Ultrafast Infrared Spectroscopy Investigation of DNA-Ligand Interactions

By Robby Fritzsch

Department of Physics
University of Strathclyde

A thesis submitted in the requirement for the degree of Doctor of Philosophy

2019

‘This thesis is the result of the author’s original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.’

Signed: 

Date: 01.05.2020

1 Table of Contents

2 List of Abbreviations

2D-IR	two-dimensional infrared
A	adenine
ANOVA	analysis of variance
BBO	bariumborate
C	cytosine
CW	continuous wave
DFG	difference frequency generation
DFT	density functional theory
ds/ssDNA	double/single stranded deoxyribonucleic acid
ESA	excited state absorption
FID	free induction decay
FT-IR	Fourier-transform infrared absorption
FWHM	full width half maximum
G	guanine
GSB	ground state bleach
HPLC	high pressure liquid chromatography
MCT; HgCdTe	mercury cadmium telluride
MgO:PPLN	magnesiumoxide-doped, periodically poled lithium niobite
MVA	multivariate analysis
Nd:YAG	neodymium-doped yttrium aluminium garnet
NMR	nuclear magnetic resonance
OD	optical density units
OPA	optical parametric amplification
OPO	optical parametric oscillator
PA	polyamide
PCA	principal component analysis
PTFE	polytetrafluoroethylene
SiC	silicon carbide
T	thymine
TFA	trifluoroacetic acid
Ti:S	titanium doped sapphire
TR-IR	time-resolved infrared
TRIS	tris(hydroxymethyl)aminomethane
TRMPS	time-resolved multiple probe spectroscopy
UV/Vis	ultraviolet and visible

3 Abstract

Two-dimensional infrared spectroscopy (2D-IR) is a label-free analysis method that can provide structural insight into biological processes by uncovering the coupling between the intrinsic, vibrational modes of a molecule. Technological advancements in the last decade have drastically reduced acquisition times for 2D-IR spectra from several hours down to seconds, enabling researches to perform 2D-IR experiments with unprecedented samples sizes and complexity. This thesis applies 2D-IR spectroscopy as a high-throughput analysis method to study non-covalent binding of small molecules to minor grooves of DNA duplex structures. Minor groove binding to DNA is relevant in the development of potential new therapeutics that target specific DNA sequences to influence a biological process. The conception of a 2D-IR screening experiment will therefore allow for an assessment of sequence-specific ligand interactions across multiple different binding sites.

Minor groove binding of two different types of ligands are investigated in this thesis. Application of the 2D-IR method in the context of screening is first assessed in a proof-of-concept by studying the interaction of a well-established DNA stain, Hoechst 33258, across a set of different dsDNA sequences. In the following chapter, 2D-IR spectroscopy is applied again to study the interaction of three new types of hairpin polyamide ligands, screened across a set of different minor grooves. The final chapter revisits the Hoechst 33258 system to investigate the impact of minor groove binding at non-equilibrium conditions using a laser-induced temperature jump. This chapter is part of preliminary work to develop a temperature-jump, 2D-IR experiment in the future.

4 Introduction

4.1 Abstract

This chapter identifies the study of DNA-ligand binding as a crucial step in the development of new pharmaceutical candidates and establishes time-resolved, infrared spectroscopy as a fast analysis tool to investigate such processes. The structural and temporal information provided by these methods combined with ever-decreasing acquisition times due to improved instrumentation creates an opportunity to conduct high throughput screening experiments. Two-dimensional infrared (2D-IR) spectroscopy is used to study dsDNA-ligand interactions across large spectroscopic datasets in this thesis. Key fundamentals of optical and vibrational spectroscopy are reviewed to explain the concept of 2D-IR spectroscopy. A short overview on the infrared response of DNA is given and its relevance to study minor groove binding is explained. The chapter ends with an outlook on perturbative methods to study non-equilibrium processes over large timeframes as an additional tool for understanding ligand-binding.

4.2 DNA and its Interaction with Minor Groove Binders

Deoxyribonucleic acid (DNA) is one of the most fundamental bio-macromolecules and is the molecular basis for heredity as it stores the genetic information for a plethora of biological structures and processes in specific sequences of four different monomers, the nucleotides[1]. Nucleotides consist of a ribose sugar derivative with a nucleobase at the 1' position and a phosphate unit at the 5' position (see bottom of **Fig. 4.1**). The phosphate unit acts as a linker between nucleotides and binds to the next monomer at the 3' position, giving the macromolecule an intrinsic directionality. Nucleobases in DNA are comprised of two purines, adenine (A) and guanine (G), and two pyrimidines, thymine (T) and cytosine (C). Naturally occurring DNA usually consists of two strands with complementary sequences that align in antiparallel orientation to form the iconic, double helical structure[2]. The two strands are connected via Watson-Crick base pairing. G and C form three, while A and T form two hydrogen bonds, giving the G-C base pair in general a higher stability[3]. Each phosphate unit of the backbone is negatively charged and therefore highly sensitive to the ionic strength of the solvent. Parameters like salt concentration or pH value will therefore change the structure of the double helix in order to adjust its exposure to the solvent[4]. DNA under physiological conditions will form right handed B-Form DNA with a characteristic major and minor groove (grey lines), which are both solvated by water. The unique structure of double stranded DNA implies that the same side groups of the nucleobases always appear in the same groove, irrespective of sequence (as shown in **Fig. 4.1**). This has important implications for spectroscopy which will be discussed in section 4.3.5.

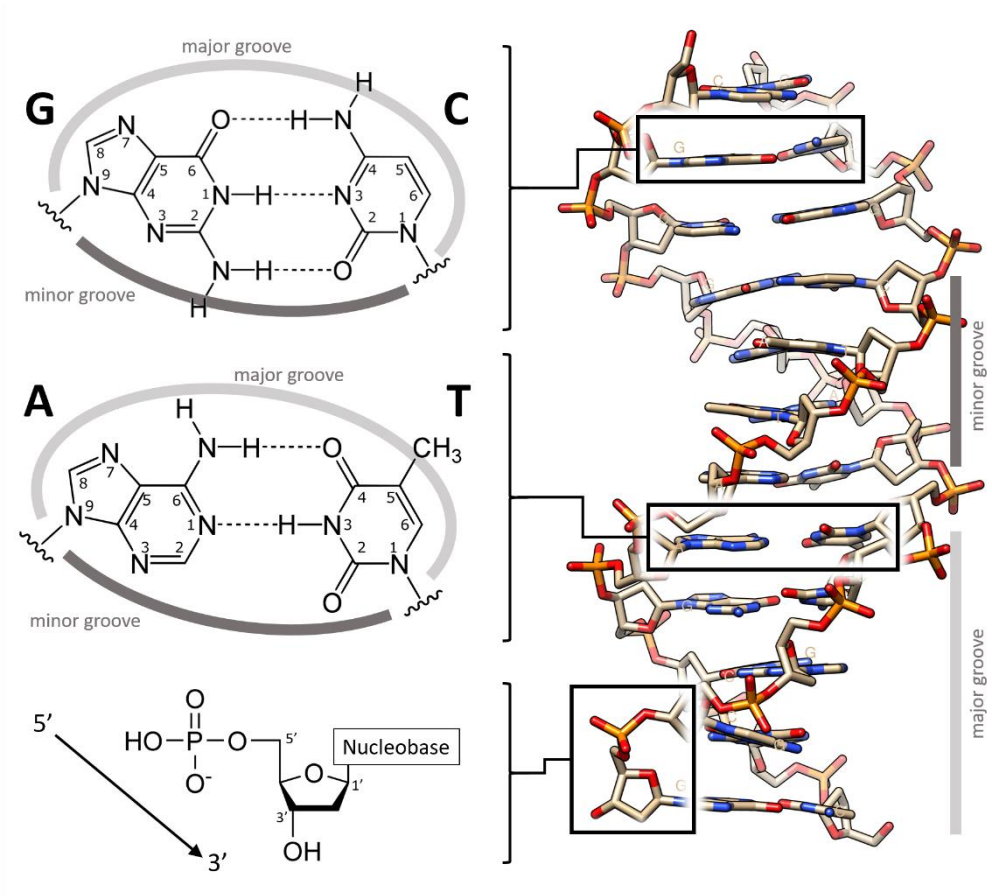


Fig. 4.1 *Structure of double stranded DNA. Crystal structure 1BNA by Drew et al.*[5]

This unique molecular structure serves purpose for two fundamental processes. The order of the nucleobases along the strands allows the storage of biological information as a sequential code that can be transcribed into biologically active structures like RNA (ribonucleic acids), to express its encoded function. The storage as a pair of complementary sequences on the other hand makes replication of genetic data straightforward, as both strands can be utilised as templates for duplication. So whether it is replication that passes genetic information on to future generations, or transcription that acts as a first step in the execution of inherited biological functions; both rely on an error free recognition of the dsDNA structure. Macromolecular parameters like supercoiling or bending of the helix as well as intimate changes to the structure of a single base pair in the sequence have a profound effect on processing the encoded genetic information[6]. Any successful interaction with the DNA is based on recognition of a specific structural element of the DNA, an active site, that proteins can interact with. Structural changes to the DNA can initiate a biological process, such as in the case of topoisomerases to generate an active site by unwinding supercoiled DNA[7,8].

Structural changes can also be induced to prevent the expression of a biological active DNA sequence, a gene, such as in the development of minor-groove binders as therapeutic agents[9–12]. Originally derived from naturally occurring molecules such as Netropsin and Distamycin A[13,14], minor groove binders target a specific sequence to either occupy an active site directly at the minor groove or change the local DNA structure to prevent the recognition of a transcription protein on the major groove[15–17]. Such transcription factor antagonists often rely on non-covalent binding such as the formation of hydrogen bonds to functional groups from nucleobases and are tailored for a high selectivity towards specific DNA sequences. They often contain planar subunits that form crescent shapes with enough flexibility to follow the curvature of the minor groove and require a certain hydrophobicity to efficiently replace the solvation water inside the groove.

Alniss *et al.* has recently published a review paper[18] that classifies minor groove binders according to their thermodynamic driving forces and makes a distinction between entropy- and enthalpy-driven ligands. The replacement of water in the minor groove, the release of counter-ions from the dsDNA backbone and removal of non-polar ligand-structures from a polar solvent environment all contribute to a gain in entropy and direct the ligand into the minor groove. Hoechst 33258 (see **Fig. 4.2a**) is a well-established fluorescent dye and minor groove binder with antibacterial activity[19] that binds due to a gain in entropy[18,20]. The formation of hydrogen bonds as well as attractive Van-der-Waals forces on the other hand can lead to a favourable enthalpy upon binding. An example for enthalpy-driven binders are hairpin polyamides (**Fig. 4.2b**) with anti-tumour activity[21] that form sequence-specific hydrogen bonds with the nucleobases[18,22].

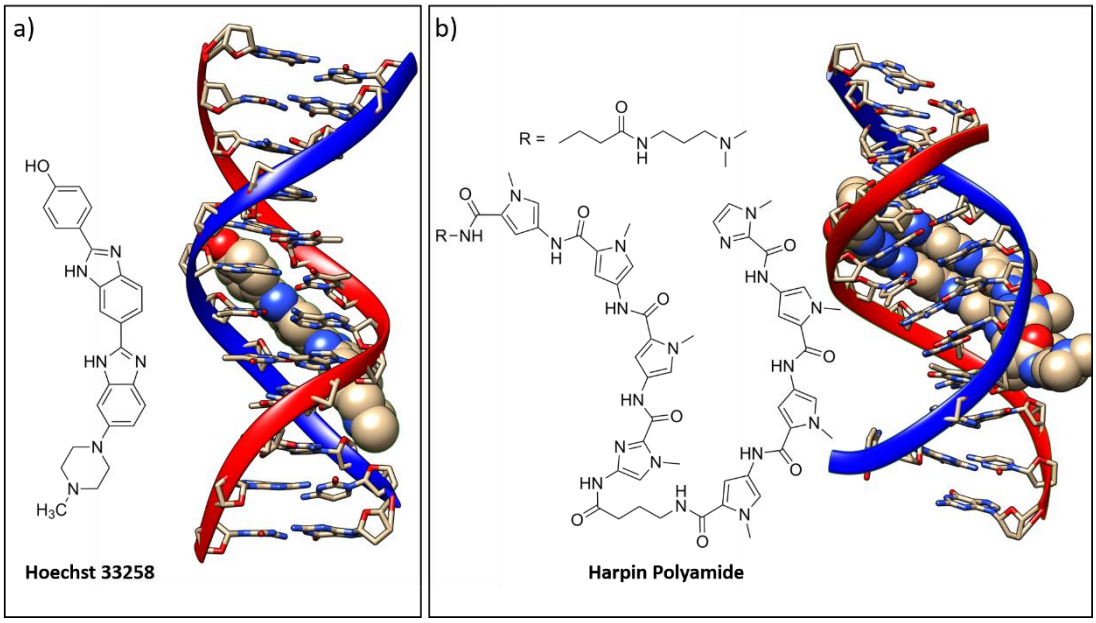


Fig. 4.2 *Minor groove binding of Hoechst 33258 targeting 5'-AATT-3' in* ***a)*** *and binding of a hairpin polyamide targeting 5'-TGTACA-3'. Crystal structure 264D in a) by Vega et al.*[23] *NMR structure 5OE1 in* ***b)*** *by Padroni et al.*[24]

One objective of this thesis is to understand the molecular driving forces behind these ligands in order to guide future development work towards more selective molecules with therapeutic effectiveness. Minor groove interactions can be studied with a variety of different methods to assess either selectivity or structure. DNA footprinting[25,26], DNA melting experiments[24,27] or biophysical assays[28,29] aim to quantify binding affinities or association constants for a range of different ligand-dsDNA combinations to assess whether the sequence recognition is selective and can compete with natural transcription factors. Such large scale experiments often lack in-depth, structural information and are therefore combined with complementary methods such as X-ray crystallography[30,31] and multi-dimensional nuclear magnetic resonance (NMR) spectroscopy[24,32,33], which can in turn be used as starting point for molecular dynamics (MD) simulations[29,32,34]. These labour-intensive, structural methods provide a high level of detail and spatial information that is imperative to structural biology questions. Crystallisations and calculation tasks however can take weeks to complete, so that a complete structural analysis is often reserved for only the most promising dsDNA-ligand combination. Additionally, the experimental data used for MD simulations is often based on static structures from crystals or averaged NMR responses in solution without inherent information about dynamics.

Time-resolved infrared spectroscopy can provide additional insight to biological processes like ligand binding, as it offers structural information on a wide range of accessible timescales down to femtoseconds to resolve ultrafast processes such as hydrogen bond formation[35–37]. Similar to multi-dimensional NMR experiments, a defined sequence of ultrashort infrared laser pulses (as opposed to radiofrequency pulses in NMR) interrogates the sample and captures the connectivity of different vibrational modes within a molecule. The selection of pulse frequencies, polarisations, and timings allow the conception of different pulse-experiments, one of which being two-dimensional infrared (2D-IR) spectroscopy[38]. As will be shown later, 2D-IR experiments can generate a 2D spectrum where cross-peaks between different vibrational modes can identify spatial proximity of molecular groups. The development of those signals over time characterises vibrational energy pathways that allow conclusions to be made about the molecular environment of the oscillating functional group. 2D-IR spectroscopy is label-free, non-destructive, requires small sample volumes and works with liquid samples.

Recent technological advancements have significantly reduced acquisition times of 2D-IR spectra from several hours down to seconds, allowing for increased data output and larger experimental datasets[39–42]. This creates the opportunity to develop a new screening tool to study a broad range of minor groove interactions while simultaneously obtaining structural information in a relatively short amount of time. The larger data throughput also requires development of appropriate analysis tools that can scale with the experiment.

The main objective of this thesis is to study the interaction of minor groove binders with dsDNA using time-resolved infrared spectroscopy and utilising its high data throughput to develop a screening experiment that is scalable with the dataset. After introducing the required spectroscopy methods and instrumental setups in the introduction and methods chapter, chapter 6 will introduce an analysis method in order to study a 2D-IR dataset of Hoechst 33258 interacting with different dsDNA sequences. Minor groove binding of Hoechst 33258 has been studied extensively and is used as a proof-of concept to show how a screening experiment might be implemented with 2D-IR spectroscopy. Chapter 7 extends this approach to a screening experiment of several hairpin polyamide ligands interacting with a set of sequences. The last chapter introduces a time-resolved temperature-jump experiment to study the interaction of Hoechst 33258 with dsDNA upon melting. This experiment focusses on dsDNA melting

dynamics from nanoseconds to milliseconds in order to assess how minor groove binding affects unfolding of the dsDNA. The experiment is also used for preliminary measurements in developing a 2D-IR, temperature-jump experiment that can cover a large temporal window.

All DNA sequences mentioned in this thesis are noted in the usual convention starting from the 5'-position to the 3'-position. 5'-GCAT-3' is simplified in the text as GCAT. Letters within a sequence notation either represent the four nucleobases of DNA directly (guanine - G, cytosine - C, adenine - A, thymine - T), or represent a possible subset of the four (S - either G or C, W - either A or T, N - any of the four). Base pairing is denoted by a dash between nucleobases, such as G-C, or A-T pairs, while a base step is denoted as letters without spacing, like a GT step. Notations such as A/T indicate that either A or T structures are being referred to (similar to W).

4.3 Infrared Spectroscopy

4.3.1 Fundamentals of Infrared Spectroscopy

Infrared spectroscopy is a technique commonly used to study the characteristic absorption profile of molecules in the infrared region of the electromagnetic spectrum. Absorption of infrared light induces molecular vibrational motions that are indicative of functional groups and structural elements within the studied molecules. Any heteronuclear bond within a molecule represents a dipole that can be driven by an oscillating electric field at a characteristic resonance frequency. Carbonyl groups for example will absorb light in a different region of the infrared spectrum than C-H bonds and each molecular structure will therefore generate a unique infrared absorption spectrum. The immediate environment around each bond will further affect its resonance frequency and can therefore act as a label-free probe for intermolecular interactions.

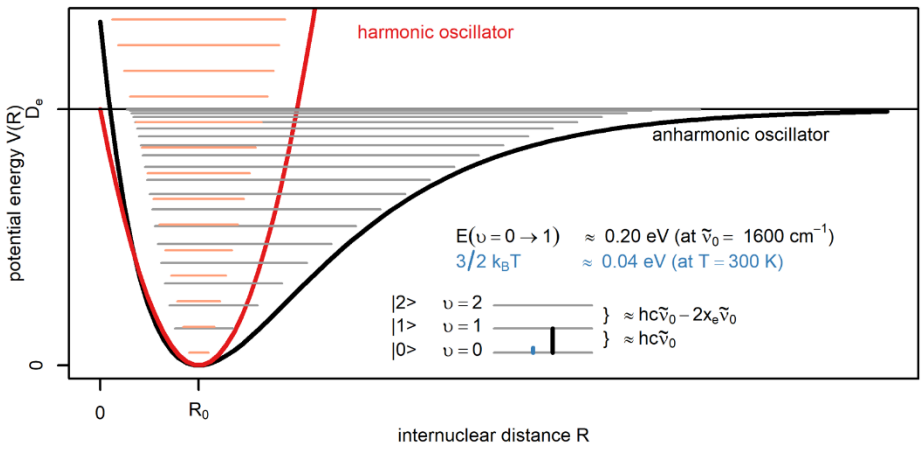


Fig. 4.3 *Potential energy curves of a harmonic (red) and anharmonic (black) oscillator. Anharmonicity causes the spacing of energy levels to reduce with rising quantum number υ. Common infrared absorption features such as a carbonyl stretching mode at 1600 cm^{-1} are caused by a transition from the vibrational ground state to the first excited state. The energy required for this transition (0.2 eV) is significantly higher than the mean kinetic energy of a gas particle at room temperature (0.04 eV).*

It is possible to describe molecular vibrations in a semi-classical approach[43,44]: The length, R_0, of a covalent bond in a diatomic molecule is defined by the minimum of a potential energy curve, $V(R)$, between the two atoms. Any displacement, $R - R_0$, from the initial bond length will increase the potential energy and close to R_0, the energy can be approximated by a parabola (see red curve in **Fig. 4.3**). In this case, the molecular bond acts just like a mechanical spring and undergoes harmonic oscillation due to the symmetry of the parabola. However, the repulsion of the positive nuclei on one side and the dissociation energy of the bond on the other side will change the shape of the curve and introduce an anharmonicity. The potential energy for a diatomic molecule is described empirical by the Morse potential (black curve):

$$V(R) = D_e \cdot \left[1 - e^{-\beta(R-R_0)}\right]^2, \quad \text{with } \beta = \left(\frac{m_{\text{eff}} \cdot \pi c \cdot \tilde{\nu}_0^{\,2}}{\hbar D_e}\right)^{1/2}. \tag{4.1}$$

D_e defines the depth of the well, and β gives the curvature of the potential. The curvature β is furthermore dependent on the effective mass of the two atoms, m_{eff}, the depth of the well, D_e, the speed of light, c and the resonance frequency, $\tilde{\nu}_0$, given in wavenumbers. The potential energy of a molecular bond can only reach discrete vibrational energy levels which can be calculated by solving the Schrödinger equation for the Morse potential. The permitted, vibrational energy levels are:

$$E(\upsilon) = \left(\upsilon + \frac{1}{2}\right) hc\tilde{\nu}_0 - \left(\upsilon + \frac{1}{2}\right)^2 x_e hc\tilde{\nu}_0, \quad \text{with } x_e = \frac{\beta^2\hbar}{2m_{\text{eff}} \cdot \omega} = \frac{\tilde{\nu}_0}{4D_e}. \tag{4.2}$$

The first term of this equation denotes the solution for a harmonic oscillator with equidistant energy levels and is usually enough to explain the most common transitions within linear infrared spectroscopy. A photon with the energy $E = h\nu = hc\tilde{\nu}_0$ is resonant with the transition from the ground state $\upsilon = 0$ into the first excited state $\upsilon = 1$, causing the bond length to oscillate around its equilibrium length R_0. This 0–1 transition is the so-called fundamental vibration and accounts for most of the observed bands in a mid-infrared spectrum. While other modes such as 0–2 transitions (overtones) can also be observed using near-infrared spectroscopy, their probability is much lower. The transition probabilities are summarised in the selection rules and evaluate whether the symmetries of initial and final state are compatible for a transition. The selection rule for a harmonic oscillator is $\Delta\upsilon = \pm 1$ and indicates the low probability for overtones (*forbidden* transitions). The resonance frequency $\tilde{\nu}_0$ is, in analogy to a spring, depending on a force constant, k_f, the effective mass m_{eff} of the two nuclei and can be described by

$$\tilde{\nu}_0 = \frac{1}{2\pi c} \cdot \left(\frac{k_f}{m_{\text{eff}}}\right)^{1/2}, \quad \text{with } \frac{1}{m_{\text{eff}}} = \frac{1}{m_1} + \frac{1}{m_2}. \tag{4.3}$$

A larger force constant k_f and a larger reciprocal effective mass will both increase the resonance frequency of the oscillating bond. An O-H bond with the small mass of a proton on one side will have a higher $1/m_{\text{eff}}$ value than an O-D bond with a heavier deuterium according to equation (4.3). This will in turn increase the resonance frequency $\tilde{\nu}_0$ and cause the absorption bands for H_2O to be at higher wavenumbers than for D_2O.

The second term of equation (4.2) accounts for the anharmonicity of the potential energy curve with x_e being the anharmonicity constant. As the Morse potential takes a dissociation energy into consideration at which the bond will break, the spacing of the vibrational energy levels will get smaller and smaller towards the dissociation energy. So while the equidistant spacing of a harmonic oscillator suggests that every energy level can be reached by the same resonance frequency, an anharmonic oscillator will require slightly smaller resonance frequencies at higher initial energy levels for a transition. This property has not only implications on the appearance of infrared absorption spectra but it is crucial for 2D-IR spectroscopy. Each signal within a 2D-IR spectrum will be based on a peak pair separated due to anharmonicity. In the harmonic

approximation, these signals would completely overlap and simply cancel each other out.

Molecules with more than two nuclei have several different vibrational modes where each can be described by its own, individual potential energy curve and discrete energy levels. For N atoms in a non-linear molecule there are 3N-6 possible, vibrational modes with unique resonance frequencies that are affected by the characteristics of the bonds involved and the individual, electrostatic environment surrounding them. An oscillating dipole in close proximity can therefore perturb the electrostatic environment of another mode and alter its resonance frequency. The two vibrational modes in this case are coupled. Coupling can occur either mechanically through bond interactions or electrostatically through space as outlined in section 4.3.3.3 and is sensitive to the intimate, structural relationship of the coupled modes. Knowledge about the immediate structure at the binding site is key to understand binding processes but conventional infrared spectroscopy is unable to resolve coupling between modes. Off-diagonal features in 2D-IR spectra on the other hand provide a straightforward way of investigating coupling patterns and make it possible follow their development over time.

4.3.2 Susceptibility and Linear Absorption Spectroscopy

In order to explain the fundamental theory of 2D-IR spectroscopy, it is necessary to first give a brief overview over different linear and non-linear, optical phenomena[45], then explain the process of conventional, linear absorption with the tools acquired and finally expand the theory to non-linear processes as employed in 2D-IR spectroscopy[38].

As light interacts with any dielectric medium, its electric field $E(t)$ will induce a time-dependent, macroscopic polarisation $P(t)$ in the material:

$$\frac{P(t)}{\varepsilon_0} = \chi^{(1)}E(t) + \chi^{(2)}E^2(t) + \chi^{(3)}E^3(t) + \cdots + \chi^{(n)}E^n(t), \tag{4.4}$$

where ε_0 is the permittivity of free space and χ is the electric susceptibility. Equation (4.4) gives a general overview of the different possible processes when light interacts with matter:

$\chi^{(1)}$ is the linear susceptibility and encompasses all linear, optical processes such as absorption or dispersion. Infrared light from incoherent sources such as black body emitters will induce an atomic charge separation even at low intensities and therefore

generate a transition dipole moment on the molecular scale, giving rise to a macroscopic polarisation. As there is one incoming and one outgoing electric field, these phenomena are referred to as two-wave-mixing processes.

$\chi^{(2)}$ is the second order term of the susceptibility and has a much smaller magnitude than the first order term. $\chi^{(2)}$-processes will only occur in non-centrosymmetric materials (i.e. on bulk surfaces) and include frequency-doubling, difference frequency generation and optical parametric generation. These processes are commonly used to convert light from one wavelength into another by using birefringent crystals and are used in 2D-IR spectroscopy to generate mid-infrared laser pulses. These are so-called three-wave-mixing processes (for example one incoming field and two outgoing fields in case of optical parametric amplification).

$\chi^{(3)}$ is the third order term and has an even lower magnitude than $\chi^{(2)}$ (relative magnitudes: $\chi^{(1)}:\chi^{(2)}:\chi^{(3)} \approx 1:10^{-8}:10^{-16}$)[46]. Only coherent light sources with high intensities such as lasers are able to induce noticeable third-order-polarisations. Two-photon absorption, pump-probe spectroscopy or 2D-IR spectroscopy are all four-wave-mixing processes that can be studied in common, isotropic media (i.e. bulk solutions).

Conventional, infrared absorption only uses the linear term of the polarisation, such that equation (4.4) reduces to $P(t) = \varepsilon_0 \chi^{(1)} E(t)$. The electric field of a light source can be approximated by a simple cosine function $E(t) = E_0 \cos(\omega t)$, where the frequency is given as angular frequency ω. An incident electric field at the resonance frequency $\omega = \omega_{01}$ can interact with the dipole of a molecule and give rise to a transition dipole moment, μ_{01}, between the ground state $|0\rangle$ and the first excited state $|1\rangle$ (written in bra-ket notation):

$$\mu_{01} = \int_{-\infty}^{\infty} \psi_0^* \cdot \hat{\mu} \cdot \psi_1 dx = \langle 0|\hat{\mu}|1\rangle \approx \frac{d\mu}{dx}\langle 0|\hat{x}|1\rangle, \tag{4.5}$$

where x is the coordinate of the vibrating bond and $\hat{\mu}$ is the electric dipole moment. The derivative $\frac{d\mu}{dx}$ describes the change of the static dipole as the bond is being stretched and compressed and the term $\langle 0|\hat{x}|1\rangle$ gives rise to the aforementioned selection rules. As light interacts with the sample, an ensemble of molecules will be excited to a linear combination of eigenstates $|0\rangle$ and $|1\rangle$ and the polarisation of the sample will be proportional to:

$$P(t) \propto +\mu_{01}^2 \sin(\omega_{01} t). \tag{4.6}$$

Note how the macroscopic polarisation of the sample in equation (4.6) contains a sine function rather than a cosine compared to the incident, electric field. The polarisation is 90° phase-shifted compared to $E(t)$. According to Maxwell's equations, the now oscillating charges emit an electromagnetic wave with its phase shifted by another 90° with respect to the polarisation. The total phase shift of the emitted signal is therefore 180° compared to the incident light and will destructively interfere with it. This will lower the overall light being detected after passing through the sample and generate the absorbance. This process explains the term two-wave-mixing process, as there is one incoming and one emitted field interacting with the sample.

From a quantum mechanical perspective, it is the coherent superposition of eigenstates $|0\rangle$ and $|1\rangle$ that generates the macroscopic polarisation. If the incident light was coming from a pulsed light source (or was simply turned off), the coherent superposition of eigenstates generated during the pulse would slowly lose its phase relationship over time once the pulse has ended. This leads to a loss in polarisation over time and the emitted field from the sample would eventually decay with a dephasing time, T_2. Additionally, any population that was created in the first excited state would eventually relax back to the ground state with the vibrational lifetime, T_1, contributing to the dephasing rate $1/T_2$ via:

$$\frac{1}{T_2} = \frac{1}{2T_1} + \frac{1}{T_2^*}. \tag{4.7}$$

T_2^* is defined as pure dephasing time that is caused by rapid fluctuations of the environment. The dephasing rate in equation (4.7) affects the natural (or homogeneous) linewidths observed in the infrared spectrum with faster dephasing leading to broader absorption features. Additionally, identical molecules in different static, molecular environments will oscillate at a range of different resonance frequencies which will further broaden the infrared features observed. This inhomogeneous broadening will contribute to the loss in macroscopic polarisation and cannot be distinguished from homogeneous broadening by linear infrared spectroscopy alone. 2D-IR spectroscopy is however able to differentiate the two as will be explained below.

The incident light pulse in this case can be described by $E(t) = E'^{(t)}\cos(\omega t)$, with $E'^{(t)}$ being the pulse envelope. The evolution of the emitted field from the sample over time is defined as the molecular response function, $R(t)$. The response function is characteristic to the sample and is only retrievable from the polarisation $P(t)$ as a convolution with the envelope of the laser pulse, $E'^{(t)}$:

$$P^{(1)}(t) = \int_0^\infty dt_1\, E'(t - t_1) \cdot R^{(1)}(t_1). \tag{4.8}$$

The indices in $R^{(1)}$ and $P^{(1)}$ indicate a first order response and polarisation. The processes of this linear absorption experiment can be further visualized by so-called double-sided Feynman diagrams to keep track of the vibrational energy within the system throughout the experiment. The experiment may seem straightforward in the case of linear absorption, but the diagrams will become helpful for more complex pulse sequences later on. The diagrams illustrate the evolution of the density matrix, which describes the statistical average of the interacting ensemble of modes. The density matrix is proportional to the outer product $|n\rangle\langle m|$ of the eigenbasis n with its complex conjugate, m, and can indicate, whether the ensemble average is in the ground state, $|0\rangle\langle 0|$, or the first excited state, $|1\rangle\langle 1|$. Configurations like $|1\rangle\langle 0|$, or $|0\rangle\langle 1|$ signify so-called coherences, where the ensemble average is in a coherent superposition of the eigenstates $|0\rangle$ and $|1\rangle$. The electric field of a light pulse can act on either the ket (left) or the bra (right) side of the density matrix and promotes or de-excites that state by 1.

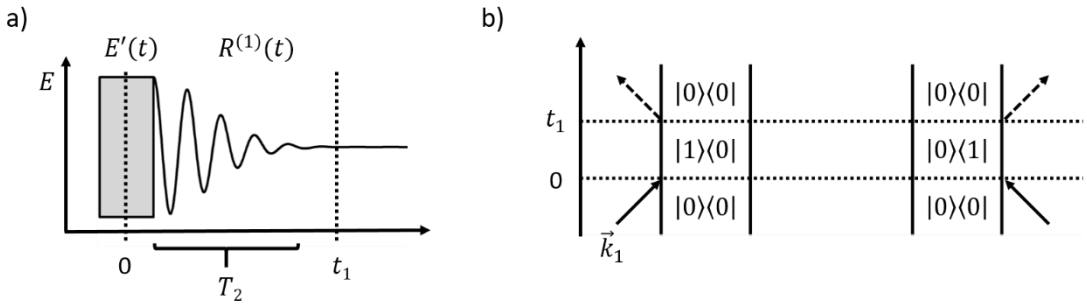


Fig. 4.4a) *Depiction of a conventional infrared absorption experiment with a pulsed light source. An electric field with a square shaped pulse envelope $E'(t)$ interacts with a sample at $t = 0$. This generates a response from the sample that decays with T_2. This characteristic response $R^{(1)}(t)$ is only observable from the macroscopic polarisation of the sample $P^{(1)}(t)$ over time. This process can be represented by double-sided Feynman diagrams in **b)** with an incoming electric field at $t = 0$ and an emitted field at $t = t_1$. Two pathways are possible as the incident field can interact with either side of the density matrix.*

The linear absorption experiment with a pulsed light source can be described by two Feynman diagrams as shown in **Fig. 4.4b**. The diagrams are read from bottom to top. At

time-zero, the electric field of the incident electric field acts on either the ket or the bra side of the density matrix and promotes the ensemble average from the ground state to a coherence. This coherence will have a dephasing time T_2 (assuming no further inhomogeneous broadening) and generate an emitted field from the same ket or bra side while relaxing back to its ground state. The emitted field is detected as part of the macroscopic polarisation and leads to an absorbance peak in the linear IR spectrum.

4.3.3 Two-Dimensional Infrared Spectroscopy

Two-dimensional spectroscopy is based on a pump-probe-type experiment, where a molecule is excited with a laser pulse at one frequency and then probed with a second laser pulse shortly after to study how the vibrational energy pumped into the system has evolved during that time. By varying the frequencies of the pump light and the probe light in this experiment, it is possible to generate a two-dimensional spectrum that uncovers the vibrational, energetic pathways of the molecule. The timing between the pump and the probe pulse can be scanned additionally to obtain the evolution of the two-dimensional spectrum over time. The probe pulse is usually a femtosecond pulse with a broad spectral bandwidth that can be dispersed on to an array detector. The array detector can collect the probe frequency axis of the 2D-IR spectrum within a single measurement, which is denoted as the horizontal axis of the 2D-IR spectra in this thesis. The pump axis (vertical axis) can be obtained either in frequency-domain or time-domain as outlined in **Fig. 4.5**. In the first case, a single pulse is used to excite the sample which is modulated in frequency to collect the pump frequency axis of the 2D-IR spectrum row by row as shown in **Fig. 4.5a**.

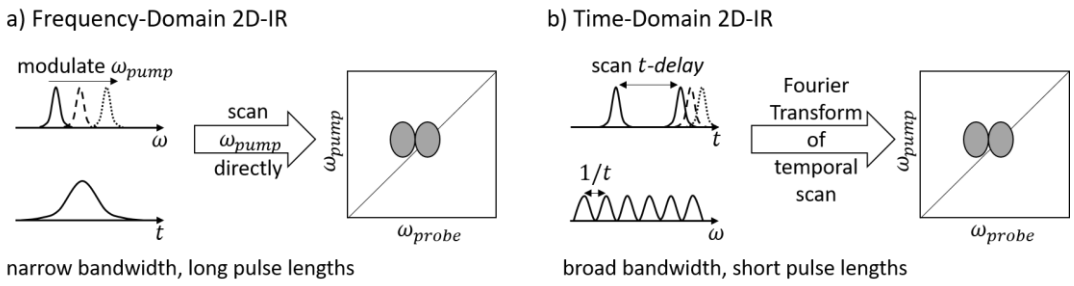


Fig. 4.5a) *Acquisition of a 2D-IR spectrum in frequency-domain using a single pump pulse with narrow bandwidth.* ***b)*** *Time-domain 2D-IR uses a pair of ultrashort pulses to excite several resonance frequencies simultaneously. Scanning different pump-pump delays enables collection of the pump axis in time with subsequent Fourier-transformation to obtain frequencies. The probe frequency axis is usually resolved using an array detector.*

This method requires tuneable, infrared light with a narrow spectral bandwidth to resolve the pump axis. A narrow bandwidth inevitably leads to long pulse durations

due to the time-bandwidth product of transform-limited pulses. This ultimately limits the achievable time-resolution as well as the spectral resolution as will be shown below. Biological samples in particular require a high temporal resolution in order to resolve their characteristic, short vibrational lifetimes, T_1 of few picoseconds[36,38,47]. It is therefore more common to collect the pump axis in time-domain using a pair of ultrashort pump pulses as shown in **Fig. 4.5b**. A pulse pair with a defined time delay will excite several frequencies at once and by scanning the delay between the two pulses it is possible to collect a 2D-IR response in time-domain. The response is then Fourier-transformed to obtain the pump frequency axis for the 2D-IR spectrum. The practical implementation of this method is outlined in more detail in chapter 5.3. The spectral resolution $\Delta\omega$ of this method is defined by the temporal step size Δt by the which the delay is incremented:

$$\frac{\Delta\omega}{2\pi} = \frac{1}{N\Delta t} \quad , \tag{4.9}$$

where N is the number of measurements. This method enables the use of spectrally broad, femtosecond short pulses as pump light for improved spectral- and time-resolution and is therefore preferred in contemporary 2D-IR spectrometers over the frequency-domain approach. The generation of the 2D-IR response in time-domain is explained below.

4.3.3.1 Time-domain 2D-IR spectroscopy

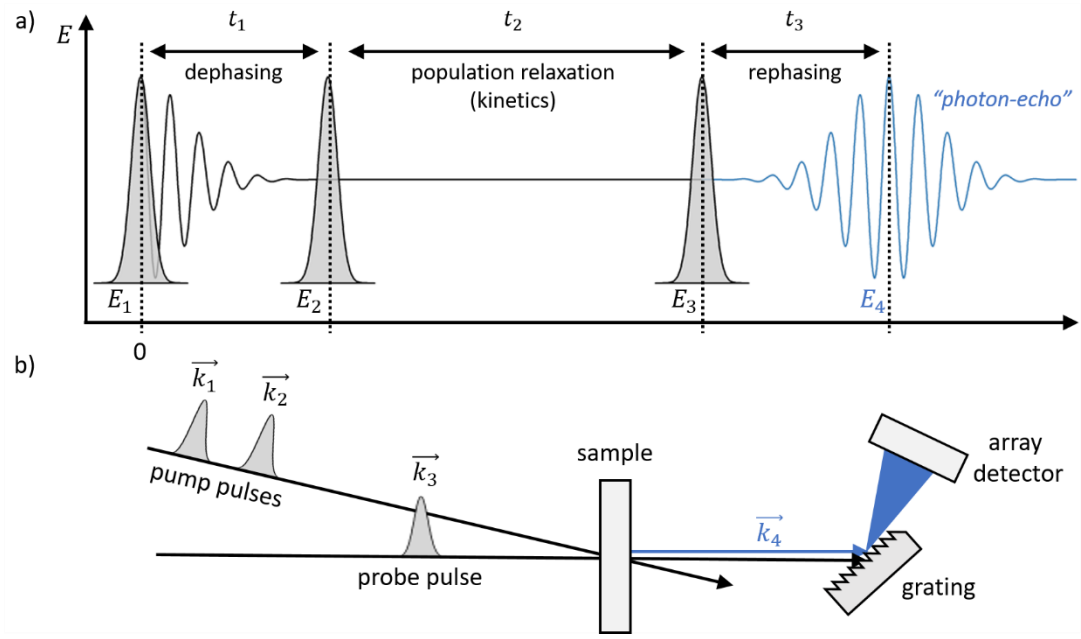


Fig. 4.6a) *Pulse sequence in a time-domain 2D-IR experiment. Two pump-pulses excite the sample and a third pulse probes the excited modes to generate a macroscopic polarisation.* ***b)*** *In pump-probe geometry, both pump-pulses are collinear and the response function of the sample is emitted in the same direction as the probe pulse, self-heterodyning the signal. A spectrometer disperses the signal on to an array detector to obtain the probe frequency. The pump-frequency is obtained after Fourier-transformation of the scanned t_1-times.*

2D-IR spectroscopy utilises the third order term $\chi^{(3)}$ of the susceptibility of the sample. As mentioned earlier, this term is much smaller than the linear susceptibility and can be regarded as a four-wave-mixing process. This terminology can be understood when looking at a 2D-IR experiment in the time-domain (see **Fig. 4.6a**): The electric field of two pump pulses, E_1, E_2, and a probe pulse, E_3, interact with the sample and generate an emitted field, E_4. The third-order polarisation of the system is again a convolution of the three incident laser fields with the desired third-order response function $R^{(3)}$ that carries the 2D-IR signal:

$$P^{(3)}(t) = \int_0^\infty dt_3 \int_0^\infty dt_2 \int_0^\infty dt_1 \, E_3(t-t_3)E_2(t-t_3-t_2)E_1(t-t_3-t_2-t_1) \cdot R^{(3)}(t_3,t_2,t_1). \tag{4.10}$$

The observable, macroscopic polarisation from such an experiment is therefore much more complex than in linear spectroscopy. Long pulses will obscure the response function $R^{(3)}$ and cause lineshape broadening of the features observed in the 2D-IR spectrum. The time-domain 2D-IR approach mitigates this problem by using femtosecond pulses throughout for E_1, E_2 and E_3.

A four-wave-mixing process follows the preservation of momentum of all four fields, so that the directionality of the emitted field, given by the wave vector $\vec{k}_4$, is defined by the three incident fields:

$$\pm\vec{k}_1 \pm \vec{k}_2 \pm \vec{k}_3 = \vec{k}_4. \tag{4.11}$$

Careful alignment of the laser pulses together with a well-defined phase relationship makes it possible to obtain a response function in a specific direction after the sample. This method is called phase matching and is used in non-linear methods to separate the response function of interest from the overall polarisation of the material. Even though there are other setups like the box-CARS geometry as outlined below, this thesis focusses on spectrometers with pump-probe geometry (see **Fig. 4.6b**). In this setup, both initial pump pulses are collinear so that $\vec{k}_1 = \vec{k}_2$. High throughput screening experiments require fast data collection and pump-probe geometry lends itself to pulse shaping (see chapter 5.3.2) which improves acquisition times and ease of data collection. Using this particular geometry, the response functions of six different energetic pathways are attainable, which can be explained by the Feynman diagrams outlined in figure **Fig. 4.7a-f**.

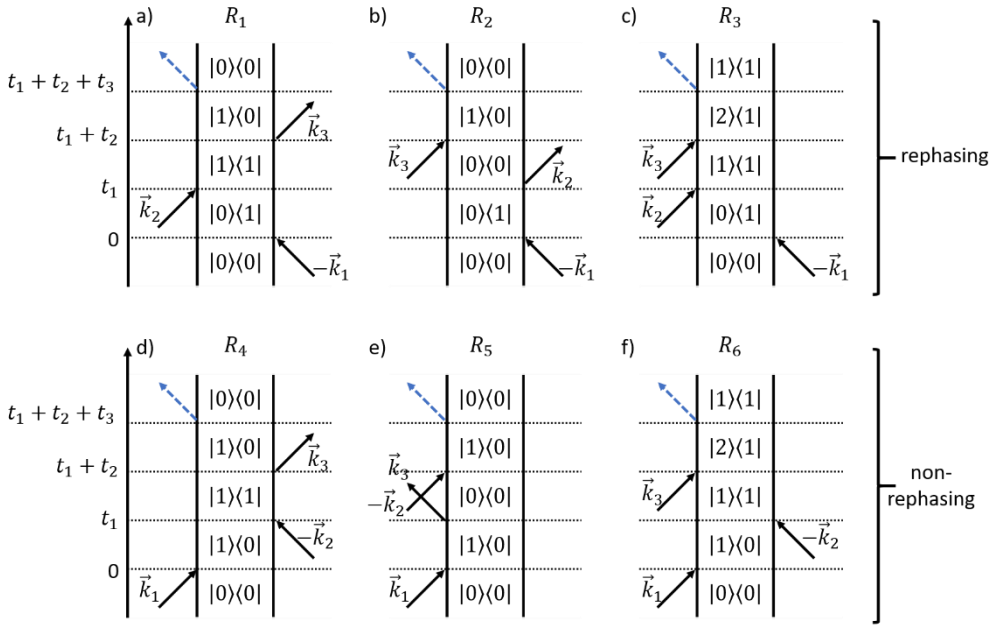


Fig. 4.7 *Feynman diagrams for six different response functions obtainable with pump-probe-geometry.*

In a 2D-IR experiment, the first laser pulse $\vec{k}_1$ will act on either the ket or the bra side of the density matrix and generate a coherence state like in the linear case. An ensemble

of molecules will be in a coherent superposition of states that dephases over time. The second laser pulse $\vec{k}_2$ will generate a population state that either repopulates the ground state $|0\rangle\langle 0|$, or populates the first excited state, $|1\rangle\langle 1|$, with lifetime T_1. The third pulse $\vec{k}_3$ will then transform the ensemble once again in a coherent superposition of states from which it will finally emit an electric field $\vec{k}_4$ with the desired, third-order response function $R^{(3)}$.

Both coherence states during t_1 and t_3 are affected by homogeneous and inhomogeneous dephasing which define the lineshape broadening of the features observed in 2D-IR. If the coherence states during t_1 and t_3 change from a higher quantum state from one side of the density matrix to the other (as in **Fig. 4.7a** to **c**), the inhomogeneous dephasing during t_1 will invert its direction and rephase during t_3 which generates a photon-echo. Inhomogeneous dephasing is caused by different resonance frequencies that remain constant for an individual oscillator throughout the experiment. A faster oscillation will travel further during t_1 than a slower oscillation. When the direction is reversed after $\vec{k}_3$, all oscillations retain their respective frequencies and thus require the same time to return to their initial position. The macroscopic polarisation will therefore peak at $t_3 = t_1$, when all oscillations are rephased (see blue trace in **Fig. 4.6a**). This is similar to a spin-echo experiment in nuclear magnetic resonance (NMR) spectroscopy, where a 180° pulse inverts the direction of the dephasing spins and causes them to rephase to a net magnetic moment[44]. In optical spectroscopy it is the rephasing of the transition-dipole-moments in the response functions R_1, R_2, R_3 that causes a photon echo. These are so-called rephasing pathways. If the coherence states during t_1 and t_3 keep the higher quantum state on the same side of the density matrix (as in **Fig. 4.7d** to **f**), the dephasing continues in the same direction after $\vec{k}_3$ and no photon-echo is generated. These non-rephasing pathways are insensitive to inhomogeneous broadening but will still emit a detectable macroscopic polarisation.

Rephasing pathways are emitted in the phase matching direction of $-\vec{k}_1 + \vec{k}_2 + \vec{k}_3$ while non-rephasing pathways are emitted in the $\vec{k}_1 - \vec{k}_2 + \vec{k}_3$ direction. Collection of the rephasing pathways alone will return a 2D-IR spectrum that contains lineshapes with absorptive as well as dispersive contributions. As a result, each spectral feature is broadened and distorted by phase-twists. In order to obtain a fully absorptive 2D-IR spectrum, the responses from rephasing and non-rephasing pathways need to be added

together which will cancel phase twists and narrow the linewidths. In pump-probe phase-matching geometry, both pump pulses are collinear ($\vec{k}_1 = \vec{k}_2$) so that rephasing and non-rephasing responses are emitted simultaneously in the same direction. As a result, purely absorptive 2D-IR spectra are readily collected in this geometry.

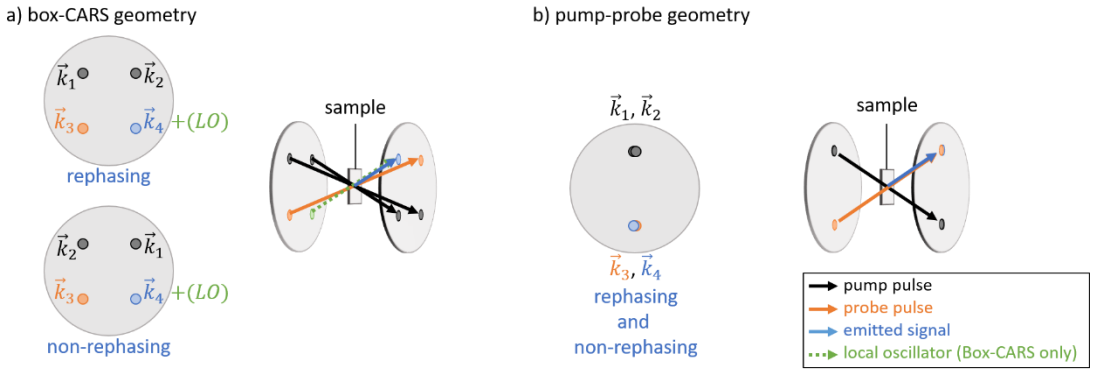


Fig. 4.8 a) *Box-CARS and* ***b)*** *pump-probe phase-matching geometry for generating 2D-IR signals. Switching pump pulses $\vec{k}_1$ and $\vec{k}_2$ in box-CARS enables collection of rephasing and non-rephasing pathways separately while pump-probe geometry allows detection of purely absorptive spectra in one measurement.*

2D-IR spectra have traditionally been acquired in box-CARS geometry, where the direction of the four pulses forms a square shape (as shown in **Fig. 4.8a**). This allows more control over the response functions detected, as switching the time-order of the two pump pulses $\vec{k}_1$ and $\vec{k}_2$ will either emit only rephasing or only non-rephasing pathways in the direction of the detector. Moving the two pump beams closer together illustrates how pump-probe geometry will capture both pathways simultaneously. Box-CARS geometry however requires an additional carrier pulse (local oscillator) that heterodynes the emitted 2D-IR response in order to be able to detect it while pump-probe geometry is self-heterodyned due to collinearity of the signal with the probe pulse $\vec{k}_3$.

Although technically being a photon-echo experiment with three individual laser pulses, the 2D-IR signal generation can also be explained using pump-probe terminology for a better understanding (especially since pump-probe geometry is employed). The first two laser pulses are considered pump-pulses that populate a first excited state of the system. The third pulse interrogates the excited system with a much lower energy than the pump pulses and drives the coherence that generates the emitted field $\vec{k}_4$ with the desired response-function. It is therefore referred to as the probe-pulse. The residual probe-pulse after the sample is collinear with the emitted field (pump-probe geometry) and acts as a local oscillator that self-heterodynes the

signal in order to conserve phase information of the signal and enable detection with commercially available infrared detectors. This signal can be dispersed with a grating on an array detector to generate the probe frequency axis of the 2D-IR spectrum. The time-delay between the two pump pulses, t_1, can be scanned and Fourier-transformed to obtain the pump-frequency axis of the 2D-IR spectrum. It is furthermore possible to change the delay between the second pump and probe pulse, t_2, to study the population relaxation and obtain a stack of time-resolved 2D-IR spectra.

4.3.3.2 Structure of the 2D-IR spectrum

The generation of a 2D-IR spectrum from experimental data is outlined in more detail in chapter 5.3.3. The purely absorptive 2D-IR spectrum of a single oscillator such as a carbonyl stretch vibration contains a negative signal characteristic to the 0–1 transition and a positive signal characteristic to the 1–2 transition of the oscillator as shown in **Fig. 4.9a** Red indicates negative, blue indicates positive values throughout. Vertical and horizontal axes of the 2D-IR spectra in this thesis are referred to as pump and probe axes (following pump-probe terminology) and denoted in wavenumbers. The coherences generated after $\vec{k}_1$ will dictate the frequency position of the signals along the pump-frequency (vertical position in the spectrum) and $\vec{k}_3$ will define positions on the probe-frequency axis (horizontal position). All pathways in **Fig. 4.7** show a coherence after $\vec{k}_1$, where the ensemble oscillates between the ground state and the first excited state. Every signal generated in this system will therefore be positioned at the same resonance frequency of the 0–1 transition along the pump-frequency axis in **Fig. 4.9a**. Four out of the six pathways (R_1, R_2, R_4, R_5) have a coherence after $\vec{k}_3$ that will emit a field with the frequency of the same 0–1 transition. These response functions are responsible for the negative signal positioned along the diagonal in the 2D-IR spectrum. The remaining two pathways (R_3, R_6) show a coherence after $\vec{k}_3$ that emit a field with the frequency resonant to the 1–2 transition of the oscillator and therefore generate a positive signal positioned at lower energies along the probe axis. Using pump-probe terminology, the two signals would be regarded as ground state bleach (GSB – negative) and excited state absorption (ESA – positive signal) respectively. In a pump-probe experiment, the GSB appears negative due to stimulated emission from the excited state (in analogy with R_1, R_4) and a reduced absorption from a depleted ground state (R_2, R_5, bleaching), while the ESA is positive due to absorption from the excited state (R_3, R_6).

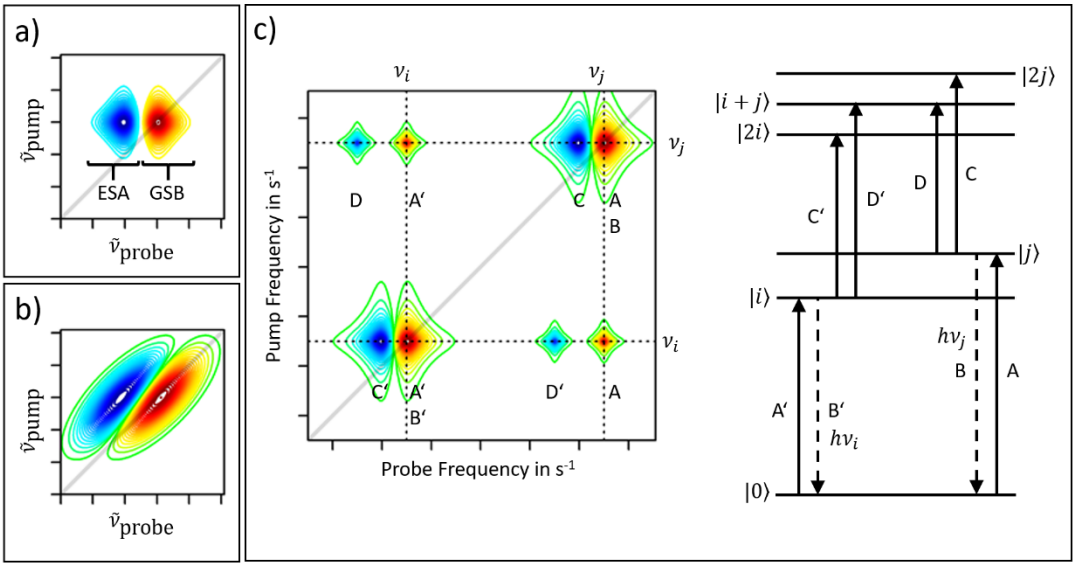


Fig. 4.9a) *Purely absorptive 2D-IR spectrum of a single oscillator.* ***b)*** *2D-IR spectrum of an ensemble of molecules with slightly different resonance frequencies.* ***c)*** *2D-IR spectrum of two coupled oscillators and corresponding energy diagram. The vertical pump axis is resolved by scanning t_1, the horizontal probe axis is resolved by the spectrometer and array detector. Red indicates negative-, blue indicates positive values.*

A single oscillator will return a single peak pair with Lorentzian peak shapes due to homogeneous broadening (see **Fig. 4.9a**). However, an ensemble of molecules in solution will usually experience different molecular environments that affect the individual resonance frequency of each molecule. This will lead to an inhomogeneous lineshape broadening of the 2D-IR signal along the diagonal and generate elliptic, gaussian peaks (**Fig. 4.9b**). As molecules diffuse through different environments over time, their 2D-IR lineshape in will become less elliptic. This so-called spectral diffusion can be quantified by time-resolved 2D-IR measurements to study the dynamic environment of the oscillators.

