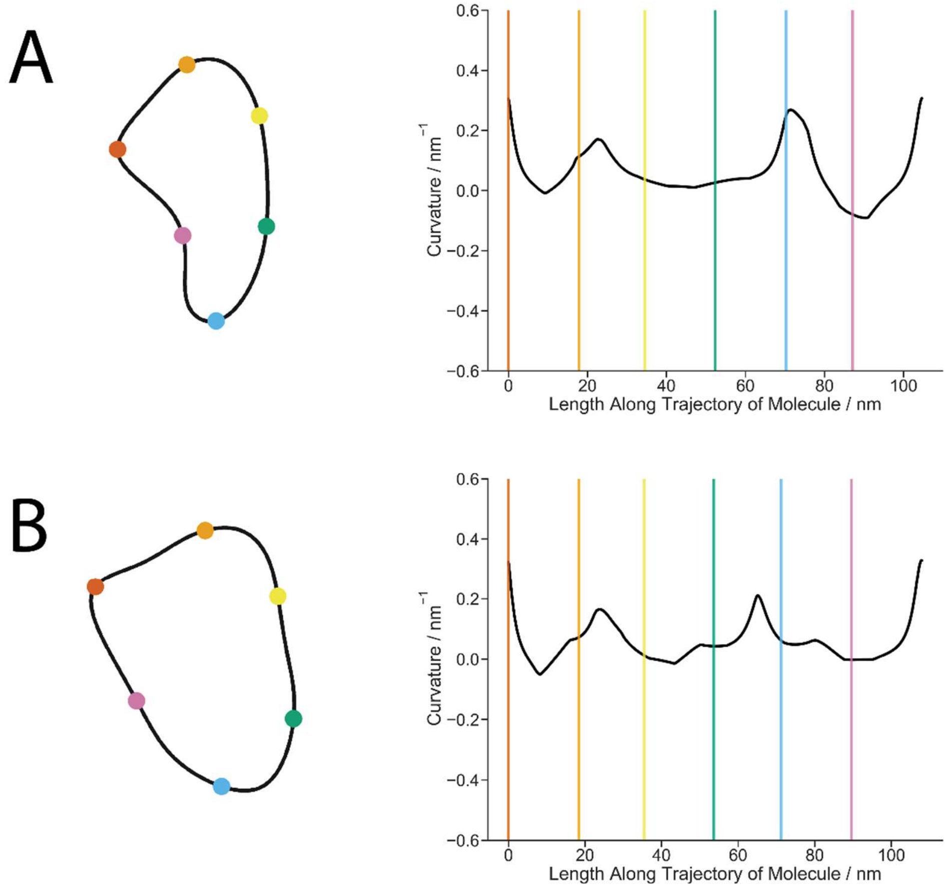


Figure 5.7 Comparison between the trajectory (left) and the curvature (right) calculated using the Python function 'numpy.gradient' for two circular molecules, A and B. Reference points marked on the trajectory of the molecules correspond to vertical lines of the same colour on the curvature plots. This new method of curvature measurement has eliminated the erroneously high values from the previous calculation based on the approximation of an infinitesimal arc (see Figure 5.2), while bends on the molecules are still accurately represented by peaks on the curvature plots.

The assessment was then extended to linear molecules. As shown in Figure 5.8 below, there were no abnormalities at the start and end points of the curvature plots. This suggests that calculating the curvature through the 'numpy.gradient' function is an improvement to the first method of approximating an infinitesimal arc.

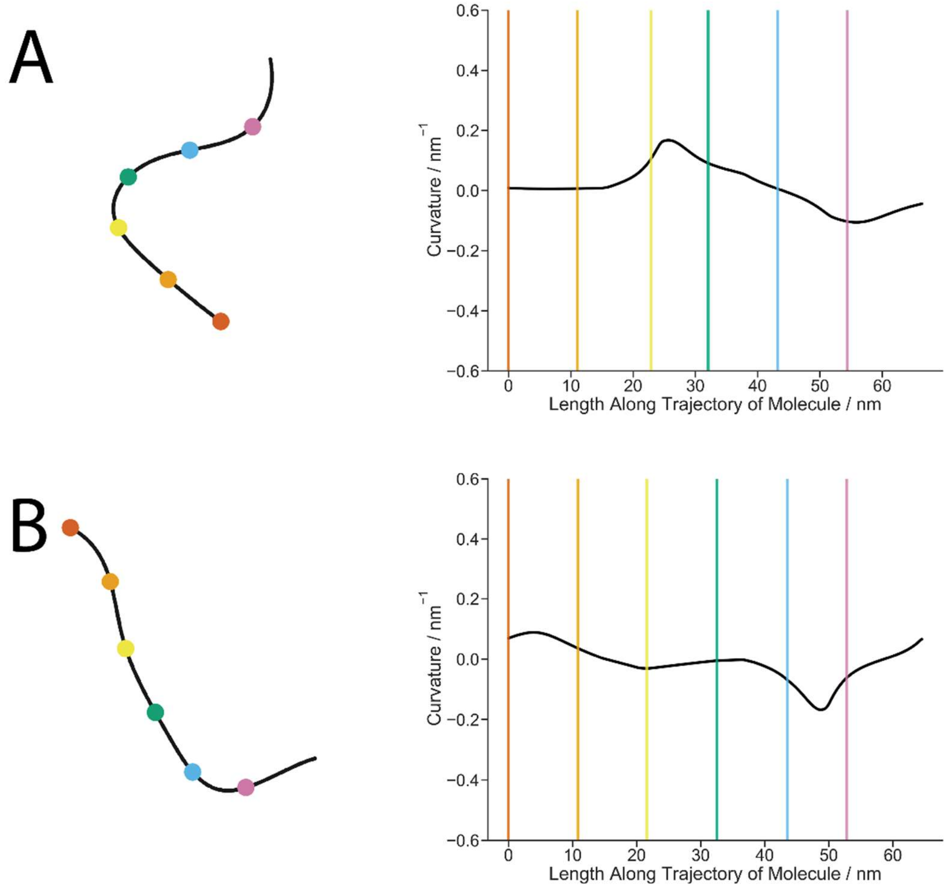


Figure 5.8 Comparison between the trajectory (left) and the curvature (right) calculated using the Python function 'numpy.gradient' for two linear molecules, A and B. Reference points marked on the trajectory of the molecules correspond to vertical lines of the same colour on the curvature plots. Unlike the previous method of measuring curvature by approximating an infinitesimal arc (see Figure 6.3), this new approach does not lead to abnormal values at the start and end of the curvature plots.

5.1.3.2.2 Testing on regular shapes

While qualitative observation showed that curvature plots obtained using the 'numpy.gradient' function were representative of the shape of the molecule and did not contain irregularities, a more quantitative assessment is needed to ensure the accuracy of this measurement method. Such an assessment could be achieved by applying the measurement algorithm to molecules with regular shapes and known curvature values.

Circles, ellipses and parabolas were generated to simulate circular and linear molecules. This allowed the direct quantitative comparison between the measurement and the theoretical curvature value, which is always the reciprocal of the radius for circles, and can be calculated through known formulas for ellipses and parabolas.

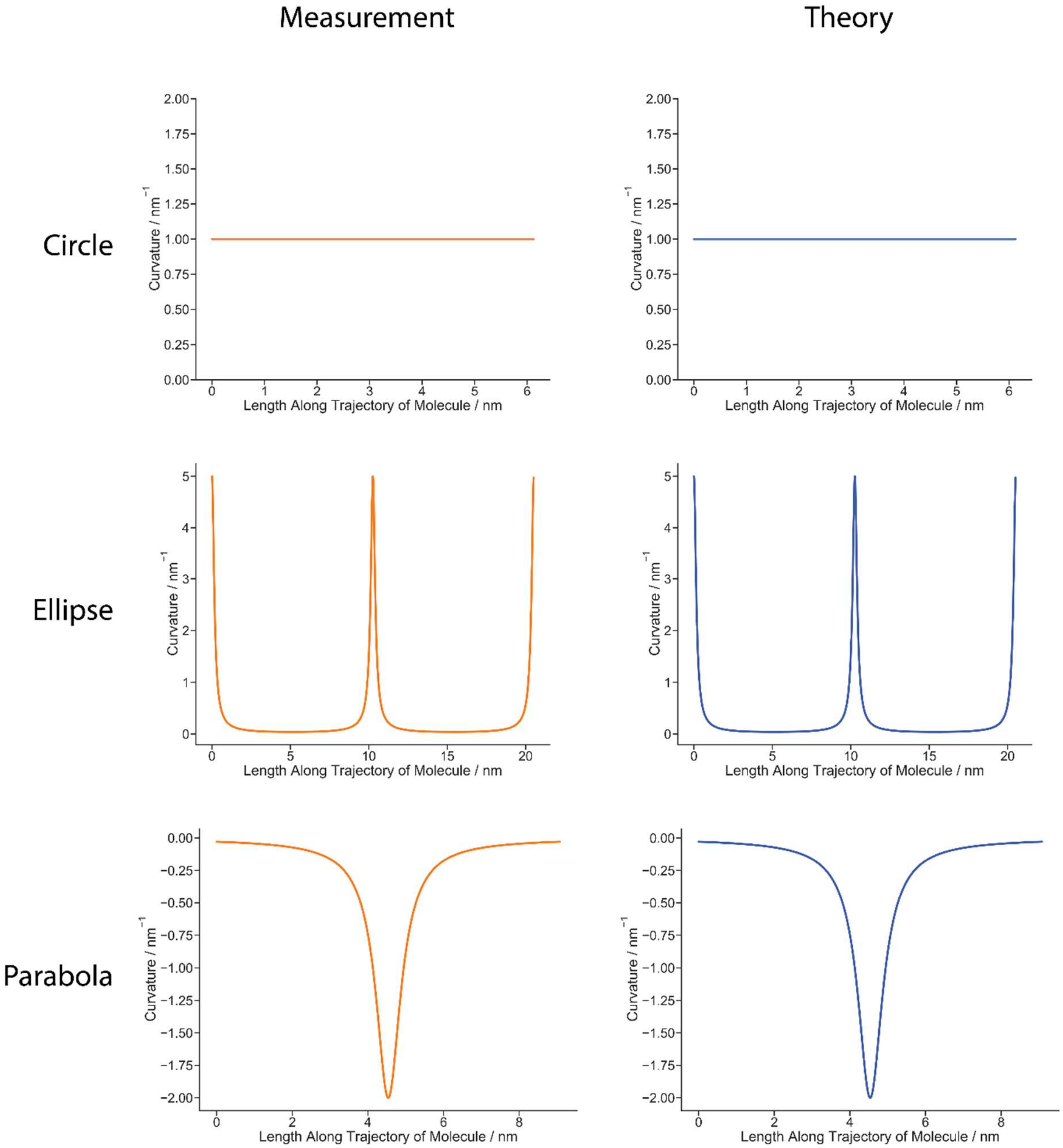


Figure 5.9 Comparison between the curvature plots obtained by measurement (left, in orange) on simulated molecules and by theoretical calculation (right, in blue). Three examples of simulated molecules with regular shapes are shown, including a circle with a radius of 1 nm, an ellipse with a semi-major axis of 5 nm and a semi-minor axis of 1 nm, and the parabola $y=x^2$ with x ranging from -2 nm to 2 nm. For all three molecules, the curvature values measured using the 'numpy.gradient' function are in complete agreement with those expected from theory.

As shown in Figure 5.9 above with three examples of simulated molecules, the curvature measurement fully matched with expected values. Both qualitative and quantitative tests have now proven that, by obtaining the first and second order gradients via the Python function 'numpy.gradient', it is possible to accurately measure the curvature at every point of a linear or circular DNA molecule, provided that tracing is successful.

It is worth noting that testing on simulated data is currently limited to circles, ellipses and parabolas due to difficulties in tracing more complex shapes. In the future, when the tracing algorithm is improved to be functional on molecules that contain crossings, the testing of a new measurement method can be expanded to include simulated molecules with the shape of the number 8, which can be generated by combining two circles with varying radii. This will allow simulated molecules to be more representative of real data while exhibiting known theoretical values for the curvature at each point and other potential measurements.

5.1.4 Challenges of applying the curvature measurement to real data

5.1.4.1 Limitations in accuracy due to tracing

While this project has successfully developed a curvature measurement algorithm that quantitatively characterises the bending extent at every point on a molecule, the applicability of this algorithm on real data is dependent on the reliability of tracing, as discussed in Section 5.1.1. In the specific case of investigating the conformational changes of DNA induced by NDP52, accurate tracing proves to be challenging due to the presence of looping and bridging. Although some DNA molecules are merely bent by NDP52 and do not exhibit untraceable shapes, this is unlikely to be sufficient for the curvature to be a viable parameter in the characterisation of NDP52-DNA interaction, because it is not yet possible to automatically separate molecules that have only gone through bending from molecules that have formed crossings.

However, the curvature measurement can be employed to study conformational changes caused by other proteins such as HU, which is known to only bend DNA without leading to crossings. This will be discussed in more detail in Sections 5.3 and 5.4.

5.1.4.2 Positive and negative curvature values

The automatic measurement algorithm is designed to not only produce local curvature values, but also provide information on the general compactness of an entire molecule. The latter is accomplished through the calculation of the maximum curvature, the mean curvature and the variance of curvature.

It is worth noting that a DNA molecule may contain bends in opposite directions, as shown in Figure 5.10 below, and consequently, the curvature values can be either positive or negative. This difference in sign between curvature measurements from different points needs to be taken into consideration when quantifying the overall shape of a molecule.

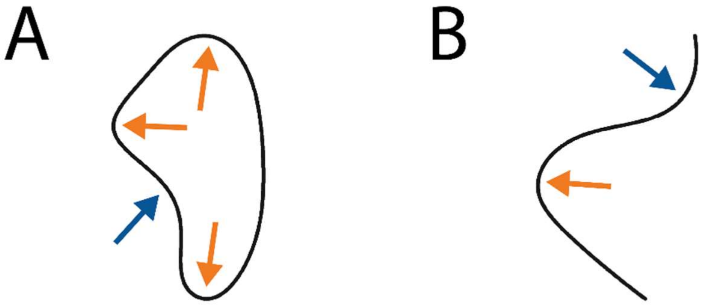


Figure 5.10 Bends on DNA can correspond to positive and negative curvature measurements. On the example circular (A) and linear (B) molecules, if positive curvature values are attributed to bends indicated by orange arrows, negative curvature values will be assigned to bends marked with blue arrows. This difference in sign affects the calculation of parameters such as mean curvature.

To determine the maximum curvature on a given molecule, the signs should be ignored and the absolute values should be used, so that the compactness of different molecules can be directly compared, with a higher maximum curvature indicating a more compact shape. The calculation of the mean curvature is, however, less straightforward since there are two possible approaches: in the first approach, the measurement at each point is converted into its absolute value, and the mean of all the absolute values is calculated; in the second approach, the mean value is calculated with the positive and negative signs taken into account, and the result is converted to its absolute value in the end. For example, if a series of points have the curvature measurements of -1, -2, -1, 0, 1, the first approach will lead to a mean curvature of (|-1| + |-2| + |-1| + |0| + |1|) / 5 = 1, while the calculation will be |((-1) + (-2) + (-1) + 0 + 1) / 5| = 0.6 for the second approach.

The former mean curvature value is better at quantifying the frequency and extent of bends on a molecule, but the latter is a more accurate representation of compactness. As illustrated in Figure 5.11 below, an S-shaped molecule contains more bends than a C-shaped molecule, but the C-shaped molecule is more compact. This distinction is relevant in certain situations, where C-shaped molecules are formed due to protein interaction, while S-shaped molecules are caused by natural bending occurring at random points and in random directions, which is unrelated to the protein; in this case, the second method of calculation is more useful, because it produces a mean curvature value close to 0 for the S-shaped molecule, with bends in opposite directions corresponding to curvature values of opposite signs and cancelling out each other; the first approach would lead to a high mean curvature value for the same molecule, reflecting the presence of bends but failing to distinguish between intrinsic bending and protein-induced compaction.

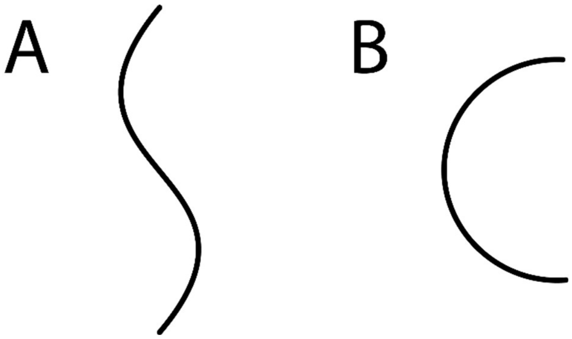


Figure 5.11 Schematic illustrations of an S-shaped (A) and a C-shaped (B) molecule. If a protein is known to induce compaction in DNA, it is more likely to result in molecule B, while molecule A may be caused by unrelated natural bending. Consequently, when calculating the mean curvature, it is useful for positive and negative measurements from different parts of molecule A to cancel out each other, so that the result is close to 0 for molecule A but much higher for molecule B, which reflects the effect of the protein rather than intrinsic bending.

The choice between the two approaches therefore depends on the needs of the specific project: the first approach is preferred when characterising only the extent of bending, while the second approach should be chosen if the directionality of the bends and the compactness of the molecule matter. For the rest of this chapter, the result from the first approach will be referred to as the 'mean absolute curvature', while the value calculated using the second approach will be referred to as the 'absolute mean curvature'.

Similarly, when calculating the variance of curvature on a molecule, two different approaches can be considered: the first determines the variance based on absolute curvature, while the second acknowledges the difference between positive and negative curvature values; the results from these two approaches are referred to as 'variance of absolute curvature' and 'variance of curvature' respectively.

5.2 Assessing the viability of characterising DNA through curvature

5.2.1 Overview of the assessment

It has been demonstrated that the bending extent at each point of a traceable DNA molecule can be accurately quantified through curvature measurement, although the reliability of this method is limited when complex shapes and crossings are involved. It is important to establish the severity of this limitation for the NDP52 project, so that we can find out whether the automatically measured curvature values can still be used to characterise certain conformational changes on DNA, or they are rendered non-viable due to tracing difficulties and only serve as a proof of concept.

Section 4.4 has shown that both the choice of immobilising agent and interaction with NDP52 can affect the shape of DNA, and that the resulting conformational changes can be quantified by measuring the maximum bounding distance. Because of this, curvature-related measurements are considered viable if they are able to capture the same changes and produce results that are in line with maximum bounding distance measurements; if this is not the case, it can be concluded that the curvature is not yet a suitable parameter for the NDP52 project.

We first assess whether conformational changes related to immobilisation agents are well represented by curvature measurements. This will be discussed in Section 5.2.2, which compares images of DNA immobilised with Ni^{2+}, Mg^{2+}-Ni^{2+} exchange and PLO; the dataset without NDP52 treatment is used so as to avoid confounding factors. Subsequently, the ability of curvature measurements to capture conformational changes induced by NDP52 interaction will be evaluated in Section 5.2.3, which focuses on the low-concentration PLO dataset since it is the most suitable for probing the effects of NDP52, as confirmed in Section 4.4.3.1.