Larger molecules and mixtures in solution contain different oscillators that can be coupled mechanically through-bonds or electrostatically through-space as mentioned before. Coupling will extend the potential energy curves of two oscillators i and j to a joined, two-dimensional potential energy surface with shared vibrational quantum states. **Fig. 4.9c** exemplifies a 2D-IR spectrum of two coupled oscillators together with a depiction of the vibrational energy states of that system. Depending on the strength of coupling, the shared quantum states will still maintain characteristics from the two uncoupled oscillators, so that the former individual, first excited states of each mode are now represented by $|i\rangle$ and $|j\rangle$. The former, second excited states are now

represented by $|2i\rangle$ and $|2j\rangle$ respectively and a combination band $|i+j\rangle$ is present between them.

The 2D-IR spectrum of the two coupled oscillators can be explained by using the coherence between the shared ground state $|0\rangle$ and $|j\rangle$ after pump pulse $\vec{k}_1$ as a starting point. This will give rise to all 2D-IR features observed at a pump frequency of ν_j. The depletion of the shared ground state after $\vec{k}_1$ will make fewer oscillators available for additional coherences between $|0\rangle$ and $|j\rangle$ (transition A) or between $|0\rangle$ and $|i\rangle$ (transition A') after the probe pulse $\vec{k}_3$. This *bleaching* leads to two negative features; one on the diagonal (A) at ν_i and one at a lower probe frequency of ν_i (A'). The two pump pulses can cause a population state $|j\rangle\langle j|$ after $\vec{k}_2$, from which three transitions B, C and D are accessible. B and C correspond to stimulated emission and excited state absorption, similar as with a single oscillator and lead to a negative peak on the diagonal (B) and a positive peak at slightly lower probe frequency next to the diagonal (C). The distance between both features in the spectrum is characteristic to the anharmonicity of the individual oscillator j. More importantly, a transition D to the combination band $|i+j\rangle$ is possible which leads to a positive peak at a probe frequency even lower than ν_i. The distance that separates the off-diagonal peak pair is referred to as off-diagonal anharmonicity and rises with coupling strength. If the combination band was simply the sum of transition frequencies of A and A', then the off-diagonal peak pair would overlap and completely cancel each other, as in the case of no coupling.

4.3.3.3 Vibrational Coupling

Coupling through space arises due to electrostatic interactions between transition dipole moments of different modes. Through-space coupling extends the linear equation (4.5) for the transition dipole moment by a quadratic term that mixes the displacement of two individual modes and removes their independence from another[48]. This additional cross-term represents an electrical anharmonicity that contributes to the visibility of the combination bands D and D' in **Fig. 4.9c**. The strength of the through-space coupling can be estimated by the "transition dipole coupling"-model that treats each transition dipole as a single vector. The relative angle between vectors will affect the relative intensities of the coupled modes in the 2D-IR spectrum and their distance from another will contribute to the off-diagonal peak-pair separation (i.e. distance between D and A'). The dipole approximation provides a useful model system for long-range interactions[49], where the separation between the dipoles is large

compared to the transition dipole vector lengths. Excitation in real molecules will however induce the displacement of a charge density in the molecule rather than a simple dipole. In these cases, more sophisticated models such as the "transition density derivation distribution"-model could be used to estimate the extent of through-space coupling.

Additionally, mechanical coupling through bonds can also occur in molecules. The restoring force of an oscillator in a larger molecule will be influenced by additional bonds to that oscillator and induce an anharmonicity in the potential energy surface. This is in analogy to the empirical Morse potential found for diatomic molecules and can be generalised to a polyatomic molecule by a Taylor expansion for the potential energy around the equilibrium geometry:

$$V = V(0) + \sum_i \left(\frac{\partial V}{\partial x_i}\right)_0 x_i + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 V}{\partial x_i \partial x_j}\right)_0 x_i x_j + \frac{1}{6}\sum_{i,j,k}\left(\frac{\partial^3 V}{\partial x_i \partial x_j \partial x_k}\right)_0 x_i x_j x_k + \cdots, \tag{4.12}$$

where x_i, x_j, x_k are different vibrational displacements in the same molecule. At equilibrium ($x_i = 0$), the first two terms are zero. The quadratic term shows a cross-term of two different vibrational displacements ($x_i x_j$, with $i \neq j$) in the same molecule. This cross-term accounts for collective, harmonic oscillations of different atoms in a molecule, known as normal modes. Transformation from atomic displacements x_i to normal coordinates Q_i allows to represent the complex motions in a molecule as independent, harmonic oscillations, each with a parabolic potential energy curve. The cubic term in equation (4.12) however removes the independence of the normal modes again and introduces a cross-term that leads to an anharmonic, multidimensional potential energy surface. This again gives rise to the appearance of combination bands as a result of non-equidistant energy levels and the visibility of the combination bands D and D' in **Fig. 4.9c**. Mechanical coupling can affect the spacing between levels and therefore affect the resonance frequencies in case of strong coupling. Geometry optimisation models such as density functional theory allow an estimation of the mechanical coupling strength.

These interactions lead to a complex vibrational energy landscape for large molecules. The vibrational modes of the nucleobases in dsDNA are coupled within a single

nucleotide, between nucleotides along the strand and across strands, forming a coupled network that is sensitive to structural changes such as base stacking or base pairing. Structural changes that weaken or strengthen a molecular bond such as hydrogen bonding will further affect the resonance frequency of the corresponding vibrational mode and shift its peak pair towards higher or lower wavenumbers.

By scanning the time delay t_2 between pump and probe pulse, it is possible to observe how the vibrational energy induced by the pump dissipates over time. The population in the excited states after $\vec{k}_2$ will return to the ground state and the corresponding signals in the 2D-IR spectrum will decay. 2D-IR signals from nucleobases in the centre of dsDNA structures show a short relaxation within 1 ps[37,47,50]. Off-diagonal features generated from coupling interactions will decay with the same rate as their corresponding diagonal features. The vibrational lifetime of each mode is dependent on its individual molecular environment to allow dissipation of the energy. Energy from the nucleobases can also be transferred to vibrational modes of other structures including the phosphate backbone[47], which will be apparent as additional off-diagonal peaks between both modes that rise and decay over time. The temporal response of each cross-peak thus gives an indication whether two modes are coupled, or an energy transfer process occurs between them.

4.3.4 Infrared Spectroscopy of DNA

DNA structures have been studied by several infrared methods, including linear absorption[51–53], transient pump-probe[54–56] and two-dimensional[57–60] infrared spectroscopy. Research has focussed on different regions of the infrared spectrum such as NH/NH_2 and OH stretching modes[61] above 3000 cm^{-1}, carbonyl and ring-modes from nucleobases[62–64] around 1600 cm^{-1} or phosphate vibrations from the backbone[47,65,66] around 1000 cm^{-1}. The region between 1550 and 1700 cm^{-1} is of particular interest for studying sequence recognition, as each nucleobase shows characteristic spectral features in this window. In aqueous solutions, these features are usually not observable due to a large absorption peak of water (H_2O bending) at 1640 cm^{-1}. Although it is possible to collect 2D-IR data from aqueous solutions at high sample concentrations using thin liquid films (< 6 µm)[67], high humidty[37], or gel phases[68], the most practical solution is replacement of the solvent with deuterated water, which shifts the solvent peak to 1250 cm^{-1}.

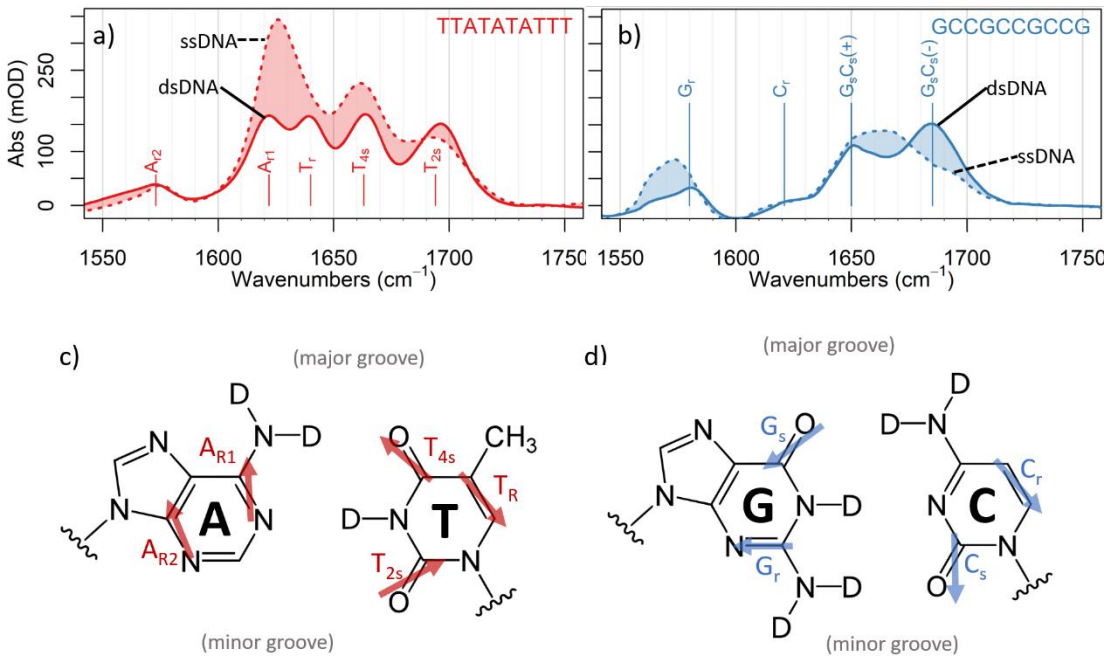


Fig. 4.10 *Conventional, linear infrared absorption spectra of 5'-TTATATATTT-3' in* ***a)*** *and 5'-GCCGCCGCCG-3' in* ***b)*** *at 10mM concentration in 100mM Tris buffer. Baseline features from the solvent were subtracted. Solid lines show the response of dsDNA at room temperature, dashed lines show the same sample 14°C above the melting point in a) and 30°C above the melting point b), characterising single stranded DNA.* ***c)*** *and* ***d)*** *qualitatively depict transition dipole moments for all four, deuterated, non-base-paired nucleobases as reported by Lee et al.*[58]

Fig. 4.10a and **b** show infrared absorption spectra of two dsDNA sequences, 5'-TTATATATTT-3' and 5'-GCCGCCGCCG-3' (and corresponding complementary strand) at room temperature (solid lines). Each of the two Watson-Crick base pairs is represented by a set of coupled, vibrational modes with characteristic resonance frequencies that can be used to differentiate A-T from G-C structures spectroscopically. However, the exact spectral response is defined by the specific geometry in the dsDNA. A perturbation to the structure such as melting of the double strand will dissociate the base pairs, remove base stacking interactions and reorganise the solvation shell, which has a noticeable effect on the infrared absorption spectrum. Dashed lines in **Fig. 4.10** show the spectrum of the same sequences above the melting point, characterising two separate single strands rather than a DNA double strand. A comprehensive study of DNA absorption features in this spectral region was carried out by Lee and co-workers in 2006. In a series of four computational papers, they investigated how base pairing, stacking interactions, coupling, solvation, deuteration, isotope labelling or DNA conformation (A-, B-, Z-Form) affect carbonyl- and ring modes of the nucleobases using density functional theory (DFT) calculations. The location of the transition dipoles for the four deuterated, non-base-paired, nucleobases is qualitatively illustrated in **Fig.**

4.10 as originally published by Lee[58]. Upon base pairing, the transition dipole moments of the nucleobase structures will change, while remain fairly localised to the corresponding functional groups, apart from carbonyl stretches G_s of guanine and C_s of cytosine. Strong coupling of these modes in the G-C base pair leads to two delocalised, collective modes $G_sC_s(+)$ and $G_sC_s(-)$ with symmetric and anti-symmetric stretching of both carbonyls in phase.

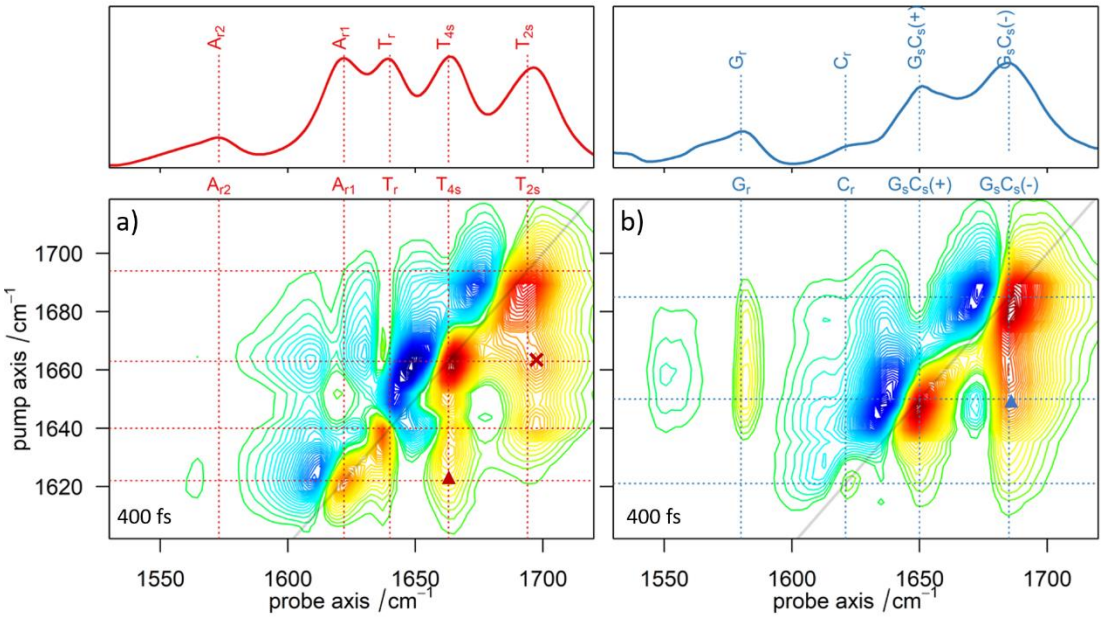


Fig. 4.11 *2D-IR spectra of both dsDNA sequences at a pump-probe time delay of 400 fs. Absorption features from linear absorption measurements align with negative signals observed with 2D-IR spectroscopy along the diagonal, indicating 0-1 transitions.*

The 2D-IR spectrum of both dsDNA sequences at room temperature is shown in **Fig. 4.11**. Each absorption peak observed with linear infrared absorption spectroscopy appears as a peak pair with a negative signal (0-1 transition) positioned on the diagonal and a positive signal (1-2 transition), horizontally offset due to anharmonicity in the 2D-IR spectrum. Off-diagonal features in **Fig. 4.11a** mostly indicate coupling between modes within the same nucleobase (i.e. T_{2s} with T_{4s}, red cross) as well as across base pairs (i.e. T_{4s} with A_{r1}, red triangle). 2D-IR melting experiments indicated that cross-peaks between modes from A and T structures (red triangle) disappear during unfolding[50]. The T_{2s} mode at 1692 cm^{-1} in **Fig. 4.11a** represents a particularly useful marker band for minor groove binding, as the corresponding C2-carbonyl group of T points towards the minor groove of dsDNA structures. Perturbations to this band are used in chapter 6 to identify minor groove binding of Hoechst 33258[69,70]. The amplitude of the A_{R1} ring mode at 1623 cm^{-1} has proven sensitive to base stacking

interactions[55] and is used as marker band for unfolding processes in chapter 8. 2D-IR spectra of pure G/C-structures such as in **Fig. 4.11b** were previously studied regarding coupling between and along dsDNA strands[62,71], indicating that the cross-peaks in **Fig. 4.11b** (blue triangle) originate from coupling across strands. While off-diagonal features from coupling are dominant in this spectral region and show vibrational relaxation on the order of 1 ps, energy transfer has also been observed from A/T-structures to modes of the phosphate backbone further down at 1080 cm^{-1} [47].

Peng *et al.* has studied all four dsDNA nucleotides individually using 2D-IR spectroscopy and quantified all observable, intramolecular coupling constants between 1500 and 1750 cm^{-1} for each nucleotide[63]. While such a detailed analysis allows the calculation of relative orientations of transition dipoles, a longer dsDNA oligomer with mixed contributions from G-C and A-T base pairs will lead to a convoluted response in the 2D-IR spectrum where cross-peaks from multiple coupled modes overlap. This convolution of peaks is already visible in the off-diagonal region of relatively simple sequences in **Fig. 4.11**. Melting experiments have alleviated this problem to some extent and identified cross-peak patterns sensitive to base pairing[50,72], but a mixed sequence containing both G-C and A-T base pairs still remains challenging to analyse in detail.

An additional complexity arises from possible coupling along the strand. The 2D-IR response between the consecutive dsDNA sequence 5'-GGGGGCCCCC-3' and alternating dsDNA sequence 5'-GCGCGCGCGCGCGCG-3' was investigated using 2D-IR spectroscopy and identified significant coupling between nearest neighbour base pairs[71]. Unique coupling parameters were identified for GG/CC steps as opposed to GC and CG steps due to a different overlap of nucleobases and dsDNA geometry in each case. As shown, a coupling model for both sequences could be developed by rigorous analysis of experimental 2D-IR data via model structures from DFT calculations. The combination of all four base pairs however leads to ten different nearest neighbour structures at varying geometries, which would all need to be parameterised by a coupling model for a comprehensive analysis. Such a *bottom-up* approach may not be practical for large screening applications, as each dsDNA-ligand interaction will additionally induce a unique perturbance to the coupled network of vibrational dsDNA modes. The following chapters will rather focus on relative differences across samples to obtain an insight how the 2D-IR response of key dsDNA-ligand combinations differentiates from the unspecific bulk response.

4.3.5 Combination with Temperature Jump Experiments

The temporal resolution of 2D-IR spectroscopy also allows investigation of ultrafast dynamic processes and was previously used to examine interactions at the dsDNA-water interface[37,56,61,73]. The spectral region above 3000 cm^{-1} covers OH-stretching modes characterising the hydration shell of the dsDNA as well as NH/NH_2-stretching modes at the nucleobases, which are directly involved in Watson-Crick base pairing via hydrogen bonding. Studies at different hydration levels suggested short vibrational lifetimes of NH/NH_2-stretching modes of 300 fs and an increased rigidity of the inner hydration shell (using anisotropy measurements of the OH-stretch) surrounding the phosphate groups. 2D-IR experiments in the region of 1000 cm^{-1} indicated structural fluctuations of the phosphate backbone with a lifetime of 300 fs using lineshape analysis of the phosphate bands on the diagonal of the 2D-IR spectrum. These experiments identified dsDNA as highly dynamic structures affected by solvent fluctuations on the order of femtoseconds. The solvated phosphate backbone acts as an efficient sink for excess energy from the dsDNA to the solvent network[37].

These processes are supported by efficient intramolecular energy transfer processes within the dsDNA[47] and contribute to the short vibrational lifetimes observed in the nucleobase region at 1550-1700 cm^{-1}. The short lifetimes limit the timescales observable with 2D-IR spectroscopy to a few picoseconds. In order to examine DNA-ligand interactions over a longer period of time, it is therefore necessary to introduce a perturbation that disrupts the equilibrium of the system. Temperature-Jump (T-Jump) experiments excite vibrational modes of the solvent water, which relax within 1-2 ps and dissipate the pumped energy to other modes[74], initialising additional process such as dsDNA unfolding. These experiments can be implemented with a variety of pump sources that evolved from electronic discharges with microsecond time-resolution[75] to the application of laser pulses that allow studies down to the picosecond timeframe[56]. Monitoring of the perturbed system can be achieved using fast optical methods such as time-resolved UV[76], infrared[77] or Fluorescence[78] spectroscopy, although other methods are possible as well if slower timescales are sufficient[79].

This perturbation technique has recently been utilised in infrared spectroscopy to follow dsDNA unfolding over timescales that reach far beyond the vibrational lifetimes of the probed modes[80]. In these methods, the oligomer is held just below the melting point and a laser pulse resonant with an absorption band of the water generates a rise

in temperature high enough to initiate dissociation of the double strand. A second infrared pulse then probes the response of the vibrational modes and the pump-probe time-delay between the pulses is scanned in order to obtain kinetic information up to milliseconds. Minor groove binding has a stabilising effect on the dsDNA structure and its unfolding mechanism[81]. This is used in melting experiments to estimate the binding affinity of a ligand to a certain dsDNA sequence. Understanding the molecular interactions involved in the stabilisation could shed further light on a ligands selectivity and is therefore studied in the last chapter of this thesis using Laser-induced temperature jumps. This work also serves as preliminary data in developing a T-Jump 2D-IR experiment in the future. Such an experiment would allow acquisition of detailed structural information over a large temporal window, opening up new possibilities to study nonequilibrium processes between a ligand and the minor groove.

Previous 2D-IR studies on dsDNA focussed on a bottom-up approach analysing the vibrational response of individual dsDNA sequences in detail and characterising all observed features with respect to calculated model spectra. Recent 2D-IR experiments investigated interactions of dsDNA with a minor groove binder (Hoechst 33258) for the first time[69], verifying 2D-IR spectroscopy as a viable method to study dsDNA-ligand interactions. Using the existing literature as a starting point, this thesis aims to study minor groove binding via 2D-IR in a top-down approach. Using multivariate analysis techniques, relative differences across a large sample sets are differentiated and categorised. The spectroscopic differences that distinguish samples from another are then analysed to identify how the vibrational modes of dsDNA is affected by the interaction with a ligand. Chapter 6 will introduce an analytical tool in a proof of concept with the well-established Hoechst 33258 ligand to assess over 2000 2D-IR spectra efficiently. Chapter 7 will apply the technique in a more relevant context investigating the 2D-IR response of three different ligand candidates and a reference ligand across 8 dsDNA sequences. The final chapter then assesses time-resolved infrared experiments with short acquisition times as a scalable tool to study non-equilibrium processes of dsDNA-ligand combinations.

4.4 References

1. Watson, J. D. The Structures of DNA and RNA. in *Molecular Biology of the Gene, Fifth Edition* 1–33 (Cold Spring Harbor Laboratory Press, 2003).

2. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).

3. SantaLucia, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci.* **95**, 1460–1465 (1998).

4. Guéron, M., Demaret, J.-P. & Filoche, M. A Unified Theory of the B-Z Transition of DNA in High and Low Concentrations of Multivalent Ions. *Biophys. J.* **78**, 1070–1083 (2000).

5. Drew, H. R. *et al.* Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci.* **78**, 2179–2183 (1981).

6. Sinden, R. R. *DNA Structure and Function*. *DNA Structure and Function* **282**, (Elsevier, 1994).

7. Postow, L., Crisona, N. J., Peter, B. J., Hardy, C. D. & Cozzarelli, N. R. Topological challenges to DNA replication : Conformations at the fork. *PNAS* **98**, 8219–8226 (2001).

8. Pommier, Y., Sun, Y., Huang, S. N. & Nitiss, J. L. Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. *Nat. Rev. Mol. Cell Biol.* **17**, 703–721 (2016).

9. Cozzi, P. Recent outcome in the field of distamycin-derived minor groove binders. *Farmaco* **55**, 168–173 (2000).

10. Dervan, P. B., Poulin-Kerstien, A. T., Fechter, E. J. & Edelson, B. S. Regulation of gene expression by synthetic DNA-binding ligands. *Top. Curr. Chem.* **253**, 1–31 (2005).

11. Barrett, M. P., Gemmell, C. G. & Suckling, C. J. Minor groove binders as anti-infective agents. *Pharmacol. Ther.* **139**, 12–23 (2013).

12. Withers, J. M. *et al.* DNA Minor Groove Binders as Therapeutic Agents. in *Comprehensive Supramolecular Chemistry II* 149–178 (Elsevier, 2017). doi:10.1016/B978-0-12-409547-2.12561-2

13. Arcamone, F., Orezzi, P. G., Barbieri, W., Nicoletta, V. & Penco, S. Distamycin A. Isolation and structures of the antiviral agent distamycin A. *Ann. Gazz Chim Ital* **97**, 1097–1115 (1967).

14. Finlay, A. C., Hochstein, F. A., Sobin, B. A. & Murphy, F. X. Netropsin, a New Antibiotic Produced by a Streptomyces. *J. Am. Chem. Soc.* **73**, 341–343 (1951).

15. Ohyama, T. *DNA Conformation and Transcription*. (Landes Bioscience, Eurekah.com, 2005).

16. Pabo, C. O. & Sauer, R. T. Transcription Factors: Structural Families and Principles of DNA Recognition. *Annu. Rev. Biochem.* **61**, 1053–1095 (1992).

17. White, C. M., Heidenreich, O., Nordheim, A. & Beerman, T. A. Evaluation of the Effectiveness of DNA-Binding Drugs To Inhibit Transcription Using the c-fos Serum Response Element as a Target †. *Biochemistry* **39**, 12262–12273 (2000).

18. Alniss, H. Y. Thermodynamics of DNA Minor Groove Binders. *J. Med. Chem.* acs.jmedchem.8b00233 (2018). doi:10.1021/acs.jmedchem.8b00233

19. Ranjan, N. *et al.* Selective Inhibition of Escherichia coli RNA and DNA Topoisomerase I by Hoechst 33258 Derived Mono- and Bisbenzimidazoles. *J. Med. Chem.* **60**, 4904–4922 (2017).

20. Han, F., Taulier, N. & Chalikian, T. V. Association of the Minor Groove Binding Drug Hoechst 33258 with d(CGCGAATTCGCG) 2 : Volumetric, Calorimetric, and Spectroscopic Characterizations †. *Biochemistry* **44**, 9785–9794 (2005).

21. Kurmis, A. A., Yang, F., Welch, T. R., Nickols, N. G. & Dervan, P. B. A pyrrole-imidazole polyamide is active against enzalutamide-resistant prostate cancer. *Cancer Res.* **77**, 2207–2212 (2017).

22. Pilch, D. S. *et al.* Binding of a hairpin polyamide in the minor groove of DNA: sequence-specific enthalpic discrimination. *Proc. Natl. Acad. Sci.* **93**, 8306–8311 (1996).

23. Vega, M. C. *et al.* Three-dimensional crystal structure of the A-tract DNA dodecamer d(CGCAAATTTGCG) complexed with the minor-groove-binding drug Hoechst 33258. *Eur. J. Biochem.* **222**, 721–726 (1994).

24. Padroni, G., Parkinson, J. A., Fox, K. R. & Burley, G. A. Structural basis of DNA duplex distortion induced by thiazole-containing hairpin polyamides. *Nucleic Acids Res.* **46**, 42–53 (2018).

25. Neidle, S. DNA minor-groove recognition by small molecules (up to 2000). *Nat. Prod. Rep.* **18**, 291–309 (2001).

26. Swalley, S. E., Baird, E. E. & Dervan, P. B. Effects of γ-Turn and β-Tail Amino Acids on Sequence-Specific Recognition of DNA by Hairpin Polyamides. *J. Am. Chem. Soc.* **121**, 1113–1120 (1999).

27. Amirbekyan, K. *et al.* Design, Synthesis, and Binding Affinity Evaluation of Hoechst 33258 Derivatives for the Development of Sequence-Specific DNA-Based Asymmetric Catalysts. *ACS Catal.* **6**, 3096–3105 (2016).

28. Han, Y.-W. *et al.* Binding of hairpin pyrrole and imidazole polyamides to DNA: relationship between torsion angle and association rate constants. *Nucleic Acids Res.* **40**, 11510–11517 (2012).

29. Aman, K., Padroni, G., Parkinson, J. A., Welte, T. & Burley, G. A. Structural and Kinetic Profiling of Allosteric Modulation of Duplex DNA Induced by DNA-Binding Polyamide Analogues. *Chem. - A Eur. J.* (2018). doi:10.1002/chem.201805338

30. Chenoweth, D. M. & Dervan, P. B. Structural Basis for Cyclic Py-Im Polyamide Allosteric Inhibition of Nuclear Receptor Binding. *J. Am. Chem. Soc.* **132**, 14521–14529 (2010).

31. Spink, N., Brown, D. G., Skelly, J. V. & Neidle, S. Sequence-dependent effects in drug-DNA interaction: the crystal structure of Hoechst 33258 bound to the d(CGCAAATTTGCG)2 duplex. *Nucleic Acids Res.* **22**, 1607–12 (1994).

32. Bostock-Smith, C. E., Harris, S. A., Laughton, C. A. & Searle, M. S. Induced fit DNA recognition by a minor groove binding analogue of Hoechst 33258 : fluctuations in DNA A tract structure investigated by NMR and molecular dynamics simulations. *Nucleic Acids Res.* **29**, 693–702 (2001).

33. Alniss, H. Y. *et al.* Recognition of the DNA Minor Groove by Thiazotropsin Analogues. *ChemBioChem* **15**, 1978–1990 (2014).

34. Furse, K. E. & Corcelli, S. A. The Dynamics of Water at DNA Interfaces : Computational Studies of Hoechst 33258 Bound to DNA. *J. Am. Chem. Soc.* **130**, 13103–13109 (2008).

35. Hunt, N. T. 2D-IR spectroscopy: ultrafast insights into biomolecule structure and function. *Chem. Soc. Rev.* **38**, 1837–1848 (2009).

36. Ghosh, A., Ostrander, J. S. & Zanni, M. T. Watching Proteins Wiggle: Mapping Structures with Two-Dimensional Infrared Spectroscopy. *Chem. Rev.* **117**, 10726–10759 (2017).

37. Szyc, Ł., Yang, M., Nibbering, E. T. J. & Elsaesser, T. Ultrafast Vibrational Dynamics and

Local Interactions of Hydrated DNA. *Angew. Chemie Int. Ed.* **49**, 3598–3610 (2010).

38. Hamm, P. & Zanni, M. *Concepts and Methods of 2D Infrared Spectroscopy*. (Cambridge University Press, 2011). doi:10.1017/CBO9780511675935

39. Greetham, G. M. *et al.* A 100 kHz Time-Resolved Multiple-Probe Femtosecond to Second Infrared Absorption Spectrometer. *Appl. Spectrosc.* **70**, 645–653 (2016).

40. Luther, B. M., Tracy, K. M., Gerrity, M., Brown, S. & Krummel, A. T. 2D IR spectroscopy at 100 kHz utilizing a Mid-IR OPCPA laser source. *Opt. Express* **24**, 4117 (2016).

41. Donaldson, P. M., Greetham, G. M., Shaw, D. J., Parker, A. W. & Towrie, M. A 100 kHz Pulse Shaping 2D-IR Spectrometer Based on Dual Yb:KGW Amplifiers. *J. Phys. Chem. A* **122**, 780–787 (2018).

42. Shim, S., Strasfeld, D. B., Ling, Y. L. & Zanni, M. T. Automated 2D IR spectroscopy using a mid-IR pulse shaper and application of this technology to the human islet amyloid polypeptide. *PNAS* **104**, 14197–14202 (2007).

43. Wedler, G. *Lehrbuch der Physikalischen Chemie*. (Wiley-VCH, 2007).

44. Atkins, P. & Paula, J. de. *Physical Chemistry*. (Oxford University Press, 2009).

45. Rullière, C. *Femtosecond Laser Pulses*. (Springer, 2005).

46. Suresh, S., Ramanand, A., Jayaraman, D. & Mani, P. Review on theoretical aspect of nonlinear optics. *Rev. Adv. Mater. Sci.* **30**, 175–183 (2012).

47. Hithell, G. *et al.* Long-Range Vibrational Dynamics Are Directed by Watson–Crick Base Pairing in Duplex DNA. *J. Phys. Chem. B* **120**, 4009–4018 (2016).

48. Atkins, P. & Friedman, R. *Molecular Quantum Mechanics*. (Oxford University Press, 2005).

49. Moran, A. & Mukamel, S. The origin of vibrational mode couplings in various secondary structural motifs of polypeptides. *Proc. Natl. Acad. Sci.* **101**, 506–510 (2004).

50. Hithell, G. *et al.* Ultrafast 2D-IR and Optical Kerr Effect Spectroscopy Reveal the Impact of Duplex Melting on the Structural Dynamics of DNA. *Phys. Chem. Chem. Phys.* **19**, 10333–10342 (2017).

51. Taillandier, E. & Liquier, J. [16] Infrared spectroscopy of DNA. in **211**, 307–335 (1992).

52. Banyay, M., Sarkar, M. & Graslund, A. A library of IR bands of nucleic acids in solution. *Biophys. Chem.* **104**, 477–488 (2003).

53. Wood, B. R. The importance of hydration and DNA conformation in interpreting infrared spectra of cells and tissues. *Chem. Soc. Rev.* **45**, 1980–1998 (2016).

54. Towrie, M. *et al.* ps-TRIR covers all the bases – recent advances in the use of transient IR for the detection of short-lived species in nucleic acids. *Analyst* **134**, 1265 (2009).

55. Doorley, G. W. *et al.* Tracking DNA Excited States by Picosecond-Time-Resolved Infrared Spectroscopy: Signature Band for a Charge-Transfer Excited State in Stacked Adenine–Thymine Systems. *J. Phys. Chem. Lett.* **4**, 2739–2744 (2013).

56. Liu, Y., Guchhait, B., Siebert, T., Fingerhut, B. P. & Elsaesser, T. Molecular couplings and energy exchange between DNA and water mapped by femtosecond infrared spectroscopy of backbone vibrations. *Struct. Dyn.* **4**, 044015 (2017).

57. Lee, C., Park, K.-H. & Cho, M. Vibrational dynamics of DNA. I. Vibrational basis modes and

couplings. *J. Chem. Phys.* **125**, 114508 (2006).

58. Lee, C. & Cho, M. Vibrational dynamics of DNA. II. Deuterium exchange effects and simulated IR absorption spectra. *J. Chem. Phys.* **125**, 114509 (2006).

59. Lee, C., Park, K.-H., Kim, J.-A., Hahn, S. & Cho, M. Vibrational dynamics of DNA. III. Molecular dynamics simulations of DNA in water and theoretical calculations of the two-dimensional vibrational spectra. *J. Chem. Phys.* **125**, 114510 (2006).

60. Lee, C. & Cho, M. Vibrational dynamics of DNA: IV. Vibrational spectroscopic characteristics of A-, B-, and Z-form DNA's. *J. Chem. Phys.* **126**, 145102 (2007).

61. Greve, C. & Elsaesser, T. Ultrafast Two-Dimensional Infrared Spectroscopy of Guanine–Cytosine Base Pairs in DNA Oligomers. *J. Phys. Chem. B* **117**, 14009–14017 (2013).

62. Krummel, A. T., Mukherjee, P. & Zanni, M. T. Inter and Intrastrand Vibrational Coupling in DNA Studied with Heterodyned 2D-IR Spectroscopy. *J Phys Chem B* **107**, 9165–9169 (2003).

63. Peng, C. S., Jones, K. C. & Tokmakoff, A. Anharmonic vibrational modes of nucleic acid bases revealed by 2D IR spectroscopy. *J. Am. Chem. Soc.* **133**, 15650–15660 (2011).

64. Hithell, G. *et al.* Effect of oligomer length on vibrational coupling and energy relaxation in double-stranded DNA. *Chem. Phys.* in Press (2018). doi:10.1016/j.chemphys.2017.12.010

65. Siebert, T., Guchhait, B., Liu, Y., Costard, R. & Elsaesser, T. Anharmonic Backbone Vibrations in Ultrafast Processes at the DNA–Water Interface. *J. Phys. Chem. B* 150710141420002 (2015). doi:10.1021/acs.jpcb.5b04499

66. Guchhait, B., Liu, Y., Siebert, T. & Elsaesser, T. Ultrafast vibrational dynamics of the DNA backbone at different hydration levels mapped by two-dimensional infrared spectroscopy. *Struct. Dyn.* **3**, 1–15 (2016).

67. Hume, S. *et al.* Measuring proteins in H2O with 2D-IR spectroscopy. *Chem. Sci.* **10**, 6448–6456 (2019).

68. Rezende Valim, L. *et al.* Identification and Relative Quantification of Tyrosine Nitration in a Model Peptide Using Two-Dimensional Infrared Spectroscopy. *J. Phys. Chem. B* **118**, 12855–12864 (2014).

69. Ramakers, L. A. I. *et al.* 2D-IR Spectroscopy Shows that Optimized DNA Minor Groove Binding of Hoechst33258 Follows an Induced Fit Model. *J. Phys. Chem. B* **121**, 1295–1303 (2017).

70. Fritzsch, R. *et al.* Rapid Screening of DNA–Ligand Complexes via 2D-IR Spectroscopy and ANOVA–PCA. *Anal. Chem.* **90**, 2732–2740 (2018).

71. Krummel, A. T. & Zanni, M. T. DNA vibrational coupling revealed with two-dimensional infrared spectroscopy: Insight into why vibrational spectroscopy is sensitive to DNA structure. *J. Phys. Chem. B* **110**, 13991–14000 (2006).

72. Sanstead, P. J., Stevenson, P. & Tokmakoff, A. Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved through Infrared Spectroscopy. *J. Am. Chem. Soc.* **138**, 11792–11801 (2016).

73. Yang, M., Szyc, Ł. & Elsaesser, T. Femtosecond two-dimensional infrared spectroscopy of adenine-thymine base pairs in DNA oligomers. *J. Phys. Chem. B* **115**, 1262–1267 (2011).

74. Steinel, T., Asbury, J. B., Zheng, J. & Fayer, M. D. Watching hydrogen bonds break: A

transient absorption study of water. *J. Phys. Chem. A* **108**, 10957–10964 (2004).

75. Kubelka, J. Time-resolved methods in biophysics. 9. Laser temperature-jump methods for investigating biomolecular dynamics. *Photochem. Photobiol. Sci.* **8**, 499 (2009).

76. Ma, H., Wan, C., Wu, A. & Zewail, A. H. DNA folding and melting observed in real time redefine the energy landscape. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 712–6 (2007).

77. Popp, A., Scheerer, D., Chi, H., Keiderling, T. A. & Hauser, K. Site-Specific Dynamics of β-Sheet Peptides withDPro-Gly Turns Probed by Laser-Excited Temperature-Jump Infrared Spectroscopy. *ChemPhysChem* **17**, 1273–1280 (2016).

78. Ansari, A., Kuznetsov, S. V. & Shen, Y. Configurational diffusion down a folding funnel describes the dynamics of DNA hairpins. *Proc. Natl. Acad. Sci.* **98**, 7771–7776 (2001).

79. Gal, M., Zibzener, K. & Frydman, L. A capacitively coupled temperature-jump arrangement for high-resolution biomolecular NMR. *Magn. Reson. Chem.* **48**, 842–847 (2010).

80. Sanstead, P. J. & Tokmakoff, A. Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization. *J. Phys. Chem. B* **122**, 3088–3100 (2018).

81. Wilson, W. D., Tanious, F. A., Fernandez-Saiz, M. & Rigl, C. T. Evaluation of Drug–Nucleic Acid Interactions by Thermal Melting Curves. in *Drug-DNA Interaction Protocols* 219–240 (Humana Press). doi:10.1385/0-89603-447-X:219

5 Methods

5.1 Abstract

This chapter outlines the practical implementation of the spectroscopy and analysis methods employed in this thesis. Generation of infrared absorption, 2D-IR and temperature-jump induced pump-probe spectra are outlined in order to give a fundamental overview how the discussed spectral features were obtained. The data analysis section explains how large spectroscopic datasets can be approached using principal component analysis and regression techniques. A more detailed discussion of relevant details can be found in the corresponding chapters as needed.

5.2 Linear Infrared Absorption Spectrometry

All linear infrared absorption spectra shown in this thesis were obtained from liquid samples using commercially available FT-IR absorption spectrometers in transmission mode, such as the Bruker Vertex or the ThermoFisher Nicolet series. These spectrometers follow the methodology outlined in **Fig. 5.1a** and use a black-body emitter like SiC or tungsten as an incoherent light source with a broadband infrared emission spectrum that is collimated to probe the sample[1]. The broad, black line in **Fig. 5.1a** illustrates the beam path of the infrared light. In order to obtain the frequency axis of the spectrum using only a single-point detector, the probe light is sent through a Michelson interferometer before it interacts with the sample. The beam is split into two arms by a beam splitter, with one arm at a fixed path length and one arm having a variable path-length due to a moving mirror. The two arms are reflected back, recombined at the beam splitter, and the difference in path-length generates an interference signal that is sent through the sample. Each position of the moving mirror generates a different subset of frequencies that probes the sample. The mirror oscillates continuously during a measurement and in order to ascertain its position, a reference laser is sent through the interferometer, which is detected by a separate, small, reference detector after the sample (not shown). The fixed wavelength of the reference beam generates an interference pattern that is maximised each time the path length of the moving arm is at a multiple of the reference wavelength. This leads to an electronic signal at the reference detector that can be used to trigger the infrared detector to ensure that the infrared absorption data is acquired at equidistant mirror positions.

The infrared light is detected (after collimation) by a single-point detector made from infrared-sensitive materials such as InGaAs or HgCdTe and returns an interferogram, $I(t)$, as shown in **Fig. 5.1b**. The maximum intensity of this interferogram is found at the position where the relative path length difference of the two arms is zero. This signal is usually pre-processed to improve the spectral resolution, the signal-to-noise-ratio and to remove data artefacts using methods like zero-padding, window functions or phase correction before performing a Fourier-transformation on the data to obtain a power spectrum, $I(\tilde{\nu})$.

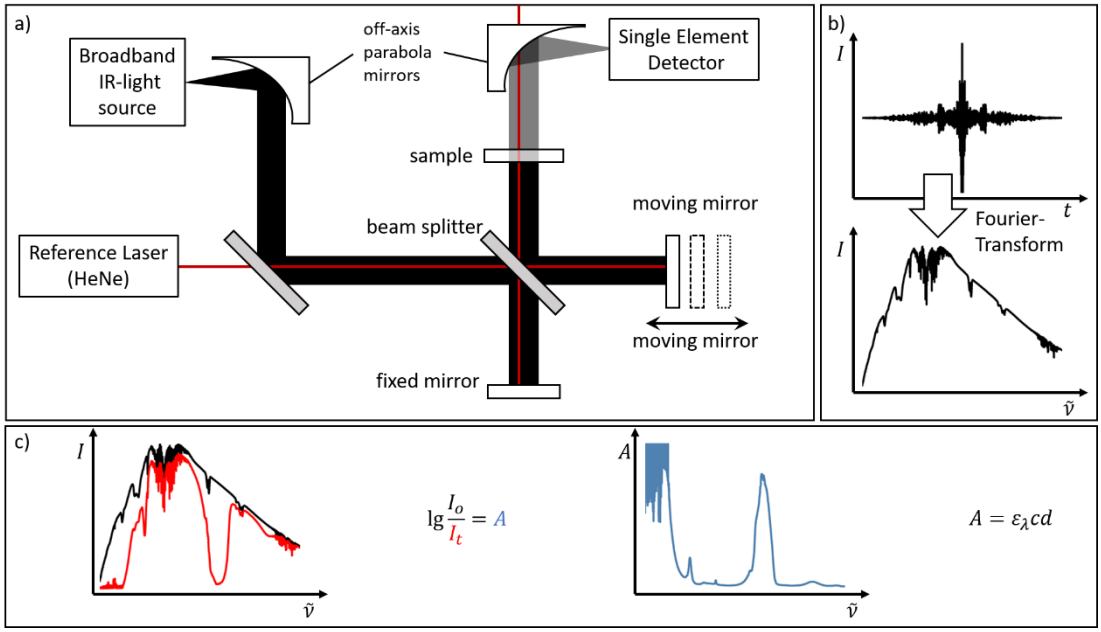


Fig. 5.1a) *Schematic setup of a transmission FT-IR spectrometer using a Michelson interferometer with HeNe reference Laser.* ***b)*** *Fourier-transformation of the detected interferogram leads to a spectral response.* ***c)*** *Comparison of the detected signal intensity without sample (black trace, I_0) and observed signal intensity with sample (red trace, I_t) leads to an absorption spectrum (blue trace). An absorbance value of 1 is equivalent to 10% transmission of light.*

In order to obtain an absorbance spectrum, a reference measurement without sample, $I_0(\tilde{\nu})$, has to be collected as well as a measurement with sample, $I_t(\tilde{\nu})$, as shown in **Fig. 5.1c**. The absorbance spectrum, $A(\tilde{\nu})$, is then calculated according to:

$$A = \lg \frac{I_0}{I_t} = \varepsilon_\lambda cd. \tag{5.1}$$

In solutions, the absorbance A of an analyte at a specific wavenumber $\tilde{\nu}$, is dependent on its wavelength-dependent, molar extinction coefficient ε_λ, its concentration c and the thickness of the sample d, according to the Lambert-Beer-Law shown in equation (5.1). The Lambert-Beer-Law is valid for homogeneous solutions with negligible scattering at low analyte concentrations and can start to deviate at high absorbance values[2]. All infrared measurements analysed in this thesis were therefore taken at absorption values below 1, in order to prevent non-linear dependencies.

5.3 2D-IR Spectrometry in Time-Domain Using Pulse Shaping

5.3.1 Instrumental Setup

The molecular structures studied in this thesis typically show vibrational lifetimes below 1 ps. In order to excite and temporally resolve such a fast process, pump and probe pulses require pulse-widths on the order of femtoseconds. Generating

femtosecond pulses in the infrared region at energies of a few µJ is technologically demanding and requires an elaborate experimental setup with different stages of pulse generation, amplification and conversion[3]. While ultrashort pulse generation of infrared light is an active research topic on its own[4–6], a number of commercial solutions are available to researchers as off-the-shelf products that simplify the instrumental setup[7–9].

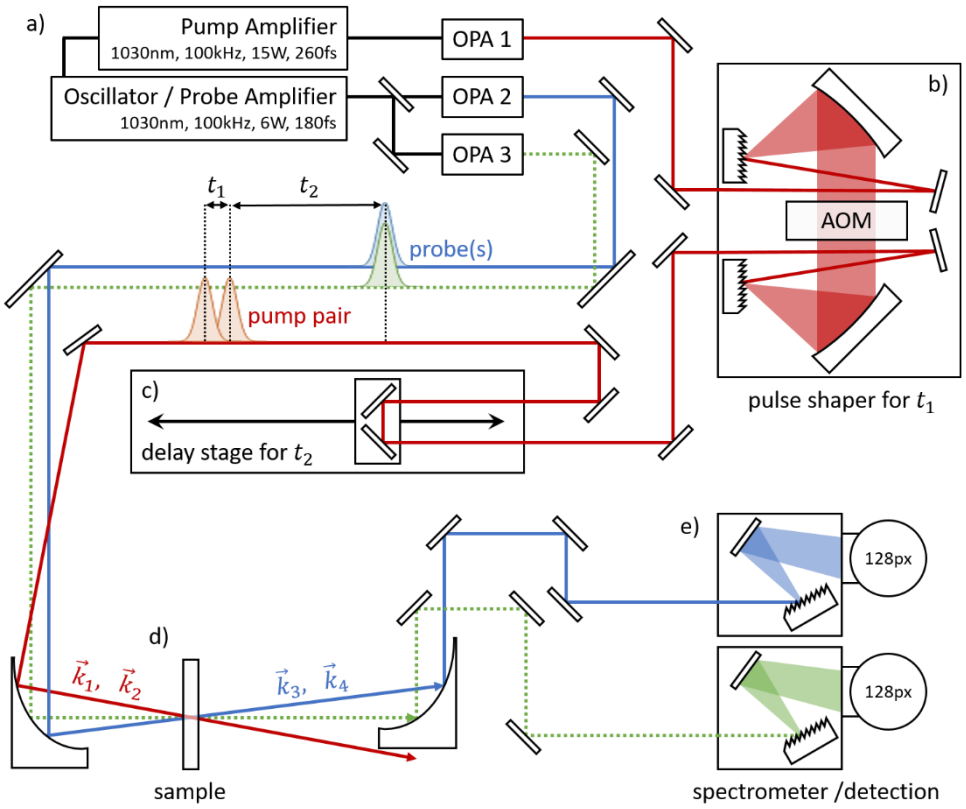


Fig. 5.2 *Instrumental setup of the LIFEtime spectrometer at RAL[10,11]. 2D-IR requires only one probe pulse (beam path in blue). Two separate probe beams (blue, green) allow acquisition of two different 2D-IR spectra at the same time, for example at different probe regions in the spectrum. The beam path of the pump pulse(s) is indicated in red. A pulse shaper is used to generate a pump pulse pair and is explained in more detail in section 5.3.2.*

All 2D-IR experiments in this thesis were performed on the same spectrometer, the LIFEtime instrument at the Rutherford Appleton Laboratories (RAL)[10,11]. The instrumental layout is shown in **Fig. 5.2**. In this setup, a seed pulse at 80 MHz repetition rate drives two regenerative amplifiers (PHAROS, Light Conversion) with Yb:KGW gain medium (Ytterbium doped potassium gadolinium tungstate) to obtain 1030 nm light at a repetition rate of 100 kHz. Light from the 15 W amplifier has a pulse duration of 260 fs and is converted by optical parametric amplification and difference frequency generation in OPA1 (ORPHEUS-HP) to infrared light that is tuneable from 4.5-10 µm at

1.7-0.7 µJ respectively. This output is used for the pump pulses in the 2D-IR experiment (red beam path in **Fig. 5.2**). The output from the 6 W amplifier has a pulse width of 180 fs and is split 50/50 to OPA2 and OPA3 (ORPHEUS-ONE) to generate two, independently tuneable probe pulses from 4.5-15 µm at 0.2–0.3 µJ respectively (blue and green beam paths). A usual 2D-IR experiment only requires one probe pulse. This setup however enables acquisition of two 2D-IR experiments simultaneously at different spectral probe regions. This is used in chapter 7, where one probe pulse covers the spectral region from 1410-1590 cm^{-1} and the second probe pulse is set to 1590-1720 cm^{-1}, allowing for greater coverage of the infrared spectrum.

Once infrared pulses are generated, the pump pulse is sent to a pulse shaper in **Fig. 5.2b** to produce the pump-pulse-pair needed for 2D-IR spectroscopy[12]. The pulse shaper is explained in more detail in section 5.3.2 and is used as a substitute to a Michelson interferometer in time-domain 2D-IR spectroscopy. It produces two collinear pulses separated by the pump-pump delay t_1. The t_1-delay is used to scan the coherences generated after the first pump pulse in time-domain, which define the positions of the 2D-IR signals along the pump axis of the 2D-IR spectrum. A delay stage after the pulse shaper is used to adjust the relative timing between the second pump pulse and the probe, t_2. This delay is scanned to study the 2D-IR response over time for extraction of vibrational lifetimes and energy transfer kinetics.

Pump and probe pulses are focussed with an off-axis parabola to the same spot in the sample in **Fig. 5.2d** to perform the 2D-IR experiment and generate the polarisation that contains the characteristic, third-order response function. As both pump pulses are collinear ($\vec{k}_1 = \vec{k}_2$), the 2D-IR experiment is acquired in pump-probe geometry and the emitted signal ($\vec{k}_4$) is collinear with probe pulse ($\vec{k}_3$), self-heterodyning the signal. The probe beam with the 2D-IR signal is then sent into a spectrometer and dispersed on to an HgCdTe array detector with 128 elements. The spectral response is calibrated with known absorption lines such as atmospheric water, so that detector pixels can be converted to the probe frequency axis of the 2D-IR spectrum.

This approach collects the probe axis of a 2D-IR spectrum in frequency-domain by dispersion on to an array detector and obtains the pump axis in time-domain by changing the t_1-delay sequentially in each pulse cycle. This generates an individual interferogram for each pixel on the detector, which is Fourier-transformed to obtain a pump frequency axis.

5.3.2 Generating Collinear Pump Pulses via Pulse Shaping

Classic interferometric methods use motorised translation stages to scan the t_1-delay for the pump axis. This requires high accuracy and reproducibility of the stage position to prevent artefacts and systematic errors. Each movement to a new position also requires some time during which the spectrometer cannot collect any data. Any difference in the orientation of the two pump pulses due to small misalignment of the interferometer will further affect the relative intensities of the rephasing and non-rephasing pathways and distort the 2D signals. A pulse shaper alleviates these problems as it generates the pulse pair by amplitude and phase modulation of a single input pulse using an acousto-optic modulator[11,13,14].

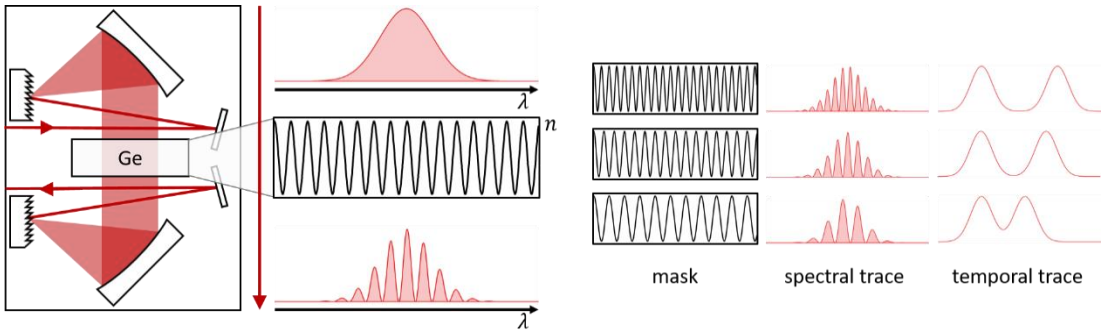


Fig. 5.3 *A mid-infrared pulse shape disperses the pump pulse on to a large Ge crystal. An acoustic wave travelling through the crystal perpendicular to the pulse modulates the refractive index of Ge and generates a programmable grating. This is used to selectively mask parts of the input pulse to generate pulse pairs with well-defined phase relationship.*

The operation principle of a mid-infrared pulse shaper is shown in **Fig. 5.3**. The input pulse is dispersed by a grating and a cylindrical mirror on to a Germanium crystal, where each frequency is focussed to a different position. A piezoelectric transducer generates an acoustic wave in the crystal, travelling perpendicular to the beam path. This changes the local refractive index of germanium at each position and selectively deflects parts of the pulse. The shape of the acoustic wave (mask) is programmable by an arbitrary waveform generator so that the spectrum of an incoming infrared pulse can be modulated to represent an interference pattern, equivalent to a pulse pair in time-domain (see **Fig. 5.3**). The dispersed, modulated pulse shape is then compressed again by another cylindrical mirror and grating to obtain the initial beam diameter. The generated pulse pair is perfectly collinear and is separated by a delay t_1, which is programmable via the waveform generator. The pulse shaper operates at the same repetition rate as the amplifier so that each pump pulse can be modulated with a different mask. This makes it possible to perform a full t_1-scan within a few

milliseconds to generate a complete 2D-IR spectrum with equal contribution from rephasing and non-rephasing pathways. The reduction in acquisition time makes the t_1-scan less sensitive to laser intensity fluctuations and phase drift over time.

5.3.3 Generating a 2D-IR spectrum

Once the pulse sequence has interacted with the sample and the probe light is acquired by the detector, the 2D-IR spectrum needs to be constructed from the collected data. Both pump pulses generate an additional transient absorption signal superimposed with the 2D-IR response. Scattered light from any of the pulses can reach the detector and introduce additional artefacts. In order to extract the 2D-IR response and remove unwanted additional features, a method called *phase cycling* is used[15]. When performing the 2D-IR experiment, the relative phase $\Delta\varphi$ of the two pump pulses is shifted by 180° in every other measurement, causing the 2D-IR signal to flip sign, while scatter and transient absorption features remain. Subtraction of the two signals increases the 2D-IR response, while removing transient signals and suppressing artefacts. This approach extracts the 2D-IR response directly from the collected data and is illustrated in **Fig. 5.4a-c**.