5.2.2 Assessing the viability of curvature measurements by characterising the effects of immobilising agents on untreated DNA

5.2.2.1 Assessing the viability of characterising untreated DNA through maximum curvature

While the local conformation of DNA can be characterised through curvature measurement at each point of the molecule, the resulting plots do not allow for straightforward comparison between different molecules. This issue can be addressed by using the maximum curvature, which facilitates inter-molecular comparison of local conformational changes by quantifying the extent of the sharpest bend on each molecule. To assess the suitability of this parameter for the characterisation of conformational changes related to immobilising agents, we compared maximum curvature measurements of untreated DNA immobilised with Ni^{2+}, Mg^{2+}-Ni^{2+} exchange and PLO.

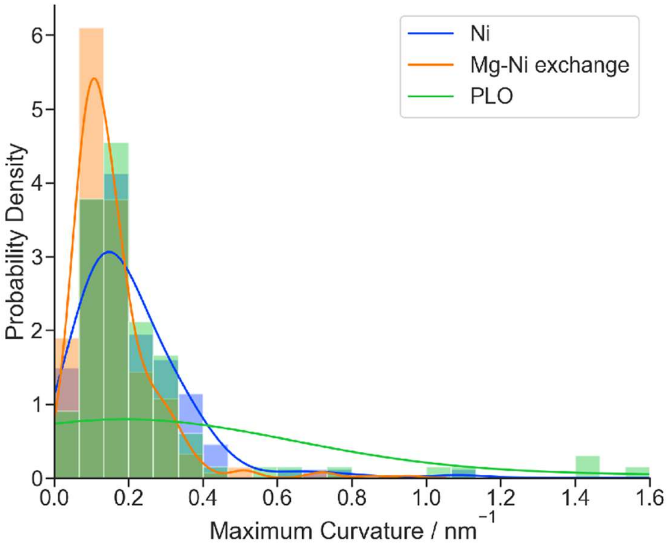


Figure 5.12 KDE plots and normalised histograms comparing the maximum curvature distributions of untreated DNA imaged with Ni^{2+} (blue, peak and standard deviation 0.15 ± 0.21 nm^{-1}, N = 132, number of images = 7), Mg^{2+}-Ni^{2+} exchange (orange, 0.11 ± 0.11 nm^{-1}, N = 418, number of images = 5) and PLO (green, 0.19 ± 1.13 nm^{-1}, N = 102, number of images = 19) as immobilising agents. 4 molecules with exceptionally high maximum curvature values (11.05 nm^{-1}, 2.35 nm^{-1}, 2.07 nm^{-1} and 1.94 nm^{-1}) are not shown on the graph but included in the calculation. Note that the sample size is slightly smaller compared to the distribution of maximum bounding distance from the same dataset

(N = 134, 419 and 103 for Ni^{2+}, Mg^{2+}-Ni^{2+} exchange and PLO respectively; see Figure 4.40 in Section 4.4.3.2); this is because the tracing algorithm used in the NDP52 project is dependent on an external program [58] *that fails to recognise molecules too close to the image border, and the maximum bounding distance measurement is unaffected since it does not require tracing.*

According to Figure 5.12, the distributions of maximum curvature are similar for DNA immobilised with Ni^{2+} and Mg^{2+}-Ni^{2+} exchange, though when PLO is used, the distribution is wider and there are more data points with higher values, indicating greater conformational variety and an increased presence of highly bent molecules. This appears to be in accordance with previous observations and measurements that PLO leads to more compaction in DNA.

Upon identifying and inspecting molecules with the highest maximum curvature measurements, it was however revealed that they had been incorrectly traced, as shown in Figure 5.13 below. In addition, to evaluate the general tracing quality, the rest of the molecules were also manually inspected and no obvious tracing errors were found.

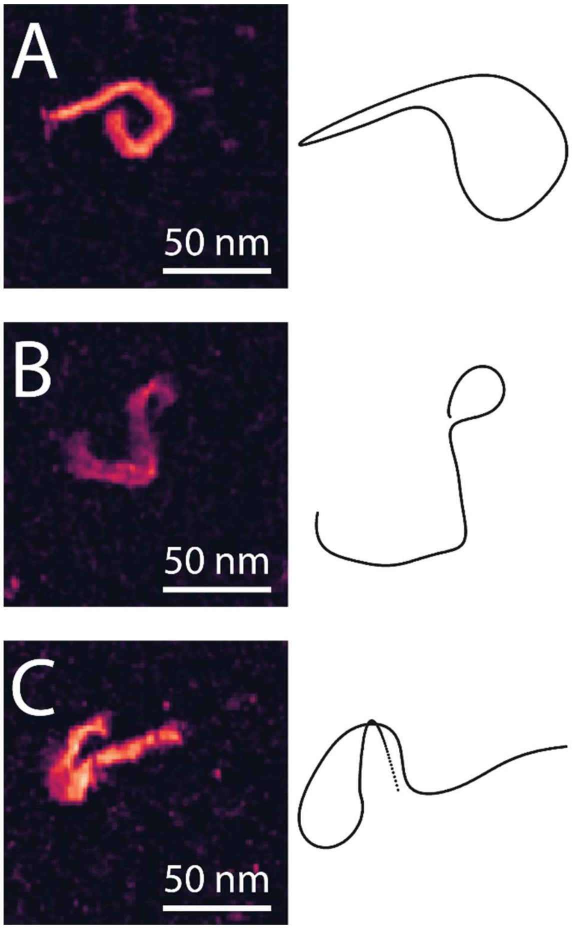


Figure 5.13 Comparison between the AFM images (left; see Figure 4.21 for height scale) and incorrectly produced traces (right) for three molecules with the highest maximum curvature measurements of 11.05 nm^{-1} (A), 2.35 nm^{-1} (B) and 2.07 nm^{-1}(C). These molecules were all immobilised with PLO, and they skewed the distribution of maximum curvature in Figure 5.12.

Since the erroneous tracing affects molecules immobilised with PLO in particular, it is likely that the apparent increase in maximum curvature when PLO is used is not directly representative of DNA conformation, but rather caused by a higher proportion of molecules with tracing difficulties. To confirm this possibility, we compared the distributions of maximum curvature for the PLO dataset before and after removing the three wrongly traced molecules from Figure 5.13.

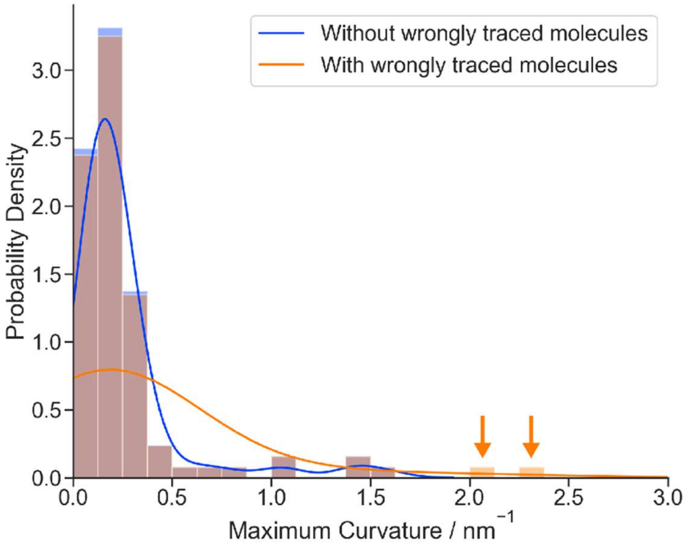


Figure 5.14 KDE plots and normalised histograms comparing the maximum curvature distributions of PLO-immobilised DNA with (orange, peak and standard deviation 0.19 ± 1.13 nm^{-1}, N = 102, number of images = 19) and without (blue, 0.16 ± 0.28 nm^{-1}, N = 99, number of images = 19) the three wrongly traced molecules (2.9% of all molecules) shown in Figure 5.13. The wrongly traced molecule with the highest maximum curvature measurement (11.05 nm^{-1}) is not shown in the graph but included in the calculation. The other two wrongly traced molecules are indicated with orange arrows.

Figure 5.14 above demonstrates that the distribution became much narrower following the removal of the three molecules with the highest maximum curvatures. This further proves that the greater conformational variety of the PLO dataset suggested by Figure 5.12 is due to unsuccessful tracing, and does not reflect the actual bending extent of DNA molecules. Consequently, the automatic measurement of maximum curvature is currently unsuitable for the characterisation of conformational differences related to immobilising agents, since it is not viable to always manually identify molecules with tracing errors.

5.2.2.2 Assessing the viability of characterising untreated DNA through mean curvatures

Conformational changes at specific localities can theoretically be represented by the maximum curvature, but this method is currently unreliable because incorrectly traced molecules exhibiting irregularly high curvature values can skew the distribution; as the mean curvature measurements are less affected by extreme values, we attempted to find out whether they could be employed to characterise global conformational differences related to immobilising agents.

As shown in Figure 5.15 below, both the mean absolute curvature and the absolute mean curvature (see Section 5.1.4.2 for their difference) of DNA molecules immobilised with Ni^{2+}, Mg^{2+}-Ni^{2+} exchange and PLO were compared.

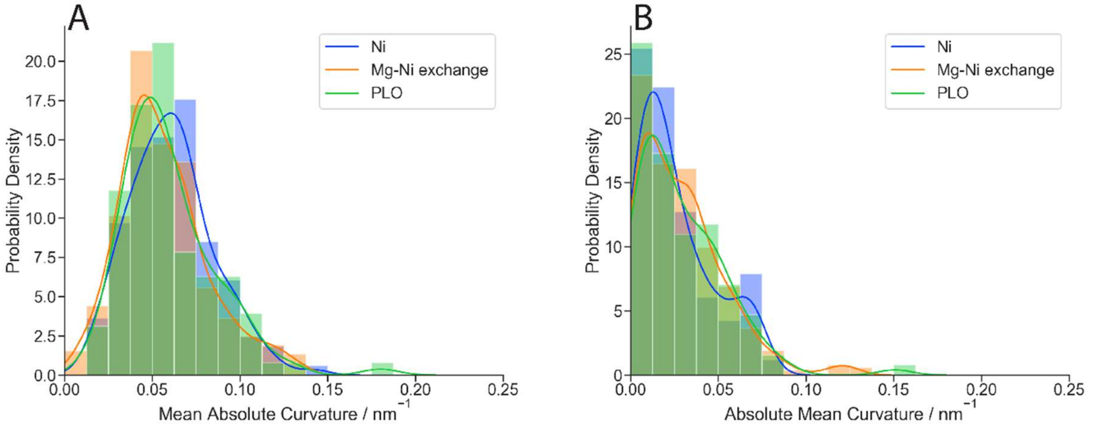


Figure 5.15 KDE plots and normalised histograms comparing the mean absolute curvature (A) and absolute mean curvature (B, see Section 5.1.4.2 for the difference between the two mean curvatures) distributions of untreated DNA imaged with Ni^{2+} (blue, peak and standard deviation 0.06 ± 0.02 nm^{-1} for mean absolute curvature, 0.01 ± 0.02 nm^{-1} for absolute mean curvature, N = 132, number of images = 7), Mg^{2+}-Ni^{2+} exchange (orange, 0.05 ± 0.03 nm^{-1} for mean absolute curvature, 0.01 ± 0.02 nm^{-1} for absolute mean curvature, N = 418, number of images = 5) and PLO (green, 0.05 ± 0.03 nm^{-1} for mean absolute curvature, 0.01 ± 0.02 nm^{-1} for absolute mean curvature, N = 102, number of images = 19) as immobilising agents.

For all three immobilising agents, the mean absolute curvature exhibits peaks at higher values than the absolute mean curvature. This is in line with expectations, since the absolute mean curvature is calculated by averaging over positive and negative measurements and is therefore closer to zero. However, neither parameter is able to capture the conformational differences between molecules immobilised with PLO and metallic ions; a possible explanation is that, the most drastic changes in global conformation are not successfully represented in the traces, while bends that are preserved in the tracing process are not sharp or frequent enough to have an impact on the mean curvature measurements.

5.2.2.3 *Assessing the viability of characterising untreated DNA through variances of curvature*

The maximum curvature quantifies the most significant change in local conformation, while the mean curvatures are associated with the global conformation. To complement these parameters and characterise the extent to which the conformation at different localities of the same molecule differ from each other, the variance and absolute variance of curvature (see Section 5.1.4.2 for their difference) are measured.

We compared both variances of curvature obtained from untreated DNA immobilised with Ni^{2+}, Mg^{2+}-Ni^{2+} exchange and PLO, so that the ability of these parameters to capture immobilisation-related conformational differences can be assessed.

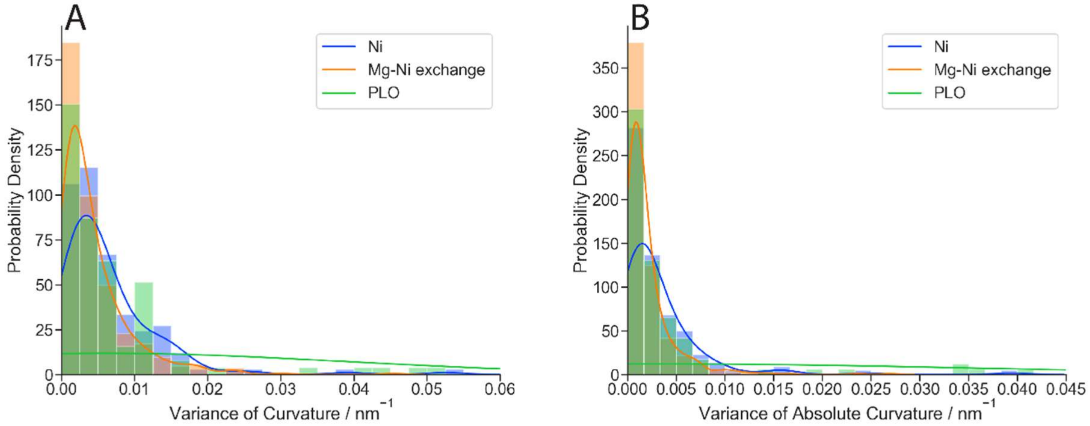


Figure 5.16 KDE plots and normalised histograms comparing the variance of curvature (A) and variance of absolute curvature (B, see Section 5.1.4.2 for the difference between the two variances) distributions of untreated DNA imaged with Ni^{2+} (blue, peak and standard deviation 0.003 ± 0.007 nm^{-1} for variance of curvature, 0.002 ± 0.005 nm^{-1} for variance of absolute curvature, N = 132, number of images = 7), Mg^{2+}-Ni^{2+} exchange (orange, 0.002 ± 0.005 nm^{-1} for variance of curvature, 0.001 ± 0.003 nm^{-1} for variance of absolute curvature, N = 418, number of images = 5) and PLO (green, 0.006 ± 0.081 nm^{-1} for variance of curvature, 0.003 ± 0.080 nm^{-1} for variance of absolute curvature, N = 102, number of images = 19) as immobilising agents. Molecule A from Figure 5.13 (variance of curvature 0.820 nm^{-1}, variance of absolute curvature 0.811 nm^{-1}) is not shown on the graphs but included in the calculation.

Similar to the case of maximum curvature, the distributions for the PLO dataset are wider and peak at higher values. To establish whether this is also due to erroneous tracing and artificial bends, we compared the distributions of variances for DNA immobilised with PLO with and without the three wrongly traced molecules shown in Figure 5.13.

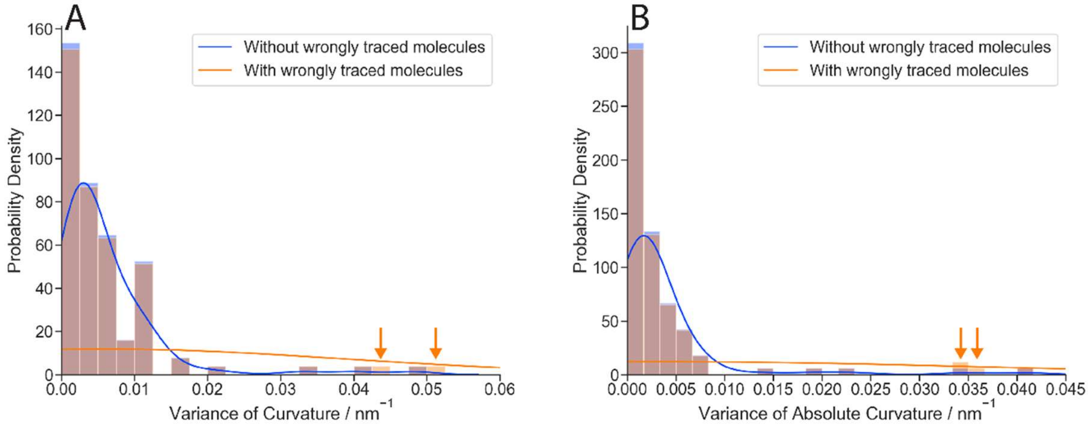


Figure 5.17 KDE plots and normalised histograms comparing variance of curvature (A) and variance of absolute curvature (B, see Section 5.1.4.2 for the difference between the two variances) distributions of PLO-immobilised DNA with (orange, peak and standard deviation 0.006 ± 0.081 nm^{-1} for variance of curvature, 0.003 ± 0.080 nm^{-1} for variance of absolute curvature, N = 102, number of

images = 19) and without (blue, 0.003 ± 0.007 nm^{-1} for variance of curvature, 0.002 ± 0.006 nm^{-1} for variance of absolute curvature, N = 99, number of images = 19) the three wrongly traced molecules (2.9% of all molecules) shown in Figure 5.13, among which molecule A (variance of curvature 0.820 nm^{-1}, variance of absolute curvature 0.811 nm^{-1}) is not shown on the graphs but included in the calculation, while molecules B and C are indicated with orange arrows.

As shown in Figure 5.17 above, the position and width of the peaks for the variances of curvature are heavily affected by extreme values from wrongly traced molecules. These parameters are therefore also limited by tracing difficulties and do not yet serve as viable metrics for the characterisation of conformational changes related to immobilising agents.

5.2.2.4 Conclusions of assessment

Among the parameters assessed, the maximum curvature and the variances of curvature initially appeared to capture conformational differences between DNA immobilised with metallic ions and with PLO; however, this was subsequently revealed to be caused by tracing errors with certain molecules in the PLO dataset, and did not reflect the actual shape of DNA. The mean curvatures, while less affected by such errors, also failed to represent global conformational differences related to immobilisation. In summary, the curvature-based parameters cannot be employed to characterise immobilisation-induced conformational differences until the tracing algorithm is improved.

5.2.3 Assessing the viability of curvature measurements by characterising the effects of NDP52 on DNA

5.2.3.1 Identifying molecules with tracing errors that lead to extreme curvature values

It has been demonstrated in Section 5.2.2 that the ability of curvature-related parameters to capture the effects of immobilisation agents on DNA conformation is hindered by inaccurate tracing on some molecules. Here we further investigate whether this issue also affects the characterisation of conformational changes caused by NDP52 interaction. To achieve this, it is necessary to first identify molecules with major tracing errors that can affect the distribution of measured parameters by introducing artificially sharp bends and extreme curvature values.