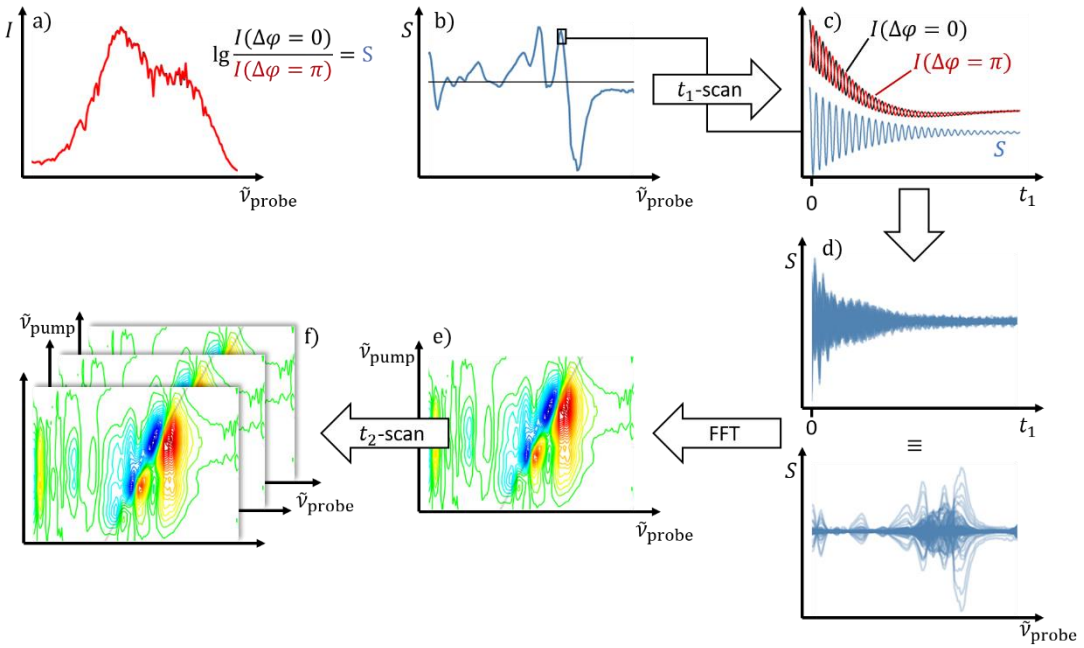


Fig. 5.4a) *To extract the 2D-IR response from the detected signal, two measurements shown in black and red are obtained using a phase cycling methodology.* ***b)*** *The extracted 2D-IR response S in blue represents data at a single t_1-time.* ***c)*** *Phase cycling removes transient absorption features from the data.* ***d)*** *The complete 2D-IR interferogram.* ***e)*** *2D-IR spectrum after Fourier transformation to obtain the pump frequency axis.* ***f)*** *The t_2-time can be scanned as well to obtain a set of 2D-matrices.*

When the LIFEtime spectrometer acquires a 2D-IR spectrum, the pump-pump delay t_1 is initially set to 0 fs and the intensity $I(\Delta\varphi = 0)$ is collected by the detector (black trace in **Fig. 5.4a**, almost completely overlapped by the red trace). If the pump pulse is centred at 1650 cm^{-1}, its electric field will oscillate with a period of 18 fs. So in order to shift the relative phase of the pump pulses by 180°, the pump-pump delay is set to half of the period, at $t_1 = 9$ fs in a second measurement. This leads to the reference measurement $I(\Delta\varphi = \pi)$, indicated by the red trace in **Fig. 5.4a**. Both traces are used to extract the 2D-IR signal S, following:

$$S = lg\frac{I(\Delta\varphi = 0)}{I(\Delta\varphi = \pi)} \ . \tag{5.2}$$

The extracted 2D-IR signal S is shown in **Fig. 5.4b** and represents a slice of the 2D-IR response at a t_1-time of 4.5 fs. With this approach, the t_1-time is scanned and an interferogram for every pixel on the probe detector is obtained. **Fig. 5.4c** illustrates the effect of phase cycling on the signal measured at a single detector pixel. The slow relaxation over t_1 in the black and red trace corresponds to a transient absorption signal that remains when changing the relative phase between the pump pulses. The division in equation (5.2) is identical to a subtraction of the logarithmic values, which removes the transient absorption signal and enhances the 2D-IR response (blue trace, **Fig. 5.4c**). The relaxation of the signal S over time corresponds to the dephasing rate of the coherence and usually decays within 4 ps for nucleotide vibrations. The t_1-scan is thus scanned from 0 to 4 ps to sample as much of the coherence as possible. A window function (Hamming window) to remove artefacts and zero filling to improve the frequency resolution is applied to the interferograms as explained in **Fig. 5.5**. These are standard techniques when processing interferometric data to improve the Fourier-transformed signal and are applied to 2D-IR interferograms throughout this thesis.

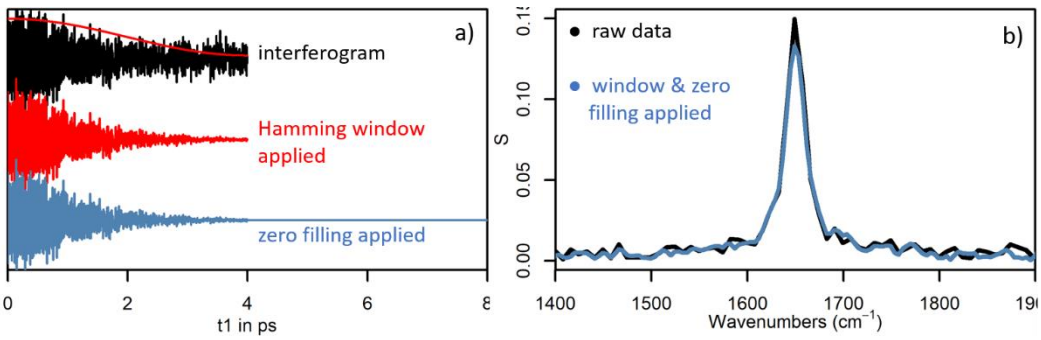


Fig. 5.5 a) *shows a raw interferogram scanned for 4 ps with a coherence that relaxes within 3 ps and considerable noise throughout the t_1-scan in black. The interferogram can be multiplied by a window function*

to suppress noise at the end of the scan where the coherence has decayed (red). After 4 ps, no signal is left and can thus be represented by a sequence of zeros to increase the scan to 8 ps (blue). ***b)*** *The Fourier-transformed signal from the pre-processed interferogram (blue) compared to the raw data shows a slightly better signal to noise ratio and an improved spectral resolution.*

The processed interferograms from the t_1-scan (illustrated in **Fig. 5.4d**) are Fourier-transformed to obtain a pump frequency axis. The 2D-IR spectrum, resolved over two frequency axes is illustrated in **Fig. 5.4e**. When collecting the data, the t_1-scan is performed several thousand times before extracting the 2D-IR response from averaged detector intensities. The number of averaging measurements, N, improves the signal-to-noise ratio, as noise typically scales with $1/\sqrt{N}$. This corresponds to an acquisition time of a few seconds for a single 2D-IR spectrum with well-resolved features, due to the 100 kHz repetition rate of the instrument. Finally, the pump-probe delay t_2 can additionally be scanned to characterise vibrational lifetime or energy transfer processes of the studied modes as indicated in **Fig. 5.4f**. This leads to a stack of 2D-IR spectra, where the temporal response of each mode can be analysed individually.

The 2D-IR spectra collected in this thesis are primarily used for screening experiments in chapters 6 and 7, and are analysed using the multivariate analysis methods outlined in section 5.5.

5.4 Temperature Jump Experiments

5.4.1 Instrumental Setup

Chapter 8 combines time-resolved spectroscopy with a laser-induced temperature jump to study the impact of ligand binding to the DNA melting dynamics. In this experiment, a laser pulse tuned to the absorption band of D_2O excites the vibrational modes of the solvent. Picosecond relaxation processes of the solvent modes dissipate the energy to other modes in the system which subsequently perturb the equilibrium of solvated dsDNA structures in the sample; not dissimilar to a macroscopic rise in temperature. To observe the structural changes following the perturbation, a broadband, infrared pulse probes the vibrational modes of the studied dsDNA. Chapter 8 utilises this technique to examine the stabilising effect of a minor groove binder with four dsDNA sequences and discusses its potential application for guiding ligand development. A long-term motivation of this project is further to develop a 2D-IR setup with preceding temperature-jump perturbation[16] and fast data acquisition for large sample throughput in the future.

The experiments outlined in chapter 8 were collected using the Ultra B setup at RAL. In this instrument, a seed pulse at 80 MHz repetition rate is regeneratively amplified (commercial oscillator and amplifier by Coherent) with a Ti:S (Titanium doped Sapphire) crystal to obtain pulses at a wavelength of 800 nm, at a repetition rate of 10 kHz and a pulse duration of <50 fs. The output from the 20 W amplifier is converted in a home-built optical parametric amplifier with BBO (Bariumborate) crystal, followed by difference frequency generation in a AgGaS2 (silver thiogallate) crystal to obtain infrared pulses with 1-3 µJ, centred at 6 µm with a spectral width of >300 cm^{-1} FWHM. This output (blue beam path) is focussed on to the sample at a spot size of ca. 50 µm in diameter. The generated femtosecond pulses are used to probe the vibrational modes of the sample and are detected in frequency domain by a spectrometer equipped with a 128-element, HgCdTe array detector. About 10% of the probe beam is separated by a beam splitter before the sample and analysed in a second spectrometer to obtain a reference measurement that accounts for shot-to-shot noise of the probe light. Both spectrometers are set to cover the same spectral region.

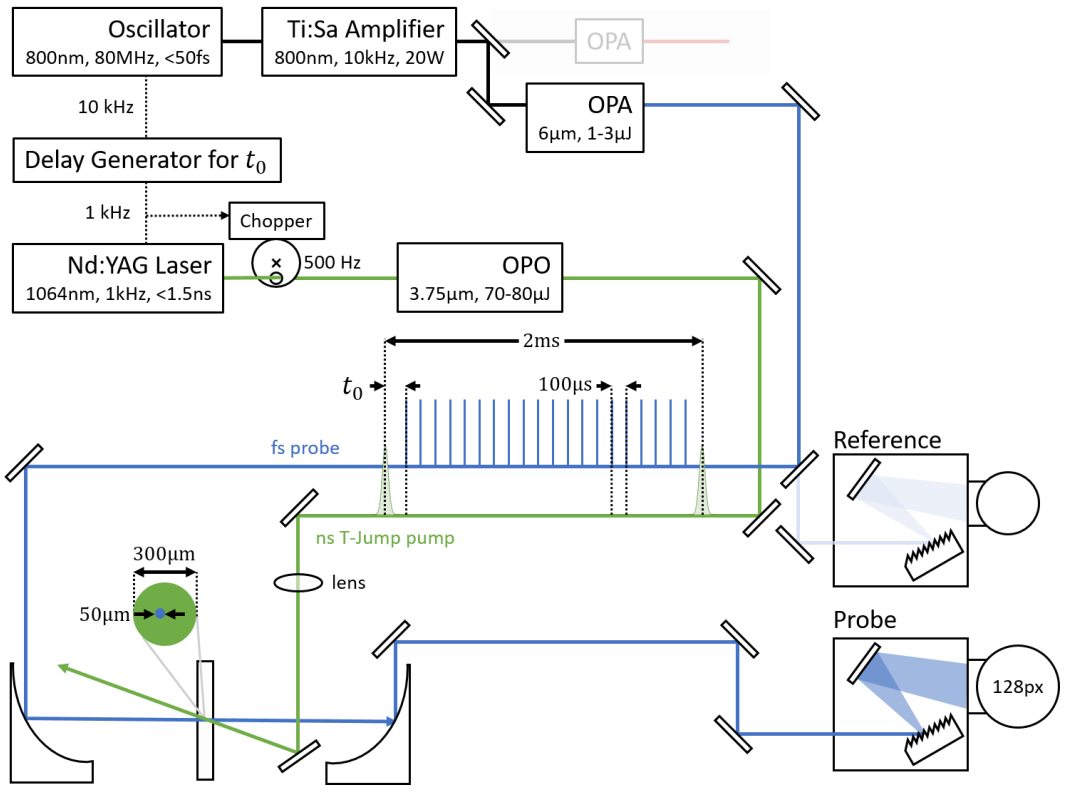


Fig. 5.6 *Instrumental setup of Ultra B at RAL to acquire time-resolved temperature jump data. Each ns-pump pulse is followed by 20 fs-probe pulses that rapidly scan the re-equilibration of the perturbed sample on microsecond timescales. Nanosecond to milisecond timescales are aquired using a delay generator that triggers the pump pulse. The Ultra B setup also contains another beam path for future 2D-IR experiments (in red), which is not used in this thesis.*

The 10 kHz trigger signal of the Ti:S amplifier is sent to a delay generator (Stanford DG645) to divide the signal down to 1 kHz and add a delay adjustable from nanoseconds to microseconds. The delayed, 1 kHz signal triggers a Q-switched, Nd:YAG (neodymium-doped yttrium aluminium garnet) laser with a pulse duration of <1.5 ns at 1064 nm. This output is used to drive a home-built optical parametric oscillator (OPO) using a MgO:PPLN (magnesiumoxide-doped, periodically poled lithium niobite) crystal in a 4 cm cavity to obtain an infrared output beam with 80 µJ pulse energy and <2 ns pulse width. The output wavelength can be tuned slightly by adjustment of the crystal oven temperature and was set to 50°C to obtain light at 3.75 µm (ca. 2650 cm^{-1}). This output is used as a ns-pump pulse to excite the symmetric O-D stretch of the D_2O solvent in the sample. The pulse train is optically chopped to lower the pump repetition rate to 0.5 kHz and increase the available recovery time for the sample to 2 ms per pump pulse. Optical chopping also enables to remove systematic, electronic noise from the detector signal, as will be discussed in chapter 8. The ns-pump beam is focussed with a lens and aligned to excite the sample from the back of the cell (due to size restrictions on the optical table). The spot size of the pump beam incident on the sample is 300 µm in diameter, significantly larger than the probe size.

5.4.2 Acquiring Pump-Probe-Data using Time-Resolved Multiple Probe Spectroscopy (TRMPS)

Pump and probe pulses operate at different repetition rates in this setup (0.5 kHz pump, 10 kHz probe). As a result, each pump pulse is followed by 20 probe pulses which interrogate the sample in intervals of 100 µs. This approach is referred to as TRMPS method (time-resolved multiple probe spectroscopy)[10,17] and enables fast data acquisition of microsecond timescales. The delay generator that triggers the ns-pump defines the delay t_{el} between the pump and the first probe pulse. If this electronic delay is set for example to $t_{el} = 0.1$ µs, multiple probe measurements are acquired at $t =$ 0.1 µs, 100.1µs, 200.1µs, … 1900.1µs. The electronic delay is then changed to the next value, so that the vibrational response of the perturbed sample can be scanned from nanoseconds to up to 2 ms using only a few pump pulses.

Just as with 2D-IR data, the pump-probe signal needs to be extracted from the transmitted probe light collected from the detector. For this purpose, the last measurement at $1900\mu s + t_{el}$ is used as a reference measurement $I_{\text{pump off}}$. This assumes, that the perturbed sample has equilibrated back to its initial state and

$I_{\text{pump off}}$ represents the unperturbed sample response. A pump-probe spectrum S, at a given pump-probe delay t is calculated using equation (5.3):

$$S = lg\frac{I_{\text{pump on}}}{I_{\text{pump off}}} - lg\frac{I_{\text{Ref, pump on}}}{I_{\text{Ref, pump off}}}\ . \tag{5.3}$$

$I_{\text{pump on}}$ is the intensity of the detected probe light after the sample, at a specific delay t. The pump-probe signal is calculated in the first term of equation (5.3) and referenced in the second term. $I_{\text{Ref, pump on}}$ is the intensity of the probe light detected before the sample, which is measured simultaneously by the separate reference detector. $I_{\text{Ref, pump off}}$ is the corresponding signal from the reference detector at $t_{pp} = 1900\mu s + t_0$.

The magnitude of the induced temperature jump is characterised by comparing the pump-probe signal of a reference sample to its temperature dependent, infrared absorption profile. This is discussed in detail in chapter 8 and will suggest a perturbation that corresponds to a sub-nanosecond rise in temperature by about 9 K. This is high enough to initiate unfolding processes of dsDNA structures and generate a pump-probe signal of the vibrational modes at the nucleotides. The effect of ligand binding on the unfolding dynamics of two dsDNA sequences is discussed in detail in chapter 8. This relatively small dataset is analysed with simple regression techniques (fitting of exponential functions) which will be introduced in the chapter as needed.

5.5 Data Analysis of Large Spectroscopic Datasets

The following two chapters will exploit the rapid data collection of the 2D-IR spectrometer to study large spectroscopic datasets, covering over 1000 2D-IR spectra across dozens of samples. Manual analysis of each individual 2D-IR spectrum is prohibitively slow and impractical for such datasets. Screening applications rather focus on relative differences between samples. Principal component analysis (PCA) is a versatile, widely-used method to extract latent variables in a dataset[18–20] and will be used in this thesis to identify similarities and differences across spectroscopic measurements. In the following sections, the concept of PCA as a tool for screening will be first introduced on a fluorescence dataset as an example with a simple, spectroscopic response. The calculation of PC scores and loading vectors is then explained using an FT-IR dataset and generalised to 2D-IR spectra. PCA is finally

compared against regression techniques to understand how the different methods are related to another and it is briefly discussed, when to apply PCA.

5.5.1 Understanding Principal Component Analysis as a Tool for Screening

Chapter 6 will study a fluorescent minor groove binder (Hoechst 33258) that is quenched in solution but fluoresces upon successful binding to the minor groove of dsDNA[21]. The fluorescence emission can therefore be used to estimate the binding affinity of different dsDNA sequences to the ligand. In this example, the fluorescence at the peak of the emission spectrum at 450 nm is studied across a set of different dsDNA-ligand combinations. This represents a simple screening experiment and leads to a dataset with a single *variable*, Y (emission at 450 nm), together with a number of *observations* for Y: Y_a, ..., Y_g (one observation per sample). A distinctively large observation for sample d in this case, as illustrated in **Fig. 5.7a** (red circle), suggests a large fluorescence and therefore successful minor groove binding.

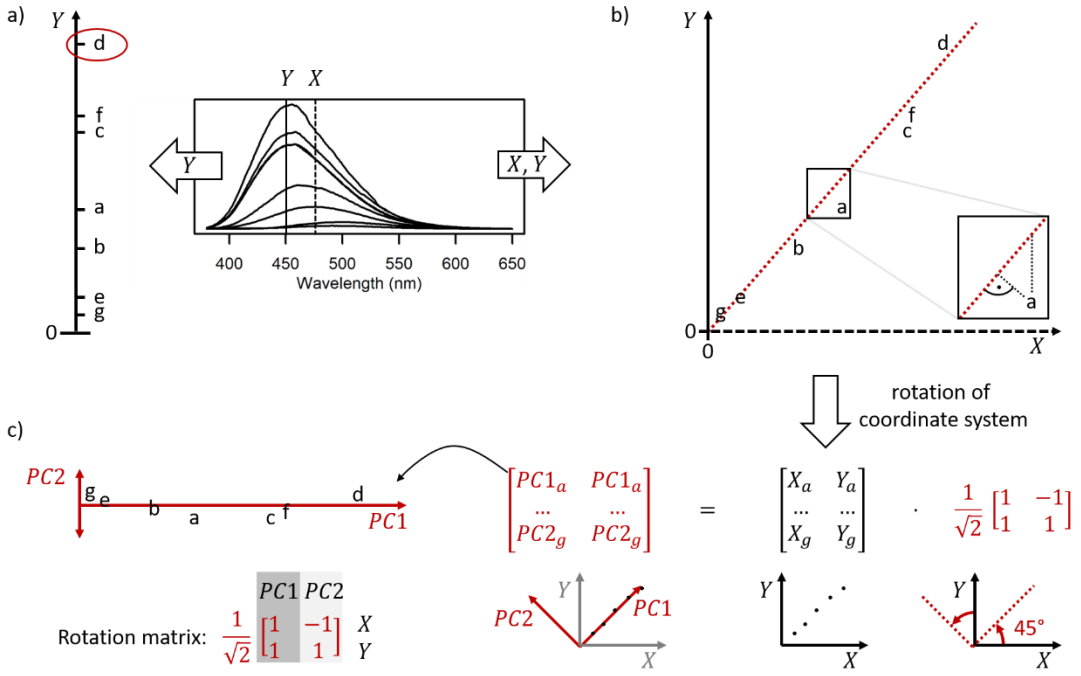


Fig. 5.7a) *Example of a univariate dataset. A single variable Y (fluorescence emission at 450 nm) is plotted with observations for samples a-g. A large observation for Y indicates a high binding affinity, such as for sample d, circled in red.* ***b)*** *A bivariate dataset. Measurements X and Y show a large covariance and seem to be affected by the same underlying process.* ***c)*** *Rotation of coordinate axes X and Y to two new axes PC1 and PC2 allows for a better representation of the underlying process that is common to X and Y. PC1 describes the main variance observed in the dataset. This is how a PCA can be visualised for a dataset with two variables.*

Although analysis of a single variable Y would be sufficient for screening the fluorescence response of each sample in this experiment, spectroscopic datasets like the emission spectrum contain much more information. It is possible to analyse the

spectrum of each sample at an additional wavelength such as 475 nm to obtain a second variable X. The emission band in the example is quite broad so that observations in X will be similar to observations in Y. Plotting observations from Y against those from X, as illustrated in **Fig. 5.7b**, shows a linear relationship; the two parameters appear to be correlated to each other. A classic, linear regression would assume that Y is linear dependent on X and would find a model function $Y(X)$ that minimises the deviation of all observations along the Y-axis (illustrated as the vertical deviation of point a from the red line in the inset of **Fig. 5.7**b). In reality however, the responses of both variables are the result of a common, underlying process (i.e. the same electronic transition). Instead of describing the dataset with a coordinate system spanned by X and Y as in **Fig. 5.7b**, it is beneficial to form a new coordinate system by rotation of X and Y. For this purpose, observations $X_a, \dots, X_g$ and $Y_a, \dots, Y_g$ are written as a $n \times 2$ – matrix and are multiplied with a 2×2 – rotation matrix (45° rotation in this example). This generates two new coordinate axes PC1 and PC2 that describe the same observations of samples a-g with a rotated coordinate system. The underlying, linear relationship common to X and Y is now described by a single axis, PC1. The variance orthogonal to PC1 is described by a second axis, PC2. This is the essence of a principal component analysis (PCA). The algorithm of a PCA rotates the first principal component PC1, so that a maximum of variance in the dataset is described. In other words, the rotation from X to PC1 minimises the distance of each observation orthogonal to the vector of PC1. This is indicated in the inset of **Fig. 5.7b** which compares the minimisation between a linear regression (vertical line) and a PCA (orthogonal line). Once PC1 is found, PC2 is rotated orthogonal to PC1 until the distance to each observation is further minimised. This approach can be extended to a multivariate dataset with p variables (i. e. p emission wavelengths instead of two, X and Y) and transforms correlated, spectral variables into orthogonal components with decreasing variance. As indicated in the plot of **Fig. 5.7c**, the first principal component PC1 alone is sufficient to describe most of the observations measured at the two emission wavelengths, which illustrates why PCA is referred to as a dimensionality reduction technique.

The values obtained for each sample after the rotation are called scores. The scores for PC1 can be interpreted just like the single emission values at 450 nm in the univariate case and will lead to the same conclusion: The DNA sequence in sample d shows the largest response to the ligand, indicating a relative selectivity compared to all other

observations. This enables to survey a large, spectroscopic dataset across all available datapoints, rather than having to identify and select an individual wavelength in advance. In addition to this, the PCA also returns a rotation matrix. The rotation matrix allows conversion of principal components to input variables and *vice versa.* The columns of the rotation matrix in **Fig. 5.7**c specify, which emission wavelength contributes to PC1, PC2 and so on. This can provide useful, spectral information in addition to the screening results obtained from the scores. As the complete emission band of Hoechst 33258 rises in case of effective minor groove binding, an analysis of the full fluorescence dataset via PCA will provide the full shape of the emission band in the corresponding column for PC1. The only significant emission band in this example will be emitted from the ligand. A PCA will therefore not be necessary, as the emission band will not be overlapped by multiple spectral features and can easily be estimated by the emission at 450 nm alone. A congested infrared spectrum on the other hand can benefit from such an analysis and help to unravel different absorption features that are affected by the same, underlying process.

For screening experiments, PCA has the advantage to give a quick overview over the principal differences between samples in a dataset and identifies the spectroscopic features that cause the differentiation. More specifically for this thesis: Samples with a similar infrared response will show common PCA scores and separate from a dsDNA-ligand combination with a more distinctive response. This separation might be indicative of a selective interaction with a ligand and can then be investigated in more detail.

5.5.2 Applying PCA to Infrared Data and extension to 2D-IR

In order to explain the practical implementation of a PCA on infrared absorption data, a dsDNA melting experiment is considered in another example. **Fig. 5.8a** shows experimental data from heating up a dsDNA sequence from room temperature to 60°C. The FT-IR absorption data is stored as $n \times p$ – matrix $\boldsymbol{D}$, with n spectra at different temperatures as rows (observations) and p wavenumbers as columns (variables). The spectrum at room temperature corresponds to the dsDNA duplex structure (blue trace) while the spectrum at high temperatures represents the response from single stranded DNA (red trace) due to unfolding. For such a dataset, it is possible to use a PCA to extract the melting curve of the DNA. Although there might be better methods to

extract the melting curve as will be discussed in section 5.5.3, this example serves as an explanation how PCA scores and PC loading vectors are calculated from $\boldsymbol{D}$.

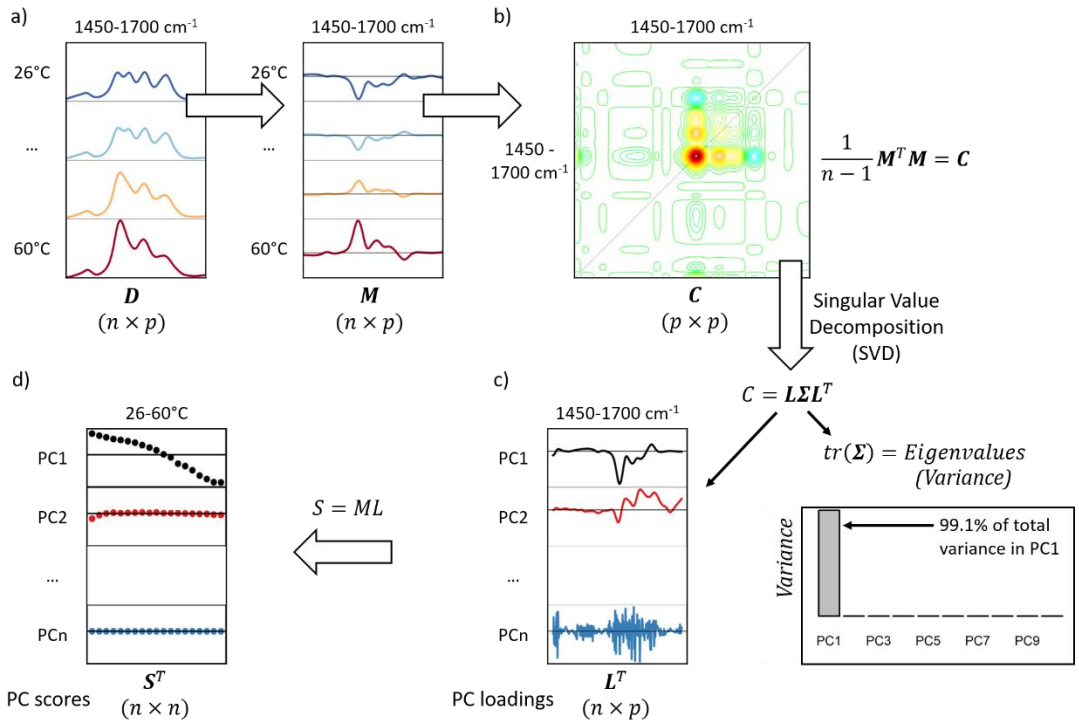


Fig. 5.8 *In order to perform a PCA on a dataset* $\boldsymbol{D}$*, the covariance matrix of* $\boldsymbol{D}$ *is calculated as* $\boldsymbol{C}$*. The orthogonal eigenvectors of* $\boldsymbol{C}$ *correspond to loading vectors of the PCA, while eigenvalues correspond to the variance of each principal component. Eigenvectors* $\boldsymbol{L}$ *and eigenvalues tr(*$\boldsymbol{\Sigma}$*) are obtained using singular value decomposition.* $\boldsymbol{L}$ *is then used to rotate the mean-centred input matrix* $\boldsymbol{M}$ *from wavenumbers to PC scores. The melting curve is retrieved in the scores for PC1.*

In the first step, the average spectrum of the dataset in **Fig. 5.8a** is calculated and subtracted from $\boldsymbol{D}$ to obtain a mean-centred dataset, $\boldsymbol{M}$. This allows calculation of covariance matrix $\boldsymbol{C}$, which indicates whether two variables (wavenumbers) change independently from each other (corresponding matrix element is zero) or show a correlation (matrix element is significant). The signs of the values in $\boldsymbol{C}$ indicate, whether two variables are correlated or anti-correlated, and the values on the diagonal of $\boldsymbol{C}$ correspond to the variance of each individual variable in the dataset $\boldsymbol{M}$. While subtraction of the mean spectrum in the first step is technically not necessary to perform a PCA, it is conceptually required to obtain a covariance matrix rather than a generic, symmetric, square matrix for $\boldsymbol{C}$. Either way, matrix $\boldsymbol{C}$ is subjected to a singular value decomposition (SVD) in order to obtain the orthogonal eigenvectors and eigenvalues of $\boldsymbol{C}$, according to equation (5.4):

$$\boldsymbol{C} = \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^{T}. \tag{5.4}$$

Using an SVD is not the only way to retrieve eigenvectors and eigenvalues from $\boldsymbol{C}$, but it is computationally efficient and utilised in software implementations to compute PCA results[23]. Matrix $\boldsymbol{L}^T$ (in **Fig. 5.8c**) contains the orthogonal eigenvectors of the covariance matrix and represents the rotation matrix discussed in the previous section 5.5.1. The SVD itself will return a $p \times p$ matrix from which the first n rows correspond to principal components PC1 to PCn. These will be referred to as loading vectors throughout the thesis. While the first two loading vectors indicate actual spectral features in the example (black and red trace), the last loading vector for PCn just resembles noise (blue trace).

$\boldsymbol{\Sigma}$ is a diagonal matrix that represents the covariance of the principal components after the rotation. All off-diagonal values are zero due to orthogonality of the eigenvectors $\boldsymbol{L}$. The values of its trace $tr(\boldsymbol{\Sigma})$, are the eigenvalues of $\boldsymbol{C}$**.** The eigenvalues represent the variance of each principal component and their sum is equal to the total variance of $\boldsymbol{C}$. The eigenvalues are decreasing with increasing principal component and the plot in **Fig. 5.8c** indicates that almost all of the variance of the example dataset is explained by the first principal component (representing 99.1% of the total variance). This suggests that the spectroscopic change in dataset $\boldsymbol{D}$ can be explained by the first principal component alone, while retaining almost all of the available information. In order to obtain the PC scores of each measurement in the dataset, the observations in $\boldsymbol{M}$ are rotated using $\boldsymbol{L}$ to obtain $n \times n$ – matrix $\boldsymbol{S}$, which is shown in **Fig. 5.8d**.

The scores of PC1 in $\boldsymbol{S}$ show a sigmoidal transition from 26°C to 60°C, from positive to negative scores (black dots). The corresponding loading vector for PC1 (black trace, **Fig. 5.8c**) identifies the spectral features that change with a sigmoidal response in dataset $\boldsymbol{M}$. It is clear from the scores that PC1 essentially describes the melting process as observed via FT-IR infrared spectroscopy and the PC scores can now be fitted to a sigmoidal function to obtain the melting point of the DNA sequence. The loading vector (first row of $\boldsymbol{L}^T$) can be further analysed to identify, which vibrational modes are affected upon unfolding of the DNA duplex.

The scores of PC2 (red dots) are much smaller than those of PC1 and only seem to deviate slightly from zero in the first few observations (spectra taken at 26-30°C). The small PC2 scores, together with the low eigenvalue observed for PC2 indicates that the contribution of PC2 to the dataset is quite small. The contribution of the last PCn is almost non-existent as the PCn scores are essentially zero across all observations (blue

dots), and the corresponding eigenvalues are virtually zero as well. The only significant process in dataset $\boldsymbol{D}$ appears to be the DNA melting process described by PC1.

Application of PCA has essentially led to a decomposition of the mean-centred dataset $\boldsymbol{M}$, to a matrix $\boldsymbol{S}$ that contains PC scores and describes different observations, multiplied by the PCA loading vectors $\boldsymbol{L}$ that contains spectral information according to:

$$\boldsymbol{M} = \boldsymbol{S}\boldsymbol{L}^T. \tag{5.5}$$

By selecting only those vectors from $\boldsymbol{S}$ and $\boldsymbol{L}^T$ that correspond to PC1, it is possible to reconstruct $\boldsymbol{M}$ using only a single principal component as outlined in the top of **Fig. 5.9a**. This is an example how PCA can be used for dimensionality reduction and to improve signal-to-noise ratios, as uncorrelated noise will appear in higher components that can be separated from the data of interest.

Analysing 2D-IR data via PCA works similarly, as a 2D-IR matrix can be concatenated to a long vector with 2D-pixels as variables instead of wavenumbers. A set of 2D-IR spectra is then represented in an analogous manner to FT-IR absorption data by a large matrix $\boldsymbol{D}$, with n spectra as rows and p pixels as columns.

5.5.3 When to Apply a PCA – Comparison to Regression Methods

A PCA returns relative differences between observations in a dataset with as few components as possible. This makes it possible to compare, separate and summarise spectral features across different samples in large, spectroscopic datasets without requiring any prior knowledge about the data. The generation of principal components purely based on covariance provides an unbiased, alternative representation of the data, which is important when evaluating the different responses in a screening experiment.

However, there are cases where there is more information about the underlying process available. In case of the FT-IR melting example, it is assumed that unfolding of the DNA duplex can essentially be described by a two-state process[22], where the drop in dsDNA concentration will directly be correlated to the gain in ssDNA concentration. While PCA identifies the melting process as the main source of variance in the dataset, it is unable to separate spectral features of dsDNA unfolding from ssDNA formation, as both follow the same sigmoidal function. The distinction based on covariance alone does not imply that an actual, physical process is depicted by the PCA. In these cases, a regression using a model function may lead to a better representation of the observed,

physical process. Two of those methods are outlined below in **Fig. 5.9b** and **c**, and are compared to the PCA approach in **Fig. 5.9a**.

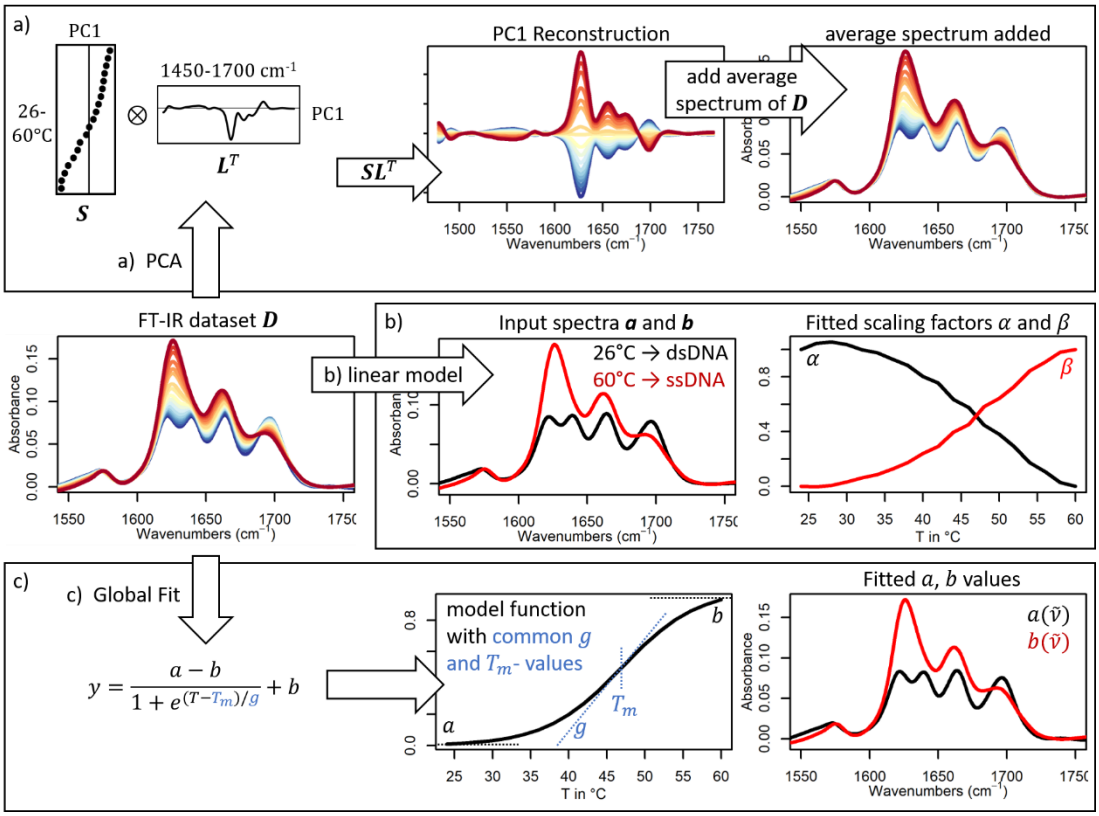


Fig. 5.9a) *A PCA on dataset* ***D*** *allows to represent the input data by a single principal component. PC1 can be used to reconstruct* ***D*** *while retaining 99.1% of the variance of the dataset.* ***b)*** *A linear model assumes that spectra at 26°C and 60°C correspond to the infrared response of dsDNA and ssDNA respectively. A linear combination of both spectra can be used to obtain scaling factors that indicate relative ssDNA and dsDNA concentrations throughout the melting experiment.* ***c)*** *A global fit using a common sigmoidal function across all wavenumbers of the dataset is used. a and b values are fitted individually for each wavenumber and return the spectral response of dsDNA and ssDNA respectively.*

In case of the FT-IR melting experiment, it can be assumed that only DNA duplexes are present in the sample far below the melting point and only single strands are present far above. The two spectra at 26°C and 60°C can therefore be used as model spectra $\boldsymbol{a}$ and $\boldsymbol{b}$ for dsDNA and ssDNA respectively (see **Fig. 5.9b**, black and red spectra). Each spectrum in the dataset can then be fitted by a linear combination of the dsDNA and ssDNA model spectra according to:

$$\boldsymbol{y} = \alpha\boldsymbol{a} + \beta\boldsymbol{b}. \tag{5.6}$$

The scaling factors α and β of such a linear model shows the consumption of dsDNA (**Fig. 5.9b**, right side, black trace) together with the formation of ssDNA (red trace), as

expected for an assumed two-state process. Such a method is applied and discussed in more detail in chapter 7.

Alternatively, it is possible to fit the temperature-dependent response of the FT-IR dataset to a sigmoidal function as illustrated in **Fig. 5.9c** using a global fit. In this approach, the absorbance change at each wavenumber is fitted using the following, sigmoidal model function:

$$y = \frac{a - b}{1 + e^{(T - T_m)/g}} + b. \tag{5.7}$$

Parameters a and b correspond to start and end values of the sigmoid and are optimised for each wavenumber individually. The sigmoidal parameters of the melting point T_m and the gradient g are shared across all wavenumbers and are optimised to a single value respectively. The dataset is thus represented by a single sigmoidal shape and the two parameters a and b, which represent two spectral responses. These are shown in the bottom right corner of **Fig. 5.9c** and correspond to the spectra of dsDNA and ssDNA respectively. A global fitting routine is used for time-resolved data in chapter 7 and is discussed there in more detail.

Both, the global fit and the linear combination approach make assumptions about the underlying process that generates the spectral response. The residuals of such fitting methods will indicate, if the used model is sufficient to give an accurate representation of the dataset or not. The choice of method always depends on the exact problem and should be decided on a case-by case basis. On a more abstract level, all three methods outlined in **Fig. 5.9c** aim to decompose a dataset $\boldsymbol{D}$ into two matrices $\boldsymbol{S}$ and $\boldsymbol{L}$, similar to equation (5.5). Matrix $\boldsymbol{L}$ contains spectral information, while matrix $\boldsymbol{S}$ aims to depict the variance across different spectroscopic measurements. Numerous additional methods exist (varimax rotation[24], principal component regression[25], alternating least squares[26], to name a few), which aim to find a decomposition $D \approx SL^T$ that gives the most comprehensive representation of the underlying processes.

With the spectroscopic and analytical methods introduced in this chapter, the focus can now be shifted to the main part of this thesis and the findings of the experiments.

5.6 References

1. Atkins, P. & Paula, J. de. *Physical Chemistry*. (Oxford University Press, 2009).
2. Mayerhöfer, T. G., Mutschke, H. & Popp, J. Employing Theories Far beyond Their Limits-

The Case of the (Boguer-) Beer-Lambert Law. *ChemPhysChem* **17**, 1948–1955 (2016).

3. Rullière, C. *Femtosecond Laser Pulses*. (Springer, 2005).

4. Pires, H., Baudisch, M., Sanchez, D., Hemmer, M. & Biegert, J. Ultrashort pulse generation in the mid-IR. *Prog. Quantum Electron.* **43**, 1–30 (2015).

5. Luther, B. M., Tracy, K. M., Gerrity, M., Brown, S. & Krummel, A. T. 2D IR spectroscopy at 100 kHz utilizing a Mid-IR OPCPA laser source. *Opt. Express* **24**, 4117 (2016).

6. Pupeza, I. *et al.* High-power sub-two-cycle mid-infrared pulses at 100 MHz repetition rate. *Nat. Photonics* **9**, 721–724 (2015).

7. Coherent Inc. RegA: High Repetition-Rate, Femtosecond Ti:Sapphire Amplifiers. *Coherent Product Catalog* 1–3 (2017). Available at: https://www.coherent.com/assets/pdf/COHR_RegA_DS_0217_1.pdf. (Accessed: 13th August 2019)

8. Coherent Inc. OPerA Solo: Ultrafast Optical Parametric Amplifier. *Coherent Product Catalog* **Rev. A**, 1–2 (2009).

9. Barret, P. H. & Palmer, M. PHAROS: High Power and Femtosecond Lasers. *Light Conversion Product Catalog* 4–9 (2019). Available at: http://lightcon.com/uploads/datasheets/PHAROS datasheets - 20190623.pdf. (Accessed: 13th August 2019)

10. Greetham, G. M. *et al.* A 100 kHz Time-Resolved Multiple-Probe Femtosecond to Second Infrared Absorption Spectrometer. *Appl. Spectrosc.* **70**, 645–653 (2016).

11. Donaldson, P. M., Greetham, G. M., Shaw, D. J., Parker, A. W. & Towrie, M. A 100 kHz Pulse Shaping 2D-IR Spectrometer Based on Dual Yb:KGW Amplifiers. *J. Phys. Chem. A* **122**, 780–787 (2018).

12. Hamm, P. & Zanni, M. *Concepts and Methods of 2D Infrared Spectroscopy*. (Cambridge University Press, 2011). doi:10.1017/CBO9780511675935

13. Shim, S., Strasfeld, D. B., Ling, Y. L. & Zanni, M. T. Automated 2D IR spectroscopy using a mid-IR pulse shaper and application of this technology to the human islet amyloid polypeptide. **104**, (2007).

14. Middleton, C. T., Woys, A. M., Mukherjee, S. S. & Zanni, M. T. Residue-specific structural kinetics of proteins through the union of isotope labeling, mid-IR pulse shaping, and coherent 2D IR spectroscopy. *Methods* **52**, 12–22 (2010).

15. Shim, S.-H. & Zanni, M. T. How to turn your pump–probe instrument into a multidimensional spectrometer: 2D IR and Vis spectroscopiesvia pulse shaping. *Phys. Chem. Chem. Phys.* **11**, 748–761 (2009).

16. Sanstead, P. J. & Tokmakoff, A. Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization. *J. Phys. Chem. B* **122**, 3088–3100 (2018).

17. Greetham, G. M. *et al.* Time-resolved multiple probe spectroscopy. *Rev. Sci. Instrum.* **83**, 103107 (2012).

18. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **2**, 559–572 (1901).

19. Gautam, R., Vanga, S., Ariese, F. & Umapathy, S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Tech. Instrum.* **2**, 8

(2015).

20. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **374**, 20150202 (2016).

21. Bazhulina, N. P. *et al.* Binding of Hoechst 33258 and its Derivatives to DNA. *J. Biomol. Struct. Dyn.* **26**, 701–718 (2009).

22. Chen, C., Wang, W., Wang, Z., Wei, F. & Zhao, X. S. Influence of secondary structure on kinetics and reaction mechanism of DNA hybridization. *Nucleic Acids Res.* **35**, 2875–2884 (2007).

23. Team, R. C. R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput.* (2014).

24. Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200 (1958).

25. Bair, E., Hastie, T., Paul, D. & Tibshirani, R. Prediction by Supervised Principal Components. *J. Am. Stat. Assoc.* **101**, 119–137 (2006).

26. Garrido, M., Rius, F. X. & Larrechi, M. S. Multivariate curve resolution–alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes. *Anal. Bioanal. Chem.* **390**, 2059–2066 (2008).

6 High-Throughput Screening of dsDNA Interactions with Hoechst 33258 using 2D-IR Spectroscopy and ANOVA-PCA

6.1 Abstract

Current 2D-IR spectrometers with short acquisition times have the potential to serve as a fast, analytical tool for screening experiments, providing structural and temporal information across large sample sizes. In this chapter, the use of 2D-IR spectroscopy as a method to screen dsDNA-ligand interactions is evaluated using the well-established minor groove binder Hoechst 33258. 2D-IR spectra of 12 different dsDNA sequences are studied in presence and absence of Hoechst 33258 to collect 2016 individual 2D-IR spectra in total. This large dataset is approached using analysis of variance combined with principal component analysis (ANOVA-PCA), which examines the spectral impact of three factors, pre-defined from the experimental design: the dsDNA sequence used, the presence of the ligand and the pump-probe waiting time of each 2D-IR spectrum. The method will identify, which 2D-IR features are affected by each factor individually, and which features arise from specific sequence-ligand-combinations. The observations compare well with published results from conventional analysis methods and show a qualitative correlation to a proxy for binding affinities for the studied dsDNA-ligand interactions. These findings confirm that 2D-IR spectroscopy is able to differentiate different dsDNA-ligand interactions from another and exemplify how a highly dimensional, spectroscopic dataset can be quickly surveyed.

This chapter contains results published in the following publication:

Fritzsch, R.; Donaldson, P. M.; Greetham, G. M.; Towrie, M.; Parker, A. W.; Baker, M. J.; Hunt, N. T.; Rapid Screening of DNA–Ligand Complexes via 2D-IR Spectroscopy and ANOVA–PCA. *Anal. Chem.* **90**, 2732–2740 (2018).

6.2 Introduction

To understand the molecular interactions behind minor groove binding, researchers rely on a variety of complementary analysis methods such as NMR[1–3], crystallography[4,5] or DNA footprinting[6], to name a few. 2D-IR spectroscopy has the potential to extend this analytical repertoire as it gained relevance in examining structure and vibrational dynamics of DNA sequences in general[7–12], and has recently proven sensitive to interactions from minor groove binding specifically[13]. The 2D-IR spectrum gives an intuitive representation of the network of the coupled, vibrational modes in the dsDNA, which are sensitive to local structural changes. The interaction with a ligand can be registered as a perturbation of the affected dsDNA modes, providing a label-free analysis method to study minor groove binding under physiological conditions in solutions. Recent advancements in instrumentation have further increased the data throughput of 2D-IR spectroscopy significantly. The use of mid-infrared pulse-shapers[14,15] and laser systems operating at pulse repetition rates of 100 kHz[16,17] have reduced acquisition times to a few seconds per spectrum, providing researchers with a structurally sensitive, high-throughput spectroscopy method. This creates an opportunity to study dsDNA-ligand interactions with 2D-IR spectroscopy across many different samples on an unprecedented scale.

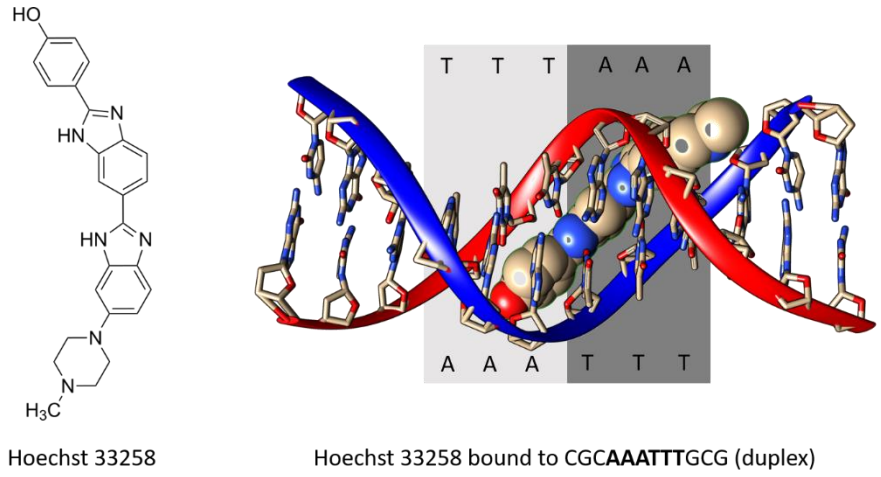


Fig. 6.1 *Chemical structure of Hoechst 33258. Crystal structure of the ligand bound to an A-tract minor groove in 5'-CGCAAATTTGCG-3'. Crystal structure 264D published by Vega et al.*[5] *.*

To test the viability of the 2D-IR method as a screening tool, a proof-of-concept experiment is conducted using a well-established ligand-dsDNA system. Hoechst 33258 (H33258) is a fluorescent dye and minor groove binder with therapeutic relevance[18,19] that targets the minor groove of A/T-rich sequences[20]. The interactions of H33258 with

dsDNA were studied extensively in the past[6,20–26]. In-place substitutions of guanines with inosine nucleobases have shown that the exocyclic N-2 amino group of guanines protrude the minor groove and sterically prevent H33258 from binding to G/C structures[21,25]. Minor groove binding of H33258 is non-covalent with a preference of "A-tract" structures, A_nT_n ($n \geq 2$, illustrated in **Fig. 6.1**), over alternating A/T motifs, $(AT)_n$. Minor grooves with A-tracts in particular are solvated by a well-defined spine of hydration[24,25], which leads to a considerable gain in entropy when replaced by H33258[27]. More importantly to this chapter, H33258 is the first minor groove binder that was studied via 2D-IR spectroscopy. The interactions of H33258 to an optimal dsDNA sequence with A-tract were compared to interactions to a suboptimal ATATAT motif. 2D-IR spectroscopy identified an up-shift in frequencies of the T_{2S} and T_{4S} carbonyl-modes of thymine, with the T_{2S}-shift being assigned to the replacement of hydrogen-bond interactions of the solvent in the minor groove by hydrogen bonding to H33258. The T_{4S}-shift was assigned to a subtle change in dsDNA propeller twists, indicative of an *induced fit* mechanism of the dsDNA structure[2,13]. The ability to resolve such structural changes from H33258 binding makes this ligand an ideal test-system for a proof-of-concept screening experiment for 2D-IR spectroscopy. The current chapter will extend these experiments to twelve dsDNA sequences with different minor grooves and collect time-resolved 2D-IR data to examine the effect of minor groove binding to the vibrational relaxation dynamics, generating 2016 2D-IR spectra in total. The observations will be compared to the previously published 2D-IR results as well as fluorescence and dsDNA melting experiments, which serve as a proxy for the expected binding affinity of each dsDNA-ligand combination.

The ability to produce large sets of spectroscopic data also requires a more rigorous analytical approach. Analysing individual 2D-IR spectra in detail can be extremely insightful but is not scalable with large sample sizes. An analytical method is required that can survey the collected, spectroscopic responses and identify individual samples with distinctive spectral features. Once identified, the 2D-IR spectrum of this sample can then be analysed in more detail. Principal component analysis (PCA) introduced in the last chapter can provide an overview of the collected dataset and enables quick differentiation of spectral responses. Screening experiments can get quite complex when (potentially) multiple ligands are interacting across a set of different DNA sequences. Time-resolved 2D-IR experiments provide an additional level of complexity by scanning the waiting time between pump and probe pulse for each of these samples

as well. An individual 2D-IR spectrum in such a case therefore depends on three different factors: the dsDNA sequence used, the ligand added to the dsDNA and the waiting time. In order to separate the effect of each of those factors and analyse them separately, analysis of variance (ANOVA) can be employed. A combination of both methods, referred to as ANOVA-PCA, was initially developed to study large mass-spectrometry datasets to analyse amniotic fluids and has subsequently been applied in a variety of scenarios[28–30] including mid-infrared absorption experiments[31]. This method will be used in this chapter to approach a large, spectroscopic dataset. Its viability as an analysis tool will be assessed and key advantages and limitations will be discussed.

6.3 Methods

6.3.1 Overview of the Experimental Design

In this study, 2D-IR spectroscopy is used to examine the interaction of twelve different dsDNA sequences with H33258, which are outlined as rows in the table of in **Fig. 6.2** (highlighted in orange). Sequences s_1 to s_7 contain A/T-rich minor groove sequences which are expected undergo minor groove binding with Hoechst 33258 to varying degrees. Sequences s_8 to s_{12} show a G/C-rich minor groove and are not expected to show minor groove binding. The ability to bind the ligand is independently assessed for each individual sequence, as discussed in section 6.3.4. The 2D-IR response of each sequence was investigated on its own, and in presence of Hoechst 33258 (columns in Fig. 6.2, blue) in a 1:1 ratio, leading to 24 different samples in total. For each of these 24 samples, a series of time-resolved 2D-IR spectra is acquired to obtain temporal information of the observed, vibrational modes of dsDNA. Spectral replicates (repeated measurements) were treated as individual data points rather than being averaged to capture the spectral variances between measurements, generating 2016 2D-IR spectra in total (as indicated by the solid points in the tables of **Fig. 6.2**).