As concluded in Section 4.4.3, PLO is the most suitable immobilising agent for the characterisation of NDP52-DNA interaction. Automatic curvature measurement was therefore performed for PLO-immobilised DNA molecules both with and without NDP52 treatment. To locate erroneously sharp bends within hundreds of molecules, the tracing quality for all molecules was manually inspected, with special attention paid to those with the highest maximum curvature values. In addition to the three molecules without NDP52 treatment that had been discussed in Section 5.2.2.1 and presented in Figure 5.13, four molecules from the NDP52-treated dataset were identified as displaying irregularly high curvature values due to tracing errors, as shown below in Figure 5.18, while obvious tracing inaccuracies were not observed on molecules with lower maximum curvature values.

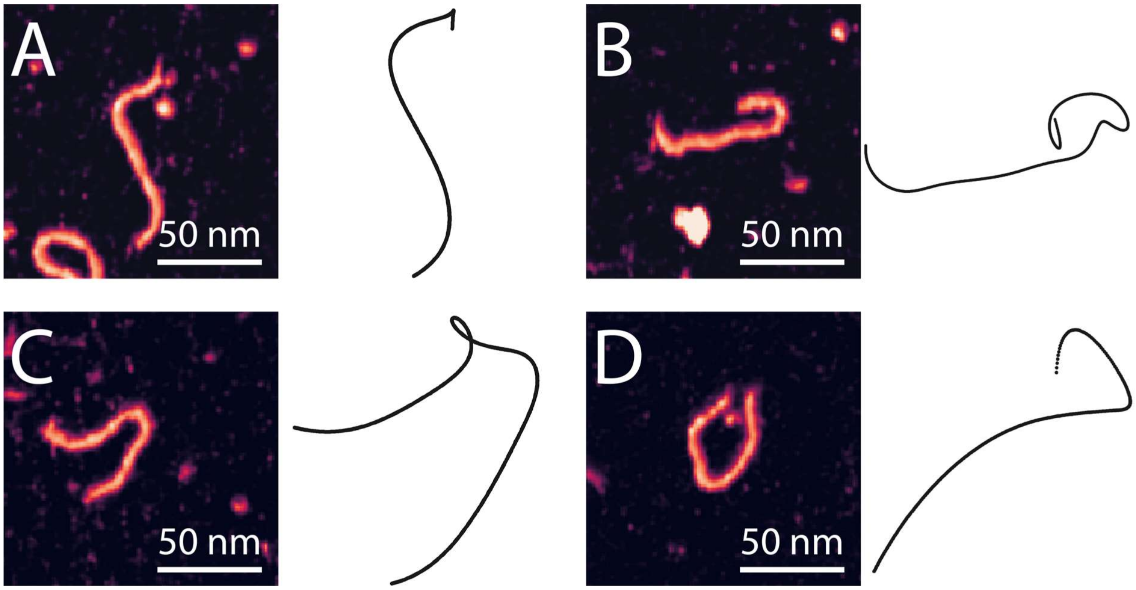


Figure 5.18 Comparison between the AFM images (left; see Figure 4.21 for height scale) and incorrectly produced traces (right) for four molecules identified to have irregularly high maximum curvature measurements of 72.65 nm^{-1} (A), 5.11 nm^{-1} (B), 3.30 nm^{-1}(C) and 1.56 nm^{-1} (D). These molecules were all immobilised with PLO and had all gone through NDP52 co-incubation. For erroneous tracing on molecules without NDP52 treatment, see Figure 5.13 in Section 5.2.2.1.

For molecules A, B and C, the traces followed their general shapes but introduced artificial bends at specific localities, resulting in incorrect curvature measurements. This is because the skeletonisation algorithm cannot distinguish between the DNA and nearby NDP52 molecules or other impurities. In the case of molecule D, it was unclear at first how a trace completely different from the DNA trajectory was produced; however, the cause of this error was subsequently revealed by observing the trace overlaid on the AFM image, as shown in Figure 5.19 below: the molecule was initially misclassified as circular instead of linear due to the presence of the protein, which was treated as part of the DNA; consequently, instead of initiating from an end of the molecule, the ordering process started from the middle and continued in the wrong direction upon encountering the protein (see Section 4.1.2.4.2); this led the ordering to end prematurely with the bottom half of the DNA discarded.

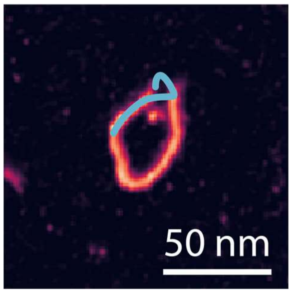


Figure 5.19 The erroneous trace of molecule D from Figure 5.18 overlaid on the AFM image (see Figure 4.21 for height scale). The tracing algorithm was unable to recognise the DNA as linear because the protein was placed next to both of its ends; this caused the ordering process to start from the middle of the DNA and to deviate from the trajectory of the molecule.

The examples of inaccurate tracing shown in Figure 5.18 above and Figure 5.13 in Section 5.2.2.1 suggest that even when there are no crossings, the reliability of the current tracing algorithm can still be limited by the presence of proteins, impurities, or background noise, since these features also present as pixels above the height threshold on an AFM image and cannot be distinguished from the DNA molecule. To resolve this issue, future projects aimed to improve tracing need to take the height of each point into account throughout the entire tracing process, starting from skeletonisation; in particular, when arranging points on the molecule into order by continuously selecting the next nearest neighbour (see Section 4.1.2.4), the 3D distance between two points should be calculated instead of the 2D distance on the X-Y plane.

5.2.3.2 *Assessing the effect of tracing errors on maximum curvature distribution*

To evaluate the effect of tracing errors on the distributions of curvature-related parameters, we compared the maximum curvature measurements before and after removing the seven molecules identified to contain extreme values from the dataset.

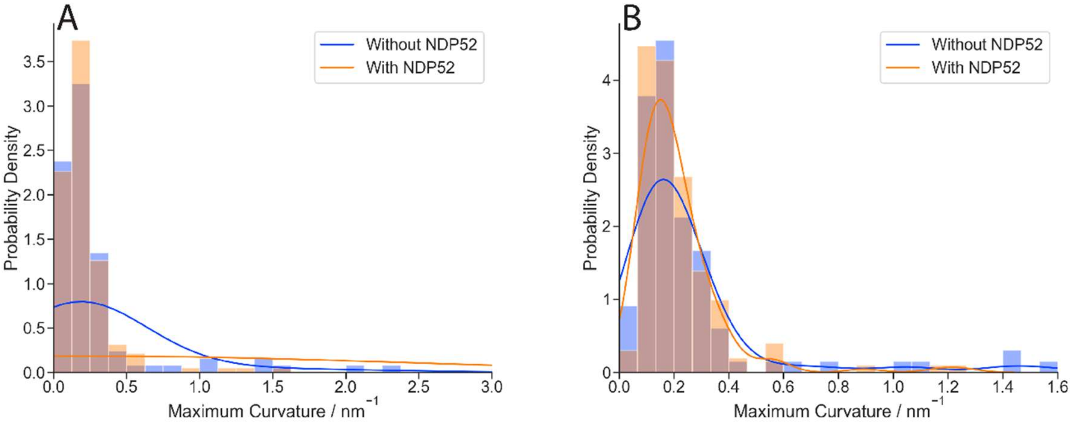


Figure 5.20 KDE plots and normalised histograms comparing the maximum curvature distributions of PLO-immobilised DNA with (orange) and without (blue) NDP52 treatment before (A; peak and standard deviation 0.19 ± 1.13 nm^{-1}, N = 102, number of images = 19 without NDP52; 0.23 ± 5.82 nm^{-1}, N = 155, number of images = 16 with NDP52) and after (B; 0.16 ± 0.28 nm^{-1}, N = 99, number of images = 19 without NDP52; 0.15 ± 0.17 nm^{-1}, N = 151, number of images = 16 with NDP52) the removal of seven molecules (2.7% of all molecules) with tracing errors from Figure 5.13 in Section 5.2.2.1 and Figure 5.18 in Section 5.2.3.1. Molecule A from Figure 5.13 and molecules A, B and C from Figure 5.18 are not shown on the graphs but included in the calculation.

As shown in Figure 5.20 above, when the erroneously traced molecules are included, the distribution of maximum curvature is much wider for DNA that has undergone NDP52 co-incubation than for untreated DNA; by comparison, when these molecules are removed, the two distributions become similar. This indicates that, comparable to the case with immobilising agents, the maximum curvature is unable to directly capture the conformational changes induced by NDP52-DNA interaction; however, since NDP52 treatment leads to the formation of complex DNA shapes and an increased presence of proteins next to DNA, it can result in more erroneous tracing, which gives rise to a wide range of curvature measurements that do not correspond to actual DNA conformation. In addition, it has been demonstrated that the tracing for NDP52-treated DNA is particularly unreliable compared to other datasets, and as a result, other curvature-related parameters including mean curvatures and variances of curvatures are also unsuitable for the quantification of conformational changes caused by NDP52.

5.2.3.3 Conclusions of assessment

NDP52 co-incubation leads to DNA crossings and causes protein molecules to be occasionally found next to DNA. Both of these factors render the tracing process difficult and result in more errors. The curvature-related measurements for NDP52-treated DNA are therefore particularly affected by inaccurate traces and irregularly high values. Because of this, the maximum bounding distance as discussed in Section 4.4.2 remains the only viable method so far to characterise NDP52-DNA interaction.

While not suitable for the NDP52 project, the curvature-based parameters can be reassessed in the future for their ability to characterise conformational changes, but this will require the use of a more appropriate dataset that do not involve complex DNA shapes and contain fewer proteins or impurities.

5.3 Developing bending angle measurement methods to characterise the effects of HU on DNA

5.3.1 Introduction to the characterisation of HU-DNA interaction

5.3.1.1 Background on HU and bending angle measurement

Due to limits of the NDP52 dataset and of the current tracing algorithm, none of the tracing-related automatic measurements discussed so far can be directly used to characterise DNA conformation. To develop and assess more methods that measure conformational changes on DNA as a result of protein interaction, a different DNA-manipulating protein needs be used to minimise tracing errors.

The HU protein known to bend bacterial DNA is suitable for this purpose, as it does not join DNA molecules together or induce crossings [152]. Furthermore, with a size of about 90 amino acids [174], HU is much smaller than NDP52 (446 amino acids [149]), and is therefore less likely to directly interfere with the DNA tracing process.

As discussed in Section 1.4.3.3, studies have suggested that HU causes DNA to bend at specific angles, possibly due to geometric restrictions of the HU-DNA complex; however, these studies do not agree on the exact value of the bending angle, with crystallography data indicating a range of 105° to 140° [101], and fluorescence resonance energy transfer measuring 70° [116]. To investigate these differences and obtain further information on the interaction mechanism between HU and DNA, our collaborators Elliot W. Chan and Agnes Noy performed molecular dynamics (MD) simulations of linear DNA molecules damaged at the centre, since previous research had shown that HU preferentially binds to damage and irregularities such as repair intermediates [153]. Their simulations showed that after interaction with HU, DNA molecules exhibited bending at the damage site as expected, but the distribution of the bending angle peaked around three different values of 100.2°, 137.0° and 161.1°. AFM imaging was subsequently carried out by Elliot W. Chan and Daniel E.

Rollins, so that these results can be verified experimentally. To facilitate the comparison with MD data, this project attempts to develop a method for the automatic measurement of DNA bending angle on AFM images.

Since the bending induced by HU always takes place at the damage site, which is located at the centre of the molecule, the automatic bending angle measurement will not be affected by the removal of the ends during the skeletonisation process (see the second half of Section 4.1.2.2.2); in addition, this new characterisation method can be adapted for future projects that also focus on conformational changes at specific sites.

5.3.1.2 AFM data provided

Elliot W. Chan and Daniel E. Rollins provided three datasets of AFM images, including undamaged DNA, damaged DNA without HU treatment, and damaged DNA co-incubated with HU. All samples were immobilised with poly-L-lysine (PLL), which was used instead of PLO due to the unavailability of the latter at the time of the experiment. The DNA molecules imaged with AFM had the same length and sequence as those involved in the simulation: the undamaged DNA was linear and contained 303 base pairs, while the damaged DNA was largely identical, with the only difference being the introduction of two flipped bases, two stacked bases and a base pair mismatch at the centre of the molecule.

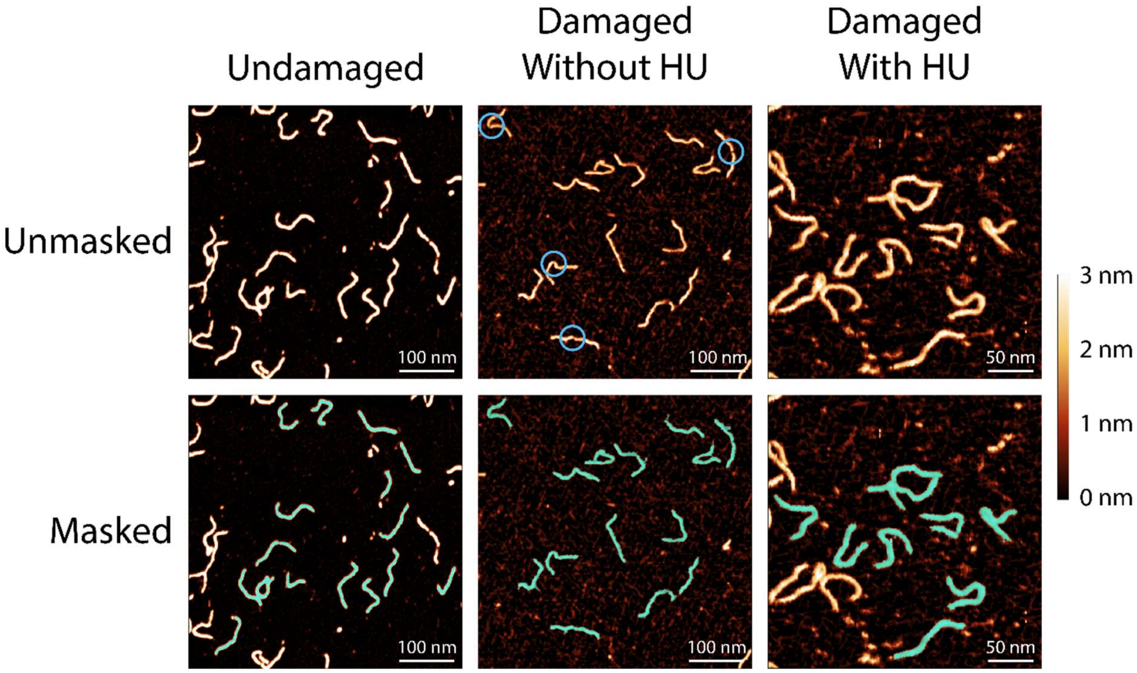


Figure 5.21 Example AFM images from three different datasets before (top) and after (bottom) the molecular identification process (masking), where DNA molecules are automatically recognised from the background and highlighted in cyan. For damaged DNA without HU treatment, the damage at the centre of the molecule can sometimes be observed, as circled in blue. The height scale attached is used for all AFM images in Sections 5.3 and 5.4.

As shown in Figure 5.21 above, HU molecules are not directly observed on AFM images since they are much smaller than DNA. This means that unlike NDP52, HU will not be mistakenly identified as part of the DNA and cause tracing errors; on the other hand, the lack of visible proteins also renders it impossible to determine through manual observation whether a DNA molecule has interacted with HU.

It is also worth remarking that, for the dataset of damaged DNA without HU co-incubation, the damage at the centre is sometimes noticeable, as circled in blue in Figure 5.21. Moreover, some molecules from this dataset exhibit a bend around the damage site. This is because while the damage does not directly bend the DNA, it causes the molecule to be more flexible. This effect needs to be taken into consideration when characterising the bending angle, so that it can be distinguished from the effects of HU.

5.3.1.3 Challenges of characterising HU-DNA interaction through bending angle

5.3.1.3.1 Addressing impurities and noise

The AFM data provided varies in quality, and some images are affected by a substantial amount of impurities and noise. While these unwanted features can be distinguished from DNA molecules during the masking stage through the adjustment of height and area thresholds, it is challenging to establish a set of parameters that can be applied to the entire dataset, because images are taken under different experimental conditions. The thresholds that work well on most of the dataset are ineffective for certain images, causing impurities to be incorrectly included in the mask, as indicated with red circles in Figure 5.22 (B) below.

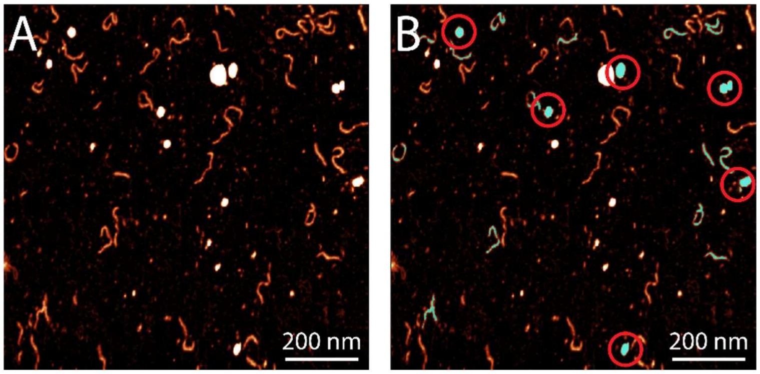


Figure 5.22 Unmasked (A) and masked (B) AFM images of damaged DNA without HU treatment, where many impurities are mistakenly selected as molecules of interest (red circles). Since experimental conditions differ from image to image, it is difficult to find a set of masking parameters that exclude noise and impurities from all data.

Since these impurities and noise cannot be completely removed by the molecular identification process, it is necessary to devise another method that excludes them from the bending angle measurement, so that the results can accurately represent the conformation of DNA molecules.

5.3.1.3.2 Excluding DNA clusters

The concentration of DNA used in the AFM imaging is higher than ideal, and as a result, molecules are frequently found next to each other forming clusters, as demonstrated by the example image in Figure 5.23 below.

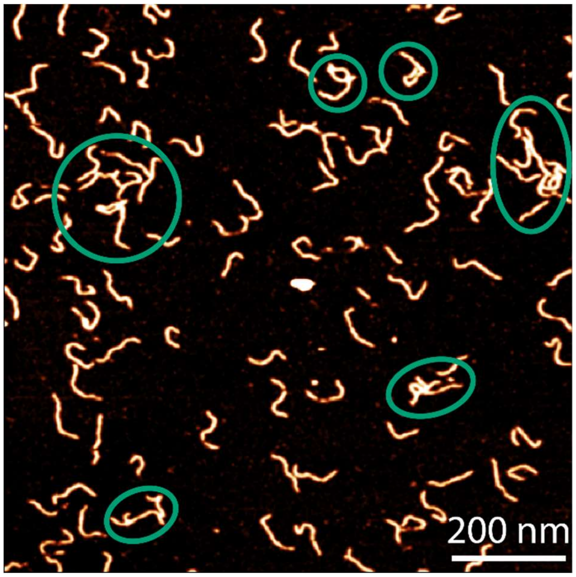


Figure 5.23 Example AFM image of undamaged DNA molecules, with clusters circled in green.

These clusters cannot be traced accurately and are unsuitable for bending angle measurement. Compared to the low-concentration PLO dataset for NDP52-treated DNA, their presence is more common on images from the HU project; it is therefore more difficult to exclude all of them at the molecular identification stage, and similar to the case of noise and impurities, an additional step is needed to remove them with more certainty from the bending angle measurement.

5.3.1.3.3 Distinguishing between natural bending and HU-induced bending

It has been discussed earlier that DNA molecules often contain bending at the central damage site even when HU interaction has not taken place. Since HU cannot be directly observed, it is not feasible to manually or automatically determine whether a bend at the centre of a molecule is induced by HU.

However, it is possible to reduce the impact of confounding factors by excluding bends located far from the centre: as circled in purple in Figure 5.24 below, these bends occur naturally and frequently but are unrelated to HU-DNA interaction, and are therefore not the focus of this project; consequently, the bending angle measurement algorithm needs to be designed so that only bends within the central portion of the molecule are considered.

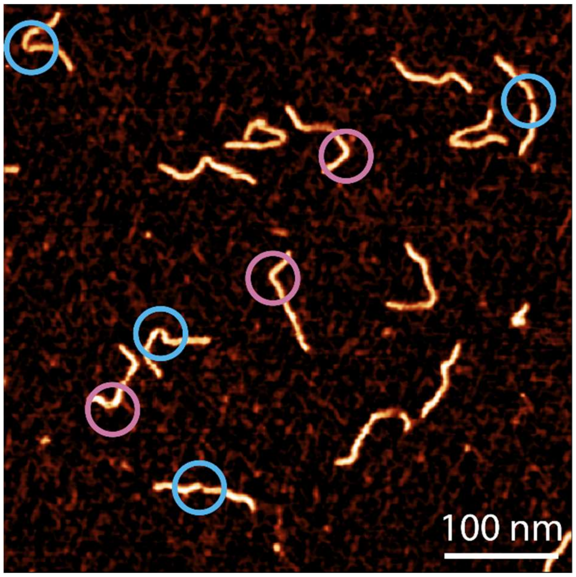


Figure 5.24 Example AFM image of damaged DNA molecules without HU treatment, where bends occur both around (blue circles) and away from (purple circles) the central damage site. The latter should be excluded from the bending angle measurement because they are unrelated to HU interaction.

5.3.2 Developing bending angle measurement methods

5.3.2.1 Filtering out unwanted molecules

To exclude impurities, noise and clusters from the bending angle measurement, identified molecules are filtered after the tracing process based on their linearity and contour length. As the DNA used in this project is linear, molecules categorised as circular (see Section 4.1.2.3) are discarded because they are likely impurities; in addition, circular molecules are unsuitable for the measurement even if they are legitimate DNA formed by linear molecules joining their end points, because it is not possible to determine whether their bends are located at the central damage site.

The contour-length-based filtering, on the other hand, is used to remove impurities and broken molecules for being too short, or clusters for being too long. The filtering thresholds are decided based on the size of the molecules used in this project, which contain 303 or 305 base pairs depending on whether they are damaged, and this corresponds to a length of 103 to 104 nm. However, data from the NDP52 project (see Figure 4.43 from Section 4.5.1.3.1) showed that tracing problems can cause the measured contour length to be lower than the theoretical value by 17%. To establish whether this is the case for the HU datasets, we compared the contour length distributions of undamaged DNA, damaged but untreated DNA, and damaged DNA that had undergone HU co-incubation.

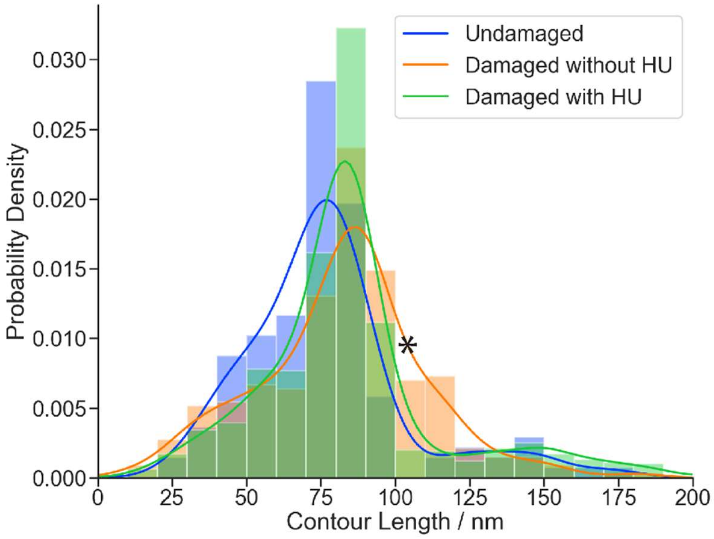


Figure 5.25 KDE plots and normalised histograms comparing the contour length distributions of undamaged DNA (blue, main peak and standard deviation 77 ± 27 nm, N = 137, number of images = 9), damaged DNA without HU treatment (orange, 86 ± 28 nm, N = 329, number of images = 23) and damaged DNA with HU treatment (green, 83 ± 32 nm, N = 990, number of images = 26). For comparison, the theoretical contour lengths for undamaged and damaged DNA molecules are 103 nm and 104 nm (black asterisk) respectively.

As shown in Figure 5.25 above, the measured contour lengths for the three datasets were 77 ± 27 nm, 86 ± 28 nm and 83 ± 32 nm, and they were lower than the values calculated from the number of base pairs by 25%, 17% and 19%. These results are in line with what was observed from the NDP52 project.

Taking the shortening effect of the tracing process into account, only molecules whose contour length is within one standard deviation from the peak value are selected for the bending angle measurement, so that impurities, broken molecules and clusters can be filtered out. This is done separately for each of the three datasets.

5.3.2.2 *Identifying the bending point*

Unlike the curvature measurement which takes place at every point of the DNA, the bending angle measurement only needs to be performed once per molecule, and this means that the first step of the measurement is to locate the bending site.

While the point with the highest absolute curvature value corresponds to the sharpest bend on a given molecule, this may be unrelated to the damage or HU interaction, as discussed in Section 5.3.1.3.3. To ensure that only bends around the damage are considered, the point of highest absolute curvature within the central 20 nm portion of a molecule is used as the site of bending angle measurement. This value is provisionally decided for testing purposes and can be subsequently adjusted if necessary.

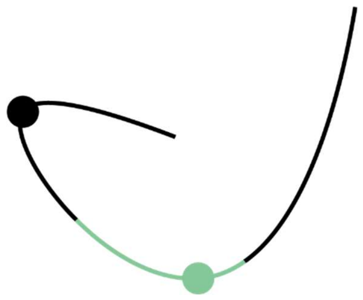


Figure 5.26 Identifying the site of bending angle measurement. The point of highest absolute curvature (green dot) within the central 20 nm portion (green curve) is chosen as the site of measurement, while any sharper bend elsewhere on the molecule (black dot) is not used because it is unrelated to the interaction between HU and the damage.

5.3.2.3 Determining the two lines that form the angle

Points around the bending site form a smooth, continuous curve as part of the splined trace, but they cannot be directly used for the bending angle measurement without transformation into two straight lines. To achieve this, the pre-existing Python function 'scipy.stats.linregress' takes the coordinates of points from two 10 nm curve segments on the left and right sides of the bending site, and calculates the gradients of two best-fit lines correspondingly, as illustrated in Figure 5.27 below. Similar to the previous step, the length of 10 nm is chosen provisionally and can be adjusted in the future based on the accuracy of the measurement.

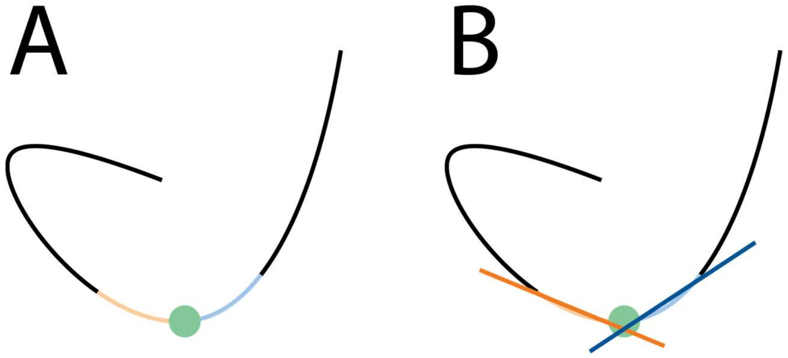


Figure 5.27 Transforming points into lines that form the bending angle. A: The measurement site (green dot) is identified on the splined trace, and points from 10 nm segments on its left (light orange curve) and right (light blue curve) sides are selected. Note that these points are represented by curves because the splined trace is smooth and continuous. B: Linear regression is performed on these two series of points, so that the lines (dark orange and dark blue) forming the bending angle can be defined.

5.3.2.4 Transforming lines into vectors and calculating the angle

The crossing of two lines leads to a pair of supplementary angles, but to ensure consistency and enable comparison between different molecules, the angle inside the bend (α in Figure 5.28 below) is always measured instead of the angle outside the bend (β in Figure 5.28). For example, a straight DNA molecule with no bends would result in a measurement of 180°, while a molecule folded in half would correspond to a bending angle approaching 0°.

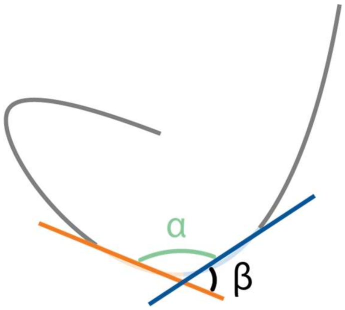


Figure 5.28 Distinguishing between the angles inside (α) and outside (β) the bend. When lines cross, two supplementary angles are formed, and the automatic measurement algorithm developed in this project chooses the angle inside the bend. To identify the correct angle, the crossing lines need to be represented by vectors.

Since the inside and the outside angles need to be distinguished from each other, the gradient alone is insufficient to define a line, and the directionality should also be taken into account. The lines produced through linear regression are therefore transformed into vectors that point outwards from the bending site. As illustrated below in Figure 5.29, for each vector, the X component is first determined using the X coordinates of the bending point and the other end point of the 10 nm segment; the Y component is subsequently calculated by multiplying the X component with the gradient.

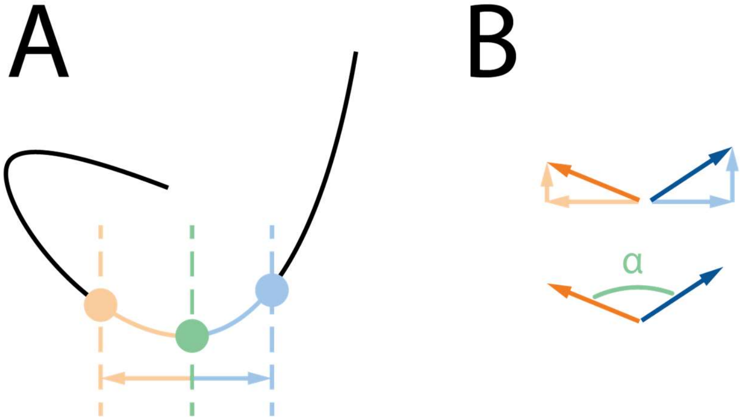


Figure 5.29 Determining the directionality of vectors that form the bending angle. A: The X component of each vector is calculated using the X coordinates of the 10 nm curve segment end points; its direction is defined as going from the bending site (green dot) to the other end point of the curve segment (light orange or light blue dot). B: The Y component is calculated by multiplying the X component with the gradient; the X and Y components are then added, and the directionality of the sum vectors (dark orange and dark blue arrows) can be used to identify the angle inside the bend (α).

Once the vectorisation is complete, the bending angle can be obtained using the formula

$$\alpha = \cos^{-1}\frac{\boldsymbol{V_a}\cdot\boldsymbol{V_b}}{\|\boldsymbol{V_a}\|\|\boldsymbol{V_b}\|} = \cos^{-1}\frac{x_a\cdot x_b + y_a\cdot y_b}{\sqrt{x_a^2+y_a^2}\cdot\sqrt{x_b^2+y_b^2}}$$

where $\boldsymbol{V_a}$ and $\boldsymbol{V_b}$ refer to the two vectors forming the angle, and x_a, x_b, y_a, y_b refer to the magnitudes of their X and Y components. The calculated angle is converted from radians to degrees at the end.

5.3.3 Identifying and addressing issues with bending angle measurement

5.3.3.1 Changing the site of measurement

As described in Section 5.3.2.2, the point of highest absolute curvature within the central 20 nm of a molecule was chosen as the site for bending angle measurement. However, it was revealed after testing on different molecules that this method could lead to unintended consequences: when a molecule did not contain any bend within its central portion, the absolute curvature typically increased slightly from one end of the 20 nm curve segment to the other; as a result, the point of highest curvature fell on an end of the segment, and the bending angle measurement was made far from the damage site.

To address this issue and ensure that the bending angle is relevant to HU interaction, the site of measurement was adjusted to the midpoint of the molecule. While this point does not always correspond to a bend, it indicates the location of the damage and of any potential HU interaction. Measurements performed at the same site for every molecule will also facilitate comparison between different datasets: for example, the effects of the damage can be characterised by comparing the bending angle distributions between undamaged and damaged molecules.

5.3.3.2 Fixing the issue with supplementary angles

5.3.3.2.1 Identifying the cause of erroneous measurements

During further testing of the automatic bending angle measurement, rare cases emerged where the measured value was the supplementary angle of the actual value. This error only occurred when at least one of the lines forming the angle was near-vertical, suggesting that the directionality may have been incorrectly determined for near-vertical lines. The cause of this was identified through a step-by-step examination of the vectorisation process: the difference in X coordinates between the site of measurement and the other end of the 10 nm segment was used to calculate the X component of the vector and assign its directionality; as illustrated below, this method worked reliably in the majority of cases, but when a near-vertical line was involved, the two end points of the curve segment had similar X coordinates, and the X component of the vector could point towards (rather than away from) the bending site, which resulted in the angle outside the bend being calculated.

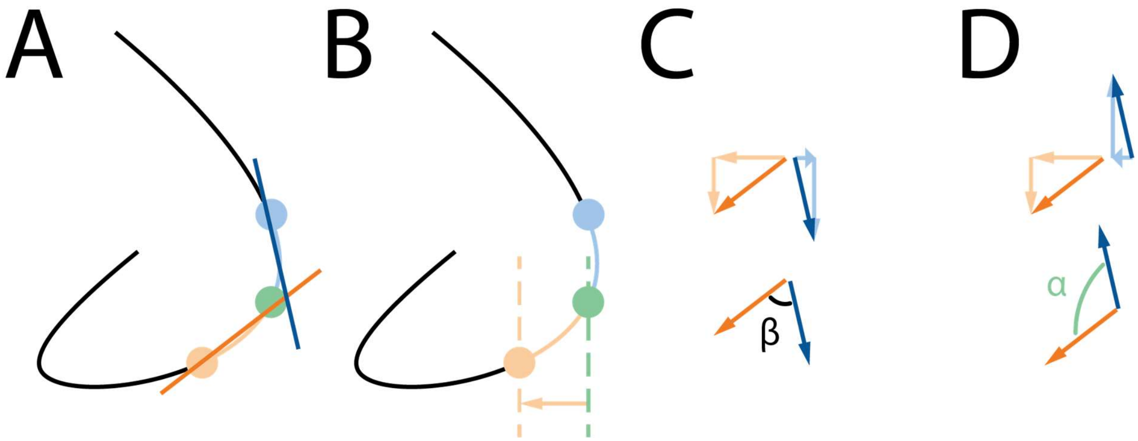


Figure 5.30 A near-vertical line can cause the wrong angle to be measured. A: The bending angle for an example molecule is formed by a regular line (dark orange) and a near-vertical line (dark blue). B: The X component of the vector corresponding to the regular line is easily determined; however, for the near-vertical line, the X coordinates of the bending site (green dot) and of the other end point (blue dot) are similar, so the X component of the vector can point towards either direction. C: If the X component of the near-vertical vector incorrectly points towards the bending site, the sum vector (dark blue arrow) will also be in the wrong direction, and the angle outside the bend (β) will be measured. D: If the X component of the near-vertical vector correctly points away from the bending site, the angle inside the bend (α) will be measured.

5.3.3.2.2 Determining directionality with both X and Y coordinates

It was first proposed to exclude angles involving at least one near-vertical line from the measurement, so that the error described above could be eliminated. This solution was not adopted as it would lead to the removal of potentially usable data. In addition, while near-vertical lines can be identified based on their high gradients, it would not be ideal to arbitrarily implement a threshold gradient value.

To enable the accurate measurement of angles involving near-vertical lines, the vectorisation process as described in Section 5.3.2.4 needs to be modified. Instead of determining the directionality of a segment solely based on the X coordinates of its two ends, the Y coordinates are also taken into consideration when necessary.

In the revised method of vectorisation, the difference in X coordinates between the two ends of a segment is only used for lines closer to the X axis than to the Y axis; these lines can be easily identified as they exhibit a gradient between -1 and 1. Conversely, the difference in Y coordinates between two ends of a segment is used if the gradient is higher than 1 or lower than -1, as this indicates that the line is closer to the Y axis. In this way, the directionality of both near-vertical and near-horizontal lines can be determined reliably.

5.3.4 Results

Automatic bending angle measurement methods suited to the needs of the HU project have been successfully developed. Issues involving the choice of the measurement site and the vectorisation of lines have been identified and addressed, and the extent of bending at the centre of linear molecules can now be reliably quantified. Due to time constraints, the length of the 10 nm curve segments on each side of the bending site (see Section 5.3.2.3) has not been modified and optimised, and this should be revisited in future projects.

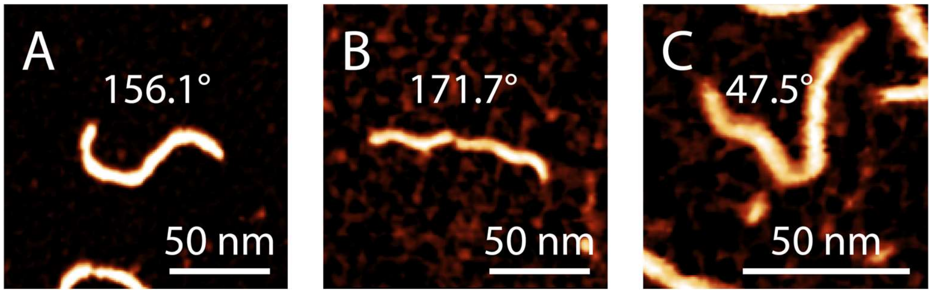


Figure 5.31 Example bending angle measurements for an undamaged DNA molecule (A), a damaged molecule without HU treatment (B) and a damaged molecule that has gone through HU co-incubation (C).