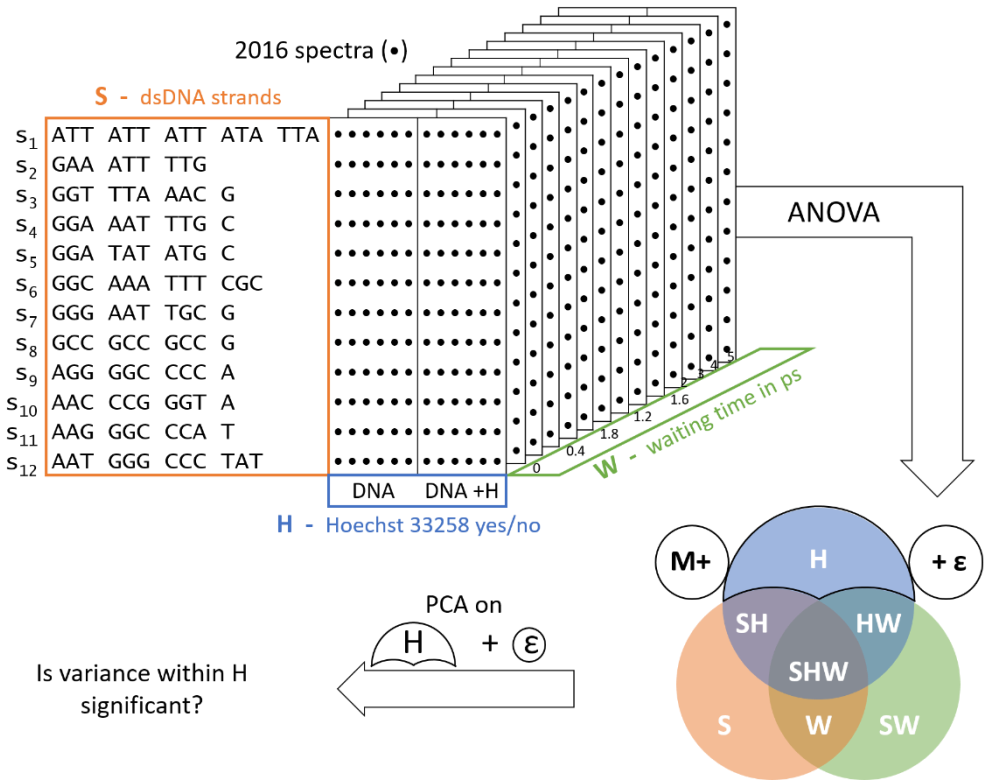


Fig. 6.2 *Experimental design and schematic representation of the ANOVA-PCA method. The set of 2D-IR spectra studied contains three main sources of variance (factors): the sequence of dsDNA,* ***S****, the presence of H33258,* ***H****, and the waiting time,* ***W****. The ANOVA method generates subsets containing the variances attributable to each factor, as well as the variance due to their interactions (e.g. (****SH****); the sequence-dependent impact of ligand binding) and the residual variance* ***ε*** *(e.g. noise). Individual subsets are compared to residuals,* ***ε****, and analysed using PCA to test for significance.*

According to the design of the experiment outlined in **Fig. 6.2**, the outcome of a 2D-IR measurement is dependent on three controlled *factors*: The dsDNA sequence used, ***S***; the presence/absence of the ligand in the sample, ***H***; and the waiting time at which the 2D-IR spectrum was acquired, ***W***. The first part of the method, the ANOVA, aims to separate the variance of the dataset according to the three confounding factors ***S***, ***H***, ***W*** into subsets. The second part, the PCA, is used to compare each subset against a residual variance (e.g. spectral noise) so that the effect of factor is tested for significance and can be studied individually. This allows for example to separate spectral features that arise from a change in dsDNA sequence from features that emerge due to interaction with the ligand. This method was chosen, as it is easily scalable to interactions of multiple ligands scanned across multiple different dsDNA sequences. Analysis can start from a simple subset of the dataset and can be gradually extended to explore the effect of each factor individually and in combination. The separation into subsets is largely based on calculating average spectra and is explained

in detail in section 6.3.6. The analysis of each individual subset will be discussed in the results chapter.

6.3.2 Materials and Sample preparation

The twelve oligomers outlined in Fig. 6.2, together with their corresponding, complementary strands were purchased as salt-free, lyophilized solids from Eurofins. TRIS base (tris(hydroxymethyl)aminomethane), NaCl, D_2O, DCl and H33258 were obtained from Sigma-Aldrich and used without further purification.

A stock solution of 40 mM was prepared for each oligomer using a deuterated TRIS-buffer (100 mM TRIS, 100 mM NaCl, pD 7.0). Solutions of complimentary strands were combined in equimolar ratios, diluted and annealed at 95°C for 5 min to form dsDNA samples. Samples containing H33258 were prepared by annealing the dsDNA in equimolar ratio with a stock solution of H33258 in TRIS buffer. The final dsDNA concentration for all samples was 10 mM.

6.3.3 Infrared Spectroscopy

Each sample was measured via linear infrared absorption spectroscopy and 2D-IR spectroscopy using a liquid transmission cell. The cell is illustrated in **Fig. 6.3** and consisted of two CaF_2 windows separated by a polytetrafluoroethylene (PTFE) spacer to achieve a sample path length of 25μm. Fourier-transformed, infrared (FT-IR) absorption spectra were acquired using a Thermo Scientific Nicolet iS10 spectrometer between 1550 and 1800 cm^{-1} at an optical resolution of 4 cm^{-1}. Six infrared absorption spectra were obtained for each sample; three before and three after the 2D-IR experiment to check for inconsistencies and characterise the variance of infrared measurements. A separate absorption spectrum of the pure buffer was used to subtract the solvent background in all FT-IR spectra.

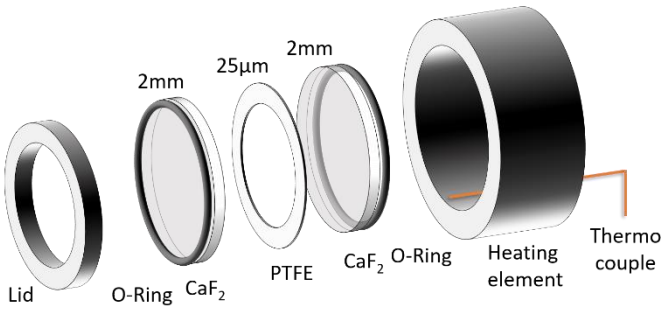


Fig. 6.3 *Illustration of a liquid transmission cell. The metal jacket of the cell contains a heating element for melting experiments. The thermocouple is located inside the jacket at the edge to the CaF_2 windows and the sample.*

2D-IR experiments were taken out using the LIFEtime instrument at the Rutherford Appleton Laboratories as outlined in chapter 5.3 [17,32]. The instrument operates at 100 kHz repetition rate and uses a mid-infrared pulse shaper for high throughput measurements. OPA1 and OPA2 were used to generate mid-infrared pulses centred at 1650 cm^{-1} for pump and probe beam, resonant with carbonyl and ring stretching modes of the nucleobases. The pump-pump delay, t_1, was scanned using the pulse shaper[14] from 0 to 4 ps at a step size of 24 fs with phase cycling methodology. Each 2D-IR spectrum was acquired for 40 s, which is equivalent to ca. 5900 individual 2D-IR measurements averaged in that time. It is noted however, that little signal-to-noise improvement was observed after 10 s, signifying that further reduction of the acquisition time may be possible in the future. The pump-probe delay, now defined as waiting time W, was scanned from 0 to 5 ps to obtain time-resolved 2D-IR spectra at 14 waiting times for each sample. Each measurement was repeated six times using the same sample in the cell to check for consistency of the measurement and to estimate the variance of each measurement. This will make it possible to estimate the significance of the 2D-IR responses as discussed in the results chapter. 2016 2D-IR spectra were acquired in total. The pure acquisition time of the complete dataset (disregarding sample preparation and changing of samples) was less than 24h, with the potential to further reduce acquisition times by a factor of 4 (10s averaging instead of 40s). In this experiment, the data acquisition was largely limited by the time it took to prepare the sample cells.

6.3.4 Estimating Binding Affinity – Melting Experiments and Fluorescence Data

Previous studies have shown that H33258 prefers A/T-rich minor grooves according to the following series with decreasing binding affinity: AATT ≫ TAAT ≈ ATAT > TATA ≈

TTAA [6,24]. In order to estimate the binding affinity of H33258 to the dsDNA sequences used in this study, the interactions were categorized into three qualitative groups by assessing results from additional fluorescence spectroscopy and dsDNA melting experiments. This grouping will be used to relate the results from 2D-IR data to observations from more conventional experiments.

Effective minor groove binding has a stabilising effect on the DNA duplex structure and raises the melting temperature T_m, of the dsDNA[33,34]. Melting experiments are often provided as complementary data when assessing different dsDNA-ligand interactions and are used as a proxy for binding affinity[35–37]. The relative change in melting temperature indicates the ligands ability to stabilise the double helix and is referred to as melting point stabilization ΔT_m in the following. The dsDNA melting experiments were performed by heating each sample in the liquid transmission cell (**Fig. 6.3**) and monitoring the FT-IR absorption spectrum. The IR absorbance of each dsDNA sequence was measured with and without H33258 at 10 mM from 20°C to 95°C in 5°C steps. The collected infrared absorption data are shown in the appendix of this chapter in **Fig. A-6.1**. From this data, the melting transition of each sample was obtained by applying a separate PCA to each of the 24 melting responses and analysing the PC1 score, as outlined earlier in chapter 5.5.2 and as published previously[12,38]. The PC1 scores attained for each sample are shown in **Fig. A-6.2** and were subjected to a sigmoidal fit to extract the melting point T_m using equation (6.1):

$$p(T) = \frac{p_{\text{start}} - p_{\text{end}}}{1 + e^{(T-T_m)/g}} + p_{\text{end}} \,. \tag{6.1}$$

The obtained PC1 scores $p(T)$ were fitted to a sigmoidal function with optimised parameters for start and end values p_{start} and p_{end} of the sigmoid, the transition point T_m and the slope g. The melting point stabilisation ΔT_m was then calculated from the melting temperatures with and without ligand for each sequence and is shown in **Fig. 6.4a**.

Hoechst 33258 has found wide applications as a dsDNA staining dye and will show a significant rise in fluorescence once bound to the minor groove[39,40]. The fluorescence of H33258 when added to dsDNA solutions is therefore another indicator that can be used to characterise the dsDNA-ligand interaction. The rise in fluorescence of H33258 due to minor groove binding was determined using a Horiba Fluorolog 3 spectrometer at an excitation wavelength of 352 nm (H33258 absorption maximum) with a quartz cuvette

of 1 cm path length. The emission spectrum was collected of each dsDNA-H33258 combination in equimolar ratio at a dilution of 7 µM for three aliquots and the emission values at 450 nm (H33258 emission maximum) were determined (see **Fig. A-6.3**). The emission values of each sample were divided by the emission of 7 µM H33258 in buffer without any dsDNA, to give the fluorescence enhancement factor as plotted in **Fig. 6.4b**.

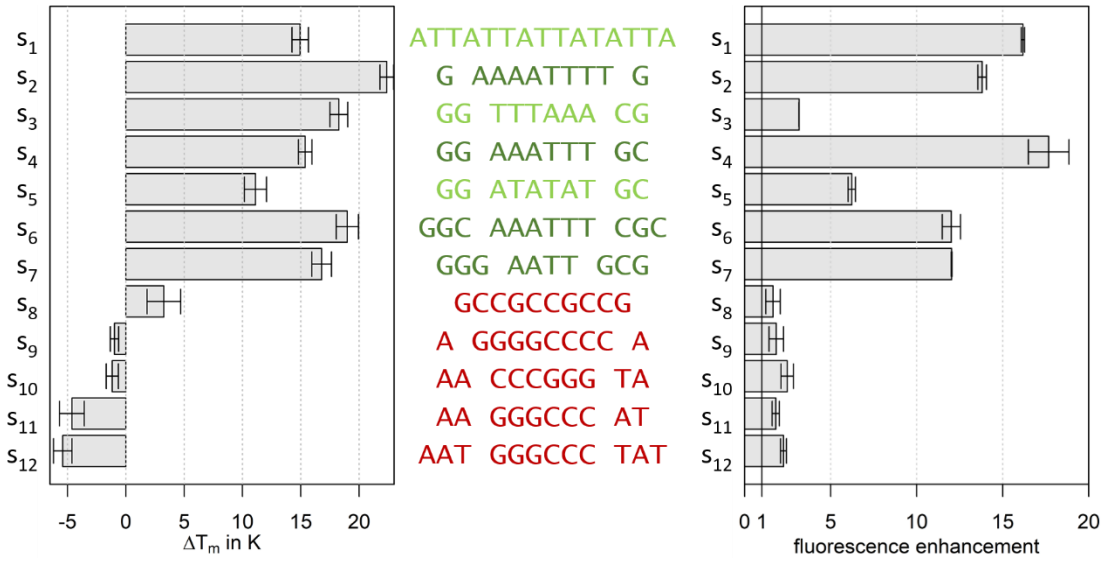


Fig. 6.4 *Estimation of the H33258 binding affinity with sequences s_1-s_{12} using melting temperature stabilisation in* ***a)*** *and fluorescence emission of H33258 in* ***b)****. The colours of each sequence in the centre indicate a qualitative estimation of the binding affinity and are used as a colour gradient in* ***Fig. 6.10*** *and* ***Fig. A-6.5****.*

The results summarised in **Fig. 6.4** largely follow the observations in the literature[6,24,41]. DNA sequences s_1 to s_7 with an A/T minor groove show a significant melting point stabilisation (>10 K) in contrast to sequences s_8 to s_{12} with G/C minor groove, with little or no stabilisation of the duplex structure. From those sequences with A/T minor groove, s_1 (5'-ATTATTATTTATTTA-3') and s_5 (5'-GGATATATGC-3') show lower ΔT_m values than sequences with an A-tract, such as s_2 (5'-GAAAATTTTG-3') and s_6 (5'-GGCAAATTTCGC-3'). Fluorescence measurements confirm the clear preference of A-tract sequences s_2, s_4, s_6 and s_7, which show a rise in fluorescence by a factor of 10 and higher. Sequences s_3 (5'-GGTTTAAACG-3') and s_5 (5'-GGATATATGC-3') show a more modest rise in fluorescence by a factor below 7.

These observations are summarised by the colour scheme from red to green indicated in **Fig. 6.4**. Combinations of H33258 with s_8 to s_{12} are coloured in red as little or no minor groove binding is observed. Combinations with A-tract sequences s_2, s_4, s_6 and s_7 lead to significant minor groove binding and are coloured in dark green. Suboptimal

minor grooves of s_1, s_3 and s_5 are indicated in light green. These three groups will be used as a qualitative proxy for binding affinities when analysing 2D-IR data.

6.3.5 Pre-Processing of 2D-IR Data

The pump and probe pulses that were used to generate the 2D-IR spectrum have a pulse duration of ca. 300 fs and 200 fs FWHM respectively. As a consequence, pump and probe pulses will overlap at short waiting times and can generate additional signals that obscure the desired 2D-IR response. 2D-IR spectra obtained at waiting times below 400 fs were therefore excluded from the dataset.

Multivariate analysis methods usually require a single input matrix that contains the complete dataset. In order to represent the collected 2D-IR data by a single matrix, each 2D-IR spectrum was concatenated into vector-form so that each 2D – pixel is treated as a separate variable. The complete data matrix $\boldsymbol{X}$ is illustrated in **Fig. 6.6a** and encompasses the 2D-IR spectra as individual rows x_n. The intensity at each pixel in the 2D-IR spectrum is represented by the columns of $\boldsymbol{X}$.

Additionally to the spectral features of interest, a PCA will also highlight any variance arising from sample thickness, concentration or laser intensity differences. To mitigate these effects, a technique called *vector-normalisation* can be applied. This technique is commonly used as a pre-processing step when analysing large, linear infrared absorption datasets[42–44]. The method treats each spectrum as an individual vector x and divides it by its magnitude $\|x\|$, to normalise it to a length of one. This is outlined in the following equation (6.2):

$$x' = \frac{x}{\|x\|}, \text{ where } \|x\| = \sqrt{{x_{(1)}}^2 + {x_{(2)}}^2 + \cdots + {x_{(p)}}^2}. \tag{6.2}$$

Vector x is an individual 2DIR spectrum written in vector-form x, with $x_{(1)}, x_{(2)}, \ldots, x_{(p)}$ being intensities observed at each pixel, p. The squared intensities in a 2D-IR spectrum are directly proportional to the variance and covariance analysed with PCA. Vector-normalisation rescales the magnitude of each spectrum to resemble the same overall variance, which ensures that each spectrum is weighted equally when a PCA is performed.

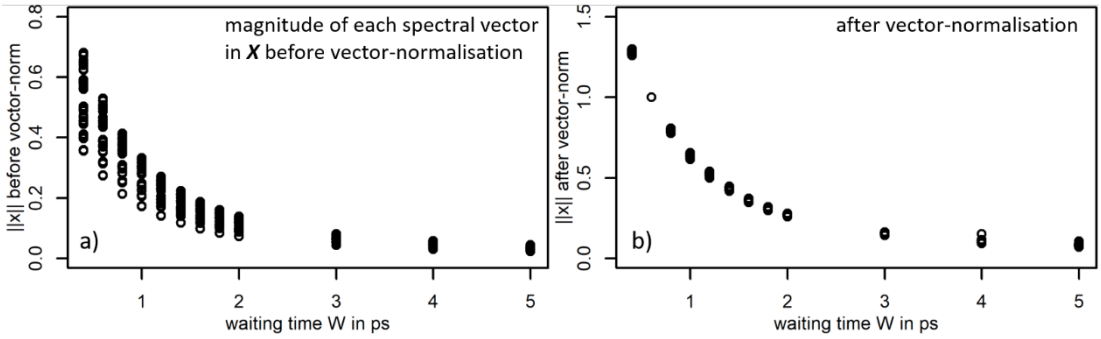


Fig. 6.5 *The magnitude* $\|x\|$ *of each spectral vector in the dataset before and after the vector-norm.*

All 2D-IR spectra collected at a waiting time of 600 fs were vector-normalised. Each vibrational response in the 2D-IR spectrum will relax over time and will therefore return a lower magnitude at late waiting times. Each individual set of time-resolved 2D-IR measurements from 0.4 to 5 ps was therefore divided by the same normalisation factor, the magnitude $\|x\|$ obtained from the corresponding 2D-IR spectrum at 600 fs. This reduced the variance across samples and across repeats, but retained the temporal response over time. The effect is indicated in **Fig. 6.5**, where the vector-magnitude of each 2D-IR spectrum in the dataset is plotted in dependency of the corresponding waiting time before- and after vector normalisation. As a result, the principal variance in the dataset will be less affected by scaling differences between spectra.

6.3.6 ANOVA-PCA: Separation of *X* into Subsets

As mentioned previously, each collected 2D-IR response in ***X*** is affected by the choice of sequence (S), by the absence or presence of the ligand (H) and finally the waiting time (W). A change in sequence will affect different spectral features than an addition of H33258 to the sample. Separating the different spectroscopic effects of these three main *factors* S, H, W is the goal of the ANOVA-PCA in its first part.

Fig. 6.6a shows dataset ***X*** with a unique identifier for each row in the matrix. Each spectrum x_n is described in terms of the sequence of the dsDNA, s_i (i = 1-12 sequences); the presence of the minor groove binder H33258, h_j (j = +/- for ligand presence/absence) and the waiting time indicated by w_k (k = 0.4-5 where the value indicates the waiting time in ps). Each sample was repeated six times, indicated by r_l (l = 1-6). Following this notation, the second replicate of sequence s_5, with H33258 present, at a waiting time of 0.8 ps, is indicated by $s_5h_+w_{0.8}r_2$ as illustrated in **Fig. 6.6a**.

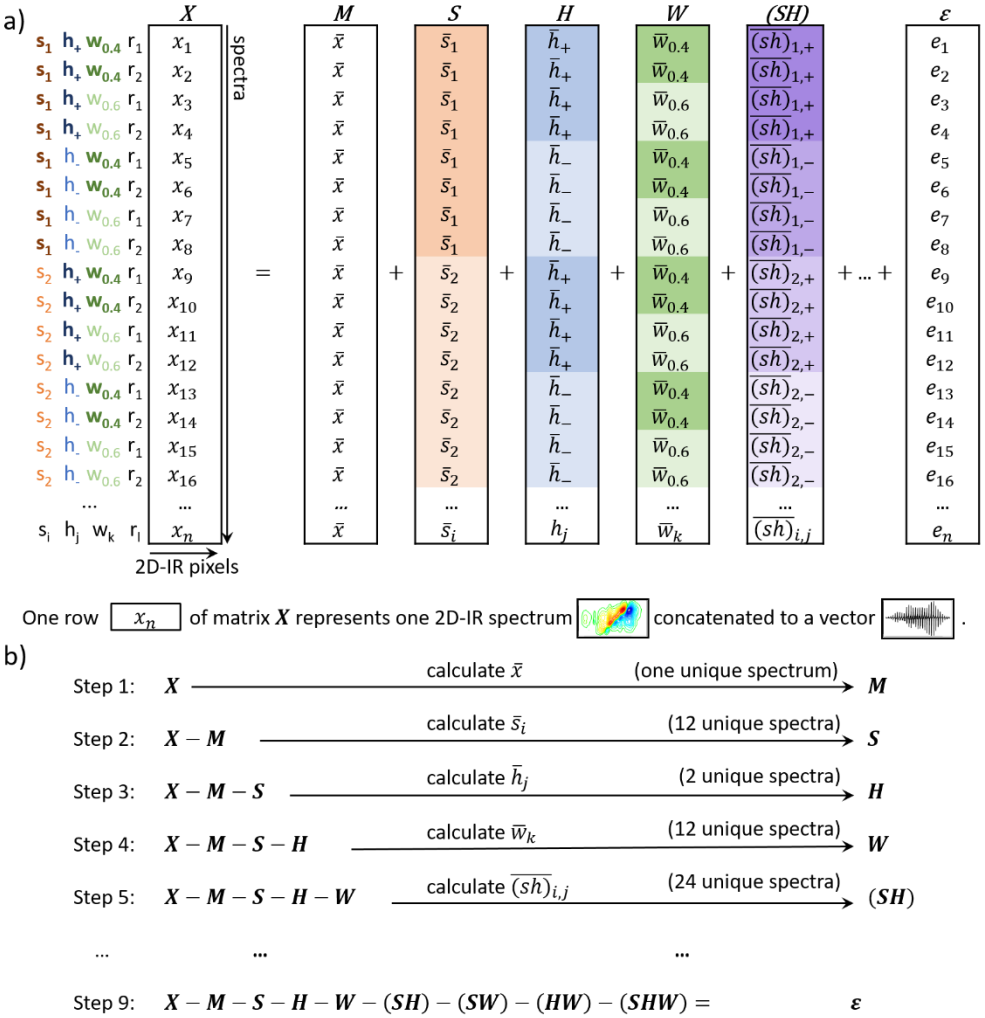


Fig. 6.6a) *Schematic representation of data matrix* ***X*** *and its decomposition into factor matrices. Blocks of colour indicate averaged spectra in each factor matrix. Each row represents one 2D-IR spectrum concatenated to a vector.* ***b)*** *Schematic structure of the step-wise subtraction to calculate factor matrices. The algorithm allocates variance from the raw data based on their origin into several new matrices with the same dimensions by calculating averages. The residual matrix contains the variance between repeats, ideally representing only instrumental noise.*

The ANOVA-PCA method is implemented as introduced in the literature by Harrington *et al.*[28] . Based on the three factors S, H and W, a series of new matrices is sequentially calculated from matrix $\boldsymbol{X}$. These matrices are then subtracted from $\boldsymbol{X}$ to obtain the residual matrix, $\boldsymbol{\varepsilon}$. This procedure allows to separate the variance of $\boldsymbol{X}$ into matrices of three main experimental factors $\boldsymbol{S}$, $\boldsymbol{H}$, $\boldsymbol{W}$; four interactions $(\boldsymbol{SH})$, $(\boldsymbol{SW})$, $(\boldsymbol{HW})$, $(\boldsymbol{SHW})$ and residual matrix $\boldsymbol{\varepsilon}$; based on known information about the dataset:

$$\begin{aligned} \boldsymbol{X} = \quad & \boldsymbol{M} + \boldsymbol{S} + \boldsymbol{H} + \boldsymbol{W} \\ & + (\boldsymbol{SH}) + (\boldsymbol{SW}) + (\boldsymbol{HW}) + (\boldsymbol{SHW}) \\ & + \boldsymbol{\varepsilon} \end{aligned} \tag{6.3}$$

The matrices in equation (6.3) are calculated in a stepwise manner as shown in **Fig. 6.6b**. In the first step, an average spectrum $\bar{x}$ is calculated from all spectra in $\boldsymbol{X}$ and a new matrix $\boldsymbol{M}$ is created. $\boldsymbol{M}$ has the same dimensions as $\boldsymbol{X}$ but contains $\bar{x}$ for every row as illustrated in **Fig. 6.6a**. Subtraction $\boldsymbol{X} - \boldsymbol{M}$ generates a residual matrix which is used in the next step as starting point. In step 2, the matrix representing the sequence *factor,* $\boldsymbol{S}$, is generated. All rows of $\boldsymbol{X} - \boldsymbol{M}$ that correspond to the same sequence s_i are averaged to give spectra $\bar{s}_i$ (i = 1-12). The twelve average spectra $\bar{s}_i$ are repeated according to their dedicated rows to form matrix $\boldsymbol{S}$. The subtraction $\boldsymbol{X} - \boldsymbol{M} - \boldsymbol{S}$ then results in a new residual matrix containing reduced variance. In step 3, $\boldsymbol{X} - \boldsymbol{M} - \boldsymbol{S}$ is used to calculate two average spectra $\bar{h}_j$, thus forming matrix $\boldsymbol{H}$**,** which accounts for the ligand *factor*. $\boldsymbol{X} - \boldsymbol{M} - \boldsymbol{S} - \boldsymbol{H}$ is then used to create matrix $\boldsymbol{W}$**,** in which all spectra with a common waiting time (w_k) are averaged. This is subtracted in turn to account for the waiting time *factor*.

These matrices represent variance from one main factor, while completely disregarding the effect of the other two factors at the same time. The spectral impact of different combinations of the three main factors is then considered by *interactions* between factors, where average spectra are calculated for each possible combination. Matrix ($\boldsymbol{SH}$) for example contains 24 unique spectra $\overline{(sh)}_{i,j}$ for twelve sequences (i), each measured with and without H33258 (j). This matrix will contain the spectral features that are specific to individual dsDNA-ligand combinations. Each interaction matrix is calculated as before and subsequently subtracted from $\boldsymbol{X}$ to further reduce the residual variance. The method is repeated until matrices for all factors and interactions are calculated and subtracted from the raw data to give residual matrix $\boldsymbol{\varepsilon}$, which contains the variance between different repeats such as the instrumental noise.

It is important to note that this method requires a balanced dataset[28], meaning that the number of spectra analysed is exactly the same for each level of each factor (i.e. each DNA sequence has been measured the same number of times). A visual representation of a balanced dataset is also shown in **Fig. 6.2**, as each row and each column in tables of this figure is represented by the same number of measurements. If this is not the case, the order of the sequential subtraction will have an impact on the results and the variance of one *factor* may end up in the matrix of another. This has implications for outliers. Three out of six repeats of one particular sample were obscured by scatter-artefacts (apparent from visual inspection) and had to be excluded from the analysis.

To maintain a balanced dataset, three repeats from all other measurements had to be removed as well, reducing the total number of spectra analysed to 864.

6.3.7 ANOVA-PCA: Studying Subsets via PCA

After the decomposition of $\boldsymbol{X}$ into subsets, each factor or interaction can be tested individually for significance via PCA. In spectroscopic terms, it means that it is possible to test, whether for example the addition of the ligand will have a meaningful impact on the 2D-IR response. The factor or interaction matrix of interest can be added to the residual matrix $\boldsymbol{\varepsilon}$, which characterises the uncertainty of the measurement. A PCA of this sum will evaluate, whether the principal variance of the sum is dominated by noise from $\boldsymbol{\varepsilon}$, or a systematic effect from the factor. If the effect of the factor or interaction is significant and outweighs the residuals, then the scores of PC1 will be dominated by the factor. Conversely, if the variance of the residuals outweighs any systematic effects from the factor, the scores of PC1 will be dominated by noise from the residuals and the factor has no significant impact on the spectrum. It is further possible to add more than one factor or interaction to the residuals in order to gradually increase the complexity of the data analysed. So if analysis of subset $\boldsymbol{H} + \boldsymbol{\varepsilon}$ established that the ligand does have a general effect on the 2D-IR spectrum, it can be specified whether this effect is sequence specific by extending the subset to $\boldsymbol{H} + (\boldsymbol{SH}) + \boldsymbol{\varepsilon}$.

Finally, the observations from 2D-IR spectroscopy were also compared to FT-IR absorption data using ANOVA-PCA in an analogous manner. The FT-IR data is only dependent on two factors, which will lead to a matrix for sequence-dependence, $\boldsymbol{S}$, a matrix for ligand presence, $\boldsymbol{H}$, and a matrix for the combination of the two, ($\boldsymbol{SH}$), following equation (6.4):

$$\boldsymbol{X} = \boldsymbol{M} + \boldsymbol{S} + \boldsymbol{H} + (\boldsymbol{SH}) + \boldsymbol{\varepsilon} \tag{6.4}$$

The ANOVA-PCA results of the FT-IR dataset were analysed in analogy to the 2D-IR results and are shown in the appendix in **Fig. A-6.4** and **Fig. A-6.5**.

6.4 Results and Discussion

6.4.1 Base Sequence Dependence

The results of the ANOVA-PCA need to be validated before applying the method in the context of a screening experiment. Matrix $\boldsymbol{S}$ is the first subset generated from the ANOVA-PCA and identifies sequence specific, spectral features. The 12 dsDNA

sequences investigated in the dataset cover a wide range from s_1 being a pure A/T sequence to s_8 as a pure G/C sequence. 2D-IR spectroscopy is sensitive to the dsDNA sequence as discussed in the introduction in chapter 4.3.4 and so the ANOVA-PCA results should be able to differentiate the response of different oligomers using subset $\boldsymbol{S}$.

Analysis of subset $\boldsymbol{S} + \boldsymbol{\varepsilon}$ via PCA lead to the PC1 and PC2 scores shown in **Fig. 6.7**. The scores plot suggests, that every spectrum falls into one of twelve clearly separated groups representing the twelve average spectra of matrix $\boldsymbol{S}$. The residuals in $\boldsymbol{\varepsilon}$ correspond to the spread of each point cloud. This represents the intra-group variance of each sequence, which is minimal compared to the inter-group variance that differentiates them. This result indicates, that each sequence returns a unique 2D-IR spectrum.

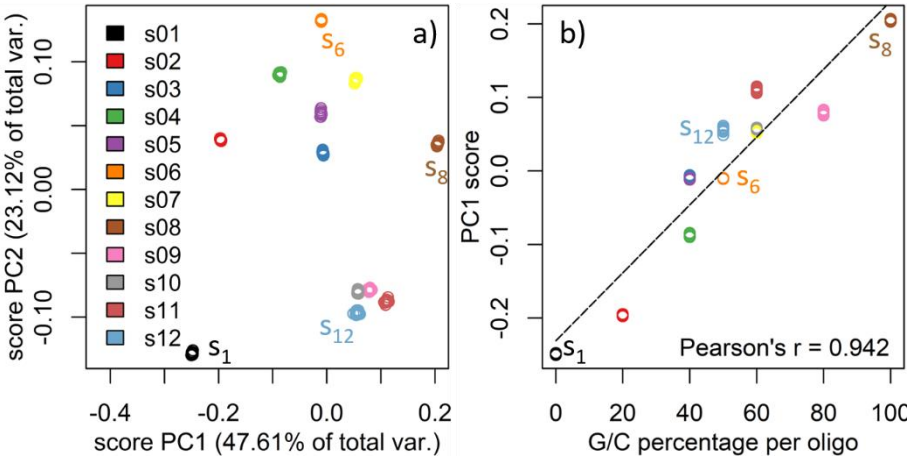


Fig. 6.7 a) *ANOVA-PCA results for dsDNA sequence dependence* $\boldsymbol{S} + \boldsymbol{\varepsilon}$*. Each point in the scores plot represents one 2D-IR spectrum and similar scores between spectra indicate common spectral features.* ***b)*** *The scores in PC1 show a good correlation to the GC percentage of the sequences measured.*

It is striking, that the responses of the pure A/T sequence, s_1 (black), and the pure G/C sequence, s_8 (brown), are spread across opposite ends of the observed PC1 scores. To check for a possible correlation, the obtained PC1 scores are plotted in **Fig. 6.7a** against the base composition of each sequence, in terms of how many nucleobases are G or C in a sequence (GC percentage). There is a clear, linear correlation between the PC1 score and the GC percentage, which is identified by a high Pearson correlation coefficient of 0.94. This is a clear indication that the 2D-IR spectra in this subset are primarily affected by the vibrational modes of the four nucleobases in this region and suggests that the PC1 score can be used to estimate the base composition of a given dsDNA sequence.

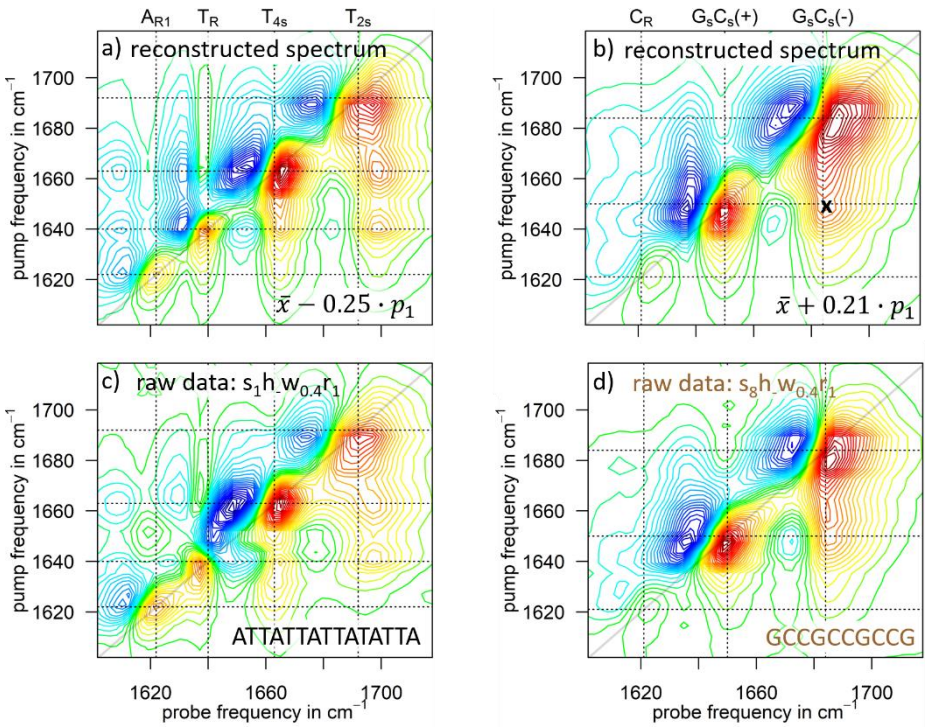


Fig. 6.8 *Top: Reconstructed spectra using minimum* ***a)*** *and maximum PC1 scores* ***b)****. Peak assignments according to literature are shown.*[7,12] *$\bar{x}$ and p_1 are vectors defined as the global average spectrum and the loading vector of PC1 respectively. Bottom: Raw 2D-IR data at 0.4 ps of* ***c)*** *DNA sequence s_1 consisting purely of A-T base pairs and* ***d)*** *sequence s_8 with only G-C base pairs.*

To identify which spectral features are characterised by PC1, the PC1 loading vector is considered next. The observed PC1 scores cover values from -0.25 for s_1 (all A/T, black) to 0.21 for s_8 (all G/C, brown). Multiplication of the minimum and maximum PC1 score with the loading vector of PC1 (p_1) respectively and addition of the global mean spectrum ($PC1\ score \cdot p_1 + \bar{x}$) leads to the 2D-IR spectra shown in **Fig. 6.8a** and **b**. This spectral reconstruction is done instead of plotting the PC1 loading vector directly, as p_1 contains relative differences rather than absolute features, making an assignment far less intuitive. These calculated spectra are compared to the raw 2D-IR spectra of s_1 (all A/T) and s_8 (all G/C) collected at a waiting time of 0.4 ps in **Fig. 6.8c** and **d**. PC1 essentially retrieves the vibrational modes of nucleobases A and T at negative PC1 scores, and the modes of G and C at positive PC1 scores. Moreover, off-diagonal features are present as well in the reconstructed spectra; notably the strong coupling between carbonyls of G and C (cross in **Fig. 6.8b**), which indicates that the inter-strand coupling present in G-C base pairs is retrieved[7]. The solid lines in **Fig. 6.8** correspond to literature values for carbonyl stretch and ring modes of dsDNA for the two Watson-Crick base pairs and agree very well with peak positions observed in the reconstructed spectra. PC1 therefore seems to retrieve the generic 2D-IR response of a G-C base pair

and a A-T base pair; irrespective of the base pair sequence along the dsDNA. The base composition therefore has the largest effect on the 2D-IR spectrum in this subset.

This becomes especially apparent in the PCA results of sequences s_6 (5'-GGCAAATTTCGC-3', orange) and s_{12} (5'-AATGGGCCCTAT-3', bright blue). Sequence s_6 has G/C structures at the ends of the sequence and an A/T in the centre, s_{12} is inverted and has A/T ends and a G/C centre. Both show a PC1 score close to zero, indicating that their base composition of 50% GC is identical. However, their PC2 score is on opposite ends of the observed PC2 range. This suggests that spectral features arising from interactions along the strand are more subtle and therefore appear at higher principal components. A similar observation was made in the literature, where interstrand and intrastrand coupling within pure G/C dsDNA strands was studied[8]. Higher principal components will provide more information to characterise each sequence in detail and could be used in a separate project to study the subtle interactions along the strand and at different positions in the dsDNA strand. For this chapter though, subset $\boldsymbol{S} + \boldsymbol{\varepsilon}$ has confirmed that 2D-IR spectroscopy is highly sensitive to the DNA sequence.

6.4.2 Spectral Change upon Addition of Hoechst 33258

In order to study ligand binding, ANOVA-PCA results need to be sensitive to small spectroscopic changes due to presence of the ligand. The next test therefore evaluates, whether the addition of H33258 to the DNA leads overall to a significant spectral change, or not. Previous experiments have shown that the concentration of H33258 used in the samples in this study is too small to be detected directly by infrared spectroscopy[13]. Any deviation in the 2D-IR spectrum upon interaction with the ligand can therefore be assigned to a change of the vibrational modes of the dsDNA. To achieve this, main factor matrix $\boldsymbol{H}$ is added to the residual matrix $\boldsymbol{\varepsilon}$ and is interrogated via PCA.

The ANOVA-PCA method calculated average spectra from the individual states in each factor. Only two different states are possible for main factor $\boldsymbol{H}$: dsDNA with or without H33258. Thus, there are only two unique spectra in $\boldsymbol{H}$: A 2D-IR spectrum averaged over all samples where the ligand is present, and another average response over all samples where the ligand is absent. The PCA of the sum $\boldsymbol{H} + \boldsymbol{\varepsilon}$ leads to a scores plot as shown in **Fig. 6.9a**, with two clusters clearly separated along PC1 and equal distance away from the coordinate origin. While the two averages in $\boldsymbol{H}$ define the separation between the centre points of both clusters, $\boldsymbol{\varepsilon}$ corresponds to the individual noise from

every collected 2D-IR measurement and generates a spread in all principal components, including PC1 and PC2. There is no overlap between both point clouds, which indicates, that there is a significant, spectroscopic difference between the two average spectra. This observation confirms that the minor groove binder has a measurable effect on the vibrational modes of the dsDNA bases. However, because $\boldsymbol{H}$ only contains responses averaged over all dsDNA strands, the subset is neither able to explain whether all dsDNA sequences interact with the binder in the same way, nor whether some of the sequences interact at all. There is simply no sequence dependency included. In other words, this test for significance only fails when the two calculated, average responses in $\boldsymbol{H}$ are smaller than the residual, instrumental noise in $\boldsymbol{\varepsilon}$.

6.4.3 Sequence Dependence of H33258 Binding

While it is clear from analysis of $\boldsymbol{H} + \boldsymbol{\varepsilon}$ that the effect of H33258 addition is measurable in the dataset; it is not yet disclosed how the response of each individual dsDNA sequence is affected by the ligand. This information is central to screening experiments, as each dsDNA-ligand combination can show a different interaction that leads to a relative preference of one combination over another. To investigate sequence-dependent impact of H33258 binding, it is necessary to include the interaction matrix ($\boldsymbol{SH}$) and perform a PCA on the sum $\boldsymbol{H} + (\boldsymbol{SH}) + \boldsymbol{\varepsilon}$. Previous studies have proven that the 2D-IR method can distinguish minor groove binding of Hoechst to two different sequences[13]. This subset will address the important question, whether the proposed method in this chapter can extend these observations to a wider set of minor grooves.

The scores plot from this subset in **Fig. 6.9b** shows individual clusters for each dsDNA sequence (indicated by colours) and further differentiates 2D-IR spectra of free dsDNA (crosses) from samples with H33258 (circles). No 2D-IR spectrum remains unperturbed when H33258 is added to the sample (no overlap between crosses and circles), which suggests that the ligand affects vibrational modes of all dsDNA sequences in the dataset. The wide spread of PC1 and PC2 scores across samples further indicates that the spectral impact of H33258 binding varies with dsDNA sequence. Higher principal components provide even more information that allow for a more detailed distinction between dsDNA responses and will contain more nuanced spectroscopic features. The observations so far suggest that H33258 interacts with each dsDNA sequence in a slightly different way and that the observed effects on the 2D-IR spectrum are complex and nuanced.

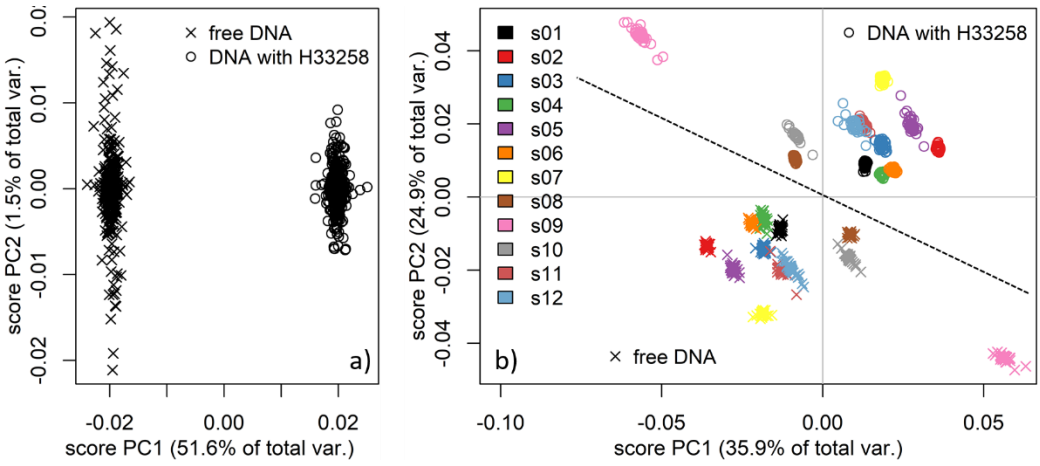


Fig. 6.9a) *PCA scores plot of subset* $\boldsymbol{H + \varepsilon}$*. Separation of ligand bound (circles) and free (crosses) DNA sequences along PC1 shows a significant change in the 2D-IR spectra when H33258 is added to the samples.* ***b)*** *PCA scores plot for subset* $\boldsymbol{H + (SH) + \varepsilon}$*. Circles show responses of bound sequences and crosses of unbound sequences. Individual sequences are separated by colour. The 2D-IR spectra of bound and unbound sequences are clearly separated into two groups by the black diagonal line.*

2D-IR spectra of free dsDNA (crosses) tend to have negative PC1 and PC2 scores and can be clearly separated from spectra of DNA with H33258 present (circles). The PC scores of each sequence essentially invert their sign when moving from the response without ligand to the response with H33258. The sign inversion can be explained from the experimental setup: Factor $\boldsymbol{H}$ only consists of two unique spectra, which were calculated from data where the average spectrum of each sequence was subtracted (for calculation of $\boldsymbol{S}$). The analyzed subset $\boldsymbol{H + (SH) + \varepsilon}$ is therefore centered with respect to the average response of each sequence, as represented by the origin of the coordinate system in **Fig. 6.9b**. PC scores for samples with ligand presence or ligand absence are thus represented by two counterparts with equal distance away from the coordinate origin and with equal PC1/PC2 ratio. The following analysis and plot in **Fig. 6.10** therefore focusses on the upper half of the scores plot (positive PC2 scores) for simplification, as the information in the lower half (negative PC2 scores) bears little additional information.

To relate the different 2D-IR responses of subset $\boldsymbol{H + (SH) + \varepsilon}$ to a proxy for binding affinities of each combination, the red-green colour scheme introduced in section 6.3.4 is used to categorise the PCA scores as shown in **Fig. 6.10**. The colour scheme is a qualitative representation of the binding affinity that could be expected for each combination, following the results from dsDNA melting experiments and fluorescence measurements. It indicates significant minor groove binding in dark green (A-tracts: s_2,

s_4, s_6 and s_7), suboptimal minor groove binding in light green (s_3 and s_5) and little to no minor groove binding in red colours (G/C-rich minor grooves, s_8 to s_{12}).

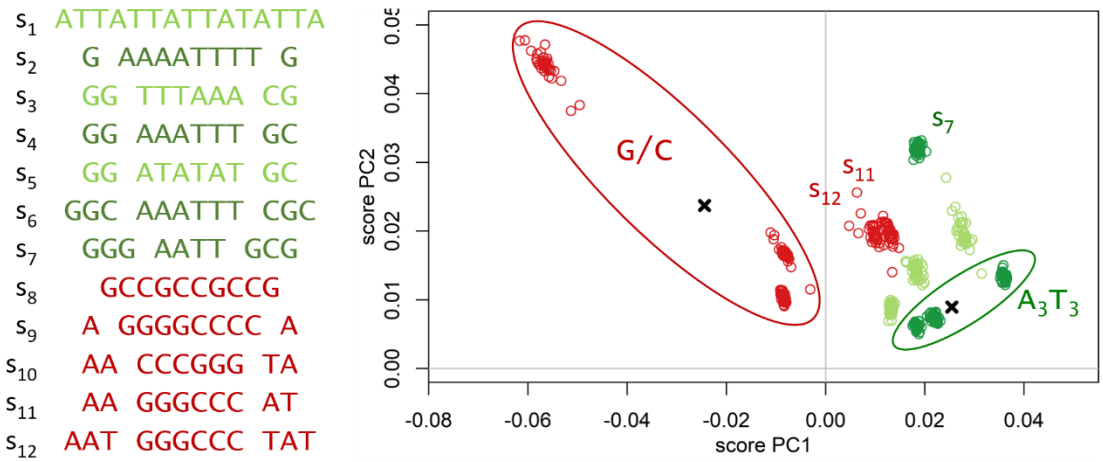


Fig. 6.10 *PCA scores plot for* $\boldsymbol{H} + (\boldsymbol{SH}) + \boldsymbol{\varepsilon}$. *Scores for free DNA spectra omitted. Colours qualitatively represent binding affinity for H33258: from red as zero binding to dark green as optimum binding (see text). Ovals show two groups of distinctive DNA-ligand interactions.*

According to **Fig. 6.10**, significant binding (dark green colours) corresponds to positive PC1 scores, while little to no minor groove binding (red colours) leads to smaller, negative PC1 scores. Sequences with well-defined A_nT_n sites of length n=3 and higher (dark green oval) are further separated by comparably low scores in PC2. The PC scores of suboptimal binding in light green fall in between the two extreme cases. These observations are important to validate 2D-IR spectroscopy as a tool for screening experiments, as they suggest a correlation of ANOVA-PCA results with conventional measurements used to estimate binding affinities and selectivity. It is likely that a sample represented by a positive PC1 and low PC2 score indicates a more efficient minor groove binder compared to a sample with negative PC1 score. The relative position of each dsDNA-ligand combination in the scores plot allows identification of similar interactions to distinguish unique 2D-IR responses from the bulk. The corresponding loading plots for PC1 and PC2 then provide the spectroscopic information that causes this separation. The distance from the coordinate origin defines the magnitude of the 2D-IR response, while the PC1/PC2 ratio defines the spectral features.

While there seems to be a general relation between the first two principal components and the proxy for binding affinity, this alone is perhaps too simplistic to completely explain the observed interactions in **Fig. 6.10**. The dsDNA oligomer s_7 (5'-

GGGAATTGCG-3') for example has a minor groove suitable for H33258 but reaches higher PC2 scores than other A_nT_n sequences. Previous studies suggest that H33258 spans approximately five base pairs along the minor groove[22], as indicated in the crystal structure illustrated in **Fig. 6.1**. The target sequence in this strand is only four base pairs long and the binder is likely to be in close proximity to a G-C base pair. This would be expected to change the spectral response upon binding in comparison to strands with a longer target sequence such as s_2, s_4 and s_6 (green oval).

The PC2 scores for H33258 interacting with oligomers 5'-AAGGGCCCAT-3' (s_{11}) and 5'-AATGGGCCCTAT-3' (s_{12}) show positive PC1 scores even though neither melting nor fluorescence experiments indicate minor groove binding. It is perhaps relevant to these observations that both s_{11} and s_{12} sequences contain A/T structures at the ends of the double strands. As the dsDNA minor groove only forms for sequence lengths of >3 bases, this motif does not represent a classic binding site for H33258. It could however be possible that the ligand interacts with these A/T rich ends rather than the inaccessible G/C cores, leading to a spectroscopic effect recognized by ANOVA-PCA as binding to an A/T-rich sequence. The negative melting point stabilization observed for s_{11} ($\Delta T_m \approx -5\ K$, **Fig. 6.4**) and s_{12} ($\Delta T_m \approx -5\ K$) seem to support this picture: While successful minor groove binding stabilizes the double helix and increases the melting point (s_1 to s_7), interactions at the end of the strands may facilitate end fraying of the double helix and decrease the melting temperature in s_{11} and s_{12}. This would be consistent with the observed deviation of their PC1 score with respect to the scores of s_8, s_9 and s_{10}.

It is important to note that observations in subset $\boldsymbol{H} + (\boldsymbol{SH}) + \boldsymbol{\varepsilon}$ do not simply result from a higher contribution of A/T structures for s_1 to s_7, or a higher contribution of G/C structures for s_7 to s_{12}. This dependency has been subtracted from the data when calculating factor $\boldsymbol{S}$ and the PC scores from $\boldsymbol{H} + (\boldsymbol{SH}) + \boldsymbol{\varepsilon}$ do not show a clear correlation to the base pair composition as the PC1 score did for subset $\boldsymbol{S} + \boldsymbol{\varepsilon}$.

The scores plot in **Fig. 6.10** suggests that there are at least two distinctive groups of binding interactions in the dataset; one being the response of DNA strands containing A_nT_n motifs ($n \geq 3$, s_2, s_4, s_6) and the other are strands with a G/C-core and a negative PC1 score (s_8, s_9, s_{10}). These are shown as green and red ovals in **Fig. 6.10** and the average PC1 and PC2 scores for the two groups are marked with black crosses. The average scores were used to reconstruct the spectral response for the two groups. The

previously discussed sign inversion of the scores implies that it is possible to generate difference spectra directly from the scores of the bound spectra. This would not be the case if more than one type of ligand was measured. In that case, the distance between scores of ligand-free to ligand-added samples would need to be calculated to produce a difference spectrum. Average scores of the first ten principal components have been used for the reconstruction to retain as much relevant data as possible. The reconstructed spectra for the two groups are shown in **Fig. 6.11**.

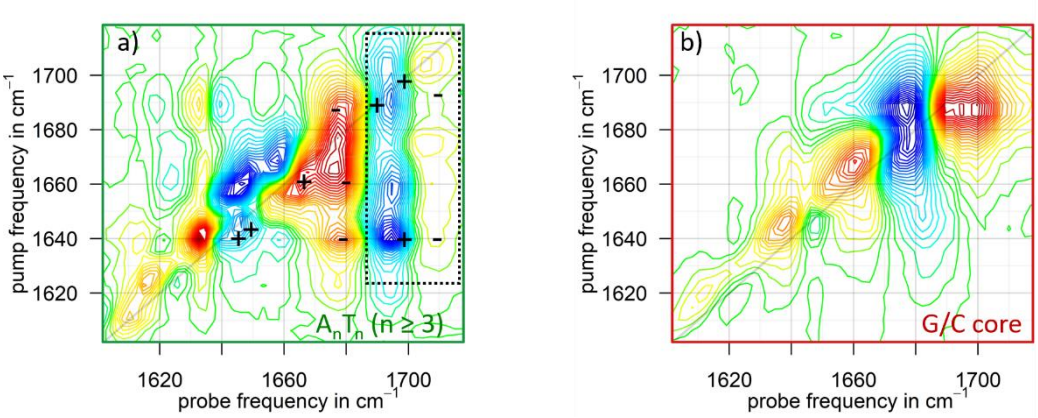


Fig. 6.11 *Reconstructed difference spectra for two distinctive DNA-ligand interactions calculated from the first ten principal components of subset* $\boldsymbol{H} + (\boldsymbol{SH}) + \boldsymbol{\varepsilon}$. ***a)*** *shows the results for the spectra included in the group identified by the green oval in* ***Fig. 6.10*** *(A_nT_n, n≥3).* ***b)*** *shows the results for the spectra included in the red oval in* ***Fig. 6.10*** *(G/C-rich minor grooves). Black +/- symbols in a) indicate position and sign of prominent bands in the difference 2D-IR spectrum relating to H33258 binding to sequence s_4 reported previously.*[13]

The group responses represent 2D-IR difference spectra, where negative signals (red) indicate a change to lower amplitudes and positive signals (blue) increase in amplitude due to ligand interactions. Both spectra in **Fig. 6.11** show a complex response. Addition of H33258 to sequences with a G/C-core (red group) affects vibrational modes along the diagonal of the 2D-IR spectrum with little change to off-diagonal features (**Fig. 6.11b**). By contrast, interactions of H33258 with A-tract sequences (green group, **Fig. 6.11a**) leads to significant changes in the off-diagonal region of the 2D-IR spectrum.

Minor groove binding of H33258 to sequence s_4 (green group) has been studied recently by 2D-IR spectroscopy using a different 2D-IR instrument with broader pump bandwidth and using phosphate buffer rather than TRIS buffer[13]. Plus and minus symbols marked in **Fig. 6.11a** show positions of prominent peaks reported in the difference 2D-IR spectra of H33258 binding to sequence s_4. These observations are in good agreement with the group response obtained from ANOVA-PCA results, again showing that the presented method is able to recover salient spectral features arising

from complex intermolecular interactions. The previous study identified a blue shift (shift towards higher energies) of the T_{2s} mode in dsDNA to above 1700cm^{-1} upon binding, resulting in a shift of the corresponding cross-peaks with T_{2S} as well. This blue shift is clearly visible in the ANOVA-PCA reconstruction in **Fig. 6.11a** (dotted rectangle), albeit with reduced amplitude due to smaller bandwidth of the LIFEtime laser (80 cm^{-1} FWHM, compared to ca. 300 cm^{-1} FWHM[13]). Understanding the complex, spectral response of the DNA to a ligand requires in-depth analysis of the vibrational modes affected, but the results show how ANOVA-PCA can highlight subtle changes due to different DNA-ligand interactions and enables analysis of large 2D-IR datasets.

The 2D-IR difference spectrum outlined in **Fig. 6.11b** characterises the group response when adding H33258 to dsDNA sequences s_8, s_9, s_{10}. The G/C structures in the minor groove of these sequences prevent H33258 from binding, which is clearly indicated by the negative results obtained from fluorescence and melting experiments for these combinations. Still, the 2D-IR spectrum of the nucleobases for this group has changed when adding H33258. The difference spectrum suggests two diagonal peak-pairs dropping in intensity at pump frequencies of 1643 and 1667cm^{-1} and one rising, diagonal peak-pair at 1687 cm^{-1} (crosses). No off-diagonal features are visible in contrast to **Fig. 6.11a**. The origin of these features is unclear. The molecular structure of H33258 is hydrophobic and tends to aggregate over time to minimise exposure to the polar solvent environment. This is one of the driving forces that causes H33258 to bind into the minor groove. Aggregation was visually noticeable when leaving an old sample of s_8 with H33258 added over night. It is possible, that the ligand also forms aggregates with dsDNA structures in these cases, which would affect the local environment of the vibrational modes observed. These observations point to a potential drawback of this method. If 2D-IR spectroscopy is sensitive enough to observe spectral changes that are unspecific to minor groove binding as well, then additional experiments such as the fluorescence measurements may become necessary to provide initial guidance for the interpretation of the 2D-IR responses observed, especially at a stage when there is little information available regarding indicative marker bands in the 2D-IR spectrum. For this experiment, the precedence of previous 2D-IR studies on H33258 already identified the T_{2S} shift, including cross-peaks, as a potential marker for minor groove binding. It is also unlikely that 2D-IR is used as the sole analysis method rather than as a complementary method.

6.4.4 Spectral Features Dependent on Waiting Time and dsDNA Sequence

The waiting time ($\boldsymbol{W}$) is the last main factor of the dataset and provides information relating to processes such as vibrational relaxation, spectral diffusion or energy transfer mechanisms[12]. These processes have only been investigated on free dsDNA structures so far, giving no precedence for dynamic effects from interactions with a ligand. The vibrational relaxations of modes at the nucleobases are fast (≤1 ps) and quite robust due to effective energy transfer processes from the nucleotide to the phosphate[45]. A perturbation of these processes due to minor groove binding would therefore be quite significant.