5.4 Characterising the effects of HU on DNA bending angle

5.4.1 Characterising the effects of damage on DNA bending angle

5.4.1.1 Measuring the bending angle of undamaged DNA

Manual observation on AFM images of damaged DNA molecules has shown that bending can occur around the damage site without HU interaction. Because of this, the effects of damage on the bending angle need to be first established before investigating changes induced by HU. To characterise the natural conformation at the centre of DNA molecules and obtain a baseline measurement for reference, the newly developed automatic bending angle measurement methods were first applied to the dataset of undamaged DNA. After filtering out molecules that are circular or of unsuitable contour length, the bending angle distribution for 103 molecules was obtained, as shown below in Figure 5.32.

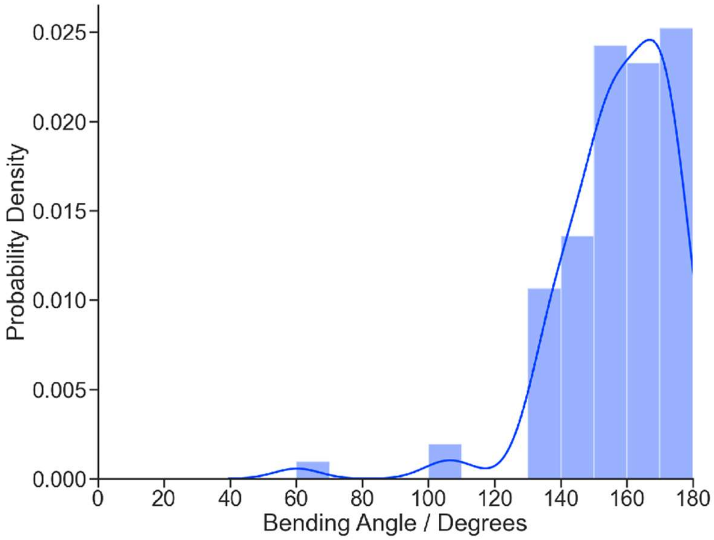


Figure 5.32 KDE plot and normalised histogram showing the bending angle distribution for undamaged DNA (peak and standard deviation 167° ± 17°, N = 103, number of images = 9).

In accordance with expectations from initial observation, undamaged DNA molecules are mostly flat at the midpoint, with bending angles ranging from 130° to 180° for the majority of the dataset. This suggests that an increase in the proportion of molecules with a measurement lower than 130° could be used as an indication of conformational changes. To ensure that the measured angles are representative of DNA shape, manual inspection was performed with special attention paid to molecules with lower bending angle values, and no obvious errors in tracing or bending angle measurement were identified.

5.4.1.2 Measuring the bending angle of damaged DNA without HU treatment

To characterise the effects of the damage on DNA conformation, bending angle measurement was performed for 225 damaged molecules without HU co-incubation that met the inclusion criteria, and the resulting distribution was compared with that of undamaged DNA.

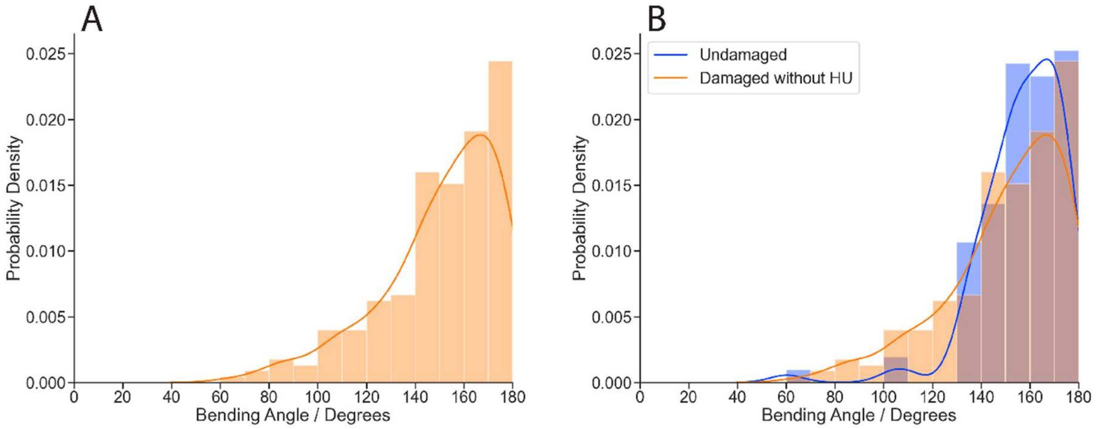


Figure 5.33 KDE plots and normalised histograms comparing the bending angle distributions of undamaged DNA and damaged DNA without HU treatment. A: The bending angle distribution of damaged DNA without HU treatment (peak and standard deviation 167° ± 24°, N = 225, number of images = 23). B: Superimposed bending angle distributions of damaged DNA without HU treatment (orange) and undamaged DNA (blue, 167° ± 17°, N = 103, number of images = 9, also shown in Figure 5.32).

As shown in Figure 5.33 above, the bending angle distribution is wider for damaged DNA. 18.7% of damaged DNA molecules exhibit a bending angle lower than 130°, compared to only 2.9% for undamaged DNA. In addition, while manual inspection of damaged DNA molecules without HU treatment revealed tracing inaccuracies caused by complex shapes, this did not result in erroneous bending angle values, because most of the incorrectly traced molecules had been excluded from bending angle measurement, and the rest contained tracing errors that were located far from the site of measurement.

Statistical analysis was carried out to quantitatively determine the extent of correlation between lower bending angles and the presence of damage, and the probabilities of the damage being present for DNA molecules falling under different bending angle thresholds were calculated using the Bayesian theorem

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

where A refers to the event of a DNA molecule containing the damage, and P(A) is equal to the proportion of damaged molecules, 225 / (225 + 103) = 0.686; B refers to the event of the bending angle being lower than the given threshold, and P(B) is equal to the number of both damaged and undamaged DNA molecules with a bending angle below the threshold, divided by the total number of molecules, 328. P(B|A) is the probability of a molecule exhibiting a bending angle that falls under the threshold, provided that the molecule is damaged, and it is equal to the number of damaged molecules under the bending angle threshold, divided by the number of all damaged molecules. P(A|B) is the probability of a molecule containing the damage, provided that its bending angle falls under the given threshold. The results of the calculation are summarised in Table 5.1 below:

Table 5.1 Probabilities of damage being present for DNA molecules under different bending angle thresholds

Bending angle range	Number of undamaged molecules (N_u)	Number of damaged molecules (N_d)	P(A) (= 225 / (225 + 103))	P(B) (= (N_u + N_d) / (225 + 103))	P(B|A) (= N_d / 225)	P(A|B)
< 100°	1	10	0.686	0.034	0.044	0.909
< 110°	3	19		0.067	0.084	0.864
< 120°	3	28		0.095	0.124	0.903
< 130°	3	42		0.137	0.187	0.933

According to the Bayesian analysis, if a DNA molecule has a bending angle below 100°, 110°, 120° or 130°, the probabilities that it has been damaged (P(A|B)) are 0.909, 0.864, 0.903 and 0.933 respectively. It should be noted that due to the difference in sample size between undamaged and damaged DNA molecules, these probabilities need to be interpreted in comparison with P(A). With that taken into account, since all four P(A|B) values are larger than 0.686, the results still provide further evidence that the damage at the centre of the DNA molecule is associated with an increased occurrence of bending angles lower than 130°. In conclusion, it can be postulated that the damage increased the flexibility of DNA, allowing it to form sharper bends more frequently compared to undamaged DNA.

5.4.2 Characterising the effects of the HU protein on DNA bending angle

5.4.2.1 Measuring the bending angle of damaged DNA with HU treatment

Since it has been established that the damage is associated with the formation of sharp bends on a small subset of molecules, it is now possible to characterise the effects of HU on DNA bending angle. To achieve this, 742 linear molecules of appropriate contour lengths were automatically selected from the dataset of damaged DNA with HU co-incubation. Their bending angles were measured and compared with the results for damaged DNA without HU treatment.

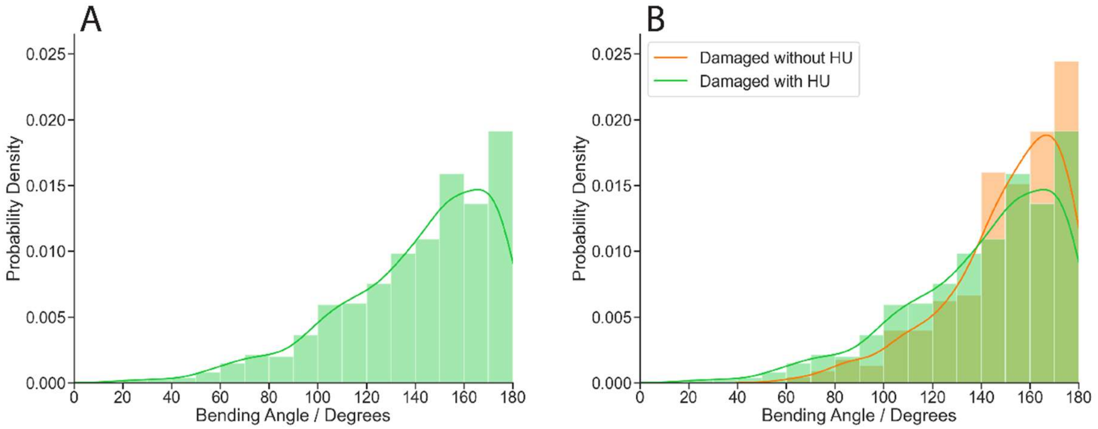


Figure 5.34 KDE plots and normalised histograms comparing the bending angle distributions of damaged DNA with and without HU treatment. A: The bending angle distribution of damaged DNA with HU treatment (peak and standard deviation 166° ± 31°, N = 742, number of images = 25). Note that the number of images is slightly lower compared to the distribution of contour length measured from the same dataset (N = 990, number of images = 26, see the 'damaged DNA with HU treatment' part of Figure 5.25 in Section 5.3.2.1); this is because one image contains no molecules that meet the selection criteria for bending angle measurement, as detailed in Section 5.3.2.1. B: Superimposed bending angle distributions of damaged DNA with (green) and without (orange, 167° ± 24°, N = 225, number of images = 23, also shown in Figure 5.33 (A)) HU treatment.

Figure 5.34 above shows that the bending angle distribution for the HU-treated dataset is wider than for damaged but untreated DNA. It can also be observed that HU co-incubation increased the proportion of molecules containing sharper bends, with 227 out of 742 (31%) having a bending angle lower than 130°. In addition, the measurements for 13 molecules (1.8%) are lower than 60°, while all measured angles were greater than 60° for DNA without HU treatment. Manual inspection of the tracing accuracy for HU-treated DNA molecules, in particular those with low bending angle values, also confirmed that the distribution reflects DNA conformation and is not caused by tracing errors.

Bayesian analysis as described in Section 5.4.1.2 above was performed again to quantitatively determine the extent of correlation between HU treatment and lower bending angles, and the results are shown below in Table 5.2. Here A refers to the event of a DNA molecule belonging to the HU-treated dataset, and P(A) is equal to the number of molecules in this dataset divided by the total number of molecules, 742 / (742 + 225) = 0.767. B refers to the event of a molecule exhibiting a bending angle lower than a given threshold, and P(B) is calculated by dividing the number of DNA molecules with a bending angle below the threshold over the total number of molecules, 967. P(B|A) is the probability of a molecule belonging to the HU-treated dataset, provided that it exhibits a bending angle below the threshold; it is equal to the number of HU-treated DNA molecules under the threshold, divided by the total number of molecules in the HU-treated dataset, 742. P(A|B) is the probability of a molecule belonging to the HU-treated dataset, provided that its bending angle falls under the given threshold, and it is calculated using the Bayesian theorem.

Table 5.2 Probabilities of HU being present for DNA molecules under different bending angle thresholds

Bending angle range	Number of molecules without HU (N_{wo})	Number of molecules with HU (N_w)	P(A) (= 742 / (742 + 225))	P(B) (= (N_{wo} + N_w) / (742 + 225))	P(B|A) (= N_w / 742)	P(A|B)
< 60°	0	13	0.767	0.013	0.018	1.000
< 70°	1	24		0.026	0.032	0.960
< 80°	3	40		0.044	0.054	0.930
< 90°	7	55		0.064	0.074	0.887
< 100°	10	82		0.095	0.111	0.891
< 110°	19	126		0.150	0.170	0.869
< 120°	28	171		0.206	0.230	0.859
< 130°	42	227		0.278	0.306	0.844

The calculation shows that for a DNA molecule exhibiting a bending angle lower than 60°, 70°, 80°, 90°, 100°, 110°, 120° or 130°, the probabilities that it belongs to the HU-treated dataset (P(A|B)) are 1.000, 0.960, 0.930, 0.887, 0.891, 0.869, 0.859, and 0.844 respectively. Despite the difference in sample size between DNA with and without HU treatment, the high values of P(A|B) in relation to P(A) suggest that HU treatment is associated with sharper and more frequent bends, and that the conformational changes observed in the HU-treated dataset cannot be attributed to the damage alone. However, comparison with data from other sources is needed to further analyse the nature of sharp bends and the mechanism of HU-DNA interaction.

5.4.2.2 Comparison with MD simulation data

5.4.2.2.1 Description of MD simulation data

As discussed earlier in Section 5.3.1.1, MD simulations conducted by our collaborators Elliot W. Chan and Agnes Noy suggested that HU interaction caused DNA to bend at specific angles. To investigate whether this is congruent with experimental data, the distribution of simulated bending angle measurements was examined.

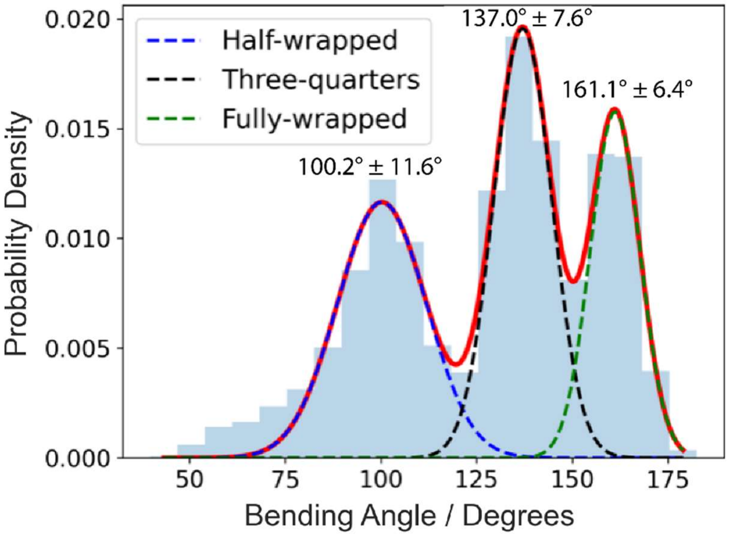


Figure 5.35 Bending angle distribution obtained through MD simulation of HU-DNA interaction. Three distinct peaks can be observed around 100.2°, 137.0° and 161.1°. Note that since this measurement uses the angle outside the bend, the peak values are equivalent to 79.8°, 43.0° and 18.9° for the inside angle used in the automatic characterisation of AFM images. Figure provided by Elliot W. Chan, who collected and analysed the MD data.

Our collaborators identified three distinct types of HU-DNA interaction, referred to as half-wrapped, three-quarters and fully-wrapped; as presented above in Figure 5.35, they correspond to peaks at 100.2° ± 11.6°, 137.0° ± 7.6° and 161.1° ± 6.4°. Since the angle outside the bend was calculated for the MD simulation data (see Section 5.3.2.4 and Figure 5.28), these values are equivalent to 79.8°, 43.0° and 18.9° for the angle inside the bend.

By comparison, among the 742 measured molecules from the HU-treated AFM dataset, only 40 or 5.4% exhibit a bending angle below 80°. There is therefore little overlap between simulated data and the experimental data, with sharp bends lower than 80° being the primary component of the former but a rare occurrence for the latter. This drastic difference calls for further analysis and explanations, which will be detailed in the following section.

5.4.2.2.2 Possible sources of the discrepancy

5.4.2.2.2.1 Proportion of DNA affected by HU

The discrepancy between MD and AFM data can be explained by several potential factors. To start with, in the MD data, all DNA molecules were simulated to have interacted with HU; by contrast, the HU may have only bound to a small proportion of DNA molecules on the AFM images. While the automatic measurements show that HU has an overall effect on the bending angle distribution, the exact percentage of DNA molecules that have been bent by HU is unknown, especially since HU is too small and cannot be directly seen when imaged together with DNA.

5.4.2.2.2.2 Data quality and DNA concentration

As discussed in Section 5.3.1.3, the quality of some AFM images is affected by impurities and high DNA concentration. Although the automatic bending angle measurement algorithm has been designed to filter out unusable data such as DNA clusters, this could introduce a selection bias. For example, if clusters are more likely to interact with HU, DNA molecules that have been bent by the protein will be underrepresented in the bending angle measurement since clusters are disregarded. Future experiments can be performed with a lower DNA concentration and a higher HU concentration to increase the possibility of HU-DNA interaction and ensure that all DNA molecules affected by HU are included in the measurement.

5.4.2.2.2.3 Effects of PLL

All DNA molecules used in the HU project were immobilised with PLL, which is structurally and functionally similar to PLO [135], and can therefore affect DNA conformation. In particular, PLL is known to align with the lattice structure of the mica substrate, with its chains forming angles of 60° or 120° [133]. This means that the orientation and bending angle of DNA are also influenced by the mica lattice structure through the layer of PLL; consequently, certain bends measured to be around 60° and 120° may be unrelated to HU interaction. To prevent the effects of PLL from interfering with bending angle measurement, the AFM imaging needs to be re-conducted using non-polymer immobilisation agents.

5.4.3 Conclusions and future work for the HU project

By comparing the automatically measured bending angles for undamaged DNA, damaged DNA without HU treatment, and damaged DNA that has gone through HU co-incubation, we are able to confirm that both the damage and HU interaction cause DNA molecules to exhibit sharper and more frequent bending. However, the reliability of the measurement is affected by experimental factors, including the concentration of DNA and the choice of immobilising agents. Because of this, concrete conclusions cannot be drawn from the lack of agreement between AFM and MD data. In the future, AFM imaging with non-polymer immobilisation is required to facilitate a valid comparison between experimental and simulated data. In addition, other characterisation techniques can be used in conjunction with AFM to quantify the proportion of DNA molecules interacting with HU. For example, electrophoretic mobility shift assay (EMSA) [175] can be employed to detect HU-DNA complexes and measure the prevalence of HU-DNA interaction, since these complexes travel through electrophoresis gel at a slower speed than untreated DNA. By performing EMSA under different HU concentrations, the ideal HU concentration to facilitate HU-DNA interaction can also be determined and applied to AFM imaging.

5.5 Conclusions

This chapter has discussed the successful development of new methods to characterise conformational changes in DNA through the automatic measurement of curvature and bending angle. The curvature measurement was assessed to be accurate for generated regular shapes, but it cannot yet be employed to quantify the effects of NDP52 on DNA due to tracing difficulties as discussed in the previous chapter. This further highlights the importance and urgency of improving the underlying tracing algorithm: the current workflow disregards height data except during the fitting process, and it would be ideal to change this, so that proteins or impurities are distinguished based on height difference and no longer treated as part of the DNA molecule. In addition, when determining the trajectory of DNA by identifying the next nearest neighbour of a given point, the 3D distance can be calculated instead of the 2D distance.