While $\boldsymbol{S}$ and $\boldsymbol{H}$ represent categorical variables, $\boldsymbol{W}$ is continuous and therefore requires a slightly different approach for analysis. Matrix $\boldsymbol{W}$ contains 12 unique spectra; one average for each waiting time. By adding the global average $\boldsymbol{M}$ to $\boldsymbol{W}$, the subset is representing the temporal evolution of the average 2D-IR spectrum over all samples (i.e. no discrimination against $\boldsymbol{S}$ or $\boldsymbol{H}$). As every excited molecular vibration in the 2D-IR spectrum will decay back to the ground state, analysis of $\boldsymbol{M} + \boldsymbol{W} + \boldsymbol{\varepsilon}$ via PCA will characterise the principal vibrational lifetime observed across all samples and across all pump and probe frequencies in the dataset. Rather than analysing PCA scores by plotting PC1 against PC2 as done previously, the PC1 scores can be plotted against their corresponding waiting time to obtain a temporal trace. This trace can be fitted to exponential decay functions to obtain an estimate for the principal relaxation described in PC1. It is possible that minor groove binding affects the vibrational energy pathways available in the system and changes the observed dynamics in the sample. A differentiation of samples according to their temporal response could therefore provide further information about different binding mechanisms.

Analysis via PCA aims to represent a dataset with as few principal components as possible by minimising covariance, rather than extracting vibrational dynamics. In order to extract more subtle and more complex processes such as energy transfer or spectral diffusion, alternative methods such as global, exponential fitting or a line-shape analysis might be more useful than an analysis of PC2, PC3 and so on. For the purpose of evaluating ANOVA-PCA as a screening tool, only the first principal component, representing the principal vibrational relaxation of the subset will be analysed in this chapter, even if more information is available. The principal, vibrational relaxation obtained from PC1 will not be very sensitive to subtle changes

from single oscillations but provides a fast method to identify large-scale perturbations to the vibrational relaxation.

The inset of **Fig. 6.12a** shows the PC1 scores of subset $\boldsymbol{M + W + \varepsilon}$ plotted against the corresponding waiting times. This subset only contains the average temporal response across all samples. The PC1 scores form a decay that can be well-represented by a mono-exponential function with a lifetime of 0.8 ps. The corresponding loading vector of PC1 is almost identical to the global average spectrum of $\boldsymbol{M}$, as displayed in the appendix in **Fig. A-6.6** and confirms that PC1 characterises a relaxation common to all spectral features in the average spectrum. It is instructive to compare this to vibrational lifetimes previously observed in literature. 2D-IR experiments on sequence s_1 (all A/T) reported lifetimes between 0.5 ps to 1 ps for this spectral region[12]. In that paper, individual diagonal peaks of only one dsDNA sequence were fitted peak-by-peak and an average lifetime of 0.75 ps was calculated. This strongly indicates that the PC1 score following ANOVA-PCA gives an accurate representation of the principal vibrational relaxation of the dataset. It is however noted that the 2D-IR spectrum of dsDNA in the base region may be particularly amenable to this type of treatment because all modes in the spectral region studied exhibit very similar vibrational relaxation dynamics. A dataset covering modes with a wide range of vibrational relaxation times may lead to a different representation by PCA.

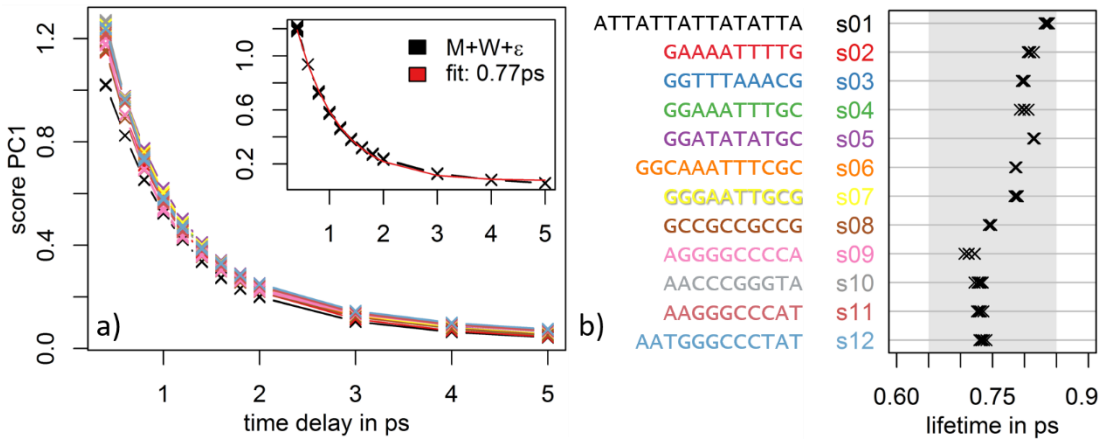


Fig. 6.12 a) inset: *PC1 score of subset* $\boldsymbol{M + W + \varepsilon}$ *plotted against waiting time.* ***a)*** *PC1 score of subset* $\boldsymbol{M + S + W + (SW) + \varepsilon}$ *plotted against waiting time.* ***b)*** *Principal, vibrational relaxation of different DNA sequences obtained from ANOVA-PCA. Mono-exponential functions were used for fitting throughout. Grey area illustrates the 200 fs FWHM of the probe pulse.*

In order to study kinetics across different sequences, the subset is extended to $\boldsymbol{M + W + S + (SW) + \varepsilon}$**.** In this subset, $(\boldsymbol{SW})$ contains sequence specific deviations from the

average relaxation in $\boldsymbol{W}$, and $\boldsymbol{S}$ represents the individual, constant offset of each sequence. A PCA is performed and the PC1 score is plotted again with respect to the waiting time in **Fig. 6.12a**, which shows individual relaxation dynamics for every sequence. PC1 is still represented by a single loading vector that resembles the global average spectrum and PC1 scores that form individual, temporal traces for each sequence. A mono-exponential fit of these decays obtains very similar lifetimes for the measured dsDNA strands between 0.7 ps and 0.85 ps (**Fig. 6.12b**). Considering the derived lifetimes more carefully indicates that G/C-rich sequences (s_8, all G/C, 0.75 ps) relax slightly more quickly than A/T-rich oligomers (s_1, all A/T, 0.83 ps). It is noted however that this difference of about 100 fs is just half the size of the FWHM of the probe pulse duration (200 fs, illustrated as grey area in **Fig. 6.12**b) and might therefore be too small to be significant.

The PCA results from this subset suggest reasonable vibrational lifetimes (0.83 ps compared to 0.75 ps[12] for s_1), which seem to be largely unaffected by the composition of the DNA. To further validate these findings, a PCA was carried out on the raw, time-resolved 2D-IR spectra of each sequence individually, instead of using ANOVA-PCA. The scores of PC1 again describe the principal vibrational relaxation of the sample and were fitted to a mono-exponential decay function. The lifetimes found by the individual PCA approach agree very well with the values obtained from the ANOVA-PCA method (see **Fig. A-6.7a** in the appendix). The ANOVA-PCA approach therefore allows for a quick survey of the principal vibrational relaxation across sequences with the caveat that mode to mode variation will not be clearly identified using PC1. The results indicate a slightly faster vibrational relaxation for G/C-rich double strands, but the magnitude of this difference is, again, comparable to the temporal resolution of the spectrometer. It is possible that the small, but consistent difference in PC1 could be an indication of a more subtle difference in the vibrational dynamics of G/C structures that cannot be recovered from the PCA approach alone.

6.4.5 Spectral Features Dependent on Waiting Time and dsDNA-Ligand Interactions

Having established that factor $\boldsymbol{W}$ can be used for a quick assessment of the principal vibrational relaxation of a sample, the corresponding subset $\boldsymbol{M} + \boldsymbol{W} + \boldsymbol{H} + (\boldsymbol{HW}) + \boldsymbol{\varepsilon}$ can be investigated, regarding an overall effect on relaxations due to interaction with H33258, irrespective of individual dsDNA sequence. The inset in **Fig. 6.13a** shows the PC1 scores obtained from this subset. Just as with analysis of $\boldsymbol{H}$, this subset only

consists of two unique temporal traces, one for dsDNA without H33258 (crosses) and one dsDNA with ligand (circles). Fitting both relaxations returns virtually identical lifetimes for free (0.77 ps) and ligand-bound DNA (0.78 ps). There is currently no published data available to validate whether H33258 minor groove binding affects the vibrational relaxation of dsDNA base modes or not. If the interaction with the ligand does affect the temporal response of the modes in this dataset, it is too subtle to be identified by the overall relaxation characterised with PC1 of this subset.

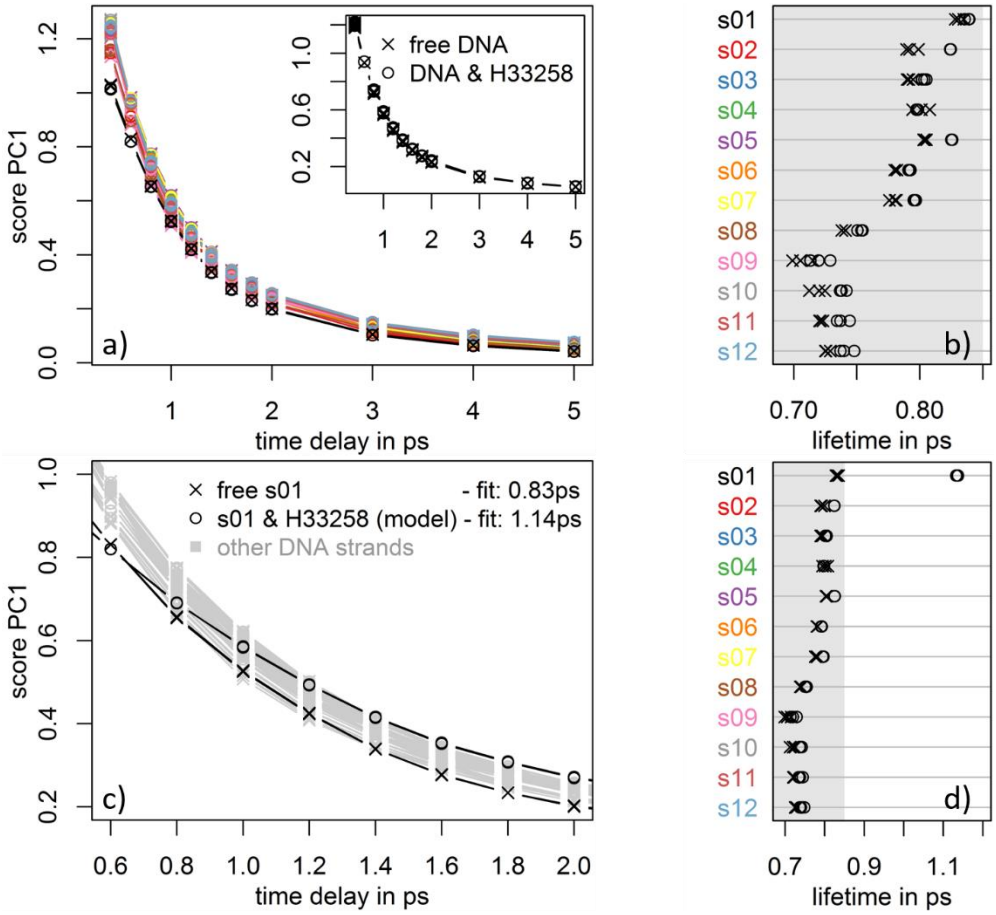


Fig. 6.13 a) inset: *PC1 score of subset* $\boldsymbol{M + W + H + (HW) + \varepsilon}$ *plotted against waiting time.* ***a)*** *main plot: PC1 score of a PCA using the complete dataset* $\boldsymbol{X}$ *plotted against waiting time.* ***b)*** *Vibrational lifetime with (circles) and without (crosses) H33258 obtained from PCA on* $\boldsymbol{X}$ *using a mono-exponential function.* ***c)*** *A PCA performed on* $\boldsymbol{X_{model}}$*. The PC1 score successfully retrieves an artificially-induced, slow, vibrational relaxation for oligomer* s_{01} *with H33258.* ***d)*** *Retrieved lifetime data from* $\boldsymbol{X_{model}}$*.*

Extending the subset to sequence specific interactions requires the addition of all remaining matrices, $\boldsymbol{S}$, $\boldsymbol{SH}$, $\boldsymbol{SW}$, $\boldsymbol{SHW}$ to completely characterise the temporal response of each sample measured. Addition of all subsets leads back to the initial input matrix $\boldsymbol{X}$, which is logical as all three factors and corresponding interactions need to be considered. A PCA on $\boldsymbol{X}$ gives rise to the PC1 scores outlined in **Fig. 6.13a**. The PC1

loading plot resembles the global average spectrum as before and fitting mono-exponential functions to the PC1 scores results in the principal vibrational lifetimes outlined in **Fig. 6.13b**. As before, lifetimes vary between 0.7 ps and 0.85 ps. The effect of H33258 on the observed lifetimes is indicated by the distance between crosses and circles in **Fig. 6.13** and show very little difference between presence and absence of the ligand. The largest difference is observed for sequence s_2 (5'-GAAAATTTTG-3') with a delay of under 50 fs, considerably smaller than the temporal resolution. The small difference observed was validated as before, by performing a separate PCA on the 2D-IR data of each individual sample and fitting the obtained PC1 score to a mono-exponential function (**Fig. A-6.7b**). Lifetimes recovered from this approach agree well with the dynamics acquired from ANOVA-PCA but are insignificant compared to the pulse duration of the probe laser.

In light of the demonstrated, relative insensitivity of DNA vibrational relaxation to sequence or ligand binding, a further test was carried out to determine the ability of ANOVA-PCA to retrieve different relaxation dynamics should they have existed in the dataset. To achieve this, a model dataset $\boldsymbol{X}_{\boldsymbol{model}}$ was generated in which the amplitudes of all spectra for s_1 in the presence of H33258 were rescaled to show a vibrational lifetime of 1.1 ps; significantly slower than observed in the collected dataset $\boldsymbol{X}$. All other responses were kept identical to $\boldsymbol{X}$. The PC1 score from $\boldsymbol{X}_{\boldsymbol{model}}$ is shown in **Fig. 6.13c** and successfully retrieves the simulated delay of 0.3 ps for the vibrational modes of s_1 when interacting with the binder (**Fig. 6.13d**). This result confirms, that the principal vibrational relaxation captured in the 2D-IR spectrum is indeed unaffected by the H33258 interaction.

6.5 Conclusions

It has been shown that the ANOVA-PCA method is able to separate a large, highly dimensional dataset into tangible subsets that can be analysed in a step by step manner with growing complexity. The separation of variance according to well-defined *factors* allows selective analysis of the information of interest and exclusion of otherwise inseparable data. Applying this method to a large 2D-IR dataset of dsDNA spectra clearly separates generic A-T base pair vibrations from generic G-C vibrations, as well as accurately revealing sequence composition. A more in-depth analysis of the sequence-dependent subset will enable the examination of subtle spectral features arising from nearest-neighbour interactions along the strand. It has also been shown

that the principal vibrational lifetime of these modes remains largely unaffected by the base composition of the double strands. This information was extracted without labour-intensive analysis of individual spectra and the results obtained from ANOVA-PCA are comparable to published results that were collected using conventional methods.

ANOVA-PCA is relevant in the development of 2D-IR spectroscopy as a screening technique to study minor groove binding. Analysis of the subset $\boldsymbol{H} + (\boldsymbol{SH}) + \boldsymbol{\varepsilon}$ identified small changes in 2D-IR spectra due to sequence-dependent dsDNA-ligand interactions. Conventional methods like fluorescence measurements or even the DNA melting point stabilization obtained from FT-IR data can only give a simple indication as to whether the ligand is binding to the target sequence or not. These methods provide very limited information about the molecular details of the binding interaction. ANOVA-PCA results extracted from 2D-IR experiments thus give a more information-rich view of how the ligand is interacting with the DNA. Using the ANOVA-PCA technique, a unique 2D-IR response for each ds-DNA-ligand interaction was observed. The PCA scores plot gives an intuitive representation of similarities and differences between samples and allows for a categorisation of samples according to common spectral responses. This is a powerful tool to summarise large numbers of dsDNA-ligand combinations and to identify few, distinctive combinations of interest.

Regarding H33258 minor groove binding, the most distinguished responses observed were summarized by two groups. Interactions with H33258 target sequence A_nT_n in the first group showed spectral changes of A/T modes in both, on and off-diagonal regions, consistent with an induced-fit type interaction[13] of the ligand. The second group of sequences with a G/C-rich minor groove showed predominately changes along the diagonal region of the 2D-IR spectrum and an absence of dominant off-diagonal changes. Additionally, a correlation of ANOVA-PCA-derived parameters with a proxy for binding affinity was shown and confirmed that the technique can pinpoint specific structural features sensitive to efficient minor groove binding. This study confirmed previous observations of 2D-IR spectroscopy on this ligands-dsDNA system and put the existing results into a broader context: Sequences with the same A_nT_n ($n \geq 3$) motif will respond with the same 2D-IR response when H33258 is bound and are clearly differentiated from interactions with incompatible minor groove. Minor groove binding of H33258 further did not affect the principal vibrational lifetime of the dsDNA, as extracted via PCA.

The high sensitivity of the 2D-IR spectrum to local changes in the molecular environment also comes with a small caveat. Simpler methods such as fluorescence experiments show a clear negative response in case of no minor groove binding. 2D-IR spectroscopy on the other hand has the ability to identify more than just minor groove binding and may recognise less specific dsDNA-ligand interactions as well. As a consequence, 2D-IR screening experiments are best used as a complementary method with other experiments; especially in cases when there is little spectroscopic precedence that would allow for a quick identification using 2D-IR marker bands alone. As outlined in the introduction, the analytical repertoire for studying DNA-ligand interactions is broad and ranges from methods with high structural resolution but low throughput such as 2D-NMR, to techniques with low structural information but high throughput such as DNA footprinting. 2D-IR experiments provide a balance of these two cases and offers a complementary analysis approach to study DNA-ligand binding from a different perspective.

Most of these observations could be reproduced using FT-IR spectroscopy, raising confidence in the analytical approach presented. The 2D-IR method in particular however offers an additional layer of spectral insight and temporal information at little or no additional time overhead for data collection.

The temporal information available from 2D-IR experiments can be analysed separately by the corresponding factors from the ANOVA decomposition. Analysis via PCA however only provides limited information regarding the principal vibrational relaxation across all observed modes in the 2D-IR spectrum. Higher principal components provide additional temporal information, but the extraction of meaningful kinetic traces is obstructed by the purely covariance-based transformation of the PCA. Alternative regression methods that impose additional assumptions on the data in form of model functions may be more effective for extraction of subtle processes such as energy transfer or spectral diffusion. The following chapter will use a global fit analysis to extract temporal information from a 2D-IR dataset.

These results exemplify, how the use of ANOVA-PCA could facilitate large-scale screening tests of ligands or drugs using time-resolved 2D-IR spectroscopy: An increasing number of sequences will lead to more densely populated score plots, providing a more nuanced analysis of interactions across sequences. The number of ligands can be extended to cover a wider range of interactions. Such an extension is

outlined in the following chapter. Provided that a ligand has a profound effect on the principal vibrational relaxation of the modes observed, ANOVA-PCA will also be able to identify these effects using the factor of the waiting time. Covering a broader spectral window will allow for a better coverage of potential spectral markers and will provide a better understanding how the inter-connected network of vibrational modes in a system is affected by perturbations. This multivariate approach can utilize the time resolution and abundance of structural information in 2D-IR spectroscopy to its full, analytical potential and could help assessing the selectivity of novel ligands by understanding the underlying binding mechanisms.

6.6 Appendix

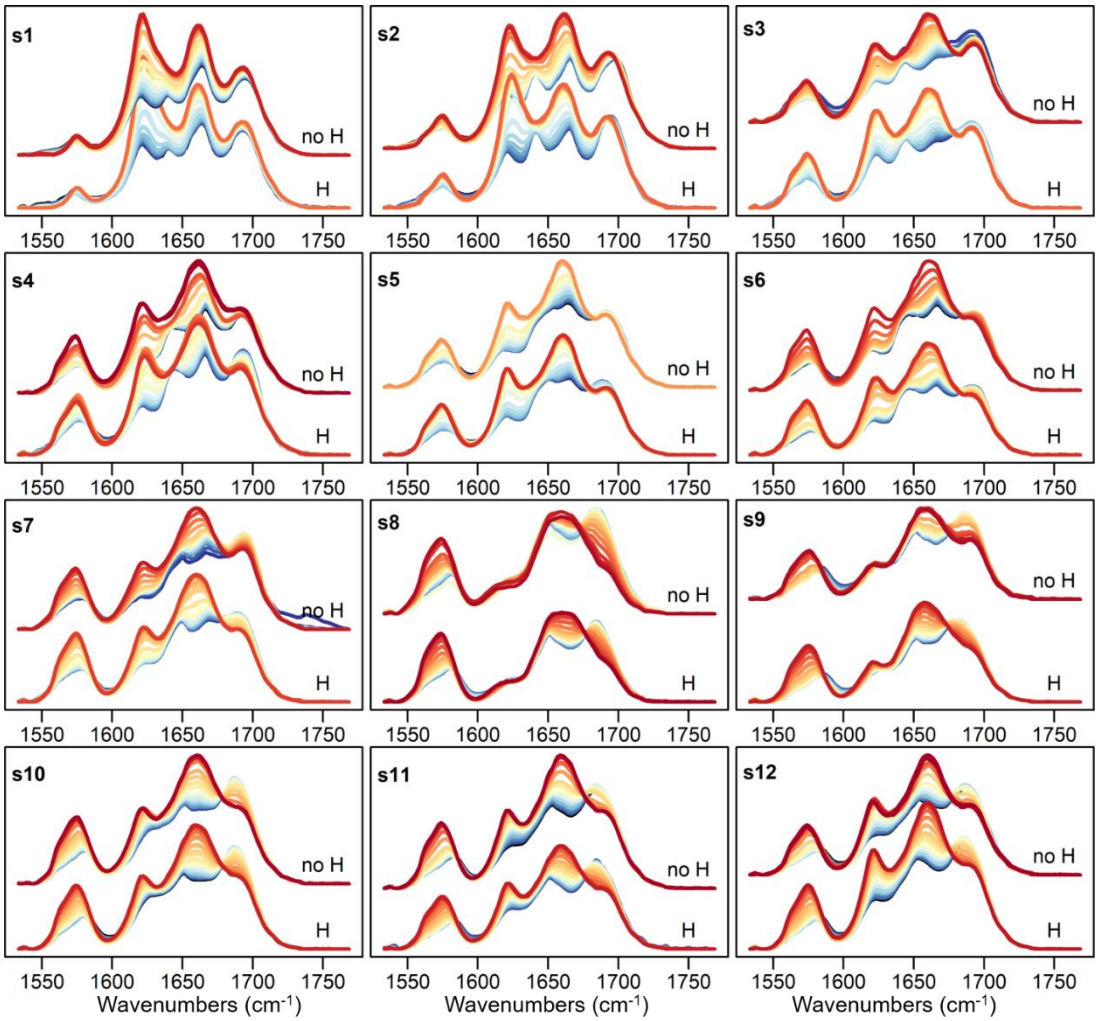


Fig. A-6.1 *FT-IR melting experiments on all 24 samples. Blue to red colours indicate temperatures from 20°C to 100°C. Responses from samples without H33258 are indicated A PCA was used to extract the*

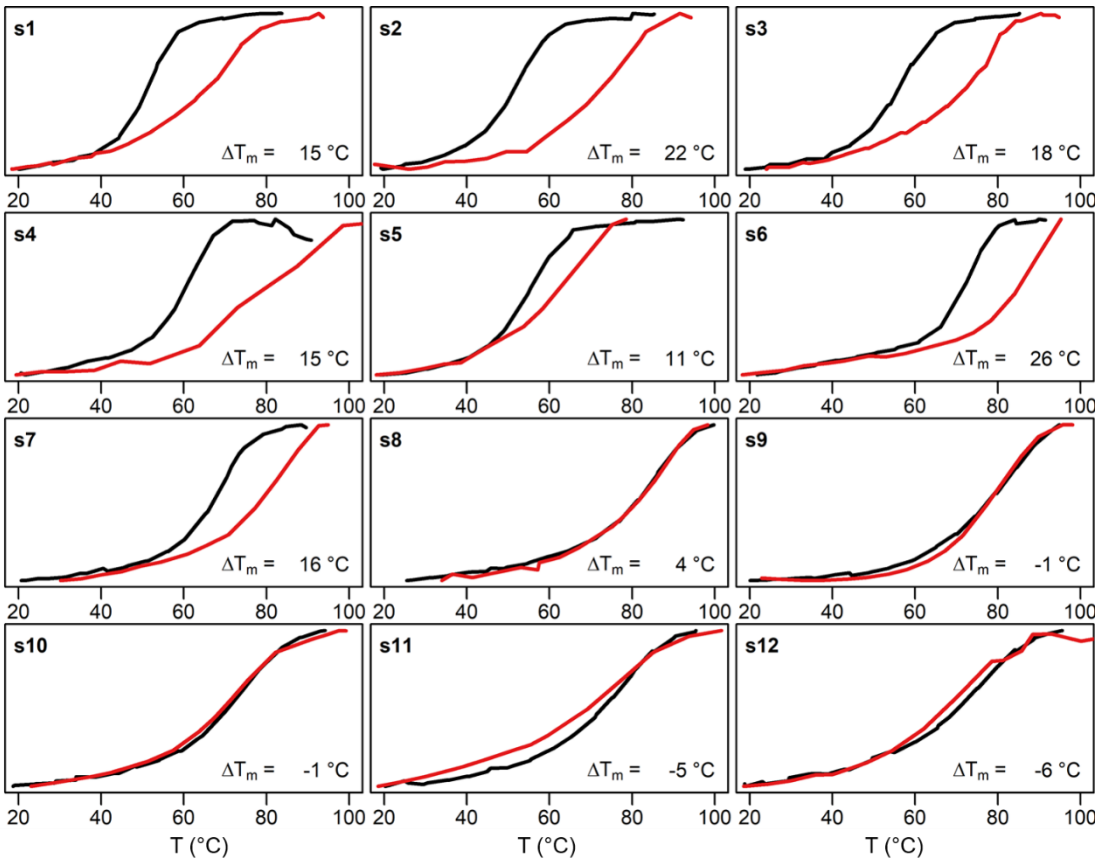


Fig. A-6.2 *PC1 scores for each of the 12 dsDNA sequences without H33258 in black and in presence of the ligand in red. To improve clarity of the plot, each sigmoid was normalised rise from zero to one. Each trace was fitted to a sigmoidal function as defined in equation (6.1) to obtain the dsDNA melting point* T_m.

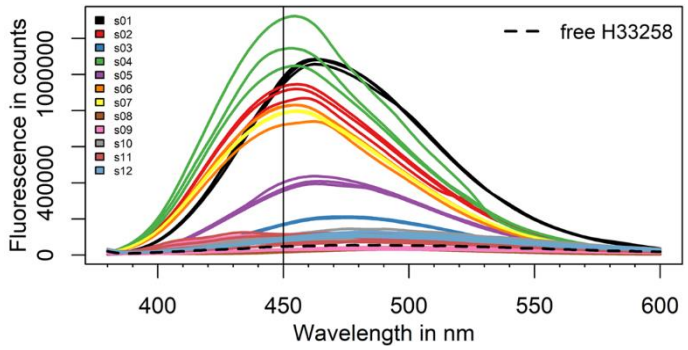


Fig. A-6.3 *Fluorescence spectra of H33258 in equimolar ratio with* s_1 *to* s_{12}. *Emission values at 450 nm are used to calculate the fluorescence enhancement factor compared to free H33258 in buffer.*

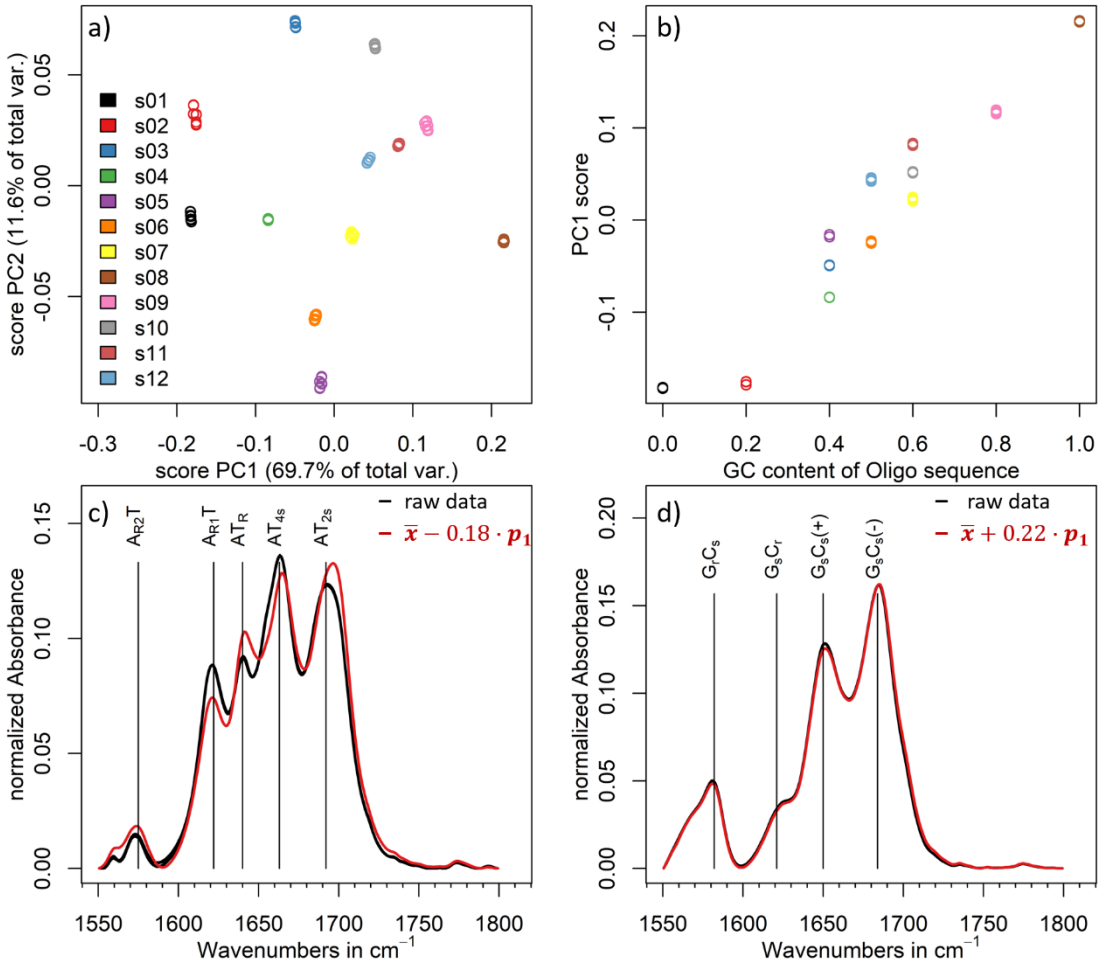


Fig. A-6.4 *ANOVA-PCA results from FT-IR absorption data. A) PCA scores plot of subset* $\boldsymbol{S} + \boldsymbol{\varepsilon}$*. b) PC1 scores show a good correlation to the GC percentage of the sequences measured. Comparison of raw FT-IR data (black) of all A/T sequence* s_1 *in c) and all G/C sequence* s_8 *in d) with reconstructed spectra (red) using the minimum and maximum PC1 score observed for this subset. The peak assignment is the same as for 2D-IR data, according to literature*[7,12]*.*

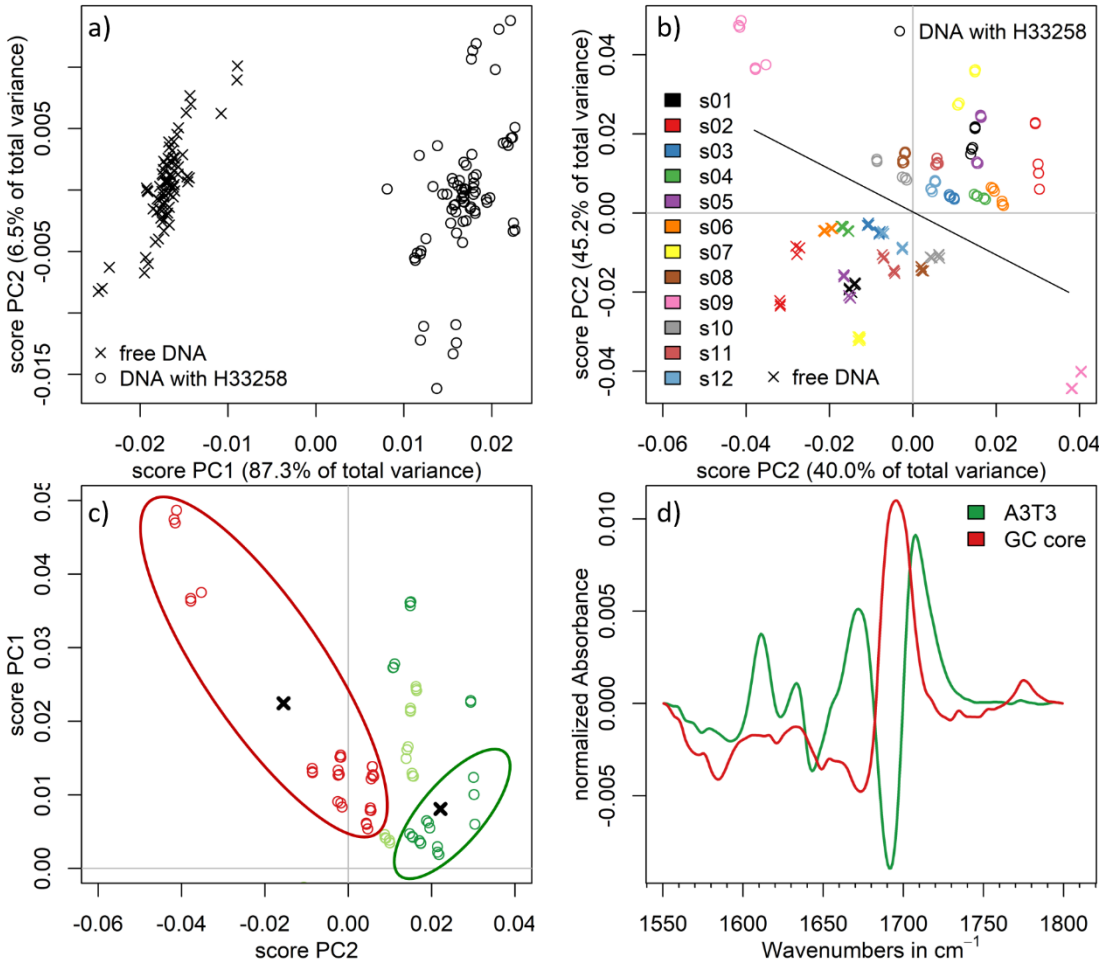


Fig. A-6.5 *ANOVA-PCA results from FT-IR absorption data.* ***a)*** *PCA scores plot of subset* $\boldsymbol{H + \varepsilon}$ *identifies a significant change in the FT-IR dataset when adding H33258 to a sequence on average.* ***b)*** *PCA scores plot for subset* $\boldsymbol{H + (SH) + \varepsilon}$*. Circles show bound sequences and crosses unbound sequences. Individual sequences are separated by colour. The bound and unbound sequences are clearly separated into two groups by the black diagonal line.* ***c)*** *The estimated binding affinity indicated by colours correlates with the PC scores of subset* $\boldsymbol{H + (SH) + \varepsilon}$*. Using the average response (black crosses) of two groups outlined in c), two difference spectra are calculated and plotted in* ***d)****, using the first 10 PCs.*

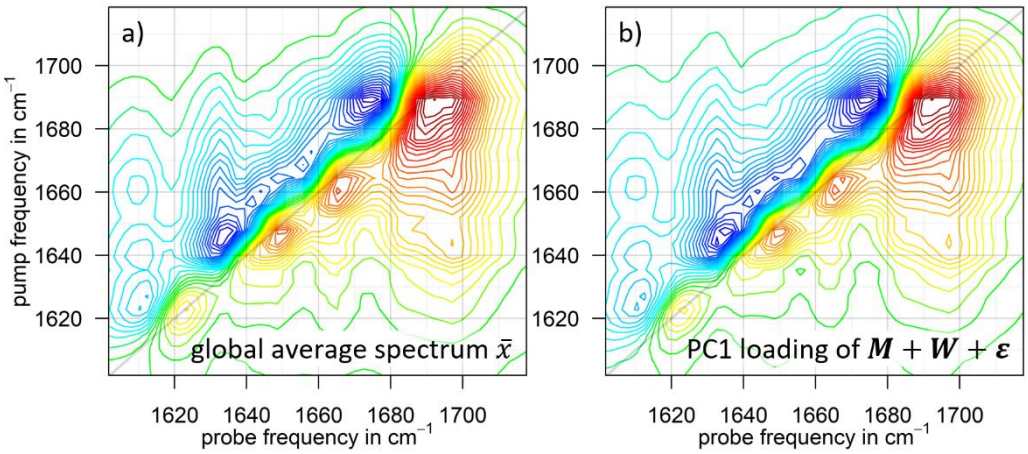


Fig. A-6.6 *The global average spectrum $\bar{x}$ in* ***a)*** *is virtually identical to the PC1 loading of subset* $\boldsymbol{M + W + \varepsilon}$ *in* ***b)****.*

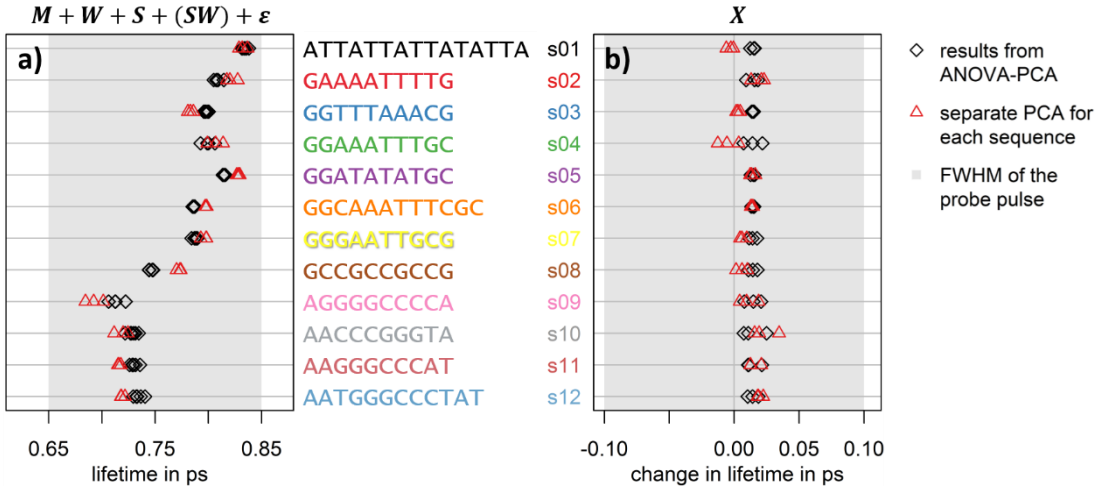


Fig. A-6.7 a) *Vibrational relaxation of different DNA sequences obtained from subset* $\boldsymbol{M + W + S + (SW) + \varepsilon}$ *using ANOVA-PCA (black symbols), compared to lifetimes obtained from PC1 scores of individual PCAs on each individual sample dataset (red triangles).* ***b)*** *Change in vibrational lifetime due to presence of H33258 (black symbols), compared to lifetime changes from individual PCA results (red triangles). Grey area illustrates FWHM of the probe pulse.*

6.7 References

1. Bailly, C. & Chaires, J. B. Sequence-Specific DNA Minor Groove Binders. Design and Synthesis of Netropsin and Distamycin Analogues. *Bioconjug. Chem.* **9**, 513–538 (1998).

2. Bostock-Smith, C. E., Harris, S. A., Laughton, C. A. & Searle, M. S. Induced fit DNA recognition by a minor groove binding analogue of Hoechst 33258 : fluctuations in DNA A tract structure investigated by NMR and molecular dynamics simulations. *Nucleic Acids Res.* **29**, 693–702 (2001).

3. Alniss, H. Y. *et al.* Recognition of the DNA Minor Groove by Thiazotropsin Analogues. *ChemBioChem* **15**, 1978–1990 (2014).

4. Quintana, J. R., Lipanov, A. A. & Dickerson, R. E. Low-temperature crystallographic analyses of the binding of Hoechst 33258 to the double-helical DNA dodecamer C-G-C-G-A-A-T-T-C-G-C-G. *Biochemistry* **30**, 10294–10306 (1991).

5. Vega, M. C. *et al.* Three-dimensional crystal structure of the A-tract DNA dodecamer d(CGCAAATTTGCG) complexed with the minor-groove-binding drug Hoechst 33258. *Eur. J. Biochem.* **222**, 721–726 (1994).

6. Abu-daya, A., Brown, P. M. & Fox, K. R. DNA sequence preferences of several AT-selective minor groove binding ligands. *Nucleic Acids Res.* **23**, 3385–3392 (1995).

7. Krummel, A. T., Mukherjee, P. & Zanni, M. T. Inter and Intrastrand Vibrational Coupling in DNA Studied with Heterodyned 2D-IR Spectroscopy. *J Phys Chem B* **107**, 9165–9169 (2003).

8. Krummel, A. T. & Zanni, M. T. DNA vibrational coupling revealed with two-dimensional infrared spectroscopy: Insight into why vibrational spectroscopy is sensitive to DNA structure. *J. Phys. Chem. B* **110**, 13991–14000 (2006).

9. Szyc, Ł., Yang, M., Nibbering, E. T. J. & Elsaesser, T. Ultrafast Vibrational Dynamics and Local Interactions of Hydrated DNA. *Angew. Chemie Int. Ed.* **49**, 3598–3610 (2010).

10. Greve, C. & Elsaesser, T. Ultrafast Two-Dimensional Infrared Spectroscopy of Guanine–Cytosine Base Pairs in DNA Oligomers. *J. Phys. Chem. B* **117**, 14009–14017 (2013).

11. Hithell, G. *et al.* Ultrafast 2D-IR and Optical Kerr Effect Spectroscopy Reveal the Impact of Duplex Melting on the Structural Dynamics of DNA. *Phys. Chem. Chem. Phys.* **19**, 10333–10342 (2017).

12. Hithell, G. *et al.* Long-Range Vibrational Dynamics Are Directed by Watson–Crick Base Pairing in Duplex DNA. *J. Phys. Chem. B* **120**, 4009–4018 (2016).

13. Ramakers, L. A. I. *et al.* 2D-IR Spectroscopy Shows that Optimized DNA Minor Groove Binding of Hoechst33258 Follows an Induced Fit Model. *J. Phys. Chem. B* **121**, 1295–1303 (2017).

14. Shim, S., Strasfeld, D. B., Ling, Y. L. & Zanni, M. T. Automated 2D IR spectroscopy using a mid-IR pulse shaper and application of this technology to the human islet amyloid polypeptide. *PNAS* **104**, 14197–14202 (2007).

15. Middleton, C. T., Woys, A. M., Mukherjee, S. S. & Zanni, M. T. Residue-specific structural kinetics of proteins through the union of isotope labeling, mid-IR pulse shaping, and coherent 2D IR spectroscopy. *Methods* **52**, 12–22 (2010).

16. Luther, B. M., Tracy, K. M., Gerrity, M., Brown, S. & Krummel, A. T. 2D IR spectroscopy at 100 kHz utilizing a Mid-IR OPCPA laser source. *Opt. Express* **24**, 4117 (2016).

17. Donaldson, P. M., Greetham, G. M., Shaw, D. J., Parker, A. W. & Towrie, M. A 100 kHz Pulse Shaping 2D-IR Spectrometer Based on Dual Yb:KGW Amplifiers. *J. Phys. Chem. A* **122**, 780–787 (2018).

18. White, C. M., Heidenreich, O., Nordheim, A. & Beerman, T. A. Evaluation of the Effectiveness of DNA-Binding Drugs To Inhibit Transcription Using the c-fos Serum Response Element as a Target †. *Biochemistry* **39**, 12262–12273 (2000).

19. Disney, M. D. *et al.* Activity of Hoechst 33258 against Pneumocystis carinii f. sp. muris, Candida albicans, and Candida dubliniensis. *Antimicrob. Agents Chemother.* **49**, 1326–1330 (2005).

20. Neidle, S. DNA minor-groove recognition by small molecules (up to 2000). *Nat. Prod. Rep.* **18**, 291–309 (2001).

21. Bailly, C., Colson, P., Henichart, J. & Houssierl, C. The different binding modes of Hoechst

33258 to DNA studied by electric linear dichroism. *Nucleic Acids Res.* **21**, 3705–3709 (1993).

22. Spink, N., Brown, D. G., Skelly, J. V. & Neidle, S. Sequence-dependent effects in drug-DNA interaction: the crystal structure of Hoechst 33258 bound to the d(CGCAAATTTGCG)2 duplex. *Nucleic Acids Res.* **22**, 1607–12 (1994).

23. Gavathiotis, E., Sharman, G. J. & Searle, M. S. Sequence-dependent variation in DNA minor groove width dictates orientational preference of Hoechst 33258 in A-tract recognition: solution NMR structure of the 2:1 complex with d(CTTTTGCAAAAG)2. *Nucleic Acids Res.* **28**, 728–735 (2000).

24. Breusegem, S. Y., Clegg, R. M. & Loontiens, F. G. Base-sequence specificity of Hoechst 33258 and DAPI binding to five (A/T)4 DNA sites with kinetic evidence for more than one high-affinity Hoechst 33258-AATT complex. *J. Mol. Biol.* **315**, 1049–1061 (2002).

25. Lankas, F. *et al.* Critical Effect of the N2 Amino Group on Structure , Dynamics , and Elasticity of DNA Polypurine Tracts. *Biophys. Chem.* **82**, 2592–2609 (2002).

26. Fornander, L. H., Wu, L., Billeter, M., Lincoln, P. & Norde, B. Minor-Groove Binding Drugs: Where Is the Second Hoechst 33258 Molecule? *J Phys Chem B* **117**, 5820–5830 (2013).

27. Alniss, H. Y. Thermodynamics of DNA Minor Groove Binders. *J. Med. Chem.* acs.jmedchem.8b00233 (2018). doi:10.1021/acs.jmedchem.8b00233

28. Harrington, P. D. B. *et al.* Analysis of variance – principal component analysis : A soft tool for proteomic discovery. *Anal. Chim. Acta* **544**, 118–127 (2005).

29. de Haan, J. R. *et al.* Gene expression Interpretation of ANOVA models for microarray data using PCA. *Bioinformatics* **23**, 184–190 (2007).

30. Sarembaud, J., Pinto, R., Rutledge, D. N. & Feinberg, M. Application of the ANOVA-PCA method to stability studies of reference materials. *Anal. Chim. Acta* **603**, 147–154 (2007).

31. Pinto, R. C., Bosc, V., Nocairi, H., Barros, A. S. & Rutledge, D. N. Using ANOVA-PCA for discriminant analysis : Application to the study of mid-infrared spectra of carraghenan gels as a function of concentration and temperature. *Anal. Chim. Acta* **629**, 47–55 (2008).

32. Greetham, G. M. *et al.* A 100 kHz Time-Resolved Multiple-Probe Femtosecond to Second Infrared Absorption Spectrometer. *Appl. Spectrosc.* **70**, 645–653 (2016).

33. Wilson, W. D., Tanious, F. A., Fernandez-Saiz, M. & Rigl, C. T. Evaluation of Drug–Nucleic Acid Interactions by Thermal Melting Curves. in *Drug-DNA Interaction Protocols* 219–240 (Humana Press). doi:10.1385/0-89603-447-X:219

34. Kellett, A., Molphy, Z., Slator, C., McKee, V. & Farrell, N. P. Molecular methods for assessment of non-covalent metallodrug–DNA interactions. *Chem. Soc. Rev.* **48**, 971–988 (2019).

35. Aman, K., Padroni, G., Parkinson, J. A., Welte, T. & Burley, G. A. Structural and Kinetic Profiling of Allosteric Modulation of Duplex DNA Induced by DNA-Binding Polyamide Analogues. *Chem. - A Eur. J.* (2018). doi:10.1002/chem.201805338

36. Padroni, G., Parkinson, J. A., Fox, K. R. & Burley, G. A. Structural basis of DNA duplex distortion induced by thiazole-containing hairpin polyamides. *Nucleic Acids Res.* **46**, 42–53 (2018).

37. Amirbekyan, K. *et al.* Design, Synthesis, and Binding Affinity Evaluation of Hoechst 33258 Derivatives for the Development of Sequence-Specific DNA-Based Asymmetric Catalysts. *ACS Catal.* **6**, 3096–3105 (2016).

38. Sanstead, P. J., Stevenson, P. & Tokmakoff, A. Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved through Infrared Spectroscopy. *J. Am. Chem. Soc.* **138**, 11792–11801 (2016).

39. Weisblum, B. & Haenssler, E. Fluorometric properties of the bibenzimidazole derivative hoechst 33258, a fluorescent probe specific for AT concentration in chromosomal DNA. *Chromosoma* **46**, 255–260 (1974).

40. Bucevičius, J., Lukinavičius, G. & Gerasimaitė, R. The Use of Hoechst Dyes for DNA Staining and beyond. *Chemosensors* **6**, 18 (2018).

41. Bazhulina, N. P. *et al.* Binding of Hoechst 33258 and its Derivatives to DNA. *J. Biomol. Struct. Dyn.* **26**, 701–718 (2009).

42. Lasch, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.* **117**, 100–114 (2012).

43. Rinnan, Å., Berg, F. van den & Engelsen, S. B. Review of the most common pre-processing techniques for near-infrared spectra. *Trends Anal. Chem.* **28**, 1201–1222 (2009).

44. Butler, H. J. *et al.* Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR spectroscopy. *Analyst* **143**, 6121–6134 (2018).

45. Hithell, G. *et al.* Effect of oligomer length on vibrational coupling and energy relaxation in double-stranded DNA. *Chem. Phys.* in Press (2018). doi:10.1016/j.chemphys.2017.12.010

7 Tailoring Polyamides: Investigating the Selectivity of Hairpin Minor Groove Binders using 2D-IR and ANOVA-PCA

7.1 Abstract

After validation of the methods, this chapter applies 2D-IR spectroscopy as a screening tool to a more relevant context. Hairpin polyamides are a class of minor groove binders that are comprised of individual building blocks, which allow the conception of sequence specific dsDNA-ligands. To improve the bioavailability of these ligands, new polyamide structures are currently being developed by modification of the G-selective building block with isopropyl side groups and using thiazole analogues. In this chapter, three new ligand candidates are screened together with a reference polyamide across a set of eight different dsDNA structures using 2D-IR spectroscopy. All four ligands are designed to target a 5'-TGWWCA-3' sequence with an expected degeneracy for W, which represents either A or T. The data is analysed via ANOVA-PCA and indicates significant perturbation of the vibrational modes of the dsDNA upon binding, possibly due to widening of the minor groove and compression of the major groove by the ligand. The new ligand candidates in particular appear to be more selective towards a 5'-TGTACA-3' motif, as combinations of sequences with TA in the centre show a unique 2D-IR response with additional cross-peaks appearing in the spectrum. Some of these cross-peaks might be indicative of PA-dsDNA interactions and appear to have an effect on the observed relaxation dynamics in the samples. These observations are amplified for ligands with the bulkier isopropyl side group substitution.

7.2 Introduction

The development of minor groove binders has led to a variety of molecular structures capable of recognising a specific double stranded DNA (dsDNA) sequence via non-covalent binding interactions[1–5]. One of these classes is the group of hairpin polyamide ligands (PA). These binders are based on heterocyclic amide units, whose modular structure can be tailored to a specific base sequence in the minor groove of dsDNA with diagnostic or therapeutic relevance[6]. A set of pairing rules was developed by Dervan *et al*[4,7–10], that generalises the selectivity behind these structures as summarised in **Fig. 7.1a**. A stacked pair of heterocycles within a hairpin polyamide, such as N-methylimidazole and N-methyl-pyrrole (Im/Py) can either selectively target a G-C base pair or a C-G base pair in the minor groove, depending on its orientation. The recognition, in a first approximation, is based on electrostatic interactions of the exocyclic N-2 amine of guanines acting as proton donors for imidazole units. A-T base pairs can be targeted and distinguished from T-A by a pair of N-methylpyrrole and 3-hydroxy-N-methylpyrrole (Py/Hp) due to hydrogen bonding from the 3-hydroxy group to the C-2 carbonyl of thymine[8,10]. A pair of N-methylpyrrole without hydroxy functionalisation (Py/Py) still shows a high selectivity for A/T structures over G/C, but is unable to distinguish T-A from A-T[11]. With this set of rules, a dsDNA motif like GTAC in **Fig. 7.1** can be recognised by a polyamide with four pairs of heterocycles (Im/Py), (Py/Py), (Py/Py), (Py/Im), linked over a γ-aminobutyric acid turn to generate a hairpin structure[9,12–14], where each heterocycle is chosen to accommodate a specific base pair.

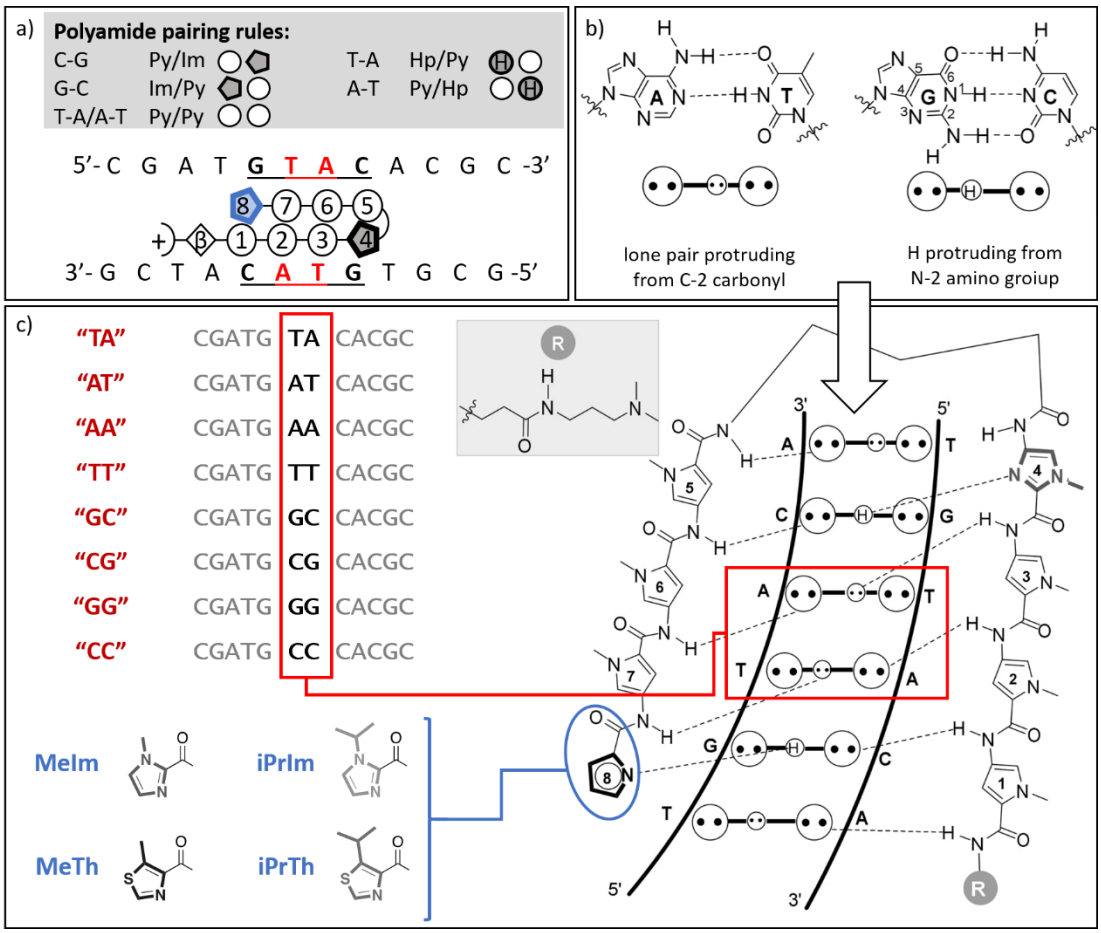


Fig. 7.1a) *Polyamide pairing rules allow the conception of sequence selective minor groove binders. A pair of stacked heterocycles (illustrated as circles and five-membered rings) targets a specific base pair.* ***b)*** *G-C and A-T base pairs viewed from the minor groove with protruding functional groups that represent possible targets for hydrogen bonding with polyamides.* ***c)*** *Minor groove binding due to hydrogen bonding as published by Aman et al.* [13]*. The current study screens four different polyamide ligands in blue across eight dsDNA sequences in red.*

Hairpin polyamides have shown biological activity[5]. These ligands are able to inhibit the transcription of estrogen[15] and androgen[14,16] receptors by occupying the GTAC minor groove. Malfunction of the androgen receptor can lead to a dysregulated expression of genes which might play a role in the development of prostate cancer[17]. A high sequence selectivity is therefore essential in their development as transcription factor antagonists in a therapeutic context. While polyamides are capable of competing with these natural transcription factors, their bioavailability is limited by poor cell permeability, especially with increasing imidazole content when targeting G-rich sequences[18,19]. This has motivated studies aiming to produce alternative, G-selective, heterocycles to extend the polyamide library and to enhance cell uptake[12,13].

In order to ensure a high sequence selectivity of new hairpin polyamide ligands, molecular interactions between dsDNA and polyamides have been investigated with multi-dimensional NMR[20,21], X-Ray crystallography[22] and molecular dynamics simulations[12,13,21]. While these methods allow for a detailed structural analysis of the interaction on an atomic level, processes like crystallisations or molecular dynamics calculations can take weeks to complete and may not be feasible to obtain for a larger set of ligand candidates screened across several different minor grooves. Classic screening methods like biophysical assays[13,23] or footprinting[11,24] on the other hand allow for a more comprehensive study of different ligand-DNA combinations, but can lack structural insight.

Two-dimensional infrared (2D-IR) spectroscopy can serve as a complementary approach to these existing methods due to its combination of structural sensitivity and comparably fast data acquisition within seconds[25–28]. 2D-IR only requires a few µl of a sample that can be analysed in solution at room temperature. The current experiment employs 2D-IR spectroscopy to study ligand binding and extends the ANOVA-PCA methodology[27,29] introduced in the previous chapter 6 to a dataset with multiple DNA sequences and four hairpin polyamides, each with significant absorption features in the infrared spectrum. The interactions with polyamides will also lead to a higher distortion of the dsDNA structure than in the previous chapter, as hairpin polyamides show significant major groove compression and hydrogen bonding with the dsDNA[12,13].

The experimental dataset features four different hairpin polyamides (see **Fig. 7.1c**, highlighted in blue), interacting with eight different dsDNA oligomers with sequence 5'-CGATG**NN**CACGC-3' (highlighted in red), where the central two nucleobases are varied systematically. The sequences are not self-complementary and are abbreviated in the text according to their central motif, with 5'-CGATG**TA**CACGC-3' (including the corresponding complementary strand) denoted as **TA**. One of the ligands was developed by Dervan *et al.*[16], contains two N-methyl-imidazole units at polyamide-positions 4 and 8 (see structure in **Fig. 7.1**), and is used as a reference structure denoted as **MeIm**. When **MeIm** forms a hairpin as illustrated in **Fig. 7.1c**, both imidazole units clamp down on the two guanines adjacent to the DNA centre within a double stranded structure of 5'-CGATGTACACGC-3'. This interaction widens the minor groove and compresses the major groove[12,13]. Pyrrole units of **MeIm** at polyamide-positions 2, 3, 6 and 7 target the central two base pairs and have a preference of A/T base pairs over G/C in the centre according to the pairing rules, but do not distinguish

T-A from A-T base pairs without further hydroxy-functionalisation[8,10,11]. The residue at the C-terminus of the four polyamides (dimethylaminopropylamine, Dp, see **Fig. 7.1**) has a preference for A/T structures[24] and will therefore align the hairpins C- and N-terminus towards the 5'end of 5'-CGATGNNCACGC-3' (binding site of the Dp residue underlined)[12,13]. The N-terminus of the polyamide will therefore align with the G of this single strand (GNNC), as opposed to the G of its complementary sequence.

The other three ligands are identical to **MeIm**, apart from an N-terminus substitution at polyamide-position 8 by 5-methylthiazole (**MeTh**), N-isopropylimidazole (**iPrIm**) and 5-isopropylthiazole (**iPrTh**) respectively, as seen in **Fig. 7.1c**. The substitution replaces imidazole with thiazole and changes the steric bulk of the alkyl residue at the N-terminus, which will affect the geometry of the interaction in the minor groove, increase the electron density available for acceptance of the N2-amino proton at the guanine and improve the hydrophobicity of the ligand, offering a better cell permeability[30,31]. It is expected that all four of the polyamide ligands will bind preferentially to sequences with *WW*-centres (**TA, TT, AA** and **AT**) as opposed to *SS*-centres (**GC, GG, CG** and **CG**) without any further distinction amongst *WW*-centres.

Previous experiments on these ligands mainly focussed on minor-groove-sequence 5'-TG**TA**CA-3' as a target and used NMR spectroscopy to assess binding. Restricted molecular dynamics simulations incorporated the results from NMR to obtain a molecular structure for the four DNA-Ligand complexes, indicating enhanced major groove compression from **MeTh** to the bulkier **iPrTh**[12], as well as a deeper penetration of the minor groove for **iPrIm** compared to **iPrTh**[13]. Another experiment utilised a biophysical assay to obtain association and dissociation rates for all four ligands[13] and found equilibrium dissociation constants in the nanomolar range or better for all polyamides binding to the 5'-TG**TA**CA-3' minor groove. Thiazole substitutions also resulted in higher dsDNA melting point stabilisations at the cost of a lower selectivity for G in 5'-TGTACA-3' compared to 5'-TTTACA-3'.

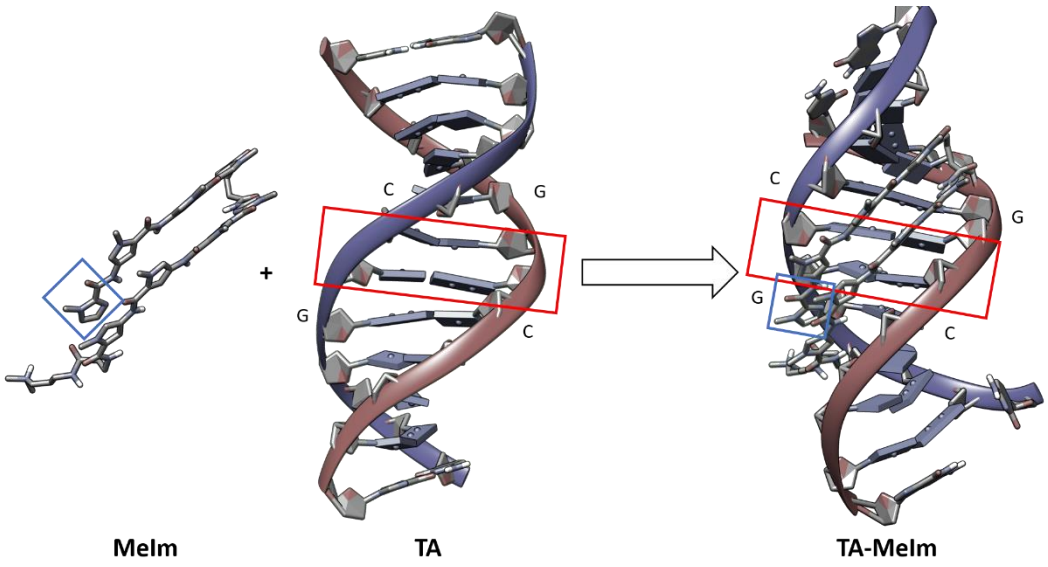


Fig. 7.2 *Proposed structures of polyamide minor groove binding calculated from NMR-restricted, molecular dynamics simulations*[12]*. Binding of the hairpin polyamide stretches the minor groove and compresses the major groove. Blue indicates the substituted polyamide position, red indicates the varied base pair positions.*

The present study keeps the critical two G-C base pairs in the same position in the minor groove and focusses on the effect of varying the two central base pairs in between (see **Fig. 7.2**, red). This keeps the interaction of the hairpin polyamide with both guanines at the ends of the binding site intact and generates a potential mismatch at the centre, which should give further insight into the selectivity of the ligands and its individual interactions to the dsDNA. Binding of the ligands to sequence **TA** is initially monitored by ^{1}H-NMR spectroscopy during sample preparation, so that full conversion of the free dsDNA can be assumed for **TA**. UV-melting experiments are performed additionally to assess the gain in dsDNA stabilisation due to binding.

7.3 Methods

7.3.1 Experimental Procedure

Access to the 2D-IR spectrometer was limited and all measurements had to be carried out at the Rutherford Appleton Laboratories within a two-week time slot. The project required preparation of all the 44 samples shown in **Fig. 7.3a** (including measurements of dsDNA and polyamides on their own). Efficient time management and experimental planning was therefore crucial to the project. The dsDNA sequences are denoted corresponding to their central motif as **TA, AT, AA, TT, GC, CG, CC** and **GG** in the following.

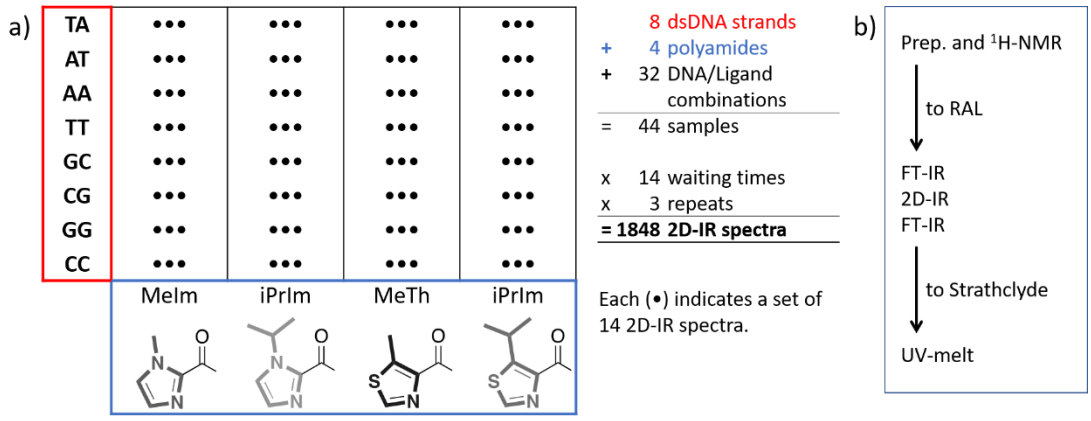


Fig. 7.3a) *Each dsDNA-PA combination is measured three times via 2D-IR. Each polyamide and dsDNA sequence were also measured on its own in buffer, leading to 44 samples in total.* ***b)*** *Samples were prepared at Strathclyde, sent to the Rutherford Appleton Laboratories (RAL) for infrared measurements and then sent back for UV-melting experiments.*

To minimise the workload at the Rutherford Appleton Laboratories, all samples were prepared two weeks in advance of the experiment at the University of Strathclyde. The details of the sample preparation are outlined below. All of the prepared samples and stock solutions were then freeze-dried, shipped to the Rutherford Appleton Laboratories for FT-IR absorption and 2D-IR measurements, freeze-dried again and shipped back for a final set of UV-melting experiments (see **Fig. 7.3b**). Freeze-drying samples is common practice when storing DNA oligomers and helped to minimise the impact of non-uniform degradation across samples throughout the experimental period. Freeze-drying also allowed to adjust the H_2O/D_2O ratio as needed for ^{1}H-NMR and infrared experiments.

7.3.2 Sample Preparation and NMR titration

16 salt free, lyophilised oligomers were purchased from Eurogentec, dissolved in 100 mM phosphate buffer (100 mM NaCl, pH 7.0, deionised water) and annealed at 95°C to obtain eight dsDNA stock solutions at 2.5 mM. Polyamides **MeIm**, **MeTh**, **iPrIm** and **iPrTh** were provided by the Burley group via solid phase synthesis and purification using reverse phase HPLC[12,13].

Each polyamide was obtained with a different purity and was therefore titrated against a solution with a known concentration of sequence **TA** using ^{1}H-NMR-spectroscopy. A D_2O/H_2O mixture of 10% D_2O was used to dilute the samples for the titration and provide a D_2O signal as a reference peak to stabilise the magnetic field strength to (*frequency lock*). Each polyamide was diluted with the solvent mixture to an approximate concentration of 12-15 mM and was added in 5 µl steps to a solution of **TA**

at 0.625 mM concentration. A 500 MHz NMR Spectrometer (Bruker) was used to collect ^{1}H-NMR spectra of the mixture for each addition of polyamide to the dsDNA sample. Aromatic protons of ligand-free dsDNA have characteristic proton shifts in the region of 7-14 ppm and formation of the dsDNA-PA complex will gradually replace these signals with a new set of peaks in that region. Complete disappearance of the unbound dsDNA peaks identified the required volume of polyamide solution for a full conversion of sequence **TA**. Repeating the titration for all four polyamides ensured that each ligand solution will yield the same conversion for sequence **TA**. The same volume of the respective polyamide was then added to the remaining seven sequences. In this approach, all 32 dsDNA-ligand combinations were generated with equal polyamide concentration across sequences.

7.3.3 Experimental setup for Infrared Measurements

Freeze-dried samples were re-dispersed in D_2O to obtain 1.25 mM solutions of pure dsDNA, pure polyamides and dsDNA-PA samples; all with a phosphate buffer and salt concentration of 50mM. Re-dispersion was done on the day of the measurement at the Rutherford Appleton Laboratories. A liquid transmission flow cell with a 50 µm polytetrafluoroethylene (PTFE) spacer between two, 2 mm thick CaF_2 windows was used for infrared measurements. Samples were injected using a gas tight Hamilton syringe (150 µL) and could be recovered after the infrared measurements were finished. The samples were washed out of the flow cell with three flushes of D_2O before loading the next sample. This did not require any additional disassembly in contrast to the temperature-controlled cell in chapter 6, which greatly reduced the sample preparation time and improved the consistency of the sample path length.

Conventional FT-IR absorption measurements were acquired using a Thermo Scientific Nicolet iS10 spectrometer at an optical resolution of 4 cm^{-1}. Each sample was measured before and after the 2D-IR measurements to check for consistency of the FT-IR absorption features. The 2D-IR experiments were performed on the LIFEtime instrument outlined in chapter 5.3, operating at 100 kHz repetition rate[28,32]. The setup uses a mid-infrared pulse-shaper[33] to generate a collinear pump-pulse pair with well-defined phase relationship, centred at 1650 cm^{-1} at a pulse width of 300 fs. The generated pump-pulse delay was scanned from zero to 3 ps in 24 fs time steps using phase cycling methodology to extract the 2D-IR response and reduce scatter artefacts in the data[34]. The samples were probed with two probe pulses centred at 1490 and

1650 cm^{-1} and pulse widths of 200 fs to cover a total spectral window from 1410 to 1720 cm^{-1}. The probe pulses were dispersed on two MCT array detectors in separate spectrometers to resolve the probe frequency axis. The probe axes were calibrated using the narrow absorption lines of atmospheric water in this region.

The loaded flow cell was transferred into the LIFEtime instrument, aligned, and three scans were taken at different sample positions on the transmission cell to confirm repeatability of the spectral response observed. Each scan consisted of a set of 14 2D-IR measurements taken at different pump-probe delays, *t*, from 0 to 5 ps to monitor relaxation dynamics of the excited molecular structures. An individual 2D-IR spectrum was averaged for 40 sec (ca. 2 million shots, just like in chapter 6), leading to a total measurement time of 30 min per sample.

7.3.4 Processing of Infrared Data

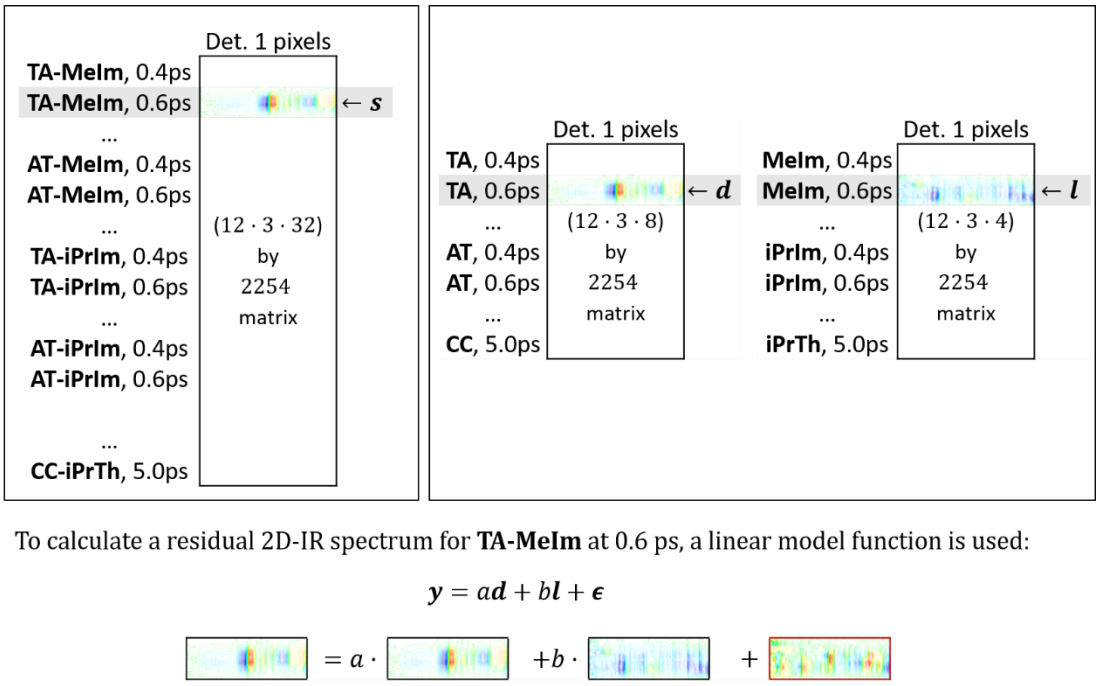
In contrast to the last chapter, the infrared absorption profile of the polyamide ligands is not negligible in this case. The pre-processing of the spectral datasets therefore had to be adjusted.

The FT-IR measurements obtained consist of absorption spectra from individual compounds in buffer (dsDNA, polyamides) and mixtures of different dsDNA-PA combinations. If the components of a mixture do not interact with each other, the spectral response of the mixture will simply be a linear combination of the individual component spectra. Any interaction that occurs in the mixture, such as ligand binding, may change the spectral response of any of the components and will therefore lead to a deviation from the linear combination. This deviation can be extracted from the data by using a multiple, linear regression with the following model function:

$$\boldsymbol{y} = a\boldsymbol{d} + b\boldsymbol{l} + \boldsymbol{\epsilon}. \tag{7.1}$$

The FT-IR spectrum obtained from a particular dsDNA-PA sample, defined as vector $\boldsymbol{y}$, is described by a sum of the corresponding dsDNA spectrum, $\boldsymbol{d}$, and ligand spectrum, $\boldsymbol{l}$, as pure compounds in buffer. Vectors $\boldsymbol{d}$ and $\boldsymbol{l}$ are multiplied by scalars a and b which are optimised in order to minimise the sum of squares obtained from the residual vector $\boldsymbol{\epsilon}$**.** The residual variance found in $\boldsymbol{\epsilon}$ contains instrumental noise and any variance that cannot be described by spectral features from any of the individual components. The residuals $\boldsymbol{\epsilon}$ will therefore contain any features that were generated as a result from combining the two components. This implies that any spectral response from non-

interacting DNA or non-interacting polyamide will be removed and will not appear in the residuals. This approach is used to remove redundant information from the dataset that is invariant to the interaction. The multiple, linear regression was performed for every dsDNA-PA spectrum of the FT-IR dataset and the residual vectors obtained were analysed via principal component analysis (PCA) in section 7.4.1 to investigate, whether certain dsDNA-PA combinations contain a similar set of spectral features.



To calculate a residual 2D-IR spectrum for **TA-MeIm** at 0.6 ps, a linear model function is used:

$$\boldsymbol{y} = a\boldsymbol{d} + b\boldsymbol{l} + \boldsymbol{\epsilon}$$

Fig. 7.4 *Each 2D-IR spectrum was concatenated into a vector and arranged into the data matrices for PA-only responses and dsDNA-only responses in* ***b)*** *and dsDNA-PA combination samples in* ***a)****. Each row in a) was fitted to a linear combination of corresponding rows in b) using model function (7.1). The same process is repeated for data from detector 2. The residuals of both fits are then arranged to a matrix* ***Y*** *and used as input for ANOVA-PCA.*

The same process was applied to the collected 2D-IR spectra. The data from the two MCT array detectors was processed in two separate, linear models and the residual vectors were combined afterwards for further analysis on a single, combined dataset. A 2D-IR spectrum is usually represented by a matrix of intensities but had to be concatenated into a single vector in order to process the data. The structure of the data is illustrated in **Fig. 7.4**, where 2D-IR spectra from detector 1 of all dsDNA-PA combinations are stored in a single matrix with individual 2D-IR pixels as columns. 2D-IR spectra for each sample were taken in three repeats at different pump-probe delays *t,* with data from 0.4 to 5 ps being used for analysis. Each of these measurements is a separate 2D-IR spectrum so that the number of rows for that matrix is $32 \cdot 3 \cdot 12 = 1152$. A dsDNA-PA combination spectrum $\boldsymbol{y}$ at a specific pump-probe delay is modelled

by the corresponding 2D-IR spectra of free dsDNA, ***d***, and free polyamide, ***l,*** at the same delay with model function (7.1). This approach is illustrated in **Fig. 7.4** and leads again to a matrix of residuals. The residual matrices from both detectors are concatenated to a single matrix, ***Y***, which is then further analysed as explained below in section 7.4.3.

The data from both detectors could alternatively be combined to a single spectrum first and then fitted with a linear model. The results of this approach are outlined in **Fig. A-7.1** and are compared to the individual fitting approach. The results obtained from both methods are quite similar. However, a separate fit on each detector data will improve the individual fitting results and remove more residuals, which will slightly improve the clarity of the spectral responses observed. The separate fitting approach was therefore used.

7.3.5 UV melting experiments

After all infrared experiments were finished, dsDNA melting experiments were performed at the University of Strathclyde using a Thermo Scientific Nanodrop 1000 UV absorption spectrometer with a heating rate of 1°C/min. The dsDNA melting of **TA**, **TT**, **AA**, **TT**, **GC** and **CC** was studied in combination with all four polyamides and without ligand at a dilution of 20 nM. Each combination was heated up three times and the sample absorption at 260 nm was acquired in dependency of the temperature $A(T)$ and fitted to a sigmoidal function according to:

$$A(T) = \frac{A_{min} - A_{max}}{1 + e^{(T-T_m)/g}} + A_{max} \ . \tag{7.2}$$

A_{min} and A_{max} are absorption values at the start and at the end of the melt, g is the gradient of the sigmoid and T_m defines the melting point. The average melting point of the three runs was used to calculate the ΔT_m difference between interacting and free dsDNA in order to characterise the melting point stabilisation.

7.4 Results and Discussion

7.4.1 FT-IR Absorption Spectroscopy

A summary of the observed FT-IR absorption spectra is shown in **Fig. 7.5**. Solid lines represent average spectra of different groups of samples as described in the plot and shaded areas indicate maximum/minimum absorbance values within each group to illustrate their variance. Each spectrum was baseline corrected by subtraction of a 50 mM phosphate buffer spectrum before plotting.

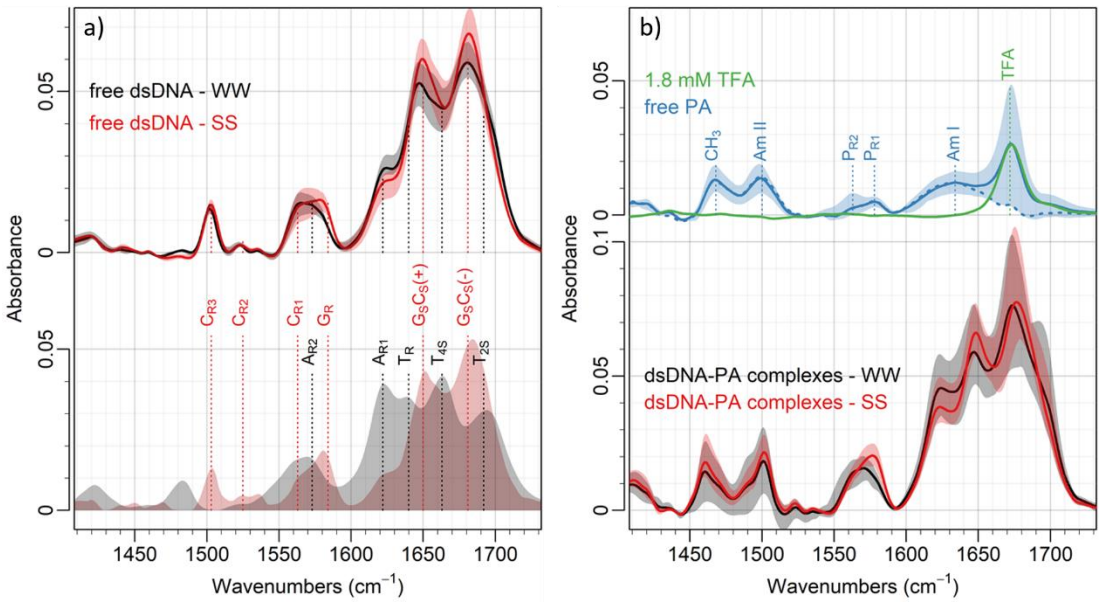


Fig. 7.5a) *Top: FT-IR absorption spectra of free dsDNA, grouped as sequences with WW central motif in black and SS motif in red. Shaded areas indicate min./max. absorption obtained at each wavenumber for each group. Bottom: Reference spectra of a pure A/T dsDNA oligomer in black and a pure G/C oligomer in red used for band assignment.* ***b)*** *Top: Average polyamide absorption spectrum in blue with band assignment according to DFT calculations and reference spectra of analogues. The large variance at 1672 cm^{-1} indicates TFA impurities from rev. phase HPLC (green). Bottom: Absorption spectra of all PA-dsDNA combinations, grouped according to the central dsDNA motif (WW in black, SS in red).*

The solid lines in the top of **Fig. 7.5a** represent the average spectrum of sequences with *WW*-centres in black (**TA**, **AT**, **AA**, **TT**) and the average of dsDNA with *SS*-centres (**GC**, **CG**, **GG**, **CC**) in red, in absence of any ligand. The spectral features observed for the free dsDNA sequences show less diversity than the set of sequences previously discussed in the Hoechst 33258 dataset in chapter 6. The 12-mer sequences used for polyamide binding differ only in the central two base pairs and otherwise share a common sequence, leading to a narrow range in base composition and therefore a small spectral variance in the infrared spectrum. The observed features follow the band assignment illustrated in the bottom of **Fig. 7.5a**, where the spectrum of a pure A/T (black) and a pure G/C (red) sequence was plotted for visual guidance[35–38]. The obtained absorption features for sequences with *SS*-centre show pronounced peaks at 1573, 1650 and 1684 cm^{-1} (red lines), corresponding to a G-ring mode (G_R) and two delocalised carbonyl-modes of C and G ($G_SC_S(+)$ and $G_SC_S(-)$) respectively. Sequences with *WW*-centre on the other hand show a slightly higher A-ring mode at 1622 cm^{-1} (A_{R1}) and a carbonyl-mode of T at 1692 cm^{-1} (T_{2S}, black lines). The region below 1550 cm^{-1} is much quieter with a single absorption at 1502 cm^{-1}, attributable to in-plane vibrations of C (C_{R3})[38] that shows little variation across sequences.

The blue trace in the centre of **Fig. 7.5b** shows the average spectrum obtained from all four different polyamides with the shaded area indicating lowest/highest absorbance values. The total synthesis of the four polyamides involves multiple reaction and purification steps which inevitably lead to a slightly different purity for each compound. This is reflected in the variance of the polyamide spectra in **Fig. 7.5b**, (blue shaded area, 1450 to 1500 cm^{-1}), which shows similar absorption features for maximum and minimum absorption traces. Reverse-phase HPLC purification of such ligands will always result in small impurities of TFA carried over from the mobile phase, as the polyamide is obtained in form of a tertiary ammonium salt. The FT-IR absorption spectrum of 1.8 mM TFA in green reveals a peak at 1672 cm^{-1} and confirms that each polyamide contains some TFA. It can be estimated from the individual absorbance of this peak that the concentration spans from almost 0 mM TFA in 1.25 mM **iPr-Th** to 2.2 mM TFA in 1.25 mM **iPr-Im**. The relative amount between TFA and polyamide in a sample will remain constant when preparing dsDNA-PA combinations. The TFA spectrum can therefore be treated as part of the polyamide response and is removed from the data when performing the linear model.

The band assignment of the polyamide absorption features is based on density functional theory calculations of structural fragments of the ligand and empirical data from the spectroscopic database of organic compounds (SDBS) of different pyrrole derivatives[39]. The eight heterocycles from each ligand will contribute ring-modes to the infrared spectrum around 1570 cm^{-1}, with N-methylated pyrroles showing CH_3-bending at 1470 cm^{-1}. The ten amide linkages in the ligands will also give rise to (deuterated) amide I and amide II vibrations at around 1634 cm^{-1} and 1500 cm^{-1} respectively. The band assignment is indicated in **Fig. 7.5b**.

The traces at the bottom of **Fig. 7.5b** show average spectra of all dsDNA-PA combinations with *WW*-centres in black and *SS*-centres in red. Upon first inspection, the spectral features of the complexes seem to be the summation of the free dsDNA spectrum with the corresponding free polyamide spectrum. The spectral region above 1550 cm^{-1} shows clear dsDNA absorption features with addition of the TFA band at 1672 cm^{-1}, while the region below 1550 cm^{-1} resembles the polyamide response. If there was no interaction between ligand and dsDNA, then the FT-IR spectra of the DNA-PA combinations in the bottom of **Fig. 7.5a** should be fully characterised by a linear combination of the two corresponding individual spectra plotted above.

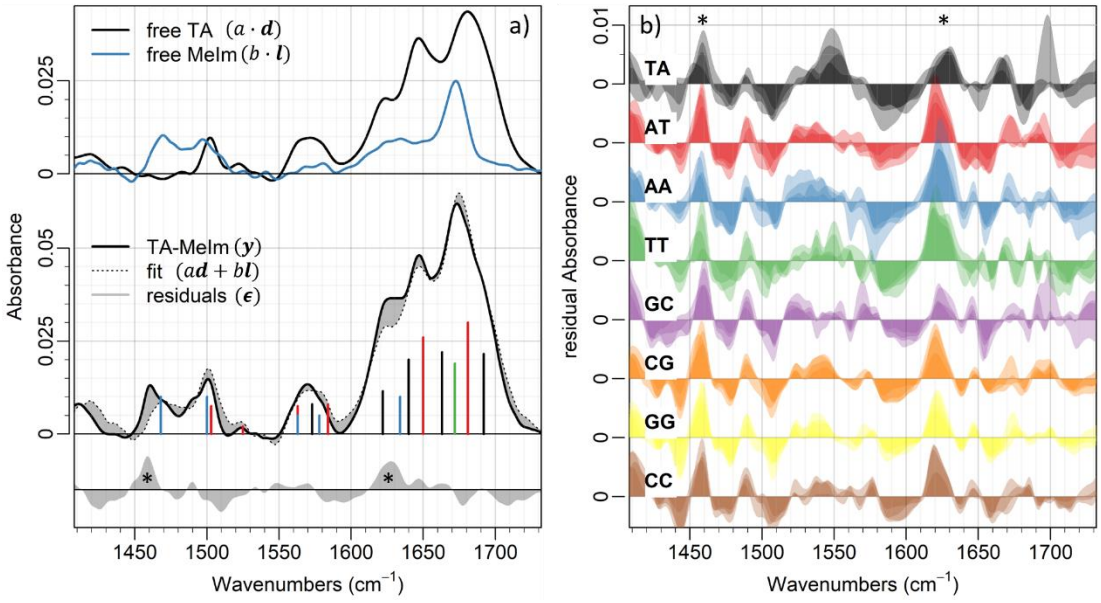


Fig. 7.6a) *Absorption spectra of **TA** (black) and **MeIm** (blue) can be used in a linear model to characterise the spectrum obtained for combination **TA-MeIm** (solid black line in the centre). The residuals of this regression (shaded grey area) are indicative of spectral feature changes due to PA-dsDNA interaction. Vertical lines indicate band assignments from A/T structures in black, G/C structures in red, polyamide in blue and TFA in green.* ***b)*** *Overview of residual spectra obtained for all 32 dsDNA-PA combinations. Colour indicates DNA sequence. Responses from different polyamide interactions are plotted on top of each other.*

Fig. 7.6a shows the result of a linear regression with model function (7.1), where spectra of sequence **TA** and ligand **MeIm** were multiplied individually (shown in the top of **Fig. 7.6a**) so that their sum (bottom, dotted line) will fit the spectrum of the **TA-MeIm** sample (solid line). The fitting result shows significant residuals across the spectrum (grey areas), indicating that the absorption bands of either the dsDNA or the ligand must have changed in the **TA-MeIm** sample. Most noticeable features in the residual spectrum include two bands at 1460 and 1620 cm^{-1} (asterisks) that indicate larger absorption features in the observed data than obtained from the model (i.e. positive residual value).

The multiple, linear regression was performed on FT-IR data of all 32 combinations and the residual spectra obtained are shown in **Fig. 7.6b**. Every dsDNA-PA combination returned a complex set of residuals, suggesting that none of the 32 samples can simply be represented as an inert mixture of dsDNA and PA. There are interactions present in all dsDNA-PA combinations which affect the vibrational modes observed to some extent. These residuals show common features across all dsDNA-PA combinations, including the rise of two bands at 1460 and 1620 cm^{-1} (asterisks) observed previously for **TA-MeIm**. For a better comparison of responses from every sample, a principal component analysis (PCA) was performed on the residuals and a scores plot with the

same colour-coding differentiating the sequences is shown in **Fig. 7.7a**. The choice of polyamide ligand is further differentiated using point symbols. The PCA was performed on a centred dataset, meaning that the global average residual spectrum was subtracted prior to the PCA. The PCA therefore separates according to relative differences between spectra and the coordinate origin of the scores plot represents the global average response.

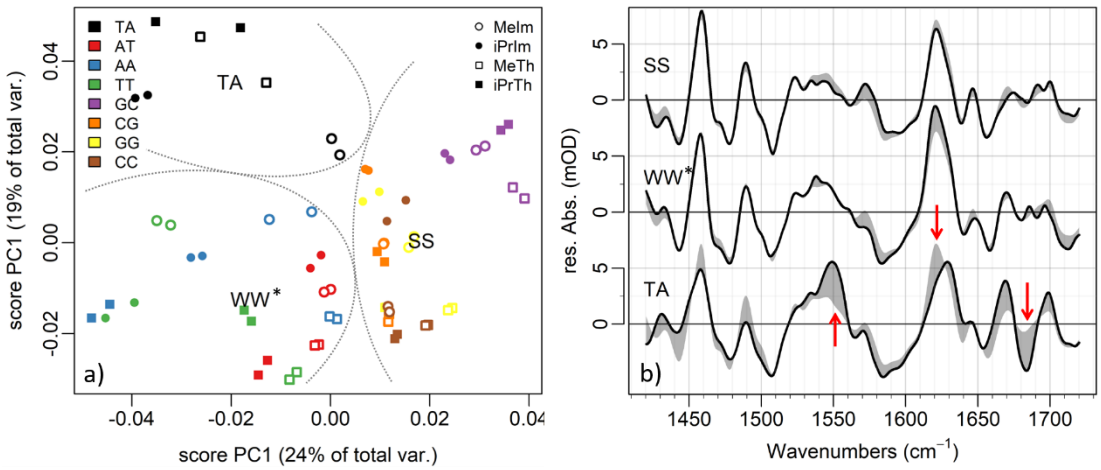


Fig. 7.7a) *PCA scores plot from FT-IR residuals. The dataset was centred prior PCA (global average subtracted). Combinations with TA sequence separate from all other combinations. Colours indicate sequence, symbols indicate polyamide used.* ***b)*** *The PCA results were averaged to three groups (TA, WW*, SS) and their spectral response is shown as black lines. Shaded areas indicate the difference of each group response from the global average residual spectrum. The TA group shows a characteristic set of residual features indicated by red arrows.*

There is a clear grouping visible across PC1, where samples with SS-sequence (**GC**, **CG**, **GG**, **CC**; purple, orange yellow, brown) have positive scores and samples with a WW-sequence (**TA**, **AT**, **AA**, **TT**; black, red, blue, green) are negative (highlighted by the dotted lines as visual guidance). There is a further separation visible across PC2 with **TA**-containing samples (black) separating from the remaining **AT**, **AA**, **TT**-containing samples (red, blue, green). This suggests a distinctive spectral response for combinations containing the **TA** oligomer and indicates a unique interaction between the polyamides and this dsDNA sequence. This is somewhat unexpected as the pyrrole pairs (Py/Py) in the centre of the polyamide have no inherent preference of a T-A base pair over A-T according to the binding rules. The **TA** combination with the reference polyamide **TA-MeIm** (black circles) shows the smallest PC2 value of all **TA** combinations, indicating that this distinction is amplified when using derivatives of the reference structure. This suggests that a substitution with an alternative heterocycle for G-recognition at the N-terminus of the polyamide enhances the differentiation of **TA**

from **TT**, **AA** and **AT**. Within this group, ligands with iso-propyl substitution (solid black symbols) tend to show larger scores, further away from the coordinate origin than ligands with methyl substitution (hollow black symbols), indicating a more dramatic deviation from the average spectral response for bulky side-groups.

Following this separation, the residuals were therefore split into three groups: SS (**GC, CG, CC** and **GG**), WW* (**TT**, **AA**, **AT** without **TA**) and TA (**TA** only). The average PC1 and PC2 scores of each group were calculated and are shown as locations of *"TA"*, *"WW*"*, *"SS"* in the scores plot. The average scores were then used to obtain the group-averaged residual spectra shown as solid lines in **Fig. 7.7b**. The shaded areas indicate the relative difference of each group response from the global average residual spectrum.

Groups SS and WW* mainly follow the global average residual spectrum with a slightly lower-than-average rise at 1620 cm^{-1} for group SS and a higher-than-average rise of the same feature for group WW*. Group TA however is characterised by a distinctive set of residual spectral features, including an enhanced rise in intensity at 1550 cm^{-1}, a lower-than-average rise at 1620 cm^{-1} and an enhanced drop at 1684 cm^{-1} (red arrows). The rise at 1620 cm^{-1} is possibly a convolution of two separate peaks with a reduced intensity of the lower frequency peak in case of the TA group. These distinctive spectroscopic features obtained for group TA suggest a unique interaction of the polyamide ligands with the **TA** sequence. A clear identification of these bands may be indicative of a structural change underpinning this apparent ligand selectivity, but the convoluted FT-IR absorption features prevent a confident band assignment. 2D-IR spectroscopy was therefore used to gain a better understanding how these residual features are coupled to other modes in order to locate them.

7.4.2 2D-IR Spectroscopy

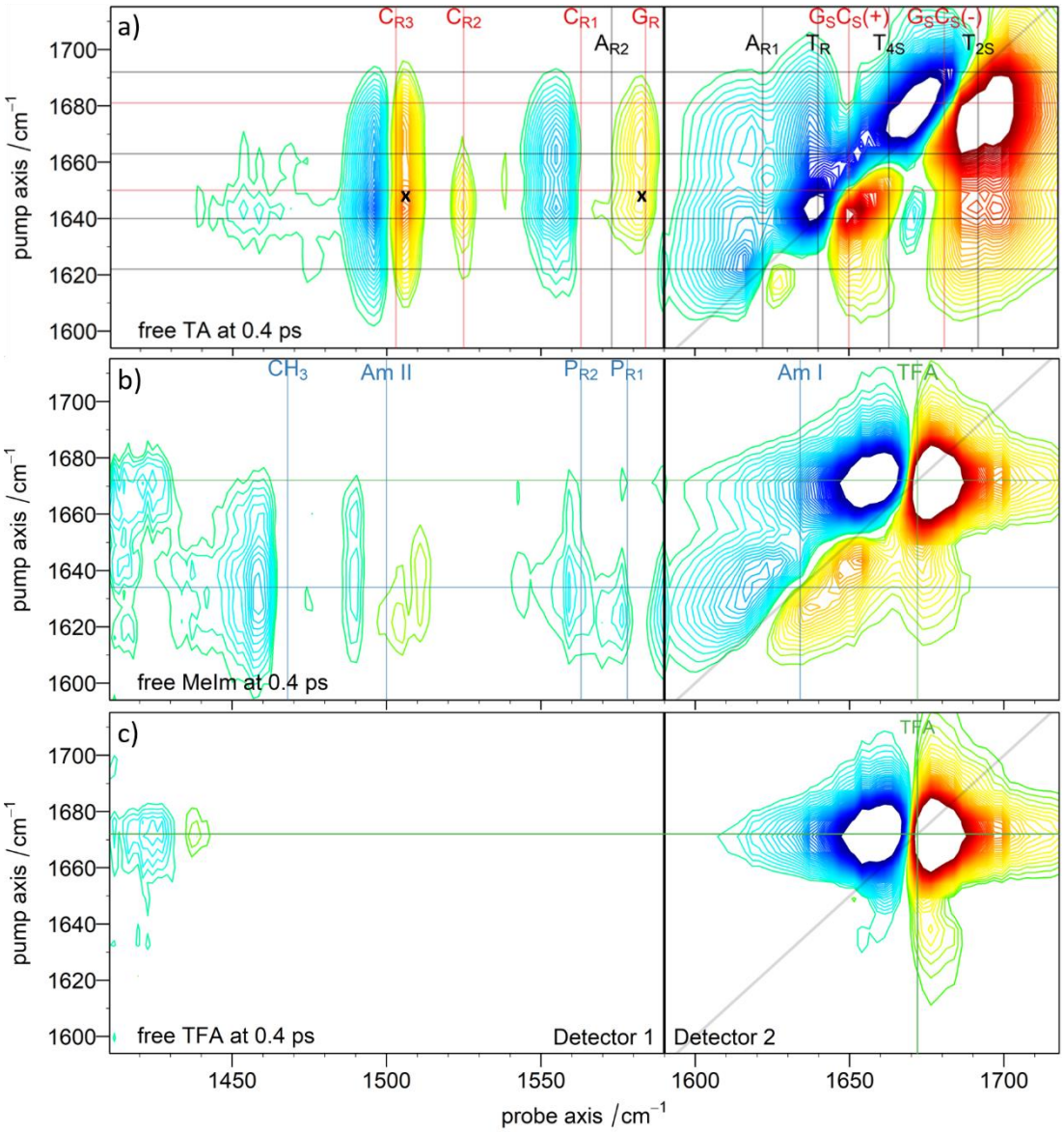


Fig. 7.8 *Representative 2D-IR spectra at 0.4 ps pump-probe delay for sequence* ***TA*** *in* ***a)****, polyamide* ***MeIm*** *in* ***b)*** *and TFA in* ***c)****. Red indicates negative signals such as ground state bleaches, blue indicates positive features such as excited state absorptions. All three spectra are shown with the same contour levels and are cut off at 70% of the maximum magnitude observed to highlight smaller features. The band assignment follows FT-IR absorption data with A/T features in black, G/C features in red and polyamide bands in blue. The strong peak-pair on the diagonal at 1672 cm^{-1} in b) and c) indicates TFA impurities.*

Fig. 7.8 shows representative 2D-IR spectra of sequence **TA** only (**a**), **MeIm** only (**b**) and TFA only (**c**), all at a pump-probe delay of t = 0.4 ps. Data from both MCT detectors was plotted on the same graph and detector 1 was used to cover the off-diagonal probe region from 1410 to 1590 cm^{-1}. Data from detector 2 was used to cover the diagonal region of the spectrum from 1590 to 1720 cm^{-1}. The vibrational modes of **TA** show significant off-diagonal features as expected for a coupled system such as dsDNA[40–42].

The band assignment from FT-IR allows identification of several cross-peaks on detector 1, arising from pairs of G/C-derived vibrational modes. The two peaks identified with crosses at (pump position in cm^{-1}, probe position in cm^{-1}) = (1650, 1582) and (1650, 1503) can be assigned to ($G_SC_S(+)$, G_R) and ($G_SC_S(+)$, C_{R3}) respectively. The 2D-IR spectrum of free **MeIm** at 0.4 ps is dominated by the TFA response at 1672 cm^{-1} and otherwise shows cross-peaks from the amide I band to amide II (1634, 1500) and amide I to CH_3-bend (1634, 1468). The 2D-IR spectrum of TFA in **Fig. 7.8c** shows a single carboxylate mode along the diagonal at 1672 cm^{-1} that shares a small cross-peak with a mode at 1433 cm^{-1}.

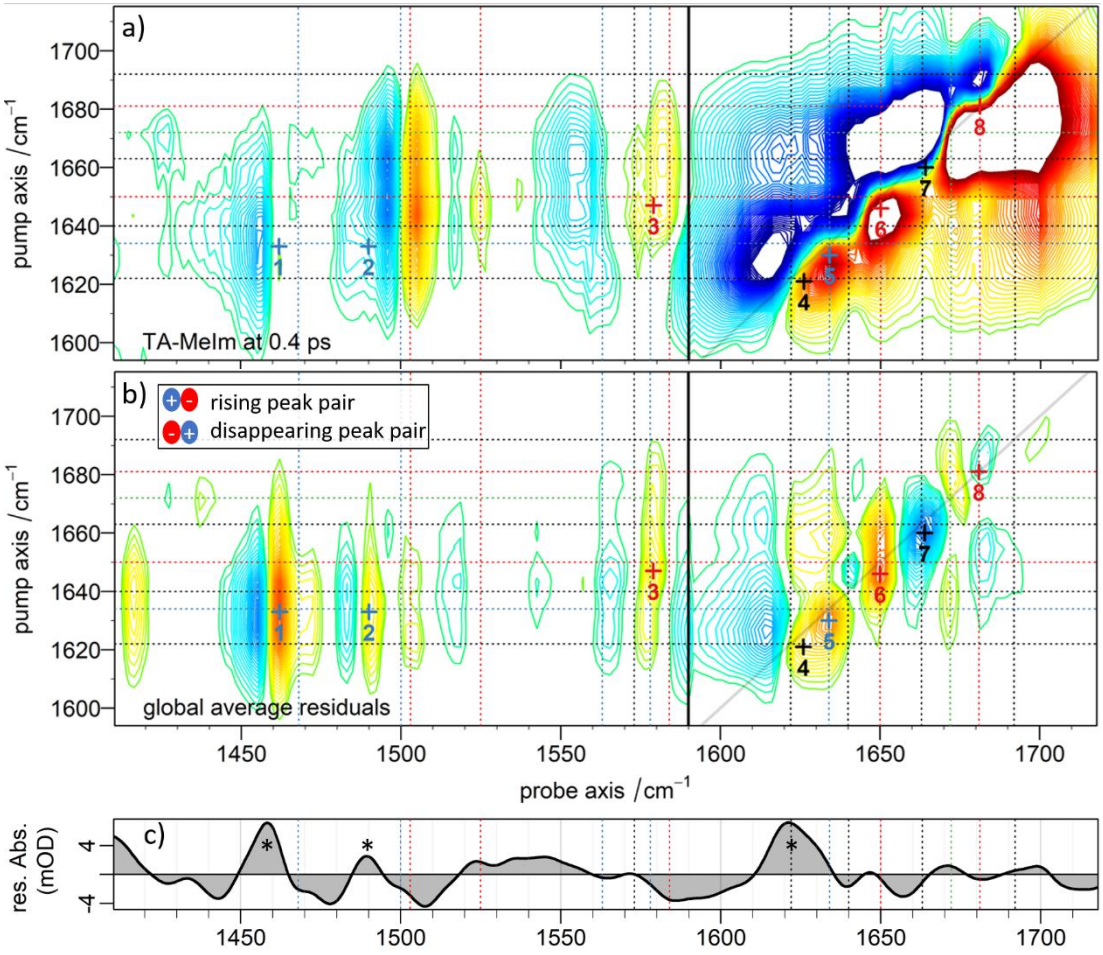


Fig. 7.9a) *Collected 2D-IR spectrum of combination* ***TA-MeIm*** *at 0.4 ps.* ***b)*** *Global average, 2D-IR residual spectrum obtained from fitting the responses of each PA-dsDNA combination. Characteristic peaks are highlighted with numbered crosses. The legend indicates how rising and disappearing peak pairs can be identified.* ***c)*** *Global average, residual spectrum obtained from fitting the FT-IR dataset. Similar residuals obtained from two different methods confirm that the linear model returns reproducible results. Dashed lines indicate absorption features of free dsDNA (A/T black and G/C red) and polyamide (blue).*

The 2D-IR spectrum of combination **TA-MeIm** at 0.4 ps is shown in **Fig. 7.9a** and closely resembles the sum of the spectra of **TA** and **MeIm** from **Fig. 7.8**. To obtain only those features that change upon mixing, the 2D-IR spectra of all dsDNA-PA combinations were analysed via multiple linear regression similar to the FT-IR absorption data using model function (7.1), as discussed in section 7.3.4. The global average residual spectrum is shown in **Fig. 7.9b** to give an indication which features remain after the fitting. Overall, the 2D-IR residuals in **b** compare very well with the global average residuals in **Fig. 7.9c**, obtained from the linear model of the FT-IR dataset. Rising off-diagonal features at probe frequencies of 1462 and 1490 cm^{-1} (points **1** and **2**) and a rising peak pair on the diagonal at 1620 cm^{-1} (points **4** and **5**) in the 2D-IR data coincide with an increase of features in the infrared absorption spectrum at the same frequencies (see asterisks in **c**). Two independent spectroscopic methods indicate residual features at the same wavenumbers, which raises confidence that the residuals returned are reproducible and indeed characterise interactions present in the mixture.

There is a clear rise of two cross-peaks at (1634,1462) and (1634,1490) (point **1** and **2**), which are about 6-9 cm^{-1} downshifted from two peak pairs visible in the free **MeIm** spectrum (**Fig. 7.8a**). The location of these peaks along the pump axis coincides well with the amide I mode of the ligand, while horizontal positions on the probe axis can be assigned to CH_3-bending and amide II modes, suggesting that these cross-peaks originate from polyamide structures. A rise of a mode at the same pump frequency of 1634 cm^{-1} on the diagonal (point **5**) is an indication for coupling between peaks at **1** and **2** with the Amide I band at **5**. The remaining spectral features at positions **3**, **4**, **6**, **7** and **8** are more ambiguous and were assigned according to the closest peak position found in the FT-IR data. The rise around point **5** seems to be elongated towards smaller wavenumbers, possibly due to presence of another rise of a peak at 1623 cm^{-1} (point **4**), coinciding with the adenine ring mode A_{R1} of dsDNA. The presence of a convoluted pair of rising peaks around 1620 cm^{-1} was identified earlier in FT-IR data in **Fig. 7.7** as well and further confirmation of this will be discussed below. Points **6** and **3** were assigned to a rise in diagonal mode $G_SC_S(+)$ and a corresponding cross-peak with G_R respectively, according to their position in the 2D-IR spectrum. The drop in intensity of peak pairs on the diagonal at point **7** and **8** were attributed to a partial loss of mode T_{4S} and mode $G_SC_S(-)$ respectively.

Fig. 7.9 shows the global average of the residual 2D-IR dataset $\boldsymbol{Y}$, but there is far more information available. In contrast to linear FT-IR absorption spectra in **Fig. 7.6**, there is no visualisation method that would allow for a quick comparison of more than a few 2D-IR spectra in parallel. Additionally, each 2D-IR spectrum evolves over the waiting time, making a traditional step-by-step analysis even more cumbersome. In order to screen the 2D-IR residuals across all dsDNA-PA samples, ANOVA-PCA was used. This method was introduced in the previous chapter and is used in this case again to separate variance according to pre-defined factors[27,29,43].

7.4.3 ANOVA-PCA of the 2D-IR Residual Dataset

An in-depth explanation how this method is implemented can be found in chapter 6. The residual 2D-IR dataset is dependent on three factors, the choice of dsDNA sequence, the choice of polyamide and the pump-probe delay. These three parameters are defined for each residual spectrum in the 2D-IR dataset, matrix $\boldsymbol{Y}$, and are used in the ANOVA-PCA to separate the variance of $\boldsymbol{Y}$ into a set of nine matrices according to:

$$\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{S} + \boldsymbol{P} + \boldsymbol{W} + (\boldsymbol{SP}) + (\boldsymbol{SW}) + (\boldsymbol{PW}) + (\boldsymbol{SPW}) + \boldsymbol{E}\,. \tag{7.3}$$

Matrix $\boldsymbol{M}$ contains the global average residual spectrum and plotting any of its rows will return the 2D-IR spectrum from **Fig. 7.9b**. Matrices $\boldsymbol{S}$, $\boldsymbol{P}$ and $\boldsymbol{W}$ are defined as main factors and contain the variance of the 2D-IR response when using different sequences in $\boldsymbol{S}$, across different polyamides in $\boldsymbol{P}$ and between different waiting times in $\boldsymbol{W}$. The following four matrices ($\boldsymbol{SP}$), ($\boldsymbol{SW}$), ($\boldsymbol{PW}$) and ($\boldsymbol{SPW}$) are defined as interaction factors and contain differences between spectra that appear for example in a certain dsDNA-PA combination across all waiting times in ($\boldsymbol{SP}$), or when using a specific ligand at a specific waiting time across all sequences in ($\boldsymbol{PW}$). The last matrix $\boldsymbol{E}$ contains any residual variance left over from this decomposition such as instrumental noise and allows a comparison to evaluate whether the variance in any of the other matrices is significant.

This set of matrices can be used to examine how polyamide, dsDNA sequence and waiting time affect the spectral response individually and in combination. It is worth reiterating at this point that the residual response from the linear model was used as input and any spectral features from free dsDNA and free polyamide have already been removed, leaving only features sensitive to the dsDNA-PA interaction. This is different to the previous chapter. By performing a PCA on the sum $\boldsymbol{S} + \boldsymbol{E}$, it is therefore possible to analyse how the 2D-IR residuals change on average with different sequences,

irrespective of the polyamide being used. A PCA scores plot of subset $\boldsymbol{S} + \boldsymbol{E}$ is shown in **Fig. 7.10a**. Just as seen with FT-IR data, there is a clear separation along PC1 between samples with *SS*-centre (positive PC1-values) from samples with *WW*-centre (negative PC1). As expected from the pairing rules, a minor groove with 5'-TGWWCA-3' returns, on average a different spectral response upon binding than a 5'-TGSSCA-3' motif. PC2 further separates samples with *WW*-centre according to **TA** < **TT** < **AA** < **AT** with rising PC2 value. In the minor groove, sequences **TA** and **AT** represent two opposing structures for polyamides, as **TA** leads to a GT-step in both single strands (see highlighted structure in the inset of **Fig. 7.10a**) and **AT** leads to two GA steps. **TT** and **AA** fall in between both structures with one GA and one GT step each. This seems to be reflected in the corresponding spectral response of PC2, as interactions with **TA** and **AT** are covering minimum and maximum observed PC2 scores respectively.