The limitations related to tracing were circumvented to some extent in the bending angle measurement, which filtered out unsuitable molecules and focused on the centre of the DNA. By measuring the bending angle for the HU datasets, this project demonstrated that damage to the double-helical structure and interaction with HU both led to an increased proportion of DNA molecules exhibiting sharp bends. An attempted comparison between results from AFM imaging and MD simulation was however shown to be not meaningful at this stage: due to experimental factors such as molecular concentration, impurities and immobilising agents, DNA molecules attached to the mica surface were under very different conditions from those in the simulation. To reduce the impact of these factors in the future, it is recommended that experimental work and image analysis take place concurrently rather than consecutively, because this will allow potential adjustments to the experimental process to be made at the earliest opportunity.

Chapter 6 Conclusions and future work

6.1 Conclusions on molecular identification

Through the adjustment of height and area thresholds, an existing molecular identification and quantification method, implemented in the open-source software TopoStats [49][61], has been adapted for a wide range of biomolecules, including DNA and proteins of different shapes. When an AFM image features multiple types of molecules, a method has been developed to selectively recognise one of them based on differences in height and area. By applying this method to experimental data, appropriate thresholding parameters have been determined to allow for the identification of single DNA molecules, DNA clusters, globular TOP1 proteins, dumbbell-shaped NDP52 proteins, NDP52 oligomers, and sub-molecular features of NDP52. In addition, existing thresholding methods and their compatibility with different datasets have been analysed. This ensures that the most suitable methods can be chosen for the molecular identification process in future projects based on the nature of the sample.

Compared to other published molecular identification methods developed for AFM datasets, the approach employed in this thesis enables a higher degree of customisation, and an easier adaptation for molecules with different shapes and sizes; in particular, it is compatible with images where multiple molecule types are present but only one type needs to be recognised.

A major challenge in automatic molecular identification lies in the variability of the dataset. For example, a single image may contain multiple protein species, DNA, aggregates, and impurities of different sizes. As a result, the height and area measurements of automatically identified features tend to form a continuum rather than distinct groups. This means that suitable height and area thresholds cannot be automatically identified through parameter sweeps, where thresholding parameters are incrementally adjusted and the resulting number of molecules is counted, with a stable number indicating a suitable range of parameters corresponding to a particular species. Given the variability of data in AFM images, and the emerging nature of this area in the field of bio-image analysis, manual assessment remains essential to the determination of molecular identification thresholds for AFM images used in this thesis.

The work in this thesis has also demonstrated that the successful identification of molecules of interest depends on, and can be limited by, the quality of the experimental data. The most common issues that present challenges to the automatic identification process are impurities and the clustering of molecules. While these issues can be partially addressed via the selection of suitable parameters, they also need to be taken into account and minimised through better experimental design.

It is possible to indirectly evaluate data quality by comparing the results of manual and automatic molecular identification. The extent of overlap between manually and automatically recognised molecules can be quantified via the Jaccard index, where a relatively high value of 0.8 or above is a good indication of sufficient data quality allowing for accurate automatic identification, while a value of around 0.5 or lower suggests that the image is of poor quality and may not be suitable for automatic analysis.

6.2 Future work for molecular identification

The current molecular identification methods do not allow for the removal of unwanted features that are too high or exhibit undesirable shapes. In addition, when performing molecular identification on a new dataset, a time-consuming trial and error process is required from the user to determine the most suitable height and area thresholds.

To increase the accuracy of the automatic identification, an additional height-based selection step can be introduced in the future, so that impurities or artefacts can be filtered out for being too high. Machine learning algorithms can also be incorporated into the molecular identification toolkit as an optional module to classify and select molecules based on shape.

It is however unrealistic to completely eliminate the need for user input in the near future, due to the variety of samples imaged in different projects. Detailed documentations and example thresholding parameters for common types of molecules can be established instead to simplify the user experience and improve the accessibility to automatic analysis methods.

These documentations should also be complemented by a guideline for experimentalists, which will contain requirements on molecular concentration and image resolution to ensure that features of interest can be accurately identified. With the recent development of the software BioAFMviewer [176], [177], which generates AFM images of proteins from crystal structures, the quality and reliability of experimental data can also be assessed through comparison with simulated data. This will allow samples that are particularly challenging for AFM imaging to be more easily recognised, so that experimental design can be improved at the earliest opportunity.

6.3 Conclusions on tracing and related measurements

As a result of the work in this thesis, the existing tracing algorithm has been improved, and it is now compatible with both linear and circular molecules. Tracing-based characterisation methods have been developed, including the automatic measurement of end-to-end distance, curvature and bending angle. A workflow has also been set up to assess the accuracy and suitability of the automatic measurement methods through comparison with manual measurements and preliminary testing on a small sample size. Compared to previously developed software with similar functionalities, the tracing workflow and related characterisation methods in this thesis present the following advantages.

Firstly, they are integrated into a single platform, where all parameters are measured within a single image analysis pipeline that can be adapted for different projects depending on the nature of the sample. Secondly, the need for manual input has been reduced: while there are configurable parameters that can be adjusted by the user, these parameters are used for the entire dataset, whereas many previous methods require the manual identification of the starting point for each individual molecule. Thirdly, the testing and application of tracing-related characterisation methods are extensively conducted on real images instead of relying primarily on simulated data; this allows us to identify problems most pertinent to real experimental data, such as the interference from immobilising agents.

The main issues with the tracing process have been identified and potential approaches to address them have been suggested, which are expanded on below.

6.4 Future work for tracing and related measurements

The current tracing algorithm is prone to errors when DNA molecules exhibit crossings or are attached to proteins. This issue needs to be addressed in the future by taking the height value of each point into account during the skeletonisation process. If an improved tracing workflow makes full use of the 3D data obtained through AFM imaging, the trajectory of DNA or other linear molecules will be more accurately followed. Points with lower height values will be prioritised for removal during skeletonisation, which will allow the skeleton to be more representative of the DNA backbone. Furthermore, when arranging points from the skeleton into order by continuously selecting the nearest neighbour as the next point on the list, the 3D distance between two points will be calculated instead of the 2D distance currently used. These changes will constitute a major upgrade compared to existing methods that transform images into binary data consisting solely of foreground and background, because proteins or impurities attached to the DNA will be distinguishable based on height difference. In addition, the tracing of complex features such as knots will also benefit from the utilisation of height information, and the trajectory of the molecule at crossings can be identified by ensuring continuity in height values. On the other hand, if accurate tracing remains challenging to achieve in the future, an algorithm can be developed to automatically evaluate tracing quality by calculating the extent of overlap between the trace and the mask. This will allow the suitability of tracing-based measurements for a given dataset to be easily determined.

Another significant limitation of the tracing process is the shortening of linear DNA molecules as pixels at both ends are erroneously deleted. Height information may also be able to prevent this from occurring, since the ends that need to be retained are higher than the periphery of the molecule that needs to be removed during skeletonisation, though further testing will be necessary to confirm the viability of this method. Alternatively, the ends that have been cut off can be re-introduced to the trace through comparison with the original molecule prior to skeletonisation.

In addition, the tracing quality is currently affected by the asymmetric nature of the skeletonisation process, where the removal of peripheral points does not take place simultaneously in all directions, but instead starts from a corner of the molecule. Consequently, if the molecule is flipped horizontally or vertically, its skeleton retains the same general shape but the specific positions of most points are changed. This can be quantified by a relatively low Jaccard index of 0.5 to 0.6 between skeletons obtained from the original and flipped versions of the same molecule. Utilising height information during skeletonisation and prioritising the removal of points with lower height values may resolve this issue, while the Jaccard index can be used to assess the reliability of new skeletonisation and tracing methods developed in the future.

6.5 Conclusions on the characterisation of DNA-protein interaction

AFM imaging and automatic analysis have been combined to provide more information on the structure of biomolecules and the mechanism of DNA-protein interaction. In particular, methods developed as part of the work in this thesis have increased our understanding on TOP1, NDP52 and HU proteins.

The structure of TOP1 has been directly visualised, and its height, size and volume have been automatically measured. The results are found to be broadly consistent with literature values, and the relevant characterisation methods can be applied to other globular proteins without visible sub-molecular structures. However, the automatic measurement of area and height for TOP1 has shown that this protein undergoes conformational changes and becomes flatter upon adsorption onto the mica surface. The same phenomenon can also be observed for NDP52, implying that it is a more general process, linked to the immobilisation of biomolecules on a surface as part of AFM sample preparation.

The predicted dumbbell-shaped structure of NDP52 has been visualised and confirmed for the first time. Automatic image analysis has been employed to quantify its size distribution and the proportion of NDP52 molecules involved in oligomerisation. The three main effects of NDP52 on DNA conformation, bending, looping and bridging, as identified during manual observation, have been characterised via the automatic measurement of maximum bounding distance and the automatic identification of bridged DNA molecules.

To uncover the effects of the protein HU on DNA, the bending angle distribution has been automatically measured and compared for undamaged DNA, damaged DNA without HU treatment, and damaged DNA that has been co-incubated with HU. The results suggest that both the damage and the HU treatment can increase the extent and frequency of the bending, though this needs to be verified in the future by eliminating confounding factors including the effects of the immobilising agent PLL.

The choice of immobilising technique has been shown to affect the conformation of DNA. Compared to metallic ions such as Ni^{2+} and Mg^{2+}, polymer-based immobilising agents tend to have a stronger effect on the bending of DNA, and are also better at preserving conformational changes on DNA induced by other molecules. This information needs to be taken into consideration during both the experimental and the analytical stages for future work on the characterisation of DNA-protein interaction through AFM imaging.

When investigating the effects of NDP52 or HU on DNA conformation, it has been noticed that not all DNA molecules interact with the protein of interest. Because of this, it is difficult to separate the effects of the protein from the effects of other factors including the immobilising agent. Bayesian analysis has been applied to address this issue and quantify the extent of correlation between protein treatment and specific DNA conformations. This approach allows infrequent but substantial conformational changes to be more systematically characterised, and should continue to be adopted in future projects to determine whether such changes are attributable to the presence of a protein or other experimental conditions.

6.6 Future outlook of automatic AFM image analysis

As stated in Section 1.2.2, this thesis is part of a wider effort to construct a comprehensive AFM image analysis platform designed for the automatic characterisation of biomolecules. The work in this project has brought us closer to this goal through the improvement of existing image analysis software and the development of new methods.

The current limitations of automatic AFM image analysis, especially those related to the tracing of linear molecules, have also been highlighted, so that they can be addressed in future projects. For example, the trajectory of DNA is difficult to determine when crossings are encountered. A successful solution to this issue will lead to more applications for automatic analysis tools: rather than relying on proxy parameters such as the aspect ratio or the maximum bounding distance, conformational changes on DNA including those induced by protein interaction or supercoiling will be characterised directly through curvature measurement, thanks to more accurate tracing for molecules exhibiting complex shapes.

It is expected that, when the relevant future work is complete, a user-friendly and versatile AFM image analysis pipeline compatible with nearly all AFM file formats will be established. It will enable the automatic characterisation of a wide range of biomolecules with minimal user input, and it will be suitable for researchers with varying experience in AFM image analysis.

References

[1] A. J. M. Wollman, R. Nudd, E. G. Hedlund, and M. C. Leake, "From animaculum to single molecules: 300 years of the light microscope," *Open Biol.*, vol. 5, no. 4, Apr. 2015, doi: 10.1098/rsob.150019.

[2] C. R. Burch and J. P. P. Stock, "Phase-Contrast Microscopy," *J. Sci. Instrum.*, vol. 19, no. 5, pp. 71–75, May 1942, doi: 10.1088/0950-7671/19/5/302.

[3] Y. Garini, B. J. Vermolen, and I. T. Young, "From micro to nano: recent advances in high-resolution microscopy," *Curr. Opin. Biotechnol.*, vol. 16, no. 1, pp. 3–12, 2005, doi: 10.1016/j.copbio.2005.01.003.

[4] S. M. Hickey *et al.*, "Fluorescence Microscopy—An Outline of Hardware, Biological Handling, and Fluorophore Considerations," *Cells*, vol. 11, no. 1, 2022, doi: 10.3390/cells11010035.

[5] P. P. Laissue, R. A. Alghamdi, P. Tomancak, E. G. Reynaud, and H. Shroff, "Assessing phototoxicity in live fluorescence imaging," *Nat. Methods*, vol. 14, no. 7, pp. 657–661, 2017, doi: 10.1038/nmeth.4344.

[6] P. A. Penczek, "Resolution measures in molecular electron microscopy," in *Methods in Enzymology*, Academic Press, 2010, pp. 73–100. doi: 10.1016/S0076-6879(10)82003-8.

[7] F. Nieto, J. J. Millán, G. G. Parreira, H. Chiarini-Garcia, and R. C. N. Melo, "Electron Microscopy: SEM/TEM," in *Handbook of Physics in Medicine and Biology*, 1st ed., R. Splinter, Ed. Boca Raton: CRC Press, 2010.

[8] X. Benjin and L. Ling, "Developments, applications, and prospects of cryo-electron microscopy," *Protein Science*, vol. 29, no. 4. Wiley-Blackwell, pp. 872–882, Apr. 01, 2020. doi: 10.1002/pro.3805.

[9] K. Bian, C. Gerber, A. J. Heinrich, D. J. Müller, S. Scheuring, and Y. Jiang, "Scanning probe microscopy," *Nat. Rev. Methods Prim.*, vol. 1, no. 1, p. 36, 2021, doi: 10.1038/s43586-021-00033-2.

[10] A. Rodríguez-Galván and F. F. Contreras-Torres, "Scanning Tunneling Microscopy of Biological Structures: An Elusive Goal for Many Years," *Nanomaterials*, vol. 12, no. 17, 2022, doi: 10.3390/nano12173013.

[11] A. Alessandrini and P. Facci, "AFM: A versatile tool in biophysics," *Measurement Science and Technology*, vol. 16, no. 6. Institute of Physics Publishing, Jun. 01, 2005. doi: 10.1088/0957-0233/16/6/R01.

[12] S. Yucel, R. J. Moon, L. J. Johnston, B. Yucel, and S. R. Kalidindi, "Semi-automatic image analysis of particle morphology of cellulose nanocrystals," *Cellulose*, vol. 28, no. 4, pp. 2183–2201, 2021, doi: 10.1007/s10570-020-03668-8.

[13] G. R. Heath *et al.*, "Localization atomic force microscopy," *Nature*, vol. 594, no. 7863, pp. 385–390, Jun. 2021, doi: 10.1038/s41586-021-03551-x.

[14] P. Soille and L. M. Vincent, "Determining watersheds in digital pictures via flooding

simulations," in *Visual Communications and Image Processing '90: Fifth in a Series*, 1990, vol. 1360, pp. 240–250. doi: 10.1117/12.24211.

[15] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans Syst Man Cybern*, vol. SMC-9, no. 1, pp. 62–66, 1979, doi: 10.1109/tsmc.1979.4310076.

[16] J. Schindelin *et al.*, "Fiji: An open-source platform for biological-image analysis," *Nature Methods*, vol. 9, no. 7. pp. 676–682, Jul. 2012. doi: 10.1038/nmeth.2019.

[17] A. B. Schroeder, E. T. A. Dobson, C. T. Rueden, P. Tomancak, F. Jug, and K. W. Eliceiri, "The ImageJ ecosystem: Open-source software for image visualization, processing, and analysis," *Protein Sci.*, vol. 30, no. 1, pp. 234–249, Jan. 2021, doi: 10.1002/pro.3993.

[18] S. R. Sternberg, "Biomedical Image Processing," *Computer (Long. Beach. Calif).*, vol. 16, no. 1, pp. 22–34, 1983, doi: 10.1109/MC.1983.1654163.

[19] M. Pachitariu and C. Stringer, "Cellpose 2.0: how to train your own model," *Nat. Methods*, vol. 19, no. 12, pp. 1634–1641, Dec. 2022, doi: 10.1038/s41592-022-01663-4.

[20] Y. Wang, X. Jin, and C. Castro, "Accelerating the characterization of dynamic DNA origami devices with deep neural networks," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-41459-w.

[21] J. Ma and B. Wang, "Towards foundation models of biological image segmentation," *Nature Methods*, vol. 20, no. 7. Nature Research, pp. 953–955, Jul. 01, 2023. doi: 10.1038/s41592-023-01885-0.

[22] M. Weigert *et al.*, "Content-aware image restoration: pushing the limits of fluorescence microscopy," *Nat. Methods*, vol. 15, no. 12, pp. 1090–1097, Dec. 2018, doi: 10.1038/s41592-018-0216-7.

[23] "What's next for bioimage analysis?," *Nat. Methods*, vol. 20, no. 7, pp. 945–946, Jul. 2023, doi: 10.1038/s41592-023-01950-8.

[24] B. A. Cimini and K. W. Eliceiri, "The Twenty Questions of bioimage object analysis," *Nature Methods*, vol. 20, no. 7. Nature Research, pp. 976–978, Jul. 01, 2023. doi: 10.1038/s41592-023-01919-7.

[25] R. F. Laine, I. Arganda-Carreras, R. Henriques, and G. Jacquemet, "Avoiding a replication crisis in deep-learning-based bioimage analysis," *Nat. Methods*, vol. 18, no. 10, pp. 1136–1144, 2021, doi: 10.1038/s41592-021-01284-3.

[26] M. D. Wilkinson *et al.*, "Comment: The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, Mar. 2016, doi: 10.1038/sdata.2016.18.

[27] X. Li, Y. Zhang, J. Wu, and Q. Dai, "Challenges and opportunities in bioimage analysis," *Nature Methods*, vol. 20, no. 7. Nature Research, pp. 958–961, Jul. 01, 2023. doi: 10.1038/s41592-023-01900-4.

[28] A. Reinke *et al.*, "Understanding metric-related pitfalls in image analysis validation," *Nat. Methods*, vol. 21, no. 2, pp. 182–194, Feb. 2024, doi: 10.1038/s41592-023-

02150-0.

[29] A. Shivanandan, H. Deschout, M. Scarselli, and A. Radenovic, “Challenges in quantitative single molecule localization microscopy,” *FEBS Lett.*, vol. 588, no. 19, pp. 3595–3602, 2014, doi: 10.1016/j.febslet.2014.06.014.

[30] X. Liu, Y. Jiang, Y. Cui, J. Yuan, and X. Fang, “Deep learning in single-molecule imaging and analysis: recent advances and prospects,” *Chem. Sci.*, vol. 13, no. 41, pp. 11964–11980, 2022, doi: 10.1039/d2sc02443h.

[31] G. Binnig, C. F. Quate, and C. Gerber, “Atomic force microscope,” *Phys. Rev. Lett.*, vol. 56, no. 9, pp. 930–933, Mar. 1986, doi: 10.1103/PhysRevLett.56.930.

[32] F. Ohnesorge and G. Binnig, “True Atomic Resolution by Atomic Force Microscopy Through Repulsive and Attractive Forces,” *Science*, vol. 260, no. 5113, pp. 1451–1456, Jun. 1993, doi: 10.1126/science.260.5113.1451.

[33] D. J. Müller, F. A. Schabert, G. Büldt, and A. Engel, “Imaging purple membranes in aqueous solutions at sub-nanometer resolution by atomic force microscopy,” *Biophys. J.*, vol. 68, no. 5, pp. 1681–1686, 1995, doi: 10.1016/S0006-3495(95)80345-0.

[34] Y. F. Dufrêne *et al.*, “Imaging modes of atomic force microscopy for application in molecular and cell biology,” *Nature Nanotechnology*, vol. 12, no. 4. Nature Publishing Group, pp. 295–307, May 01, 2017. doi: 10.1038/nnano.2017.45.

[35] K. H. S. Main, J. I. Provan, P. J. Haynes, G. Wells, J. A. Hartley, and A. L. B. Pyne, “Atomic force microscopy—A tool for structural and translational DNA research,” *APL Bioengineering*, vol. 5, no. 3. AIP Publishing LLC AIP Publishing, p. 031504, Jul. 09, 2021. doi: 10.1063/5.0054294.

[36] B. Pittenger, N. Erina, and C. su, *Quantitative mechanical property mapping at the nanoscale with PeakForce QNM*, vol. 128. 2010. doi: 10.13140/RG.2.1.4463.8246.

[37] F. Moreno-Herrero and J. Gomez-Herrero, “AFM: Basic Concepts,” in *Atomic Force Microscopy in Liquid: Biological Applications*, A. M. Baró and R. G. Reifenberger, Eds. Wiley-VCH, 2012.

[38] F. Gołek, P. Mazur, Z. Ryszka, and S. Zuber, “AFM image artifacts,” *Appl. Surf. Sci.*, vol. 304, pp. 11–19, Jun. 2014, doi: 10.1016/j.apsusc.2014.01.149.

[39] “Life Science AFM Probes - NanoAndMore.” https://www.nanoandmore.com/Life-Science-Biological-Soft-Contact-AFM-Probes (accessed Jul. 04, 2021).

[40] “Bruker AFM Probes - FASTSCAN-D.” https://www.brukerafmprobes.com/p-3816-fastscan-d.aspx (accessed Sep. 27, 2024).

[41] A. D. Bates and A. Maxwell, “DNA structure,” in *DNA Topology*, Oxford: Oxford University Press, 2005.

[42] “Bruker AFM Probes - PEAKFORCE-HIRS-F-B.” https://www.brukerafmprobes.com/p-3958-peakforce-hirs-f-b.aspx (accessed Sep. 27, 2024).

[43] D. Ricci and P. C. Braga, “Recognizing and avoiding artifacts in atomic force microscopy imaging,” in *Atomic Force Microscopy: Biomedical Methods and*

Applications, Humana Press Inc., 2004, pp. 25–37. doi: 10.1007/978-1-61779-105-5_3.

[44] N. Qi, Y. Fang, X. Ren, and Y. Wu, “Varying-gain modeling and advanced DMPC control of an AFM system,” *IEEE Trans. Nanotechnol.*, vol. 14, no. 1, pp. 82–92, Jan. 2015, doi: 10.1109/TNANO.2014.2366197.

[45] E. Anguiano and M. Aguilar, “A cross-measurement procedure (CMP) for near noise-free imaging in scanning microscopes,” *Ultramicroscopy*, vol. 76, no. 1–2, pp. 39–47, 1999, doi: 10.1016/S0304-3991(98)00074-6.

[46] “Scan Line Artefacts.” http://gwyddion.net/documentation/user-guide-en/scan-line-defects.html (accessed Nov. 18, 2021).

[47] B. W. Erickson, S. Coquoz, J. D. Adams, D. J. Burns, and G. E. Fantner, “Large-scale analysis of high-speed atomic force microscopy data sets using adaptive image processing,” *Beilstein J. Nanotechnol.*, vol. 3, no. 1, pp. 747–758, 2012, doi: 10.3762/bjnano.3.84.

[48] “MountainsSPIP® image analysis software for scanning probe microscopes.” https://www.digitalsurf.com/software-solutions/scanning-probe-microscopy/ (accessed Jun. 22, 2023).

[49] J. G. Beton *et al.*, “TopoStats – A program for automated tracing of biomolecules from AFM images,” *Methods*, vol. 193, pp. 68–79, Feb. 2021, doi: 10.1016/j.ymeth.2021.01.008.

[50] E. J. Miller, W. Trewby, A. F. Payam, L. Piantanida, C. Cafolla, and K. Voïtchovsky, “Sub-nanometer resolution imaging with amplitude-modulation atomic force microscopy in liquid,” *J. Vis. Exp.*, vol. 2016, no. 118, p. 54924, Dec. 2016, doi: 10.3791/54924.

[51] Q. Zou, K. K. Leang, E. Sadoun, M. J. Reed, and S. Devasia, “Control issues in high-speed AFM for biological applications: Collagen imaging example,” *Asian J. Control*, vol. 6, no. 2, pp. 164–178, Jun. 2004, doi: 10.1111/j.1934-6093.2004.tb00195.x.

[52] Y. F. Dufrêne and P. Hinterdorfer, “Recent progress in AFM molecular recognition studies,” *Pflugers Arch. Eur. J. Physiol.*, vol. 456, no. 1, pp. 237–245, Apr. 2008, doi: 10.1007/s00424-007-0413-1.

[53] G. R. Heath, E. Micklethwaite, and T. M. Storer, “NanoLocz: Image Analysis Platform for AFM, High-Speed AFM, and Localization AFM,” *Small Methods*, 2024, doi: 10.1002/smtd.202301766.

[54] M. Würtz *et al.*, “DNA accessibility of chromatosomes quantified by automated image analysis of AFM data,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, Dec. 2019, doi: 10.1038/s41598-019-49163-4.

[55] E. Ficarra, L. Benini, E. Macii, and G. Zuccheri, “Automated DNA fragments recognition and sizing through AFM image processing,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 9, no. 4, pp. 508–517, Dec. 2005, doi: 10.1109/TITB.2005.855546.

[56] E. Ficarra, D. Masotti, E. Macii, L. Benini, G. Zuccheri, and B. Samorì, “Automatic intrinsic DNA curvature computation from AFM images,” *IEEE Trans. Biomed. Eng.*,

vol. 52, no. 12, pp. 2074–2086, Dec. 2005, doi: 10.1109/TBME.2005.857666.

[57] A. Sundstrom *et al.*, “Image analysis and length estimation of biomolecules using AFM,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 6, pp. 1200–1207, 2012, doi: 10.1109/TITB.2012.2206819.

[58] D. Nečas and P. Klapetek, “Gwyddion: An open-source software for SPM data analysis,” *Cent. Eur. J. Phys.*, vol. 10, no. 1, pp. 181–188, 2012, doi: 10.2478/s11534-011-0096-2.

[59] C. Schmied *et al.*, “Community-developed checklists for publishing images and image analyses,” *Nat. Methods*, vol. 21, no. 2, pp. 170–181, Feb. 2024, doi: 10.1038/s41592-023-01987-9.

[60] F. Kienberger *et al.*, “Improving the contrast of topographical AFM images by a simple averaging filter,” *Ultramicroscopy*, vol. 106, no. 8–9, pp. 822–828, Jun. 2006, doi: 10.1016/j.ultramic.2005.11.013.

[61] N. Shephard, S. Whittle, M. Gamill, M. Du, and A. Pyne, “TopoStats - Atomic Force Microscopy image processing and analysis.” May 2023. doi: 10.15131/shef.data.22633528.v2.

[62] C. Han and C. C. Chung, “Advanced flattening method for scanned atomic force microscopy images,” *J. Korean Phys. Soc.*, vol. 60, no. 5, pp. 680–683, Mar. 2012, doi: 10.3938/jkps.60.680.

[63] Y. Wang, T. Lu, X. Li, and H. Wang, “Automated image segmentation-assisted flattening of atomic force microscopy images,” *Beilstein J. Nanotechnol.*, vol. 9, no. 1, pp. 975–985, Mar. 2018, doi: 10.3762/bjnano.9.91.

[64] P. Fechner, T. Boudier, S. Mangenot, S. Jaroslawski, J. N. Sturgis, and S. Scheuring, “Structural information, resolution, and noise in high-resolution atomic force microscopy topographs,” *Biophys. J.*, vol. 96, no. 9, pp. 3822–3831, May 2009, doi: 10.1016/j.bpj.2009.02.011.

[65] P. I. T. Chang, Y. Y. Song, and M. C. Hsaio, “Estimation of DNA persistence length with atomic force microscopy imaging,” 2016. doi: 10.1109/ICARCV.2016.7838753.

[66] K. Bose, C. J. Lech, B. Heddi, and A. T. Phan, “High-resolution AFM structure of DNA G-wires in aqueous solution,” *Nat. Commun.*, vol. 9, no. 1, pp. 1–9, Dec. 2018, doi: 10.1038/s41467-018-04016-y.

[67] Y. Fang *et al.*, “Solid-State DNA Sizing by Atomic Force Microscopy,” *Anal. Chem.*, vol. 70, no. 10, pp. 2123–2129, May 1998, doi: 10.1021/ac971187o.

[68] A. Sanchez-Sevilla, J. Thimonier, M. Marilley, J. Rocca-Serra, and J. Barbet, “Accuracy of AFM measurements of the contour length of DNA fragments adsorbed on mica in air and in aqueous buffer,” *Ultramicroscopy*, vol. 92, no. 3–4, pp. 151–158, Aug. 2002, doi: 10.1016/S0304-3991(02)00128-6.

[69] W. R. Bowen and T. A. Doneva, “Artefacts in AFM studies of membranes: Correcting pore images using fast fourier transform filtering,” *J. Memb. Sci.*, vol. 171, no. 1, pp. 141–147, Jun. 2000, doi: 10.1016/S0376-7388(00)00297-0.

[70] S. W. W. Chen and J. L. Pellequer, “DeStripe: Frequency-based algorithm for removing stripe noises from AFM images,” *BMC Struct. Biol.*, vol. 11, no. 1, pp. 1–10, Feb. 2011, doi: 10.1186/1472-6807-11-7.

[71] V. Kocur, V. Hegrová, M. Patočka, J. Neuman, and A. Herout, “Correction of AFM data artifacts using a convolutional neural network trained with synthetically generated data,” *Ultramicroscopy*, vol. 246, Apr. 2023, doi: 10.1016/j.ultramic.2022.113666.

[72] T. W. Ridler and S. Calvard, “Picture Thresholding Using an Iterative Selection Method,” *IEEE Trans. Syst. Man Cybern.*, vol. SMC-8, no. 8, pp. 630–632, 1978, doi: 10.1109/tsmc.1978.4310039.

[73] R. O. Duda and P. E. Hart, “Use of the Hough Transformation to Detect Lines and Curves in Pictures,” *Commun. ACM*, vol. 15, no. 1, pp. 11–15, 1972, doi: 10.1145/361237.361242.

[74] O. M. Gordon, J. E. A. Hodgkinson, S. M. Farley, E. L. Hunsicker, and P. J. Moriarty, “Automated Searching and Identification of Self-Organized Nanostructures,” *Nano Lett.*, vol. 20, no. 10, pp. 7688–7693, Oct. 2020, doi: 10.1021/acs.nanolett.0c03213.

[75] M. Schneider, A. Al-Shaer, and N. R. Forde, “AutoSmarTrace: Automated chain tracing and flexibility analysis of biological filaments,” *Biophys. J.*, vol. 120, no. 13, pp. 2599–2608, Jul. 2021, doi: 10.1016/j.bpj.2021.05.011.

[76] D. M. Bangalore *et al.*, “Automated AFM analysis of DNA bending reveals initial lesion sensing strategies of DNA glycosylases,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, Dec. 2020, doi: 10.1038/s41598-020-72102-7.

[77] T. Y. Zhang and C. Y. Suen, “A fast parallel algorithm for thinning digital patterns,” *Commun. ACM*, vol. 27, no. 3, pp. 236–239, Mar. 1984, doi: 10.1145/357994.358023.

[78] A. Bates and A. Maxwell, “Knots and catenanes,” in *DNA Topology*, Oxford: Oxford University Press, 2005, pp. 107–124.

[79] V. Uhlmann, C. Haubold, F. A. Hamprecht, and M. Unser, “DiversePathsJ: Diverse shortest paths for bioimage analysis,” *Bioinformatics*, vol. 34, no. 3, pp. 538–540, Feb. 2018, doi: 10.1093/bioinformatics/btx621.

[80] P. Dabrowski-Tumanski, P. Rubach, W. Niemyska, B. A. Gren, and J. I. Sulkowska, “Topoly: Python package to analyze topology of polymers,” *Brief. Bioinform.*, vol. 22, no. 3, pp. 1–8, May 2021, doi: 10.1093/bib/bbaa196.

[81] C. Rivetti and S. Codeluppi, “Accurate length determination of DNA molecules visualized by atomic force microscopy: Evidence for a partial B- to A-form transition on mica,” *Ultramicroscopy*, vol. 87, no. 1–2, pp. 55–66, 2001, doi: 10.1016/S0304-3991(00)00064-4.

[82] J. Marek *et al.*, “Interactive measurement and characterization of DNA molecules by analysis of AFM images,” *Cytom. Part A*, vol. 63A, no. 2, pp. 87–93, Feb. 2005, doi: 10.1002/cyto.a.20105.

[83] S. F. Konrad *et al.*, “High-throughput AFM analysis reveals unwrapping pathways of H3 and CENP-A nucleosomes,” *Nanoscale*, vol. 13, no. 10, pp. 5435–5447, Mar. 2021, doi: 10.1039/d0nr08564b.

[84] A. L. B. Pyne *et al.*, “Base-pair resolution analysis of the effect of supercoiling on DNA flexibility and major groove recognition by triplex-forming oligonucleotides,” *Nat. Commun.*, vol. 12, no. 1, p. 1053, Dec. 2021, doi: 10.1038/s41467-021-21243-y.

[85] R. E. Franklin and R. G. Gosling, “Molecular configuration in sodium thymonucleate,” *Nature*, vol. 171, no. 4356, pp. 740–741, Apr. 1953, doi: 10.1038/171740a0.

[86] J. D. Watson and F. H. C. Crick, “Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid,” *Nature*, vol. 171, no. 4356, pp. 737–738, 1953, doi: 10.1038/171737a0.

[87] M. Kaushik *et al.*, “A bouquet of DNA structures: Emerging diversity,” *Biochem. Biophys. Reports*, vol. 5, pp. 388–395, Mar. 2016, doi: 10.1016/j.bbrep.2016.01.013.

[88] A. Rich, “DNA comes in many forms,” *Gene*, vol. 135, no. 1–2, pp. 99–109, Dec. 1993, doi: 10.1016/0378-1119(93)90054-7.

[89] A. D. Bates and A. Maxwell, “DNA supercoiling,” in *DNA Topology*, Oxford: Oxford University Press, 2005.

[90] M. D. Frank-Kamenetskii and S. M. Mirkin, “Triplex DNA structures,” *Annual Review of Biochemistry*, vol. 64. Annual Reviews Inc., pp. 65–95, Nov. 28, 1995. doi: 10.1146/annurev.bi.64.070195.000433.

[91] D. E. Gilbert and J. Feigon, “Multistranded DNA structures,” *Curr. Opin. Struct. Biol.*, vol. 9, no. 3, pp. 305–314, 1999, doi: 10.1016/S0959-440X(99)80041-4.

[92] S. V. S. Mariappan, A. E. Garcia, and G. Gupta, “Structure dynamics of the DNA hairpins formed by tandemly repeated CTG triplets associated with myotonic dystrophy,” *Nucleic Acids Res.*, vol. 24, no. 4, pp. 775–783, Feb. 1996, doi: 10.1093/nar/24.4.775.

[93] V. Brázda, R. C. Laister, E. B. Jagelská, and C. Arrowsmith, “Cruciform structures are a common DNA feature important for regulating biological processes,” *BMC Molecular Biology*, vol. 12, no. 1. BioMed Central Ltd., Aug. 05, 2011. doi: 10.1186/1471-2199-12-33.

[94] A. I. H. Murchie and D. M. J. Lilley, “Supercoiled DNA and cruciform structures,” *Methods Enzymol.*, vol. 211, no. C, pp. 158–180, Jan. 1992, doi: 10.1016/0076-6879(92)11010-G.

[95] A. Jain, G. Wang, and K. M. Vasquez, “DNA triple helices: Biological consequences and therapeutic potential,” *Biochimie*, vol. 90, no. 8. Elsevier, pp. 1117–1130, Aug. 01, 2008. doi: 10.1016/j.biochi.2008.02.011.

[96] T. M. Bryan, “G-quadruplexes at telomeres: Friend or foe?,” *Molecules*, vol. 25, no. 16. Multidisciplinary Digital Publishing Institute (MDPI), Aug. 01, 2020. doi: 10.3390/molecules25163686.

[97] T. S. Hsieh, “DNA Supercoiling,” in *Encyclopedia of Biological Chemistry: Second Edition*, Elsevier Inc., 2013, pp. 154–156. doi: 10.1016/B978-0-12-378630-2.00244-9.

[98] A. T. Fenley, R. Anandakrishnan, Y. H. Kidane, and A. V. Onufriev, “Modulation of nucleosomal DNA accessibility via charge-altering post-translational modifications in

histone core," *Epigenetics and Chromatin*, vol. 11, no. 1, pp. 1–19, Mar. 2018, doi: 10.1186/s13072-018-0181-5.

[99] P. Tessarz and T. Kouzarides, "Histone core modifications regulating nucleosome structure and dynamics," *Nat. Rev. Mol. Cell Biol.*, vol. 15, no. 11, pp. 703–708, Nov. 2014, doi: 10.1038/nrm3890.

[100] J. Pfluger and D. Wagner, "Histone modifications and dynamic regulation of genome accessibility in plants," *Curr. Opin. Plant Biol.*, vol. 10, no. 6, pp. 645–652, Dec. 2007, doi: 10.1016/j.pbi.2007.07.013.

[101] K. K. Swinger, K. M. Lemberg, Y. Zhang, and P. A. Rice, "Flexible DNA bending in HU-DNA cocrystal structures," *EMBO J.*, vol. 22, no. 14, pp. 3749–3760, Jul. 2003, doi: 10.1093/emboj/cdg351.

[102] S. C. Schultz, G. C. Shields, and T. A. Steitz, "Crystal structure of a CAP-DNA complex: The DNA is bent by 90°," *Science*, vol. 253, no. 5023, pp. 1001–1007, 1991, doi: 10.1126/science.1653449.

[103] G. B. Koudelka, P. Harbury, S. C. Harrison, and M. Ptashne, "DNA twisting and the affinity of bacteriophage 434 operator for bacteriophage 434 repressor," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 85, no. 13, pp. 4633–4637, 1988, doi: 10.1073/pnas.85.13.4633.

[104] C. Oguey, N. Foloppe, and B. Hartmann, "Understanding the sequence-dependence of DNA groove dimensions: Implications for DNA interactions," *PLoS One*, vol. 5, no. 12, p. e15931, 2010, doi: 10.1371/journal.pone.0015931.

[105] R. Wing *et al.*, "Crystal structure analysis of a complete turn of B-DNA," *Nature*, vol. 287, no. 5784, pp. 755–758, 1980, doi: 10.1038/287755a0.