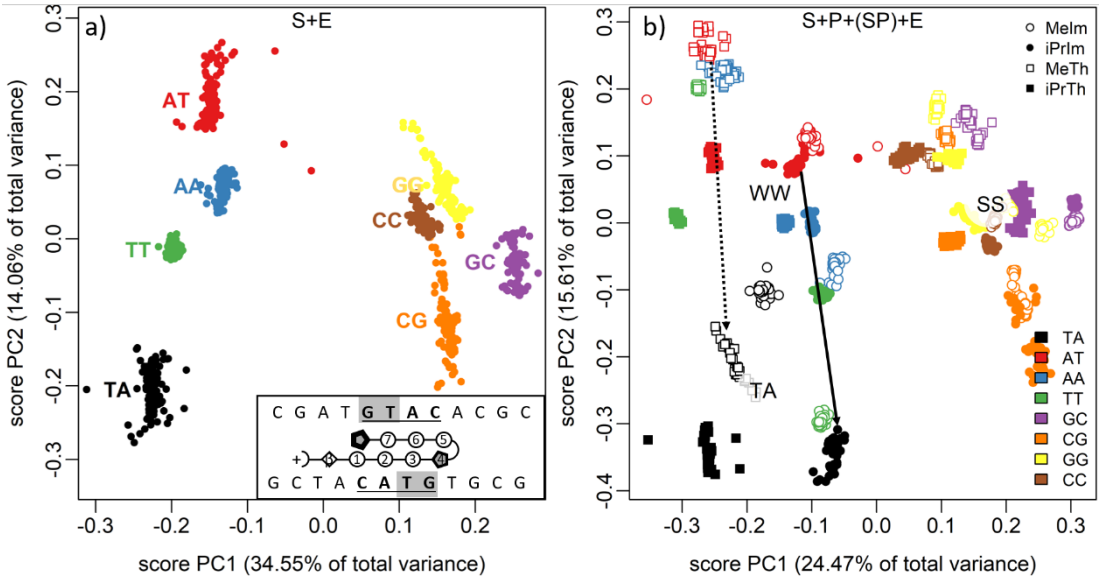


Fig. 7.10a) *PCA scores plot of subset* $\boldsymbol{S} + \boldsymbol{E}$. ***b)*** *PCA scores plot of subset* $\boldsymbol{S} + \boldsymbol{P} + (\boldsymbol{SP}) + \boldsymbol{E}$. *Colours indicates sequences, symbols indicate polyamides used.*

To observe how the choice of polyamide affects the residual 2D-IR response, the sum of $\boldsymbol{S} + \boldsymbol{P} + (\boldsymbol{SP}) + \boldsymbol{E}$ was analysed next via PCA. Since $\boldsymbol{M}$ is excluded from the subset, the coordinate origin of the scores plot will be represented by the global average residual spectrum. This PCA is comparable to the PCA on the FT-IR residuals shown in **Fig. 7.7**, as the response of each individual dsDNA-PA combination is being accounted for, while temporal information is still excluded. The scores plot is shown in **Fig. 7.10b** with different symbols indicating which polyamide is used. Comparing both PC score plots in this figure illustrates, how the initial grouping into sequences is further separated by

the choice of polyamide; now differentiating the response of all 32 samples. The four interactions of polyamides with **TA** separate further across PC2 in **Fig. 7.10b**, with the reference **TA-MeIm** (hollow, black circles) being closest to the bulk response of all other samples. The other three **TA** samples return a more distinctive 2D-IR response with **iPrIm** and **iPrTh** (solid black symbols) showing the most negative PC2 values of the dataset. This aligns with observations from FT-IR data and suggests a unique interaction of the substituted ligand candidates with the TG**TA**CA minor groove.

When comparing the residual 2D-IR response of the same polyamide ligand across different sequences (same symbols, different colours), **MeTh** (hollow squares) appears to be the most selective. Combinations of this ligand with **AT**, **TT** and **AA** group together (upper left corner of **Fig. 7.10b**) and only **TA-MeTh** clearly separates from these responses by a negative PC2 score (dotted arrow). Combinations with **iPrIm** (solid circles) show a spread of **AT** < **AA** < **TT**, again with a large separation from **TA-iPrIm** which is almost entirely described by PC2 (solid arrow). The PC scores of reference polyamide **MeIm** (hollow circles) are close to scores from **iPrIm** combinations, indicating similar spectral features, albeit with a different sequence order. **TT-MeIm** (green hollow circles) appears lower than **TA-MeIm** and falls close to other combinations with **TA** at negative PC2 scores. An early study on the A-T/T-A degeneracy of classical hairpin polyamides actually indicated a slight preference of 5'-TGTTT-3' over 5'-TGTAT-3' for two consecutive Py/Py pairs in a similar polyamide, with a 4x higher association constant for 5'-TGTTT-3'[11]. The paper also observed down to 12x higher association constants for 5'-TGTTT-3' relative to 5'-TGAAT-3' and suggested that narrowing of the minor groove caused by the GA base step may lead to this difference[44]. This provides evidence that GT and GA base stepping plays a subtle role in the selectivity of conventional hairpin polyamides and gives an indication as to why ligands with larger N-terminus such as **iPrIm** or **iPrTh** could be more susceptible to this base step.

In the previous chapter, dsDNA melting experiments were carried out to estimate binding affinities of the ligand and the results were compared with PCA scores to investigate the presence of a correlation between the parameters. Successful minor groove binding can stabilise the double stranded structure of the DNA, which increases the observed melting point. The change in melting point, referred to as the melting point stabilisation, was obtained using UV-absorption for all combinations (except

those with **CG** and **GG**) and is shown as a colour gradient in **Fig. 7.11** from +8°C to +24°C (red to green).

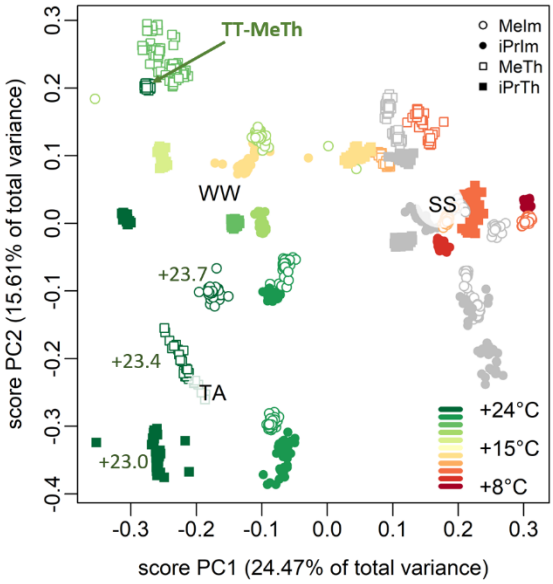


Fig. 7.11 *PCA scores plot of subset* ***S*** *+* ***P*** *+ (****SP****) +* ***E****. Colours indicates DNA melting point stabilisation upon interaction with the corresponding polyamide. Grey indicates omitted dsDNA-PA combinations.*

All of the measured melting point stabilisations are above +8°C, indicating an increase in double strand stability even in the cases of **GC** and **CC**. It is therefore possible that the ligand aligns with the minor groove in those cases to some extent and even forms a hairpin structure. The exocyclic N2-amino protons of the central two guanines will however clash with the protons of amide groups from the ligand and prevent a stronger association. Sequences with *WW*-centre show clearly higher melting point stabilisations (+14°C and more) as expected, which appear on the left side of the plot at negative PC1 values. The stabilisation does seem to increase further with falling PC2 scores, as the three most stable combinations, **TA-MeIm** (+23.7°C), **TA-MeTh** (+23.4°C) and **TA-iPrTh** (+23.0°C) all show negative PC2 values. Indeed, combinations with the **TA** sequence not only show a unique spectral response, but also return the largest melting point stabilisations in the dataset. These are followed by combinations with **TT** (**MeTh** with +22.8°C, **iPrTh** with +22.5°C and **TT-MeIm** with +22.1°C). The lowest stabilisation of all samples with *WW*-centre was observed for **AT-iPrIm** with +14.0°C at positive PC2 values. These values suggest an overall preference of **TA** over **AT** sequences that leads to higher stabilisation values with increasingly negative PC2 scores, but individual polyamide combinations show a more nuanced melting response that cannot be described by PC2 alone.

7.4.4 Spectral Analysis of Residual 2D-IR Features

The results so far have indicated a unique spectral response for polyamides interacting with the **TA** sequence. The residual 2D-IR spectra were therefore separated into three groups with SS (combinations with **GC, CG, CC** and **GG**), WW* (**TT**, **AA**, **AT** without **TA**) and TA. The PCA scores from subset $\boldsymbol{S} + \boldsymbol{P} + (\boldsymbol{SP}) + \boldsymbol{E}$ were used to calculate group-average scores, which are shown as locations of *"TA"*, *"WW*"*, *"SS"* in the plot in **Fig. 7.10b**. Groups SS and WW* are mainly described by PC1 and have PC2 scores close to zero. Group TA has a PC1 score very similar to group WW* and separates only along PC2. The group scores were then used to obtain 2D-IR residual spectra for each group shown in **Fig. 7.12**. Subset $\boldsymbol{S} + \boldsymbol{P} + (\boldsymbol{SP}) + \boldsymbol{E}$ does not contain the global average residual spectrum, which was therefore added to the group responses so that each spectrum shows the deviation of the observed spectrum from the linear combination model.

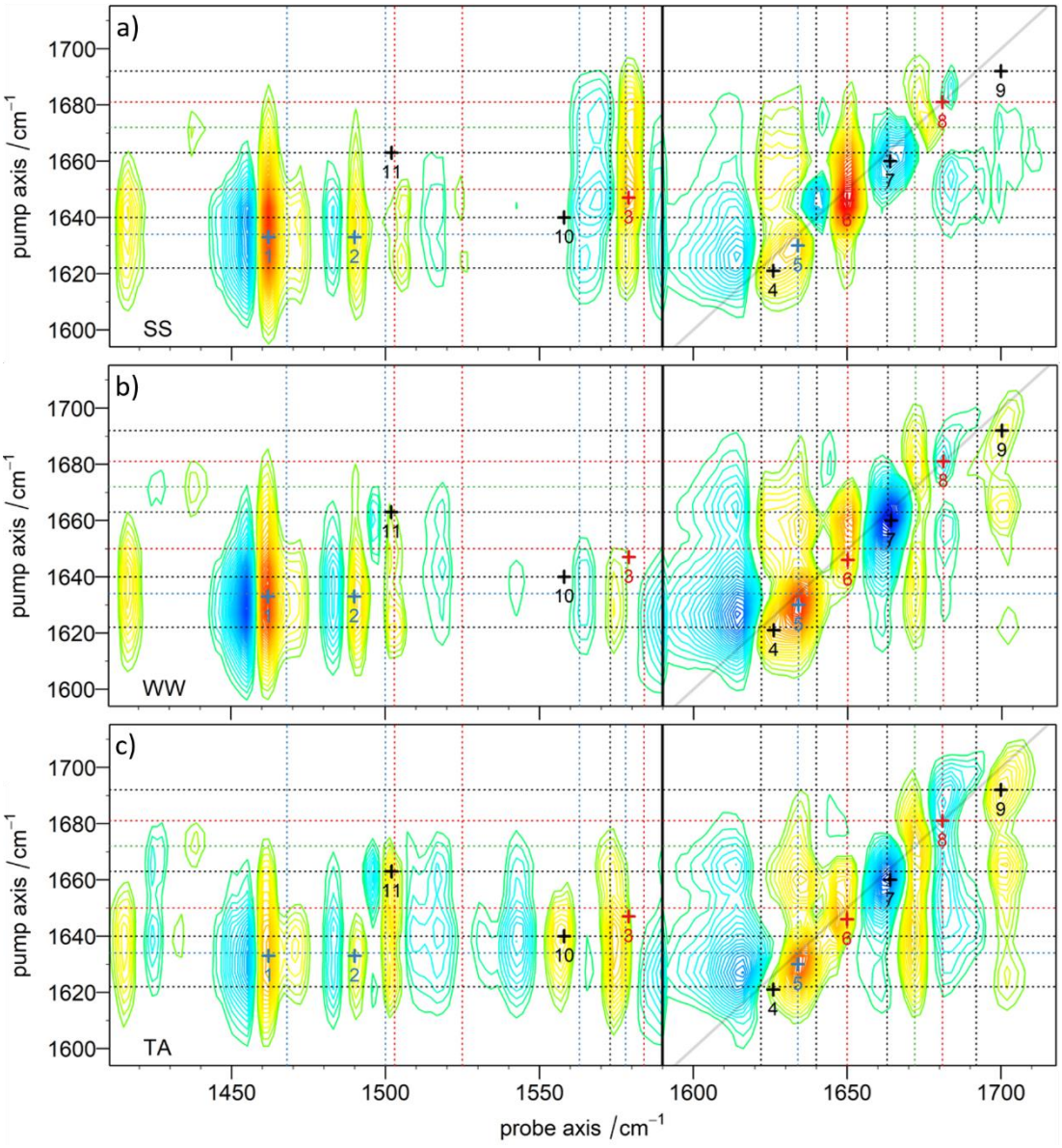


Fig. 7.12 *2D-IR spectra of group SS in* ***a)****, WW* in* ***b)*** *and TA in* ***c)****. Lines indicate bands from G/C structures in red, A/T structures in black and polyamide in blue, as observed in the individual component spectrum.*

The residual spectrum for group SS in **Fig. 7.12a** largely follows the global average residual response in **Fig. 7.9** with peaks **1-8** present in the spectrum. Spectral features **1, 2** and **5** (blue) were assigned intra-molecular cross-peaks from the amide I band (**5**) of the ligand to down-shifted N-methyl bending (**1**) and amide II (**2**) modes. The down-shift of these modes is present in all residual spectra and was also observed in FT-IR measurements. UV-melts have shown that even *SS*-sequences stabilise the dsDNA to some extent, so it is plausible if these features characterise reorganisation of the polyamide from its solvated form in buffer to a hairpin structure to fit into the minor groove. The amide linkages of a hairpin interact with the minor groove as proton

donors and have outward-facing carbonyl groups. So while interactions with inward-facing N-H bonds are expected to change upon binding, the carbonyl solvation by water should remain intact. DFT calculations on N-methylacetamide have shown that Amide II modes are sensitive to N-H hydrogen bonding and Amide I to carbonyl hydrogen bonding[45]. This may be reflected in the lower frequency observed for Amide II (point **2**) as well as an absence of a frequency shift in Amide I (point **5**).

Peaks **3, 4, 6**, **7** and **8** (black and red) originate from a structural change of the DNA double strand, as they were assigned to modes of nucleobases. Most characteristic features for group SS are rises of peaks **3** and **6**, together with an absence of off-diagonal features below the diagonal. Both peaks are assigned to G/C structures with the rising diagonal at **6** ($G_SC_S(+)$), possibly causing an increase of corresponding cross-peaks at **3** (G_R) as well. This suggests a perturbation to the vibrational energy landscape of G/C nucleobases for this group, which would be expected for a 5'-TGSSCA-3' minor groove interacting with a ligand.

The residual spectrum for group WW* in **Fig. 7.12b** shows an absence of G/C cross-peaks at **3**, and a less dominant diagonal peak **6**. This is almost certainly caused by the replacement of G/C with A/T structures in the centre. The group instead shows a drop in intensity of diagonal peak **7** (T_{4s} carbonyl) and a rise at peak **9** (T_{2s} carbonyl), together with emergence of a cross-peak pattern below the diagonal with peak **7** (T_{4s}) and **4** (A_{R1}). A comparison to the 2D-IR spectrum of free dsDNA in **Fig. 7.8a** suggests that the rising features at peak **9** and below are part of the coupled cross-peak network of A/T modes in dsDNA and have not shifted in frequency. A change in relative intensities of these coupled cross-peaks indicates a geometrical change between individual transition dipoles due to binding. This would imply a distortion of the dsDNA structure upon binding at A/T motifs. Such a perturbation is likely in case of a hairpin binding to 5'-TGWWCA-3', as the minor groove will be widened, and the relative orientation of individual nucleobases will change. Diagonal peaks **4** and **5** rise further in intensity than for group SS. Peak **4** is assigned to a rise of the A_{R1} ring-mode which has been associated with reduced base stacking for A/T structures[42,46].

The residual spectrum for group TA in c) shows even stronger cross-peaks of **9** (T_{2s}) with peaks **7** (T_{4s}) and **4** (A_{R1}) than for group WW*. The continued rise of off-diagonal features suggests a larger distortion of A/T structures at increasingly negative PC2 scores in group TA compared to WW*. Samples **TA-iPrIm** and **TA-iPrTh** returned the

lowest PC2 scores of the dataset, considerably lower than the reference combination **TA-MeIm**. Previous studies on the TG**TA**CA motif have shown that these two polyamides also cause a larger major groove compression than the reference polyamide **MeIm**[13]. The perturbance of A/T modes that distinguishes group TA might therefore be related to the structural change that drives the major groove compression.

Although A/T cross-peak intensities with T_{2s} increase for group TA, the rise of the diagonal peak **4** (A_{R1}) is less intense for group TA than in WW*. This suggests a smaller perturbation of a reporter mode on adenine when binding ligands to the minor groove of **TA**, as opposed to **AT**. In the 5'-TG**TA**CA-3' sequence, the adenines are furthest away from the guanine that is targeted by the substituted C-terminus of the ligand, which could explain the smaller, spectroscopic impact on the A_{R1} mode upon binding.

A smaller rise of peak **4** also allows for a better distinction to diagonal peak **5** (Amide I). This reveals a potential cross-peak above the diagonal between the amide I and the T_{4s} mode (cross-peak between **7** and **5** above diagonal), which could be a sign of coupling between a polyamide and a dsDNA mode. Ligand coupling with the T_{4s} mode seems counter-intuitive, as the corresponding C4-carbonyl of thymine points towards the major groove, even though these polyamides clearly bind to the minor groove of GTAC[12,13]. The vibrational modes of thymine however form a heavily coupled system[37,42], so it is possible that additional coupling to the T_{2s} mode (peak **9**) is present, but not visible due to the narrow bandwidth of the pump pulse (80 cm^{-1} FWHM). The long vertical tail of the cross-peak between **7** and **5** towards 1700 cm^{-1} could be an indication for this. Such a coupling would give spectroscopic evidence for hydrogen bonding between amide residues and the C2-carbonyl of the central thymine units. The suggested cross-peak between **7** and **5** is also visible in group WW*, although being obscured by additional cross-peaks between **7** and **4.**

The loss of a peak pair at point **8**, together with a loss of corresponding, vertical cross-peaks at a probe frequency of 1682 cm^{-1} is also observed for group TA. This loss appears to be more dominant in TA than in WW*. This has been observed previously when analysing FT-IR residuals in **Fig. 7.7b** and is assigned to a loss of the delocalised G-carbonyl mode $G_SC_S(-)$, based on its location. Cross-peaks between **6** and **8** ($G_SC_S(+)$, $G_SC_S(-)$) have been associated with strong inter-strand coupling between the carbonyls of G and C, leading to two delocalised vibrations for this base pair[36,40]. The apparent perturbance of this cross-peak indicates that inter-strand, G-C coupling is affected by

the ligand, possibly due to a competing interaction with the polyamide. Loss of the ($G_SC_S(+)$, $G_SC_S(-)$) cross-peak is observed in all three groups upon binding, although TA shows a more pronounced contribution of negative features at slightly lower wavenumbers, towards 1672 cm^{-1}. This feature might be reminiscent of a peak shift to lower frequencies of peak **8** and its cross-peaks, but a clear and definite assignment will not be possible due to overlap with the TFA peak in that region.

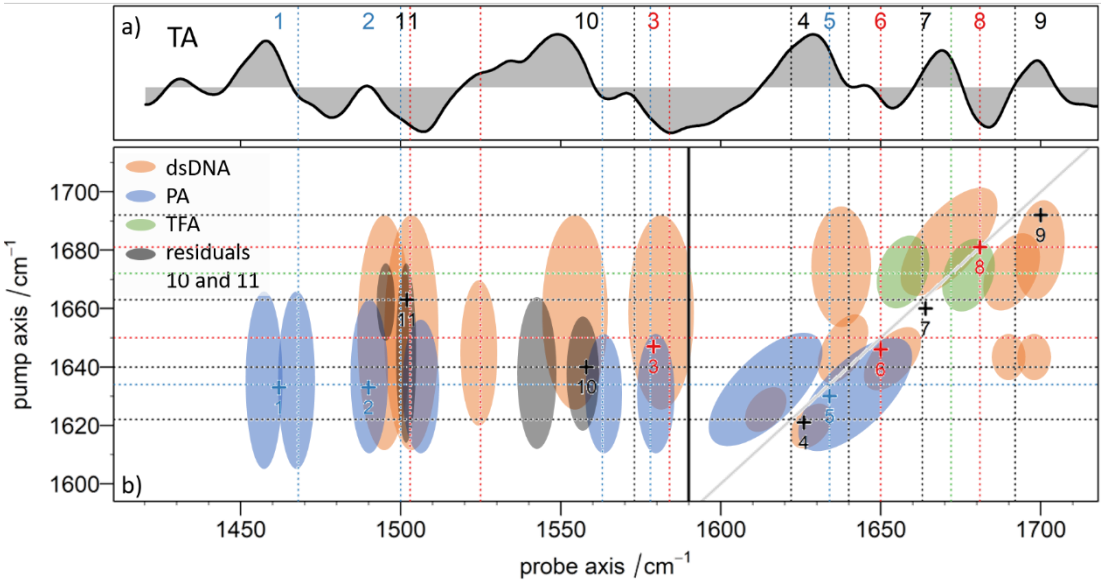


Fig. 7.13a) *FT-IR residual spectrum for group TA, copied from* ***Fig. 7.7*** *in section 7.4.1.* ***b)*** *Overlay of 2D-IR absorption features in order to assign residual peaks 10 and 11.*

The off-diagonal region between 1410 to 1590 cm^{-1} also shows two new cross-peaks at positions **10** and **11** in group TA. These features are barely visible in group WW* and are characteristic of samples with a negative PC2 score. In order to further characterise them, peak positions of **1-11** were overlapped with spectral features from free dsDNA (orange areas) and free polyamides (blue areas) as shown in **Fig. 7.13b**. The 2D-IR spectrum of polyamide in buffer consists of cross-peaks with the Amide I band at a pump wavenumber around 1630 cm^{-1}. Peak **11** is pumped at considerably higher wavenumbers than all other polyamide cross-peaks, at around 1664 cm^{-1}, coinciding with T_{4s} from the dsDNA. The probe frequency of peak **11** overlaps with the Amide II as well as the C_{R3} ring-mode of cytosine and could therefore be assigned to either one of the two. Peak **10** is even more ambiguous at a central pump frequency of 1640 cm^{-1}, coinciding with the T_R ring-mode of thymine and close to the Amide I mode. The peak is located at 1557 cm^{-1} along the probe axis, close to ring-modes of cytosine (C_{R2}) and polyamides (P_{R2}). Both features show a potential cross peak from PA to dsDNA, which

would give further evidence for an interaction between ligand and dsDNA sequence, but a conclusive band assignment is not possible.

7.4.5 Comparison of the temporal response of TA-MeIm with TA-iPrIm

The time-resolved information available from this dataset is plentiful and an in-depth analysis of all 44 samples would require a separate chapter on its own. The ANOVA-PCA approach from the last chapter considers the complete dataset, but only retrieves a principal vibrational relaxation that provides a very simplistic representation of the vibrational dynamics in the dataset.

The analysis instead focusses on characterising the dynamics of **TA-MeIm** and **TA-iPrIm**. Both combinations use the same dsDNA sequence but have shown different 2D-IR responses according to the PCA results in **Fig. 7.10b**. In order to analyse the temporal information for the two dsDNA-PA combinations, the corresponding, time-resolved, 2D-IR residuals obtained from the linear model were subjected to a global fit. The exact relaxation values returned from the residuals may be different to the kinetics observed in the raw 2D-IR response, but they allow for a relative comparison of subtle differences in the time-resolved data. A bi-exponential function was used to fit the temporal response of each 2D-IR pixel in a stack of residual 2D-IR spectra to two common decay times, τ_a and τ_b. The function was defined as:

$$y(t, \mathrm{px}) = a_{\mathrm{px}} \cdot e^{-t/\tau_a} + b_{\mathrm{px}} \cdot e^{-t/\tau_b}\,, \tag{7.4}$$

where $y(t, \mathrm{px})$ is the residual 2D-IR intensity at a specific 2D-IR pixel, px, over time, t. Parameters a_{px} and b_{px} are pre-exponential factors, which were optimised for each individual pixel. Once converged, each pre-exponential factor forms a map of 2D-IR intensities that represent spectral features with the same decay time. **Fig. 7.14** shows the result of such a global fit for **TA-MeIm**. Plot a illustrates the two decay times returned from the fit and plot b gives an indication how well the model agrees with the data. A global, bi-exponential fit may be unable to model each individual relaxation precisely, but it allows to separate faster relaxations from slower processes. Plots c and d show the intensities of pre-exponential factors a_{px} and b_{px} and indicate which features relax faster and which relax more slowly.

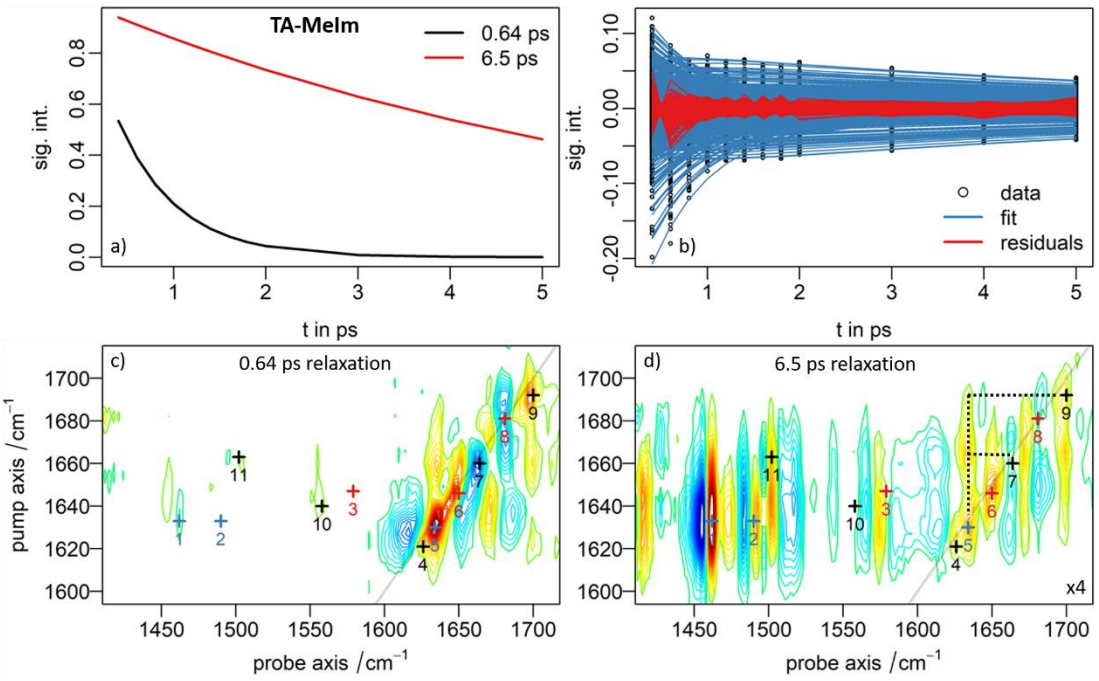


Fig. 7.14 *Global fit results for the time-resolved, 2D-IR residual spectra of* ***TA-MeIm a)*** *indicates a fast and a slow relaxation.* ***b)*** *Each pixel was fitted to a bi-exponential function with common relaxation times.* ***c)*** *and* ***d)*** *show obtained, pre-exponential factors for the fast and slow relaxation respectively. Red colours indicate negative values, blue colours positive values. The intensities in plot d) were multiplied by 4 to improve peak-visibility.*

The 0.6 ps relaxation in **Fig. 7.14a** is on the same timescale as the fast, vibrational lifetimes usually observed for dsDNA modes between 1600 and 1700 cm^{-1} [27,41]. The corresponding pre-exponential factors in **Fig. 7.14c** show that predominantly modes on the diagonal are affected, together with few off-diagonal features in the same spectral region. These modes will be excited directly by the pump pulse (centred at 1650 cm^{-1}) and appear to relax together within the first picosecond of the measurement. It is therefore possible, that the fast-relaxing residuals in **Fig. 7.14c** represent a network of coupled vibrational modes in the sample, that is being perturbed by minor groove binding. The features are clearly dominated by dsDNA modes, although the amide I mode of the polyamide will possess a similar vibrational relaxation of on the order of 1 ps[47,48] and may therefore contribute.

The second, slower relaxation of 6.5 ps has more likely emerged as a result from a preceding energy transfer process that dissipates vibrational energy less efficiently. Such a process seems evident from **Fig. 7.14d**, as energy transfer from one mode to another will enhance off-diagonal features in particular. It appears that most of the transferred energy relaxes at point **1**; the down-shifted cross-peak from Amide I to the CH_3-bending mode of the ligand. This suspected, intra-molecular energy transfer to the

methyl-groups of pyrroles (and possibly imidazoles) is the most significant feature in **Fig. 7.14d**. Additionally, the earlier suspected cross-peak between amide I of **MeIm** and T_{4S} of **TA** (cross-peaks **5** with **7**, dotted lines) is even more apparent in this plot than in previously shown data and extends further up to T_{2S}, suggesting a direct observation of PA-dsDNA interactions.

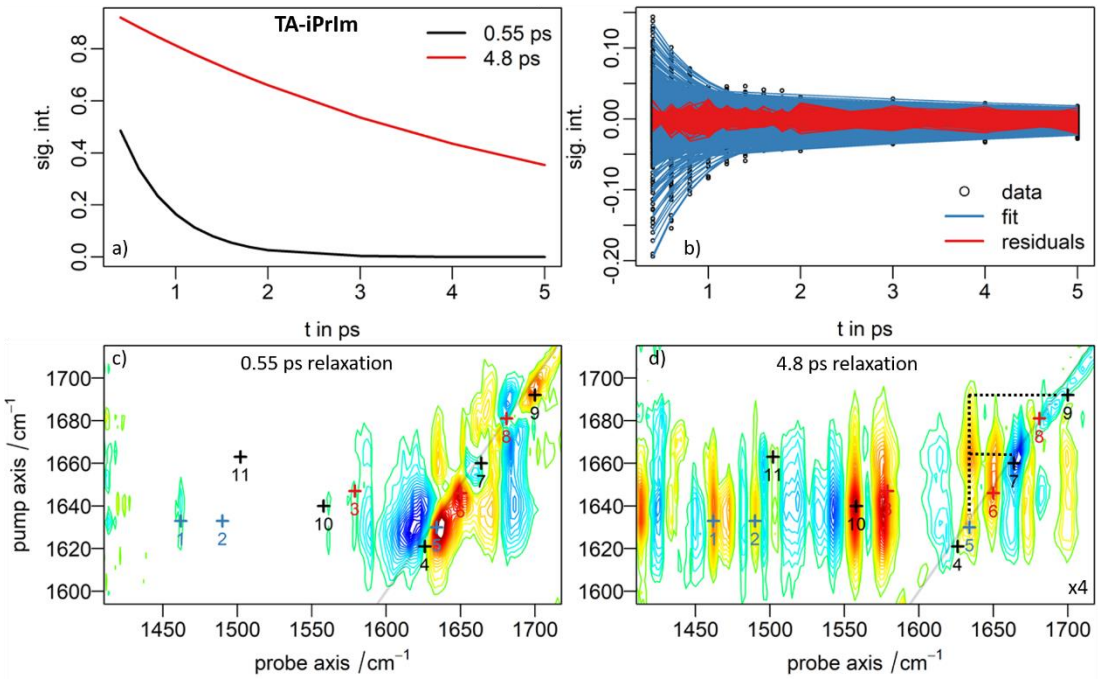


Fig. 7.15 *Global fit results for the time-resolved, 2D-IR residual spectra of* ***TA-iPrIm a)*** *indicates a fast and a slow relaxation.* ***b)*** *Each pixel was fitted to a bi-exponential function with common relaxation times.* ***c)*** *and* ***d)*** *show obtained, pre-exponential factors for the fast and slow relaxation respectively. The intensities in plot d) were multiplied by 4 to improve peak-visibility. Plots a-d are on the same scale as a-d in* ***Fig. 7.14****.*

This response is compared to a global fit of 2D-IR data for **TA-iPrIm** shown in **Fig. 7.15**. The fast decay of 0.55 ps again describes possible vibrational relaxation of the pumped diagonal features with corresponding coupling cross-peaks outlined in **Fig. 7.15c**. The fast responses of **TA-MeIm** and **TA-iPrIm** compare well with each other as points **4-9** and even cross-peaks show the same, qualitative change. The slower relaxation however differs between both samples. With 4.8 ps, the second decay is notably faster for **TA-iPrIm**, which is also visible when comparing the temporal response of the input data in **Fig. 7.14b** and **Fig. 7.15b**, as the raw residuals of **TA-iPrIm** show smaller intensities at 2-5 ps. The most dramatic difference between the two samples is observed in the pre-exponential factors obtained for the slow relaxation. Instead of the large relaxation at point **1** in **TA-MeIm**, two peaks at point **10** and close to point **3** are now the most significant features in the spectrum. This

suggests a different, predominant, energetic pathway in the slow residual dynamics of the **TA-iPrIm** complex. Both features were pumped at a central frequency of ≈1639 cm^{-1} (T_R or Am I) and show a relaxation between 1550 and 1580 cm^{-1}, a convoluted, spectral region of ring-modes from G, A or C of dsDNA and the heterocycles of polyamide. It is possible that the energetic pathway from the Am I to the CH_3-bend in **TA-MeIm** (point **1**) has been replaced to some extent by a competing pathway from the Amide I to various ring-modes in **TA-iPrIm**. Ring modes in the centre of the DNA or in the ligand will be able to dissipate energy more readily to surrounding modes than the bending modes of unpolar methyl-groups at the periphery of the ligand, which would account for the shorter relaxation time in **TA-iPrIm**. Both samples are distinguished from each other in the PCA results by a separation in PC2 due to the emergence of peak **10**. The additional energetic pathway observed for **TA-iPrIm** will therefore also play a role in the unique response that was observed for **TA-MeTh** and **TA-iPrTh,** as this peak is visible in all three cases. Analysing time-resolved information of this dataset using the ANOVA-PCA approach only returned the vibrational coupling in the first principal component and the energy transfer response could not be readily extracted without further analysis of additional, principal components.

7.5 Conclusions and Future Outlook

This chapter has extended the application of 2D-IR as a screening tool to study the selectivity of three novel minor groove binders with respect to an existing reference structure. 32 different dsDNA-PA combinations were studied via linear infrared absorption and 2D-IR spectroscopy. In a first step, redundant information, invariant to the dsDNA-ligand interaction was removed from the dataset by performing a linear model using individual component spectra. The residuals from the linear model were then analysed via ANOVA-PCA in order to study how the choice of dsDNA sequence and ligand affects the 2D-IR response and to identify common response patterns among different combinations. From these results, key dsDNA-PA interactions were selected, and the temporal response of the 2D-IR residual features was analysed using a global fit.

The screening experiment on different dsDNA-polyamide combinations has shown a significant perturbation of the observed infrared response in all 32 cases. Interaction with any of the eight sequences caused a shift of Amide II and CH_3-bending modes to lower wavenumbers for all four polyamides. The interaction also seemed to affect

inter-strand mode coupling of G with C carbonyls, as cross-peak intensities between both modes decreased upon binding (peak **8**) in all cases. Combinations with SS-centre further returned a perturbation of G/C-related vibrational modes and responded with a small, but still significant melting point stabilisation (>8°C). This suggests that even for a 5'-TGSSCA-3' minor groove, reorganisation of the polyamide to a hairpin can possibly take place and minor groove binding may occur to some extent, albeit forming a less stable complex.

However, all four polyamide ligands showed a clear preference of a 5'-TGWWCA-3' minor groove over 5'-TGSSCA-3', with melting stabilisations above +14°C and an observed perturbation of A/T-related vibrational modes for sequences with WW-centre. This observation follows the established pairing rules for polyamide binding as expected.

The screening experiment has further indicated a preference for the 5'-TG**TA**CA-3' motif in case of ligands with modified N-terminus. Polyamides **iPrIm**, **MeTh** and **iPrTh** were developed as analogues to the reference ligand **MeIm** and returned a distinctive spectral response when interacting with **TA**. These distinctive features were observed with linear infrared absorption as well as 2D-IR spectroscopy and suggest a perturbation of the coupled network of A/T modes, together with emergence of a set of possible dsDNA-PA cross-peaks. The perturbation possibly stems from widening the minor groove and compressing the major groove upon binding, as observed for these ligands in case of **TA**[13]. The features are less prominent for dsDNA motifs **AA** and **AT,** which signifies that the GT step in **TA** and **TT** at the substitution site could play a crucial role in the sequence-recognition of the polyamide analogues. UV-melting experiments returned the largest melting point stabilisations for dsDNA-PA complexes with **TA** (above +21°C) and confirmed these observations. Earlier studies have suggested a small preference of a GT over an GA base step even for the reference ligand **MeIm**[11], but the larger N-terminus in the newly developed polyamide analogues, especially in the cases with iso-propyl residue, seems to amplify this preference, possibly leading to a deviation from the established pairing rules for these ligands.

Analysis of time-resolved 2D-IR data for **TA-MeIm** and **TA-iPrIm** further indicated a competing relaxation process in case of **iPrIm**, which leads to slightly faster relaxation times than compared to **MeIm**. The data indicates that a suspected, intra-molecular energy transfer from Amide I to the CH_3-bending mode in the polyamide may be

partially superseded by a faster transfer of energy to an unassigned ring-mode between 1550-1580 cm^{-1}. Presence of an intermolecular energy-transfer from the polyamide to the DNA is indicated but cannot be completely confirmed without further evidence. This process contributes to the unique spectral response that separates the interaction of the alternative polyamide structures with the 5'-TG**TA**CA-3' minor groove from all other combinations.

The pure acquisition time of the complete 2D-IR dataset was 22 h for all 44 samples. The 2D-IR spectrometer *LIFEtime* at the Rutherford Appleton Laboratories has developed to a turn-key instrument with minimal user interference once it is aligned. The usage of a flow cell and externally prepared, freeze-dried samples allow for a greatly reduced sample preparation time on-site. These technical advancements together with scalable analysis methods such as ANOVA-PCA transform 2D-IR spectroscopy to a versatile screening tool providing detailed structural information. Automation of the sample loading procedure via stopped-flow pump systems or multi-well plates could open up the possibility to supply researchers with a remote, on-demand service for time-resolved infrared spectroscopy. The ambiguity on some of the spectral features may further be reduced by combining experimental data with simulated 2D-IR spectra of dsDNA-ligand complexes.

7.6 Appendix

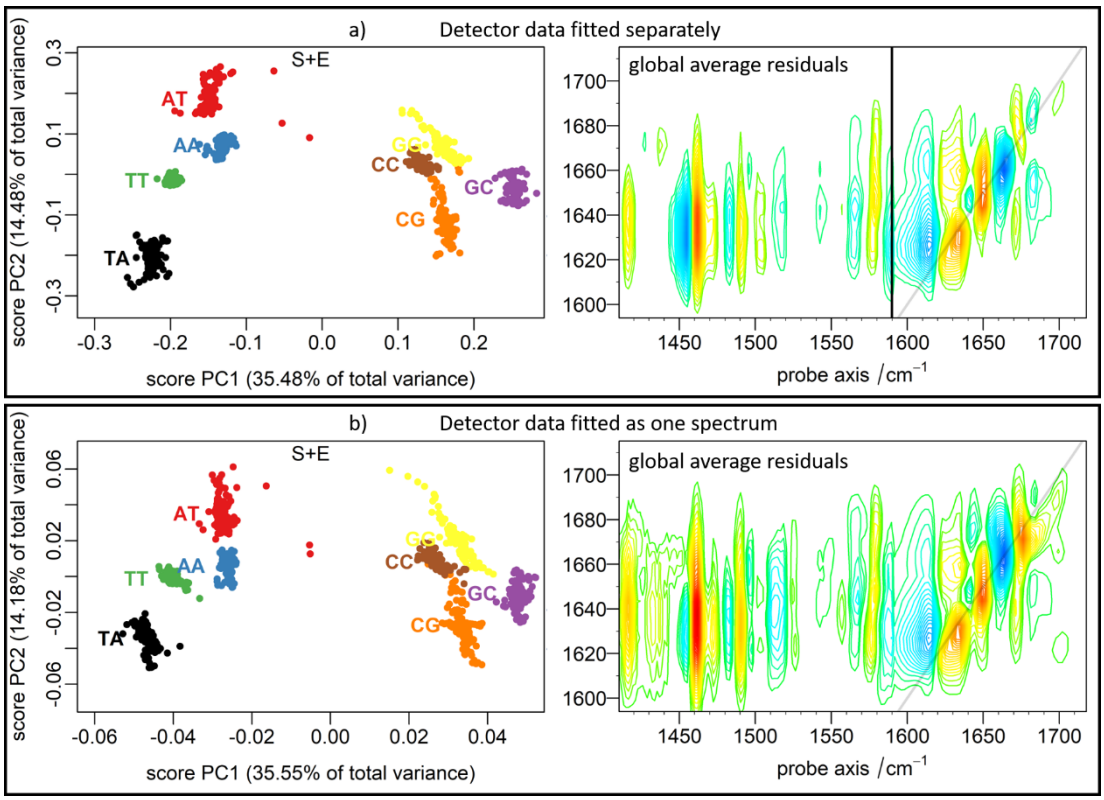


Fig. A-7.1 *When performing the linear combination outlined in section 7.3.4, the 2D-IR data collected from the two detectors can either be fitted separately and then combined to a single dataset (**a**), or combined first and then fitted with combined spectra (**b**). The corresponding ANOVA-PCA result for a subset **S+E** and the global average residual spectrum is shown for each approach and are quite similar. However, the fitting of combined spectra inevitably leads to more residuals, including residual features from TFA.*

7.7 References

1. Neidle, S. DNA minor-groove recognition by small molecules (up to 2000). *Nat. Prod. Rep.* **18**, 291–309 (2001).

2. Alniss, H. Y. Thermodynamics of DNA Minor Groove Binders. *J. Med. Chem.* acs.jmedchem.8b00233 (2018). doi:10.1021/acs.jmedchem.8b00233

3. Bazhulina, N. P. *et al.* Binding of Hoechst 33258 and its Derivatives to DNA. *J. Biomol. Struct. Dyn.* **26**, 701–718 (2009).

4. Dervan, P. B., Doss, R. M. & Marques, M. A. Programmable DNA binding oligomers for control of transcription. *Curr. Med. Chem. Anticancer. Agents* **5**, 373–387 (2005).

5. Withers, J. M. *et al.* DNA Minor Groove Binders as Therapeutic Agents. in *Comprehensive Supramolecular Chemistry II* 149–178 (Elsevier, 2017). doi:10.1016/B978-0-12-409547-2.12561-2

6. Dervan, P. B. & Bürli, R. W. Sequence-specific DNA recognition by polyamides. *Curr. Opin. Chem. Biol.* **3**, 688–693 (1999).

7. Geierstanger, B. H., Mrksich, M., Dervan, P. B. & Wemmer, D. E. Design of a G·C-specific DNA minor groove-binding peptide. *Science (80-.).* **266**, 646–650 (1994).

8. White, S., Szewczyk, J. W., Turner, J. M., Baird, E. E. & Dervan, P. B. Recognition of the four Watson–Crick base pairs in the DNA minor groove by synthetic ligands. *Nature* **391**, 468–471 (1998).

9. Dervan, P. B. & Edelson, B. S. Recognition of the DNA minor groove by pyrrole-imidazole polyamides. *Curr. Opin. Struct. Biol.* **13**, 284–299 (2003).

10. Dervan, P. B. Molecular recognition of DNA by small molecules. *Bioorg. Med. Chem.* **9**, 2215–35 (2001).

11. White, S., Baird, E. E. & Dervan, P. B. Effects of the A·T/T·A Degeneracy of Pyrrole–Imidazole Polyamide Recognition in the Minor Groove of DNA †. *Biochemistry* **35**, 12532–12537 (1996).

12. Padroni, G., Parkinson, J. A., Fox, K. R. & Burley, G. A. Structural basis of DNA duplex distortion induced by thiazole-containing hairpin polyamides. *Nucleic Acids Res.* **46**, 42–53 (2018).

13. Aman, K., Padroni, G., Parkinson, J. A., Welte, T. & Burley, G. A. Structural and Kinetic Profiling of Allosteric Modulation of Duplex DNA Induced by DNA-Binding Polyamide Analogues. *Chem. - A Eur. J.* (2018). doi:10.1002/chem.201805338

14. Chenoweth, D. M. & Dervan, P. B. Structural Basis for Cyclic Py-Im Polyamide Allosteric Inhibition of Nuclear Receptor Binding. *J. Am. Chem. Soc.* **132**, 14521–14529 (2010).

15. Nickols, N. G. *et al.* Activity of a Py-Im Polyamide Targeted to the Estrogen Response Element. *Mol. Cancer Ther.* **12**, 675–684 (2013).

16. Nickols, N. G. & Dervan, P. B. Suppression of androgen receptor-mediated gene expression by a sequence-specific DNA-binding polyamide. *Proc. Natl. Acad. Sci.* **104**, 10418–10423 (2007).

17. Kurmis, A. A., Yang, F., Welch, T. R., Nickols, N. G. & Dervan, P. B. A pyrrole-imidazole polyamide is active against enzalutamide-resistant prostate cancer. *Cancer Res.* **77**, 2207–2212 (2017).

18. Liu, B. & Kodadek, T. Investigation of the Relative Cellular Permeability of DNA-Binding Pyrrole–Imidazole Polyamides. *J. Med. Chem.* **52**, 4604–4612 (2009).

19. Nishijima, S. *et al.* Cell permeability of Py–Im-polyamide-fluorescein conjugates: Influence of molecular size and Py/Im content. *Bioorg. Med. Chem.* **18**, 978–983 (2010).

20. Alniss, H. Y. *et al.* Recognition of the DNA Minor Groove by Thiazotropsin Analogues. *ChemBioChem* **15**, 1978–1990 (2014).

21. de Clairac, R. P. L., Geierstanger, B. H., Mrksich, M., Dervan, P. B. & Wemmer, D. E. NMR Characterization of Hairpin Polyamide Complexes with the Minor Groove of DNA. *J. Am. Chem. Soc.* **119**, 7909–7916 (1997).

22. Chenoweth, D. M. & Dervan, P. B. Allosteric modulation of DNA by small molecules. *Proc. Natl. Acad. Sci.* **106**, 13175–13179 (2009).

23. Han, Y.-W. *et al.* Binding of hairpin pyrrole and imidazole polyamides to DNA: relationship between torsion angle and association rate constants. *Nucleic Acids Res.* **40**, 11510–11517 (2012).

24. Swalley, S. E., Baird, E. E. & Dervan, P. B. Effects of γ-Turn and β-Tail Amino Acids on Sequence-Specific Recognition of DNA by Hairpin Polyamides. *J. Am. Chem. Soc.* **121**, 1113–1120 (1999).

25. Hunt, N. T. 2D-IR spectroscopy: ultrafast insights into biomolecule structure and function. *Chem. Soc. Rev.* **38**, 1837–1848 (2009).

26. Baiz, C. R., Peng, C. S., Reppert, M. E., Jones, K. C. & Tokmakoff, A. Coherent two-dimensional infrared spectroscopy: Quantitative analysis of protein secondary structure in solution. *Analyst* **137**, 1793 (2012).

27. Fritzsch, R. *et al.* Rapid Screening of DNA–Ligand Complexes via 2D-IR Spectroscopy and ANOVA–PCA. *Anal. Chem.* **90**, 2732–2740 (2018).

28. Donaldson, P. M., Greetham, G. M., Shaw, D. J., Parker, A. W. & Towrie, M. A 100 kHz Pulse Shaping 2D-IR Spectrometer Based on Dual Yb:KGW Amplifiers. *J. Phys. Chem. A* **122**, 780–787 (2018).

29. Harrington, P. D. B. *et al.* Analysis of variance – principal component analysis : A soft tool for proteomic discovery. *Anal. Chim. Acta* **544**, 118–127 (2005).

30. O'Hare, C. C. *et al.* DNA sequence recognition in the minor groove by crosslinked polyamides: The effect of N-terminal head group and linker length on binding affinity and specificity. *Proc. Natl. Acad. Sci.* **99**, 72–77 (2002).

31. Anthony, N. G. *et al.* Short Lexitropsin that Recognizes the DNA Minor Groove at 5'-ACTAGT-3': Understanding the Role of Isopropyl-thiazole. *J. Am. Chem. Soc.* **126**, 11338–11349 (2004).

32. Greetham, G. M. *et al.* A 100 kHz Time-Resolved Multiple-Probe Femtosecond to Second Infrared Absorption Spectrometer. *Appl. Spectrosc.* **70**, 645–653 (2016).

33. Shim, S., Strasfeld, D. B., Ling, Y. L. & Zanni, M. T. Automated 2D IR spectroscopy using a mid-IR pulse shaper and application of this technology to the human islet amyloid polypeptide. *PNAS* **104**, 14197–14202 (2007).

34. Hamm, P. & Zanni, M. *Concepts and Methods of 2D Infrared Spectroscopy*. (Cambridge University Press, 2011). doi:10.1017/CBO9780511675935

35. Lee, C., Park, K.-H. & Cho, M. Vibrational dynamics of DNA. I. Vibrational basis modes and couplings. *J. Chem. Phys.* **125**, 114508 (2006).

36. Lee, C. & Cho, M. Vibrational dynamics of DNA. II. Deuterium exchange effects and simulated IR absorption spectra. *J. Chem. Phys.* **125**, 114509 (2006).

37. Lee, C., Park, K.-H., Kim, J.-A., Hahn, S. & Cho, M. Vibrational dynamics of DNA. III. Molecular dynamics simulations of DNA in water and theoretical calculations of the two-dimensional vibrational spectra. *J. Chem. Phys.* **125**, 114510 (2006).

38. Banyay, M., Sarkar, M. & Graslund, A. A library of IR bands of nucleic acids in solution. *Biophys. Chem.* **104**, 477–488 (2003).

39. Spectral Database for Organic Compounds (SDBS); NMR. *Spectral Database for Organic Compounds (SDBS); SDBS No.: HSP-03-379* Available at: http://riodb01.ibase.aist.go.jp/sdbs/. (Accessed: 8th August 2015)

40. Krummel, A. T., Mukherjee, P. & Zanni, M. T. Inter and Intrastrand Vibrational Coupling in DNA Studied with Heterodyned 2D-IR Spectroscopy. *J Phys Chem B* **107**, 9165–9169 (2003).

41. Hithell, G. *et al.* Effect of oligomer length on vibrational coupling and energy relaxation in double-stranded DNA. *Chem. Phys.* in Press (2018). doi:10.1016/j.chemphys.2017.12.010

42. Hithell, G. *et al.* Ultrafast 2D-IR and Optical Kerr Effect Spectroscopy Reveal the Impact of Duplex Melting on the Structural Dynamics of DNA. *Phys. Chem. Chem. Phys.* **19**, 10333–10342 (2017).

43. Sarembaud, J., Pinto, R., Rutledge, D. N. & Feinberg, M. Application of the ANOVA-PCA method to stability studies of reference materials. *Anal. Chim. Acta* **603**, 147–154 (2007).

44. Dickerson, R. E., Goodsell, D. S. & Neidle, S. '...the Tyranny of the Lattice...'. *Proc. Natl. Acad. Sci.* **91**, 3579–3583 (2006).

45. Myshakina, N. S., Ahmed, Z. & Asher, S. A. Dependence of Amide Vibrations on Hydrogen Bonding. *J. Phys. Chem. B* **112**, 11873–11877 (2008).

46. Doorley, G. W. *et al.* Tracking DNA excited states by picosecond-time-resolved infrared spectroscopy: Signature band for a charge-transfer excited state in stacked adenine-thymine systems. *J. Phys. Chem. Lett.* **4**, 2739–2744 (2013).

47. Feng, C. J. & Tokmakoff, A. The dynamics of peptide-water interactions in dialanine: An ultrafast amide i 2D IR and computational spectroscopy study. *J. Chem. Phys.* **147**, 1–11 (2017).

48. Tan, J., Zhang, J., Li, C., Luo, Y. & Ye, S. Ultrafast energy relaxation dynamics of amide I vibrations coupled with protein-bound water molecules. *Nat. Commun.* **10**, 1010 (2019).

8 Investigating Minor Groove Binding of Hoechst 33258 in dsDNA using Time-Resolved, Temperature-Jump IR Spectroscopy

8.1 Abstract

Minor groove binder Hoechst 33258 is revisited in this chapter using time-resolved infrared spectroscopy at non-equilibrium conditions. A ns-pulse is used to generate a rise in solvent temperature in the sample that initiates unfolding processes of DNA duplex structures. The observed melting dynamics are used to study the stabilising effect of minor groove binding on different dsDNA sequences. Hoechst 33258 binding to ds(5'-GCAAATTTCC-3') initially results in a significant reduction of pre-melting disruption at the minor groove on nanosecond timescales, which in turn leads to end-fraying of the G/C-ends. Binding also causes a dramatic delay of unfolding dynamics on microsecond timescales across the strand by a factor of 10. Similar observations are made for ligand binding to the suboptimal minor groove of ds(5'-GCATATATCC-3'), which shows a smaller impact to the disruption on nanoseconds and a smaller delay of unfolding processes by a factor of 5. These results indicate how effective ligand binding influences the dynamic behaviour of base pairs beyond the binding site and provides a new analytical approach to study minor groove binding in solutions on a broad range of timescales.

This chapter contains results published in the following publication:

Fritzsch, R.; Greetham, G. M.; Clark, I. P.; Minnes, L.; Towrie, M.; Parker, A. W.; Hunt, N. T. Monitoring Base-Specific Dynamics during Melting of DNA–Ligand Complexes Using Temperature-Jump Time-Resolved Infrared Spectroscopy. *J. Phys. Chem. B* **123**, 6188–6199 (2019).

8.2 Introduction

The minor groove of double-stranded DNA (dsDNA) has proven to be an effective target site in the development of clinically relevant DNA binders[1–3]. Small molecules can be tailored to recognise the minor groove of a specific DNA sequence and effective binding to the dsDNA can render a particular gene inaccessible for biological processes like transcription or replication. Derivative structures of Hoechst 33258 have shown antibacterial activity as well as bacterial topoisomerase I inhibition[4–6]. Hoechst 33258 targets "A-tract" sequences (A_nT_n, n≥2) with two or more consecutive adenines and thymines along the strand[7–10]. The ligand is believed to bind in an induced fit model, where subtle adjustments to the propeller twists of the base pairs ensure a non-covalent but tightly bound DNA-Ligand-complex[11,12]. Effective binding into the minor groove stabilises the dsDNA structure, which manifests itself as an increase in the melting point of the dsDNA[13]. Molecular dynamics simulations of such a process using covalently bound ligands have observed an increased level of refolding events in presence of a ligand[14], although the exact, underlying molecular mechanism of this stabilisation is still unclear. A detailed understanding of the melting process at the binding site can therefore deliver key insights for the successful development of new minor groove binders with high selectivity.

As outlined in the beginning of this thesis (see sections 4.3.5 and 5.4), a nanosecond, infrared laser pulse tuned to an absorption feature of the solvent can be used to perturb the thermal equilibrium of a sample and induce a temperature jump (T-Jump) of several degree Celsius[15–18] (9°C in this case). This perturbation can trigger unfolding processes of dsDNA structures, which can be followed over time using infrared spectroscopy[19]. Sanstead *et al.* have recently used this approach to investigate melting dynamics of free oligomers with G-C base pairs at different positions within an otherwise *all-A/T* sequence[20]. The vibrational modes affected by the T-Jump contain structural information that characterises the stability at a particular position within the double strand with nanosecond to millisecond time resolution. By comparing the T-Jump response of free dsDNA with the response of the corresponding dsDNA-Ligand complex, it is possible to extract where and when the double helical structure is stabilised due to interaction with a ligand.