[106] A. McPherson and J. A. Gavira, "Introduction to protein crystallization," *Acta Crystallogr. Sect. F Struct. Biol. Commun.*, vol. 70, no. 1, pp. 2–20, Dec. 2014, doi: 10.1107/S2053230X13033141.

[107] M. Lukin and C. de los Santos, "NMR structures of damaged DNA," *Chemical Reviews*, vol. 106, no. 2. American Chemical Society, pp. 607–686, Feb. 2006. doi: 10.1021/cr0404646.

[108] P. Bendel and T. L. James, "Structural and dynamic differences between supercoiled and linear DNA from proton NMR," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 80, no. 11, pp. 3284–3286, 1983, doi: 10.1073/pnas.80.11.3284.

[109] R. P. Barnwal, F. Yang, and G. Varani, "Applications of NMR to structure determination of RNAs large and small," *Arch. Biochem. Biophys.*, vol. 628, pp. 42–56, Aug. 2017, doi: 10.1016/j.abb.2017.06.003.

[110] T. A. M. Bharat *et al.*, "Structure of the immature retroviral capsid at 8Å resolution by cryo-electron microscopy," *Nature*, vol. 487, no. 7407, pp. 385–389, Jul. 2012, doi: 10.1038/nature11169.

[111] Y. Cheng, "Single-particle Cryo-EM at crystallographic resolution," *Cell*, vol. 161, no. 3, pp. 450–457, Apr. 2015, doi: 10.1016/j.cell.2015.03.049.

[112] Q. Wu, S. Liang, T. Ochi, D. Y. Chirgadze, J. T. Huiskonen, and T. L. Blundell, "Understanding the structure and role of DNA-PK in NHEJ: How X-ray diffraction and cryo-EM contribute in complementary ways," *Prog. Biophys. Mol. Biol.*, vol. 147, pp. 26–32, Oct. 2019, doi: 10.1016/j.pbiomolbio.2019.03.007.

[113] T. W. Guo *et al.*, "Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex," *Cell*, vol. 171, no. 2, pp. 414–426, Oct. 2017, doi: 10.1016/j.cell.2017.09.006.

[114] R. F. Thompson, M. Walker, C. A. Siebert, S. P. Muench, and N. A. Ranson, "An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology," *Methods*, vol. 100, pp. 3–15, May 2016, doi: 10.1016/j.ymeth.2016.02.017.

[115] M. D. Barkley and B. H. Zimm, "Theory of twisting and bending of chain macromolecules; analysis of the fluorescence depolarization of DNA," *J. Chem. Phys.*, vol. 70, no. 6, pp. 2991–3007, Mar. 1979, doi: 10.1063/1.437838.

[116] K. Wojtuszewski and I. Mukerji, "HU binding to bent DNA: A fluorescence resonance energy transfer and anisotropy study," *Biochemistry*, vol. 42, no. 10, pp. 3096–3104, Mar. 2003, doi: 10.1021/bi0264014.

[117] B. Klejevskaja *et al.*, "Studies of G-quadruplexes formed within self-assembled DNA mini-circles," *Chem. Commun.*, vol. 52, no. 84, pp. 12454–12457, 2016, doi: 10.1039/c6cc07110d.

[118] B. Dumat, A. F. Larsen, and L. M. Wilhelmsson, "Studying Z-DNA and B-to Z-DNA transitions using a cytosine analogue FRET-pair," *Nucleic Acids Res.*, vol. 44, no. 11, p. e101, Jun. 2016, doi: 10.1093/nar/gkw114.

[119] J. M. Fogg *et al.*, "Exploring writhe in supercoiled minicircle DNA," *J. Phys. Condens. Matter*, vol. 18, no. 14, Apr. 2006, doi: 10.1088/0953-8984/18/14/S01.

[120] S. Delaney, R. Murphy, and F. Walsh, "A comparison of methods for the extraction of plasmids capable of conferring antibiotic resistance in a human pathogen from complex broiler cecal samples," *Front. Microbiol.*, vol. 9, Aug. 2018, doi: 10.3389/fmicb.2018.01731.

[121] J. Yoo, D. Winogradoff, and A. Aksimentiev, "Molecular dynamics simulations of DNA–DNA and DNA–protein interactions," *Current Opinion in Structural Biology*, vol. 64. Elsevier Current Trends, pp. 88–96, Oct. 01, 2020. doi: 10.1016/j.sbi.2020.06.007.

[122] M. Li *et al.*, "Mechanism of DNA translocation underlying chromatin remodelling by Snf2," *Nature*, vol. 567, no. 7748, pp. 409–413, 2019, doi: 10.1038/s41586-019-1029-2.

[123] B. Hellenkamp *et al.*, "Precision and accuracy of single-molecule FRET measurements—a multi-laboratory benchmark study," *Nat. Methods*, vol. 15, no. 9, pp. 669–676, Sep. 2018, doi: 10.1038/s41592-018-0085-0.

[124] N. C. Stellwagen, "Accurate molecular weight determinations of deoxyribonucleic acid restriction fragments on agarose gels," *Biochemistry*, vol. 22, no. 26, pp. 6180–6185, Dec. 1983, doi: 10.1021/bi00295a022.

[125] S. Ido *et al.*, "Beyond the helix pitch: Direct visualization of native DNA in aqueous solution," *ACS Nano*, vol. 7, no. 2, pp. 1817–1822, 2013, doi: 10.1021/nn400071n.

[126] P. J. Haynes, K. H. S. Main, and A. L. Pyne, "Atomic Force Microscopy of DNA and DNA-Protein Interactions," *protocols.io*, 2020, doi: 10.1007/978-1-0716-2221-6_5.

[127] R. Henderson and P. N. T. Unwin, "Structure of the purple membrane from Halobacterium halobium," *Biophys. Struct. Mech.*, vol. 3, no. 2, p. 121, 1977, doi: 10.1007/BF00535804.

[128] A. Pyne, R. Thompson, C. Leung, D. Roy, and B. W. Hoogenboom, "Single-molecule reconstruction of oligonucleotide secondary structure by atomic force microscopy," *Small*, vol. 10, no. 16, pp. 3257–3261, 2014, doi: 10.1002/smll.201400265.

[129] Y. L. Lyubchenko and L. S. Shlyakhtenko, "AFM for analysis of structure and dynamics of DNA and protein-DNA complexes," *Methods*, vol. 47, no. 3, pp. 206–213, Mar. 2009, doi: 10.1016/j.ymeth.2008.09.002.

[130] F. Ostendorf, C. Schmitz, S. Hirth, A. Kühnle, J. J. Kolodziej, and M. Reichling, "How flat is an air-cleaved mica surface?," *Nanotechnology*, vol. 19, no. 30, Jul. 2008, doi: 10.1088/0957-4484/19/30/305705.

[131] H. G. Hansma and D. E. Laney, "DNA binding to mica correlates with cationic radius: assay by atomic force microscopy," *Biophys. J.*, vol. 70, no. 4, pp. 1933–1939, Apr. 1996, doi: 10.1016/S0006-3495(96)79757-6.

[132] C. Sissi and M. Palumbo, "Effects of magnesium and related divalent metal ions in topoisomerase structure and function," *Nucleic Acids Res.*, vol. 37, no. 3, pp. 702–711, Feb. 2009, doi: 10.1093/nar/gkp024.

[133] B. Akpinar, P. J. Haynes, N. A. W. Bell, K. Brunner, A. L. B. Pyne, and B. W. Hoogenboom, "PEGylated surfaces for the study of DNA-protein interactions by atomic force microscopy," *Nanoscale*, vol. 11, no. 42, pp. 20072–20080, 2019, doi: 10.1039/c9nr07104k.

[134] M. Bussiek, N. Mücke, and J. Langowski, "Polylysine-coated mica can be used to observe systematic changes in the supercoiled DNA conformation by scanning force microscopy in solution.," *Nucleic Acids Res.*, vol. 31, no. 22, p. e137, Nov. 2003, doi: 10.1093/nar/gng137.

[135] A. Podestà, L. Imperadori, W. Colnaghi, L. Finzi, P. Milani, and D. Dunlap, "Atomic force microscopy study of DNA deposited on poly L-ornithine-coated mica," *J. Microsc.*, vol. 215, no. 3, pp. 236–240, Sep. 2004, doi: 10.1111/j.0022-2720.2004.01372.x.

[136] G. Liu *et al.*, "Biological Properties of Poly-L-lysine-DNA Complexes Generated by Cooperative Binding of the Polycation," *J. Biol. Chem.*, vol. 276, no. 37, pp. 34379–34387, Sep. 2001, doi: 10.1074/jbc.M105250200.

[137] T. C. Marsh, J. Vesenka, and E. Henderson, "Nar00004-0164.Pdf," *Nucleic Acids Res.*, vol. 23, no. 4, pp. 696–700, 1995.

[138] B. Alberts, *Molecular biology of the cell*, Sixth edit. New York, New York: Garland Science, 2015.

[139] L. Alonso-Sarduy, G. Longo, G. Dietler, and S. Kasas, “Time-lapse AFM imaging of DNA conformational changes induced by daunorubicin,” *Nano Lett.*, vol. 13, no. 11, pp. 5679–5684, 2013, doi: 10.1021/nl403361f.

[140] J. J. Champoux, “DNA topoisomerases: Structure, function, and mechanism,” *Annu. Rev. Biochem.*, vol. 70, pp. 369–413, Nov. 2001, doi: 10.1146/annurev.biochem.70.1.369.

[141] A. Bates and A. Maxwell, “DNA topoisomerases,” in *DNA Topology*, Oxford: Oxford University Press, 2005, pp. 125–144.

[142] Y. Pommier, A. Nussenzweig, S. Takeda, and C. Austin, “Human topoisomerases and their roles in genome stability and organization,” *Nature Reviews Molecular Cell Biology*, vol. 23, no. 6. Nature Publishing Group, pp. 407–427, Feb. 28, 2022. doi: 10.1038/s41580-022-00452-3.

[143] Y. Pommier, E. Leo, H. Zhang, and C. Marchand, “DNA topoisomerases and their poisoning by anticancer and antibacterial drugs,” *Chem. Biol.*, vol. 17, no. 5, pp. 421–433, 2010, doi: 10.1016/J.CHEMBIOL.2010.04.012.

[144] V. J. Venditto and E. E. Simanek, “Cancer therapies utilizing the camptothecins: A review of the in vivo literature,” *Molecular Pharmaceutics*, vol. 7, no. 2. NIH Public Access, pp. 307–349, Apr. 04, 2010. doi: 10.1021/mp900243b.

[145] T. W. Kim and F. Innocenti, “Insights, challenges, and future directions in irinogenetics,” *Therapeutic Drug Monitoring*, vol. 29, no. 3. pp. 265–270, Jun. 2007. doi: 10.1097/FTD.0b013e318068623b.

[146] S. Mostowy *et al.*, “p62 and NDP52 proteins target intracytosolic Shigella and Listeria to different autophagy pathways,” *J. Biol. Chem.*, vol. 286, no. 30, pp. 26987–26995, Jul. 2011, doi: 10.1074/jbc.M111.223610.

[147] B. Morriswood *et al.*, “T6BP and NDP52 are myosin VI binding partners with potential roles in cytokine signalling and cell adhesion,” *J. Cell Sci.*, vol. 120, no. 15, pp. 2574–2585, Aug. 2007, doi: 10.1242/jcs.007005.

[148] J. H. Kim, H. Li, and M. R. Stallcup, “CoCoA, a nuclear receptor coactivator which acts through an N-terminal activation domain of p160 coactivators,” *Mol. Cell*, vol. 12, no. 6, pp. 1537–1549, 2003, doi: 10.1016/S1097-2765(03)00450-7.

[149] F. Korioth, C. Gieffers, G. G. Maul, and J. Frey, “Molecular characterization of NDP52, a novel protein of the nuclear domain 10, which is redistributed upon virus infection and interferon treatment,” *J. Cell Biol.*, vol. 130, no. 1, pp. 1–13, Jul. 1995, doi: 10.1083/jcb.130.1.1.

[150] C. K. Yang, H. K. Jeong, and M. R. Stallcup, “Role of the N-terminal activation domain of the coiled-coil coactivator in mediating transcriptional activation by β-catenin,” *Mol. Endocrinol.*, vol. 20, no. 12, pp. 3251–3262, Dec. 2006, doi: 10.1210/me.2006-0200.

[151] Á. dos Santos *et al.*, “Autophagy receptor NDP52 alters DNA conformation to modulate RNA polymerase II transcription,” *Nat. Commun.*, vol. 14, no. 1, pp. 1–24, May 2023, doi: 10.1038/s41467-023-38572-9.

[152] K. K. Swinger and P. A. Rice, “IHF and HU: Flexible architects of bent DNA,” *Current Opinion in Structural Biology*, vol. 14, no. 1. Elsevier Current Trends, pp. 28–35, Feb. 01, 2004. doi: 10.1016/j.sbi.2003.12.003.

[153] D. Kamashev and J. Rouviere-Yaniv, “The histone-like protein HU binds specifically to DNA recombination and repair intermediates,” *EMBO J.*, vol. 19, no. 23, pp. 6527–6535, Dec. 2000, doi: 10.1093/emboj/19.23.6527.

[154] E. W. Chan, “Mechanisms behind Protein-DNA Interactions Unveiled with Molecular Simulation and Atomic Force Microscopy,” University of York, 2023. Accessed: May 19, 2024. [Online]. Available: https://etheses.whiterose.ac.uk/34365/

[155] A. Basu *et al.*, “Measuring DNA mechanics on the genome scale,” *Nature*, vol. 589, no. 7842, pp. 462–467, Jan. 2021, doi: 10.1038/s41586-020-03052-3.

[156] “Human Topoisomerase I Assay Kits- Available from Inspiralis.” https://www.inspiralis.com/products/topoisomerases/topoisomerase-enzymes-and-assay-kits/human-topoisomerase-i/ (accessed Jul. 11, 2021).

[157] “Bruker AFM Probes - SPM/AFM probes and accessories.” https://www.brukerafmprobes.com/p-3710-msnl-10.aspx? (accessed Sep. 27, 2024).

[158] C. R. Harris *et al.*, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825. Nature Research, pp. 357–362, Sep. 17, 2020. doi: 10.1038/s41586-020-2649-2.

[159] “Data Levelling and Background Subtraction.” http://gwyddion.net/documentation/user-guide-en/leveling-and-background.html (accessed Nov. 18, 2021).

[160] “gwy.GrainQuantity.” http://gwyddion.net/documentation/head/pygwy/gwy.GrainQuantity-class.html (accessed Jul. 23, 2023).

[161] “TopoStats/ at maxgamill-sheffield/pca_and_dbscan · AFM-SPM/TopoStats · GitHub.” https://github.com/AFM-SPM/TopoStats/tree/maxgamill-sheffield/pca_and_dbscan (accessed Jul. 10, 2023).

[162] M. R. Redinbo, L. Stewart, P. Kuhn, J. J. Champoux, and W. G. J. Hol, “Crystal structures of human topoisomerase I in covalent and noncovalent complexes with DNA,” *Science*, vol. 279, no. 5356, pp. 1504–1513, Mar. 1998, doi: 10.1126/science.279.5356.1504.

[163] E. F. Pettersen *et al.*, “UCSF Chimera--a visualization system for exploratory research and analysis,” *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004, doi: 10.1002/JCC.20084.

[164] K. H. S. Main, “Targeting Twist: Single Molecule Insights into the Effect of DNA Supercoiling on Topoisomerase Interactions and Topoisomerase Inhibitor Chemotherapeutics,” University College London, 2022. [Online]. Available: https://discovery.ucl.ac.uk/id/eprint/10153041/

[165] Z. Liu, R. Meng, Y. Zu, Q. Li, and L. Yao, “Imaging and studying human topoisomerase I on mica surfaces in air and in liquid by atomic force microscopy,” *Scanning*, vol. 31, no. 4, pp. 160–166, Jul. 2009, doi: 10.1002/sca.20154.

[166] M. Argaman, S. Bendetz-Nezer, S. Matlis, S. Segal, and E. Priel, “Revealing the mode of action of DNA topoisomerase I and its inhibitors by atomic force microscopy,” *Biochem. Biophys. Res. Commun.*, vol. 301, no. 3, pp. 789–797, Feb. 2003, doi: 10.1016/S0006-291X(03)00025-1.

[167] H. Fischer, I. Polikarpov, and A. F. Craievich, “Average protein density is a molecular-weight-dependent function,” *Protein Sci.*, vol. 13, no. 10, pp. 2825–2828, Jan. 2004, doi: 10.1110/ps.04688204.

[168] A. Ioanoviciu, S. Antony, Y. Pommier, B. L. Staker, L. Stewart, and M. Cushman, “Synthesis and mechanism of action studies of a series of norindenoisoquinoline topoisomerase I poisons reveal an inhibitor with a flipped orientation in the ternary DNA-enzyme-inhibitor complex as determined by X-ray crystallographic analysis,” *J. Med. Chem.*, vol. 48, no. 15, pp. 4803–4814, Jul. 2005, doi: 10.1021/jm050076b.

[169] J. Jumper *et al.*, “Highly accurate protein structure prediction with AlphaFold,” *Nat. 2021 5967873*, vol. 596, no. 7873, pp. 583–589, Jul. 2021, doi: 10.1038/s41586-021-03819-2.

[170] M. Varadi *et al.*, “AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D439–D444, Jan. 2022, doi: 10.1093/NAR/GKAB1061.

[171] P. Virtanen *et al.*, “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

[172] C. Rivetti, M. Guthold, and C. Bustamante, “Scanning force microscopy of DNA deposited onto mica: Equilibration versus kinetic trapping studied by statistical polymer chain analysis,” *J. Mol. Biol.*, vol. 264, no. 5, pp. 919–932, Dec. 1996, doi: 10.1006/jmbi.1996.0687.

[173] “numpy.gradient — NumPy v1.26 Manual.” https://numpy.org/doc/stable/reference/generated/numpy.gradient.html (accessed Mar. 05, 2024).

[174] K. Drlica and J. Rouviere-Yaniv, “Histonelike proteins of bacteria,” *Microbiol. Rev.*, vol. 51, no. 3, pp. 301–319, Sep. 1987, doi: 10.1128/MR.51.3.301-319.1987.

[175] L. M. Hellman and M. G. Fried, “Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions,” *Nat. Protoc.*, vol. 2, no. 8, pp. 1849–1861, 2007, doi: 10.1038/nprot.2007.249.

[176] R. Amyot and H. Flechsig, “BioAFMviewer: An interactive interface for simulated AFM scanning of biomolecular structures and dynamics,” *PLOS Comput. Biol.*, vol. 16, no. 11, p. e1008444, Nov. 2020, doi: 10.1371/JOURNAL.PCBI.1008444.

[177] R. Amyot, N. Kodera, and H. Flechsig, “BioAFMviewer software for simulation atomic force microscopy of molecular structures and conformational dynamics,” *J. Struct. Biol. X*, vol. 7, p. 100086, Jan. 2023, doi: 10.1016/J.YJSBX.2023.100086.