This chapter uses minor groove binding of Hoechst 33258 as a reference system to assess the capabilities of temperature-jump experiments in time-resolved, infrared

spectroscopy as a tool to study DNA-ligand interactions. The technique provides unprecedented insight into the molecular interactions that give rise to the large minor groove stabilisation of Hoechst 33258 (as seen in chapter 6). T-Jump responses of four different dsDNA sequences were measured on their own and in a 1:1 ratio with Hoechst 33258 to study how dsDNA unfolding dynamics are affected by minor groove binding.

Two of these dsDNA sequences are *"A-tract"* sequence ds(5'-GCAAATTTCC-3') (**mAA**) and *"alternating"* sequence ds(5'-GCATATATCC-3') (**mAT**). Both structures contain the same G-C base pairs at the ends of the strand and a minor groove purely consisting of A-T base pairs. These two will be referred to as *mixed sequences* throughout the text. G-C and A-T base pairs have characteristic vibrational modes that allow for independent observation of melting dynamics at the ends of the strands and in the centre. Recent T-Jump experiments on free dsDNA have shown[20] that sequences with G-C base pairs at the end do not undergo *end-fraying* upon melting but dissociate together with A-T base pairs in the centre as one process. The A-tract sequence **mAA** contains an optimal minor groove configuration for Hoechst 33258 binding in contrast to alternating sequence **mAT** that contains a suboptimal minor groove for this ligand[8,11]. Both were previously investigated in chapter 6, where the preference of **mAA** over **mAT** is evident from the 2D-IR response.

Two more dsDNA sequences were considered in these experiments. Sequences ds(5'-TTAAATTTTT-3') (**pAA**) and ds(5'-TTATATATTT-3') (**pAT**) were used as reference samples that contain similar minor grooves to the aforementioned sequences but with a lower melting point. The intention for these oligomers was to evaluate, whether lower steady state temperatures will affect the observed melting dynamics and if the observations are reproducible. They will be referred to as *pure A/T* sequences throughout the text.

8.3 Methods

8.3.1 Sample Preparation and FT-IR melting

Oligomers were purchased as salt free, lyophilised solids from Eurogentec, while all other chemicals were purchased from Sigma-Aldrich and used without further purification. In analogy to previous chapters, a stock solution of 40 mM was prepared for each single stranded oligomer using a deuterated TRIS buffer (100 mM TRIS, 100 mM NaCl, pD 7.0). Complementary strands were mixed in equimolar ratios, diluted

and heated to 95°C for 5 min and left to anneal for about an hour. Solutions containing Hoechst 33258 were prepared by annealing the dsDNA in equimolar ratio with a 20 mM stock solution of Hoechst 33258 in TRIS buffer. The dsDNA concentration of all final solutions was 10 mM.

FT-IR melting experiments were performed on a Nicolet iS10 Spectrometer from Thermo Fisher Scientific at an optical resolution of 4 cm^{-1}. The sample temperature was adjusted by a temperature controlled liquid cell with resistive heating (the same as in chapter 6), fitted with two, 2 mm thick CaF_2 windows and a 6 μm PTFE spacer that defines the sample path length. The short path length was set deliberately to minimise non-uniform heat distribution as will be discussed further below. For each sample, a steady-state melting response was collected by acquiring FT-IR spectra at different cell temperatures from 20 °C in 2°C steps until the DNA absorption features indicated complete transformation to ssDNA.

8.3.2 Experimental Design of the T-Jump Experiment

Fig. 8.1 shows a schematic representation of the pulse sequence. First, a nanosecond T-Jump pulse (red) excites the sym. O-D stretch of the D_2O solvent in the sample at 2660 cm^{-1}. The excited solvent modes relax within 1-2 ps and dissipate the energy to other modes in the system[21]. This dissipation leads to a new thermal equilibrium in the solvent after about 5 ps (as reported for pure 10% HOD in D_2O)[21], which can be recognised as an elevated temperature (blue). This sub-ns rise in temperature will in turn perturb the local environment of the oscillators in the DNA. Prior to the arrival of the T-Jump pump pulse, the steady-state sample temperature is kept just below the melting point of the DNA so that the perturbation from the T-Jump pulse will trigger a melting process. This process is followed over time by a series of femtosecond, mid-infrared probe pulses (black) that probe the absorption features of the perturbed sample in the infrared region from 1400 cm^{-1} to 1800 cm^{-1} (following carbonyl, N-H bend and ring-modes of the nucleobases). With the repetition rate of the infrared probe laser (10 kHz) being 20 times faster than the T-Jump pulse (0.5 kHz), 20 individual spectra are acquired within a single pump cycle, covering timescales from 100 μs to 1.8 ms. Over the course of one pump cycle, the kinetic energy induced by the T-Jump pulse will dissipate out of the sample (blue trace vanishes over time), the DNA will start to refold again and the sample will return to its initial state. The rate of this re-equilibration is affected by a variety of parameters such as the volume of the perturbed

sample, the thermal conductivity of the cell windows, or the absorbance of the pumped solvent mode, and can therefore be tuned to some extent to the repetition rate of the T-Jump pulse. Finally, the last probe spectrum is used as a single *pump-off* measurement to convert the remaining 19 spectra into pump-probe measurements. By changing the temporal offset of the first probe pulse relative to the T-Jump pulse in subsequent cycles, nanosecond to millisecond timeframes can be covered.

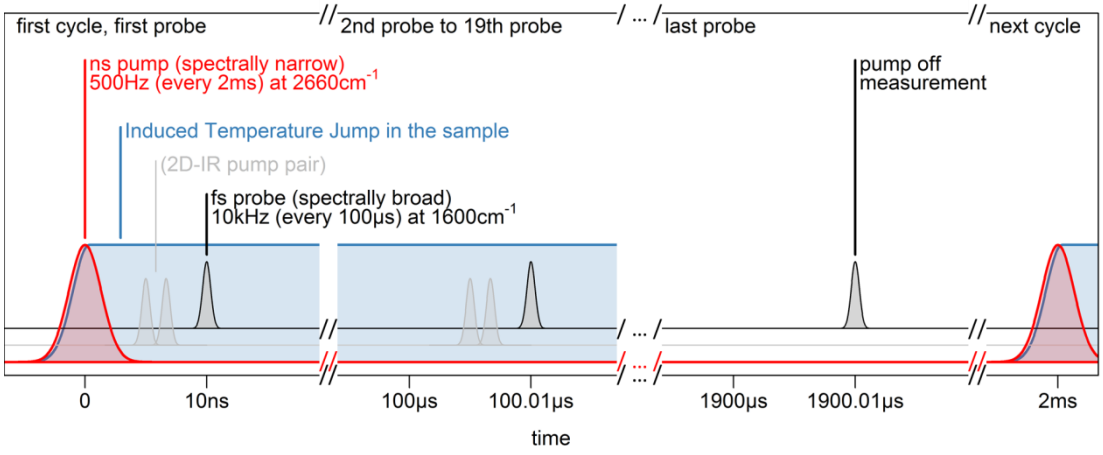


Fig. 8.1 Schematic representation of the pulse sequence used for the T-Jump experiment as they arrive at the sample. Note that the x-axis is not continuous. Additional pulses pairs in grey indicate the sequence for possible 2D-IR experiments but are disregarded for the current experiments. See below for further details.

The pulse generation and technical details of the instrumental setup are explained in more detail in chapter 5.4. The T-Jump pump has a pulse width of 4 ns, an energy of 35 µJ and operates at a repetition rate of 1 kHz, optically chopped to 500 Hz to obtain a time window of 2 ms for each measurement cycle. The centre wavelength was set to pump the high frequency side of the sym. D_2O stretch at 2660 cm^{-1}. At a sample thickness of 6 µm, this frequency shows an absorbance of 1.6 OD. These parameters were selected deliberately after sampling a selection of different path length, concentration and pump frequency values to obtain a T-Jump with the highest possible magnitude while ensuring complete equilibration before the next measurement cycle after 2 ms.

The femtosecond probe pulse has a pulse duration of 50 fs, is centred at 1600 cm^{-1} with a spectral width of ca. 300 cm^{-1} FWHM at a repetition rate of 10 kHz. The T-Jump pulse is triggered by q-switching and is synchronised to the probe pulse so that the temporal offset between pump and probe can be adjusted via delay of the electronic trigger signal. Probe pulses were dispersed by a grating and analysed by a HgCdTe array detector.

Each sample was measured at the same thickness and concentration as the respective FT-IR melting experiments (6 µm, 10mM dsDNA) and a single T-Jump response was obtained within 12 min (each electronic time delay was averaged for 5 s or 2500 pulse cycles). T-Jump measurements were recorded at different steady-state temperatures by heating up the sample with the temperature controlled, liquid cell from below the melting point in 5°C steps until a maximum T-Jump response was reached. Measurements for **mAA** and **mAT** were repeated at least twice with a freshly prepared sample to ensure reproducibility of the results. Sequences **pAA** and **pAT** were only measured once due to limited instrument access.

The experimental design presented in this chapter will set the foundation for future T-Jump 2D-IR experiments with fast repetition rates. A comparable T-Jump 2D-IR setup reported in literature operates at a repetition rate of 20 Hz (≙ 50 ms) to ensure enough time for complete equilibration of the 50 µm thick sample, leading to an acquisition time of 8 hours per sample[19,20,22]. The approach in the current experiment employs a faster cooldown of a thinner sample at the expense of a less uniform heat distribution (see trifluoroacetic acid response below). This combined with pulse shaping techniques to generate the 2D-IR response should allow for larger studies conducted in a shorter amount of time that focus on relative differences across samples rather than absolute changes within a single measurement, as often needed for drug discovery.

8.3.3 T-Jump Calibration via Trifluoroacetic Acid

In order to characterise the induced temperature jump in terms of magnitude and duration, the T-Jump response of an appropriate reference molecule was used. The reference needs to follow the immediate solvent response as closely as possible without undergoing any additional processes that convolute the spectrum. Trifluoroacetic acid (TFA) was used as a reference because of its small structure, direct interaction with the hydrogen bonds of the solvent and its simple FT-IR response in the carbonyl region as shown in **Fig. 8.2a**. The plot indicates FT-IR difference spectra of a solution of 130 mM TFA in D_2O from 20°C (reference) to 44°C from blue to red colours.

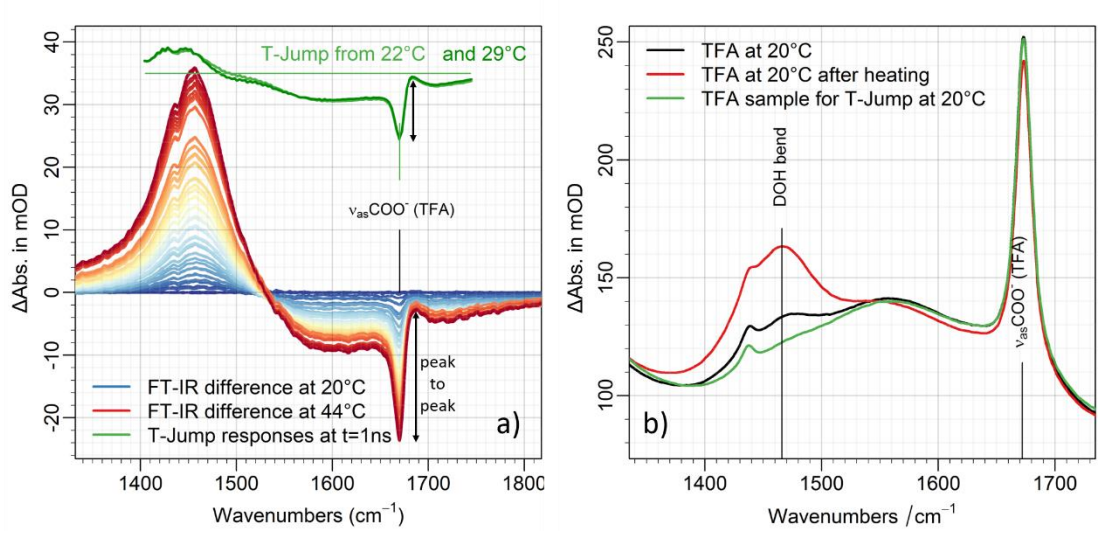


Fig. 8.2 a) FT-IR difference spectra of 130mM TFA in D_2O (from blue at 20°C to red at 44°C). Green spectra indicate the T-Jump responses of TFA at 1ns from 22°C and from 29°C (overlapped and almost identical). b) FT-IR spectra of TFA before a temperature ramp in black, after the temperature ramp in red and from a new sample that was used for the T-Jump experiments.

The asymmetric carboxylate stretch of TFA is positioned at 1672 cm^{-1} and shows a shift to higher frequencies upon heating (+16 cm^{-1}). Its change in absorbance from peak to peak as indicated in the plot can be recorded as a function of temperature to obtain a calibration for T-Jump measurements. A corresponding T-Jump response at a delay time of 1 ns in green compares very well with the observed difference spectrum from FT-IR measurements. Apart from the TFA response at 1672 cm^{-1}, **Fig. 8.2a** indicates another temperature-dependent mode at 1450 cm^{-1}. This is the DOH-bending mode of the solvent and was initially considered as an internal reference band, but subsequent heating and cooling uncovered that protonation from the atmosphere at elevated temperatures over time will increase its absorbance, making it unsuitable for accurate referencing. This is shown in **Fig. 8.2b**. Comparing the same sample at ambient temperature before and after heating to 44°C (black and red spectra in **Fig. 8.2b**) shows an increase of the DOH-bend, while the TFA signal and solvent background return to their initial values. To prevent unnecessary protonation over time, a fresh sample was prepared for the TFA T-Jump experiment and its FT-IR spectrum is shown in green in **Fig. 8.2b**. The position of the DOH bend is outside of the region of interest for the DNA experiments (1530 cm^{-1} to 1720 cm^{-1}) and will therefore not affect the analysis.

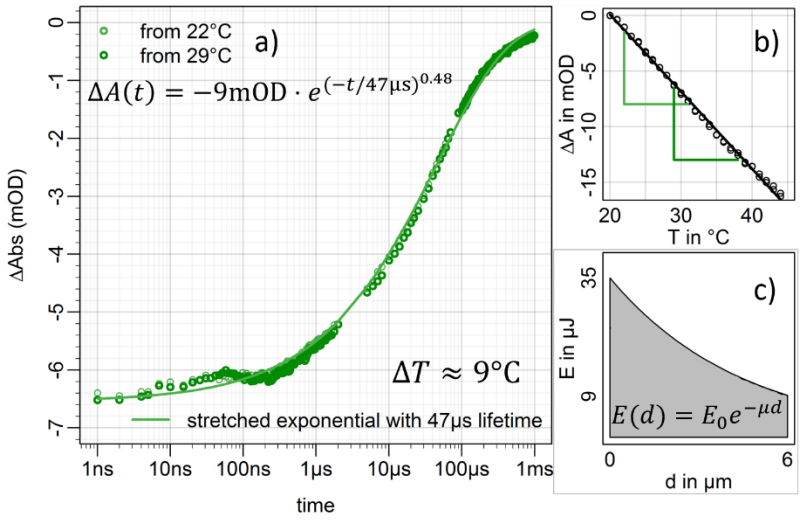


Fig. 8.3 a) T-Jump responses of the 1672 cm^{-1} TFA peak over time from 22°C and 29°C. Plotted on a log-scale for t. Fitting of a stretched exponential gives a lifetime of 50 µs. b) FT-IR change of TFA band from 20°C. Green steps indicate initial and final temperature of T-Jumps. c) T-Jump pulse energy across the sample according to the Lambert-Beer law. µ indicates the attenuation coefficient and was estimated from the given path length, d, energies, E(d=6µm) and E_0. Deviation from a linear profile indicates non-uniform absorption.

Two T-Jump measurements were recorded starting at steady state temperatures of 22°C and 29°C. Their spectral features at t = 1 ns (see green responses in **Fig. 8.2a**) are very similar to FT-IR difference spectra. Both show an increase of the DOH bend at 1450 cm^{-1} and a +14 cm^{-1} shift in frequency of the TFA band. The kinetic traces of this band at 1670 cm^{-1} are shown in **Fig. 8.3a** with essentially the same kinetics for both initial temperatures. The minimum of the TFA response at -9 mOD, equivalent to the maximum T-Jump, is reached immediately at 0 ns and then relaxes back to zero as a stretched exponential function with a lifetime of approx. 50 µs (β = 0.48). The remaining signal after 1 ms has reduced to ca. 3% of its value at t=0 ns. This change in absorbance should be directly proportional to the momentary change in temperature of the sample. The exponential relaxation over time is likely to be stretched due to inhomogeneous heating across the path length. Measurements of the pump energy right before and after the sample indicated a reduction in pulse energy from 35 µJ to 9 µJ. The loss of pulse energy due to absorption across the sample was estimated according to the Lambert-Beer-Law and is illustrated in **Fig. 8.3c**. A non-linear absorption profile will lead to a larger temperature jump at the front of the sample than at the back. The profile in **Fig. 8.3c** is still fairly linear so that an estimate of the overall T-Jump magnitude via absorbance measurements seems reasonable. FT-IR melting experiments indicate that the TFA band is changing its absorbance by -0.71±0.1 mOD/°C (linear fit shown in **Fig. 8.3b**), which relates the signals of both T-

Jump experiments to an increase in temperatures by about 9°C (from 22°C to 31°C and from 29°C to 38°C as shown in b). This value is an estimate of the overall temperature change in the sample and will be depending on the actual penetration depth of the pump pulse.

The time points in the T-Jump data for TFA stop after 1 ms, even though the pump Laser was optically chopped to 500 Hz. This is because the last 10 of the 20 probe pulses in a measurement cycle were taken as *pump-off* measurements in order to generate pump-probe spectra. This was only done for the TFA calibration data and means that every *pump-on* measurement is referenced by an individual *pump-off* measurement taken exactly 1 ms later. This method was changed when moving to experiments on DNA in order to obtain time points beyond 1 ms. For the remaining data shown in the results section, only the last one of the 20 probe pulses in a cycle was used as a pump-off reference and the first 19 were used as data points. This method should lead to slightly higher noise levels as less reference measurements are taken but should return more accurate time traces if residual dynamics still evolve within the 2nd millisecond. The total relaxation time for the sample is 2 ms in both cases; just the data treatment is different.

8.3.4 Data Pre-Processing

To improve the signal-to-noise ratio, all collected T-Jump responses were pre-processed in four steps. In the first step, shot-to-shot noise from laser intensity fluctuations was reduced by collecting a reference measurement of the probe beam before it was interacting with the sample. A fraction of the probe pulse (ca. 10%) was separated from the beam and analysed in a separate spectrometer. This reference signal was then subtracted from the data.

In the next step, electronic noise was reduced from the dataset. It was noticed that the controller unit of the nanosecond pump laser introduced electronic noise on the MCT detector that acquired the probe signal when the laser was operating. A blank measurement was taken by removing the liquid cell completely from the beam path and showed that the electronic noise was in fact systematic and followed the 1kHz repetition rate of the nanosecond laser. This was obvious from the data as the pump pulse was optically chopped to 500 Hz and the noise repeated its pattern when comparing fluctuations between time points taken in the first milisecond with time points from the second milisecond (see **Fig. 8.4a**). This noise pattern was constant, as

long as exactly the same time delays were measured. All T-Jump responses were therefore taken at the same time delays and in these cases, the blank measurement could be used to subtract and suppress the artefacts irrespective of the actual sample or steady-state temperature. This is indicated in **Fig. 8.4b**, where different samples at different cell temperatures show a similar noise pattern.

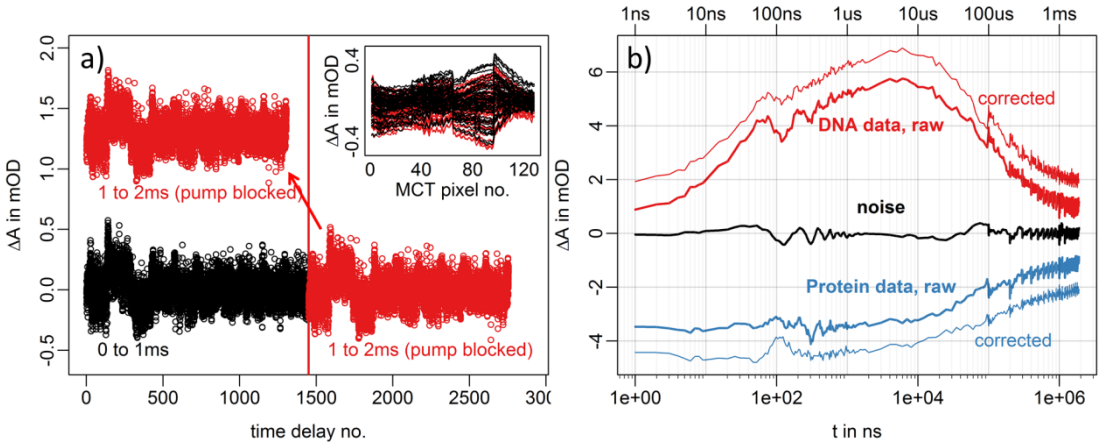


Fig. 8.4 a) Blank measurement of one MCT pixel shows electronic noise pattern that repeats itself after 1 ms (red colours). The x-axis shows data-points as if they were taken with equidistant spacing in time for clarity. Inset: Electronic noise causes step-like artefacts every 32 pixels. b) Exemplary time traces of the blank measurement (black), a DNA sample at 30°C (red) and a protein sample at 40°C (blue) cell temperature. All traces are taken from the same MCT pixel representing 1648 cm^{-1}. The thin lines indicate the same data after the noise trace was subtracted.

The next step was a reduction in data size. The pulse sequence used in these experiments resulted in a large number of data points above 100 µs, since a single measurement on nanosecond timescales would always return 19 other time delays as well, each being 100 µs apart from another (see **Fig. 8.1**). A measurement of 1 ns and 2 ns data would return 0.001 µs, 100.001 µs, 200.001 µs, …, in the first cycle and cover 0.002 µs, 100.002 µs, 200.002 µs and so on in the next. This does not only increase the total number of data points without gaining significant more information, but can lead to a bias in statistical models that favors microsecond timescales over nanosecond data. Time delays with the same three significant figures were therefore averaged to reduce the total number of timestamps to below 500.

Finally, a T-Jump response for the buffer solution was acquired. The spectral response of the buffer was then initially used as a baseline that was scaled and subtracted from the spectral T-Jump response in the DNA measurements. This method however was later on replaced by a different method as discussed in the results chapter to separate melting dynamics from a generic response to heat of the sample.

8.4 Results and Discussion

8.4.1 FT-IR Measurements

FT-IR absorption spectra were acquired for all four dsDNA samples while gradually heating the sample cell from room temperature in 2°C steps to observe the spectral impact of melting. The vibrational modes observed in these steady-state measurements should also be affected in time-resolved measurements if melting is observed.

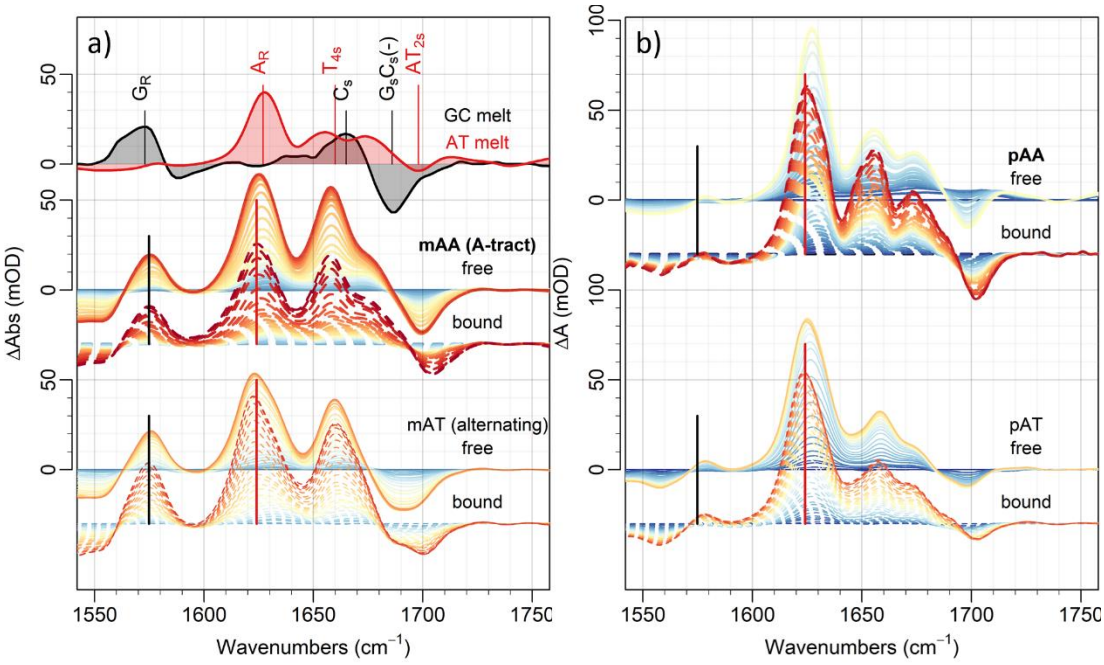


Fig. 8.5 a) Black and red traces at the top are reproduced from chapter 4.3.4 and show band assignments for FT-IR difference spectra when melting a pure G/C sequence (black) and a pure A/T sequence (red). Spectra below are FT-IR difference spectra for free (solid lines) and bound (dashed lines) sequences mAA (GCAAATTTCC) and mAT (GCATATATCC), subtracted from the corresponding absolute spectrum at 30°C. b) FT-IR difference spectra for pure A/T sequences pAA (TTAAATTTTT) and pAT (TTATATATTT), subtracted from the corresponding absolute spectrum at 20°C. Blue colours indicate low temperatures; red indicates high temperatures (up to 100°C). Black and red vertical lines indicate G/C and A/T bands respectively.

The FT-IR data is shown for mixed sequences **mAA** and **mAT** in **Fig. 8.5a** as difference spectra from the corresponding absolute spectrum at 30°C. The difference spectra for pure A/T sequences **pAA** and **pAT** are shown in **Fig. 8.5b** as the change from 20°C due to their lower melting points. The black and red traces at the top of **Fig. 8.5a** are used as guidance for band assignment and represent difference spectra of dsDNA from a pure A/T sequence and a pure G/C sequence, which were previously discussed in the introduction chapter 4.3.4. The DNA melting process typically leads to a reduction of highly coupled bands of dsDNA such as the negative $G_sC_s(-)$ feature at 1686 cm^{-1} in the G/C melting spectrum in black, together with an increase of ssDNA bands (rise of C_s at

1665 cm^{-1}, black spectrum). Spectral features above 1640 cm^{-1} tend to arise from carbonyl stretch vibrations of G, C and T of dsDNA as well as ssDNA, leading to convoluted peak shapes that are not straightforward to assign in difference spectra of mixed sequences. Previous chapter 6 identified a shift of T_{2s} in the duplex to above 1700 cm^{-1} when successfully binding Hoechst 33258 to an A/T-rich minor groove[11,23]. This seems to correspond to the presence of a negative feature above 1700 cm^{-1} in the difference spectra when adding Hoechst 33258 to all four sequences (bound cases, dashed lines), and gives a first indication of Hoechst 33258 interacting with the dsDNA. As discussed in 6, infrared absorption of Hoechst 33258 is insignificant at 10 mM concentration[11,23].

The black and red traces at the top of **Fig. 8.5a** show two spectral features below 1640 cm^{-1} with very little cross-interference and can be clearly assigned to A/T melting at 1623 cm^{-1} (red trace) and G/C melting at 1575 cm^{-1} (black trace). The peak at 1623 cm^{-1} can be assigned to an A ring-mode[19,24–27] that is suppressed in dsDNA due to base stacking interactions along the helix[26]. The rise of this mode is present in difference spectra of all four sequences shown in **Fig. 8.5** and will be used as a marker band for the stability of A/T structures such as the minor groove in the following T-Jump experiments (A/T is colour coded with red in the following figures). The peak at 1575 cm^{-1} can be assigned to a G ring-mode[24,25,28] which is suppressed in dsDNA[20,24] and is absent in difference spectra of pure A/T sequences **pAA** and **pAT** (**Fig. 8.5b**). This band will be used as a marker to track unfolding of G/C structures such as the outer ends of mixed sequences **mAA** and **mAT** (G/C is colour coded with black in the following figures).

The change in absorbance of the G and A ring-modes at 1575 cm^{-1} and 1623 cm^{-1} can be plotted against temperature (see **Fig. 8.6**) to obtain sigmoidal functions that represent the melting transition for G/C (black) and A/T structures (red) respectively. In order to emphasize the melting transition, a linear function was fitted to the first three data points as shown in the inset of **Fig. 8.6b** and subtracted from the data[19]. The linear trend observed in these experiments possibly stems from thermal expansion of the sample throughout the experiment, as it is generally observed for all spectral features. After removal of the linear component, the intensity change was normalized to one for a better comparison between datasets and based on the assumption that the DNA has fully melted during the experiment. This leads to a representation of the melting curves in **Fig. 8.6,** where the inflection point of the sigmoid is positioned at a value of 0.5,

indicating that 50% of the dsDNA has melted at this temperature. This will be defined as the melting point of the G/C and A/T structures in the following.

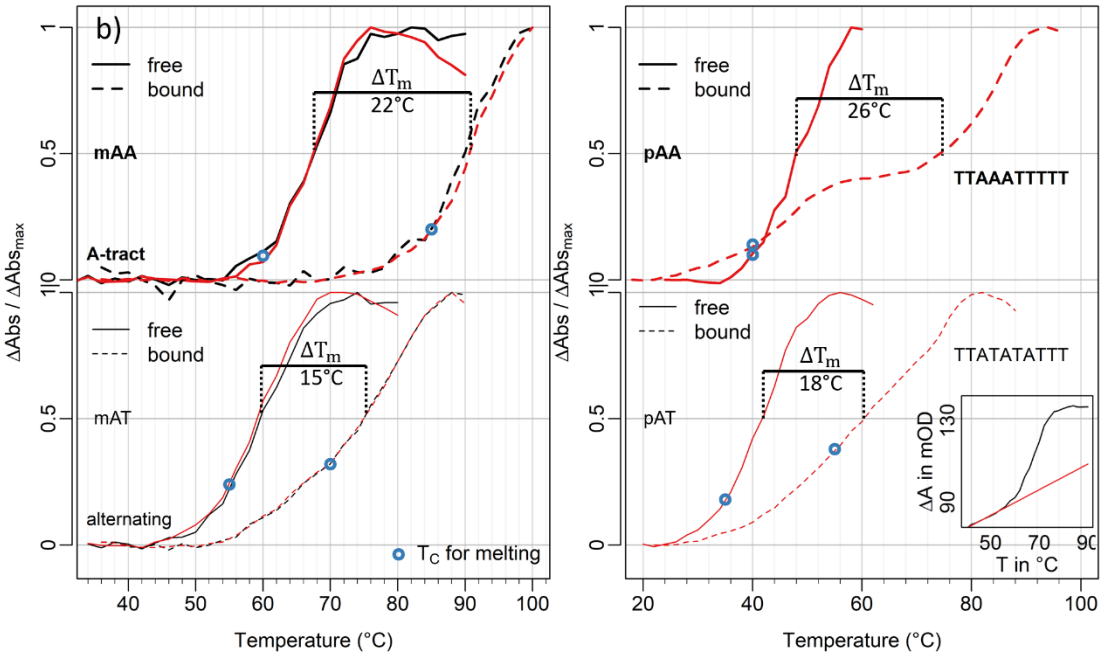


Fig. 8.6 Temperature dependent, sigmoidal functions obtained from FT-IR difference spectra at 1575 cm^{-1} (G ring-mode, black) and 1623 cm^{-1} (A ring-mode, red). a) shows mixed sequences and b) shows pure A/T sequences. A linear fit was applied to low temperature absorbance values and was subtracted from the data to emphasise the inflection point of the sigmoidal transition, as shown in the inset of b). The data was normalised to one for better visual comparison. Plots are therefore illustrative.

The melting curves observed from A and G ring modes in **Fig. 8.6a** reach their melting point at the same temperature in each experiment (black and red traces overlap at 0.5), indicating that the A/T minor groove and G/C ends in mixed sequences **mAA** and **mAT** always unfold to the same degree under equilibrium conditions. This is the case whether Hoechst 33258 is present (dashed lines), or not (solid lines). Minor groove binding of Hoechst 33258 increases the melting point for **mAA** from 68°C to 90°C and for **mAT** from 60°C to 75°C. The rise in melting point signifies a stabilisation of the double helical structure due to minor groove binding and will be defined as ΔT_m. Values for **mAA** ($\Delta T_m = +22$°C) and **mAT** ($\Delta T_m = +15$°C) highlight the ligand's preference for A-tract structures and are in good agreement to previous findings (+24°C and +16°C respectively)[11]. Minor groove binding of Hoechst 33258 is found to be mostly entropically driven[29], by replacement of a well-structured spine of hydration[30] together with subtle rearrangement of the dsDNA structure as an induced fit[11,12]. It was further shown in literature that the N-2 amino group of guanine sterically prevents Hoechst 33258 from binding as it obstructs the minor groove[31]. Pure A/T sequences **pAA** and

pAT will thus provide a less restrictive minor groove and Hoechst 33258 may bind in unexpected configurations.

Sequences **pAA** and **pAT** without a ligand (solid lines in **Fig. 8.6b**) melt at 48°C and 42°C respectively. This is about 20°C earlier than **mAA** (68°C) and **mAT** (60°C) due to their lack of G-C base pairs. Sequences with an alternating minor groove (**mAT**, **pAT**) show a slightly lower melting point than the same sequences with an a A-tract (**mAA**, **pAA**), as would be expected when looking at melting point predictions that incorporate base step parameters (dsDNA with TA steps are stabilised by a smaller free energy than dsDNA with AA steps)[32]. Binding to **pAT** (bottom of **Fig. 8.6b**) shows a large ΔT_m of +18°C with a more gradual melting curve than in the free DNA. Binding to **pAA** results in a large nominal melting point stabilisation of $\Delta T_m = +26$°C, but shows a very broad sigmoidal rise with two inflection points. This indicates a poor selectivity which could result in an ensemble of different interactions, possibly with a significant amount of free dsDNA. 5'-TTAAATTTTT-3' and its complementary sequence that are used in **pAA** have a significant portion of partial overlap, which is more prone to dsDNA mismatching during the annealing process and could explain the unexpected melting curve to some extent. The unusual melting behaviour will be discussed further below in more detail.

T-Jump experiments were performed while keeping the steady-state temperature of the sample cell at a defined value. Changing the cell temperature moves the starting point of the T-Jump experiment along the melting curve and changes the processes that are being triggered. A T-Jump of 9°C may return a generic heating response of double stranded DNA at 20°C, will trigger unfolding processes close to the melting point and returns a generic response of single stranded DNA well above the melting point. The T-Jump response of each sample was therefore acquired at a range of different cell temperatures. The blue circles in **Fig. 8.6** indicate the cell temperature for each sample, at which the magnitude of the dsDNA response was maximised in the T-Jump experiments.

8.4.2 Sub-nanosecond DNA response follows solvent relaxation

Fig. 8.7 represents a summary over several T-Jump measurements at different starting temperatures for free dsDNA **mAT** (5'-GCATATATCC-3') without ligand. The top of **Fig. 8.7a** shows spectral features obtained at a cell temperature of 25°C, well below the melting point of **mAT** ($T_m = 60$°C) and will therefore predominantly probe dsDNA

without triggering significant unfolding processes. As the T-Jump pulse perturbs the equilibrium of the sample at t=0 ns (purple spectrum), the vibrational modes of the DNA respond immediately (on sub-nanosecond timescales) with a set of spectral features that slightly rises further over the next microsecond (green) and relaxes back to zero within 1 ms (dark red).

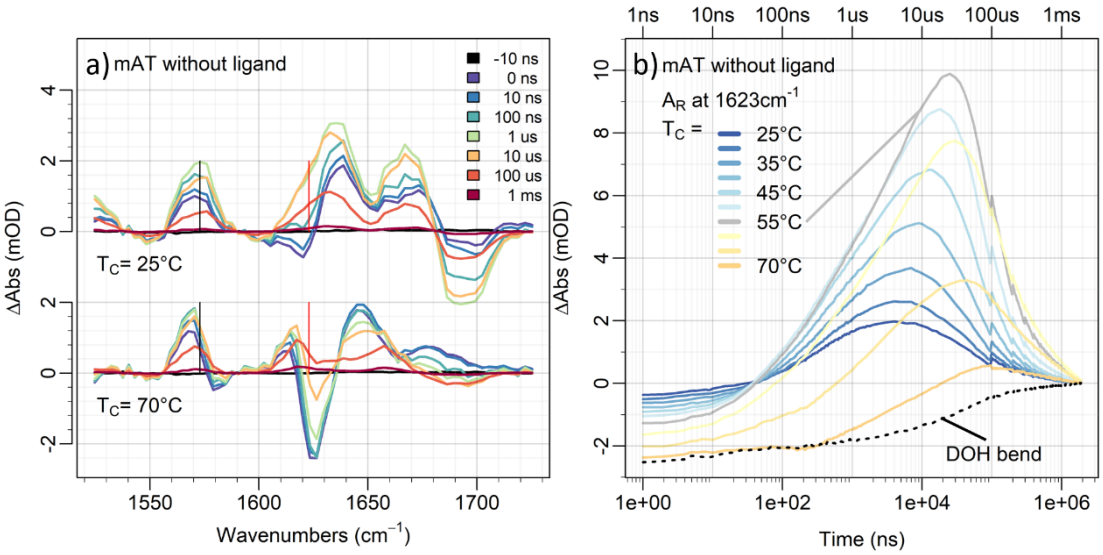


Fig. 8.7 a) Spectral T-Jump response of free mAT (GCATATATCC) at a cell temperature of 20°C (top) and 70°C (bottom). b) Temporal evolution of the T-Jump response at 1623 cm-1 with respect to different cell temperatures from 20°C (blue) to 70°C (yellow), taken in 5°C steps. Time axis is on a log scale. A maximum rise is observed at $\Delta\mathbf{A_{max}}(\mathbf{t = 30\mu s}, \mathbf{T_c = 55°C}) = \mathbf{10mOD}$ *(grey trace).*

This response changes significantly when the same T-Jump is performed at a cell temperature of 70°C (lower dataset), at the end of the melting transition, where most of the DNA is present as single strands. The T-Jump signal at 70°C starts again with an instantaneous response of at t=0 ns (purple) that immediately starts to relax back to zero over the course of the measurement. It seems that the initial response at t=0 ns (purple) is particularly sensitive to the starting temperature. These observations suggest an initial process on sub-nanosecond timescales whose spectral response is affected by the concentration of dsDNA to ssDNA structures in the sample. As both T-Jump responses in **Fig. 8.7a** evolve over time, their spectral features appear to align and start to show more similarities with each other. Comparing the responses at 100 µs (red spectra) shows two rises at 1575 cm-1 and around 1623 cm-1 (G and A ring-modes, black and red vertical lines), together with a loss of a mode at 1700 cm-1, bearing close resemblance to the FT-IR difference spectra when melting **mAT** in **Fig. 8.5a**. This is indicative of a second process on microsecond timescales that resembles the melting process previously observed with FT-IR absorption experiments. A T-Jump experiment

performed at a cell temperature just below the melting point should facilitate unfolding and therefore maximise the spectral features on microsecond timescales.

Fig. 8.7b shows the temporal response of the observed T-Jump signals at 1623 cm^{-1} (A ring-mode) for free **mAT**, at different cell temperatures from 25°C to 70°C (blue to orange). The A ring-mode shows a distinctive rise on the microsecond timeframe that is maximised at a cell temperature of 55°C (grey trace), 5°C below the melting point of **mAT**. This confirms that the T-Jump pulse triggers DNA unfolding processes within microsecond timescales and also confirms that the estimated T-Jump magnitude of 9°C obtained from TFA measurements is reasonable. The rise in T-Jump signal is followed by a relaxation at 100 µs which indicates refolding of the DNA, governed by the thermal energy dissipation out of the sample. Accurate quantification of the rise dynamics due to melting is however obstructed due the additional relaxation process at early time delays, which was identified earlier. **Fig. 8.7b** shows that an increase in cell temperature leads to an initial drop in signal at t=1 ns. This initial process needs to be characterised first, before moving on to analysing the response due to melting.

Fig. 8.7b shows that the negative signal remains well beyond 10 ns, ruling out an instrumental response due to overlap of pump and probe pulse at t=0 ns. It is far more likely that the induced rise in solvent temperature will affect the local environment surrounding the vibrational modes of the DNA. Increased thermal motion and a lower density of D_2O molecules at elevated temperatures will lead to increased hydrogen bond distances in the solvent network that solvates the DNA. This can affect vibrational modes of nucleobases in unfolded ssDNA in particular, as their exposure to the solvent network is greater than in the double helix. Reorganisation of the solvation shell typically evolves within a few picoseconds, which is instantaneous on the timescale of a nanosecond T-Jump pulse[21,33–35]. This would imply that the T-Jump response at t=0 ns is indeed dependent on the concentration of dsDNA and ssDNA and will relax together with the re-equilibration of the solvent temperature.

The first statement of this sentence can be evaluated by comparing the T-Jump response at 0 ns from different starting temperatures with FT-IR difference spectra well below and above the melting point. In these two cases, the spectral change in FT-IR experiments due to a melting process might still be present but is significantly smaller than the response to heating. **Fig. 8.8a** at the top shows FT-IR difference spectra when free **mAT** is heated by 4°C from 32°C (blue) and from 86°C (red) respectively. This is

compared to the DNA T-Jump response (same figure just below) at 0 ns at different starting temperatures from 25°C to 70°C. The low and high temperature T-Jump responses show indeed a similarity to the FT-IR difference spectra of what is likely to be the response of dsDNA to heating and ssDNA to heating. A linear combination has shown that all intermediate responses at 0 ns can be modelled by the two responses at 20°C (blue trace in the centre) and 70°C (orange trace in the centre) with few residuals (grey traces in **Fig. 8.8a**). This indicates that the immediate T-Jump DNA response is changing with dsDNA/ssDNA concentration.

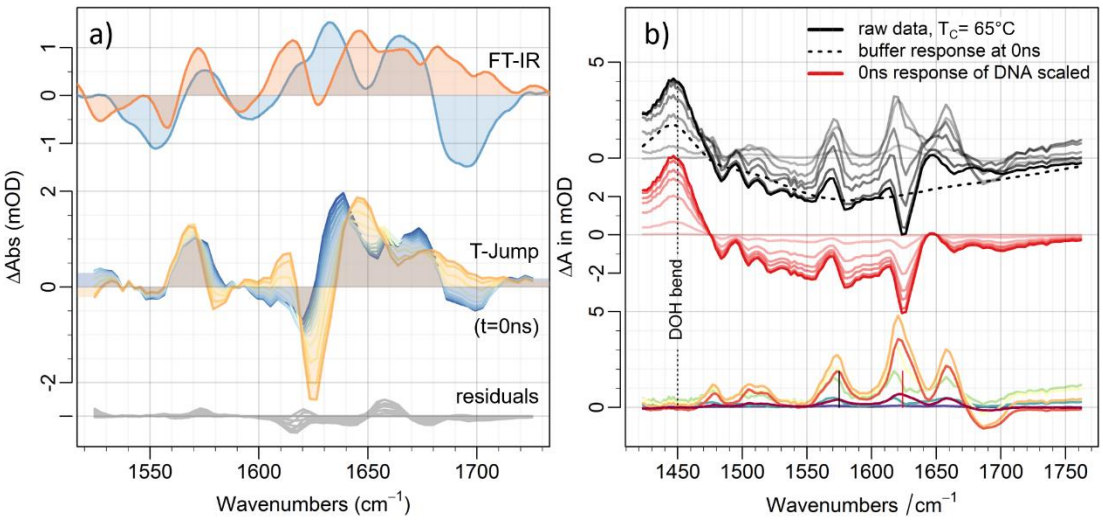


Fig. 8.8 All data taken from free mAT. a) Top: FT-IR difference spectra for a 4°C rise in steady state temperature from 32°C (blue) and 86°C (red) respectively. Centre: T-Jump responses at t=0 ns from cell temperatures between 25°C and 70°C. Colours indicate temperatures. Bottom: Residuals from a linear combination of 0 ns responses at 25°C and 70°C as mentioned in the text. b) Black: T-Jump dataset for mAT at 65°C cell temperature before baseline correction with a buffer response (dashed line shows 0 ns). Red: 0 ns response from the raw data, scaled with respect to the intensity of the DOH bend at each time delay. Bottom: Corrected T-Jump spectra with initial response removed. Actual intensities of black and red traces are twice as high as depicted in the figure.

If the re-equilibration of the 0 ns DNA response follows the solvent relaxation as stated earlier, then both processes can be described by the same temporal profile. **Fig. 8.8b** shows the T-Jump dataset of free **mAT** at a cell temperature of 65°C in black, but without prior baseline correction of the buffer response (dashed line). The DOH bend visible at 1450 cm^{-1} in the raw data was deemed unsuitable for absolute measurements earlier but can now be used for a relative representation of the momentary temperature of the solvent. By forming the outer product of the temporal response of the solvent at 1450 cm^{-1} with the spectral response at 0 ns of the DNA sample, it is possible to obtain a matrix with the same dimensions as the raw data (red dataset in **Fig. 8.8b**). All spectral features described in this dataset follow the same relaxation of

the solvent as obtained from the signal at 1450 cm^{-1}. Subtraction of these features from the raw data (black minus red spectra) leads to a new dataset (bottom spectra) that is not only free of the spectrally broad solvent response, but also contains no initial DNA response at 0 ns (purple trace). This corrected T-Jump dataset in **Fig. 8.8b** shows spectral features at late time delays (orange trace, rise at 1575, 1623, 1658 cm^{-1}, loss at 1700 cm^{-1}) that are very similar to the melting responses obtained from FT-IR data. This approach shows that the initial response from the DNA can be separated from subsequent processes which will allow a more conclusive analysis of the melting kinetics. It confirms that the initial DNA response is correlated to the solvent relaxation and supports the argument that this is a direct consequence of the solvent network reorganisation due to heating.

This approach was used in all following sections to separate the initial heating response from the data and to obtain T-Jump responses that can be discussed with regards to the DNA unfolding process. This method is sensitive to the composition of the sample, as the response at 0 ns will change depending on the amount of dsDNA and ssDNA. Any spectral changes on sub-nanosecond timescales due to heating, such as possible thermal expansion of the sample should also be removed with this method.

8.4.3 T-Jump DNA response due to unfolding

The spectral features of T-Jump experiments on mixed sequences **mAA** and **mAT** are shown in **Fig. 8.9** and compare the response of the free DNA at the top with data for the Hoechst 33258 complex at the bottom. The cell temperatures, T_c, chosen in this figure are set to 5°C below the corresponding melting point, T_m (as observed from FT-IR data), for a given sample, so that the 9°C T-Jump pulse is able to probe the main transition of the melting curve and the T-Jump response is maximised. Sequence **mAA** (A-tract) without the ligand in the top of **Fig. 8.9a** was measured at a cell temperature of $T_c = 60°C$ $(T_m - 8°C)$ and shows a gradual rise of both, G and A ring-modes (black and red vertical lines) reaching their maxima within tens of microseconds and dropping back to zero at 1-2 ms (dark red spectrum), indicating unfolding and refolding of the DNA. Sequence **mAT** without the ligand (**Fig. 8.9b**, top) shows a response very similar to **mAA** with a microsecond rise and subsequent decay of both ring-modes. The spectral features observed for **mAT** with this method are also in line with FT-IR difference spectra for **mAT** in **Fig. 8.5**.

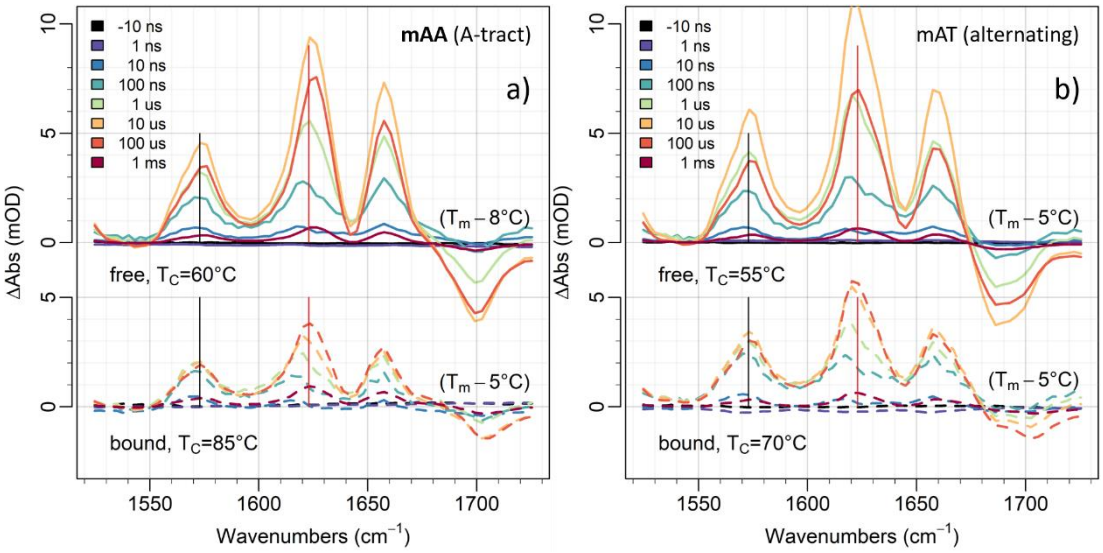


Fig. 8.9 T-Jump response of mAA in a) and mAT in b) in absence (top) and presence (bottom) of Hoechst 33258, probing the centre of the corresponding melting transition. The data was baseline corrected by subtracting the outer product of the temporal response at 1450 cm^{-1} (DOH bend) with the spectrum obtained at t = 0. Colours indicate time.

Adding the minor groove binder to these sequences stabilises the double helical structure and raises the DNA melting point just like in FT-IR experiments so that the cell temperature had to be increased accordingly (+25°C for **mAA** and +15°C for **mAT**, spectra at the bottom). The spectral features of both T-Jump responses are qualitatively comparable to their FT-IR counterparts with the only visible, characteristic difference

upon binding being the shift of the T_{2s} band around 1700 cm^{-1} to higher frequencies. To gain a better understanding of the melting process shown in **Fig. 8.9**, the response of the G ring-mode (black) and A ring-mode (red) was plotted against time in **Fig. 8.10** for both, mixed sequences.

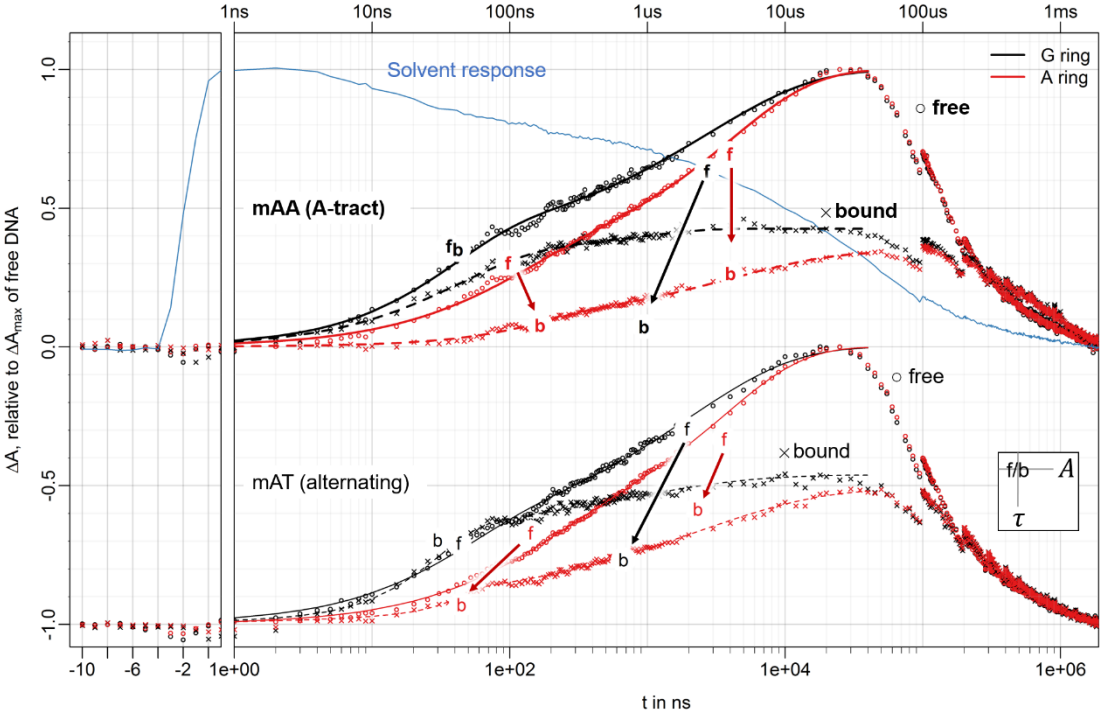


Fig. 8.10 Temporal traces of the G ring-modes in black and A ring-modes in red for mAA (GCAAATTTCC) at the top and mAT (GCATATATCC) at the bottom. All traces from free DNA were normalised to one and the responses from bound DNA were scaled accordingly to maintain the relative ratio between free and bound DNA. The blue trace indicates the temporal evolution of the DOH bend, which is directly proportional to the momentary temperature of the system. Each trace was fitted to a sum of two, stretched, exponential functions as outlined in equation (8.1). The fitted parameters of each single exponential function (τ, A) is illustrated as the (x,y)-location of points "f" and "b" throughout the plot. "f" indicates results from free DNA, "b" indicates results from DNA in presence of Hoechst 333258. Colours indicate association to G (black) and A (red) ring-modes.

Points in **Fig. 8.10** show temporal traces for the A-tract sequence **mAA** at the top and alternating sequence **mAT** at the bottom. For a better comparison between ring-modes and between sequences, traces obtained from free DNA (circles in **Fig. 8.10**) were normalised to one. In order to keep the relative ratio between data from free to bound cases the same, the signals in the corresponding bound cases were divided by the same factor as was used for the normalisation for the free cases. To quantify rise times, a sum of two exponential functions was fitted to the rising part of each temporal trace (lines represent fitted responses from 0 ns to 40 µs) and is shown in equation (8.1):

$$\Delta A(t) = -A_1 e^{\left(-\frac{t}{\tau_1}\right)^{\beta_1}} - A_2 e^{\left(-\frac{t}{\tau_2}\right)^{\beta_2}} + A_1 + A_2. \qquad (8.1)$$

Due to the normalisation, amplitudes A_1 and A_2 will sum up to one for traces from free DNA. Both exponentials had to include a stretching exponent β_1, β_2 in order to obtain a converging fit. This indicates that the kinetics observed in these traces arise from a heterogeneous ensemble of vibrational modes that all rise at slightly different times around τ_1 and τ_2. This can be expected if the T-Jump perturbation induced non-uniform heating across the sample, as was already shown earlier to be the case. The results of these fits are summarised in **Table 1** and parameters of each individual exponential function are illustrated as a point ("f" – free DNA, "b" – bound DNA) in **Fig. 8.10**. Its position along *x* indicates rise time and the position along *y* indicates its amplitude.

	A_1		A_2		τ_1 in ns		τ_2 in µs		β_1		β_2	
	G	**A**	**G**	**A**	**G**	**A**	**G**	**A**	**G**	**A**	**G**	**A**
Free mAA	0.4	0.3	0.6	0.7	36	97	2.7	4.0	0.6	0.7	0.6	0.7
Bound mAA	0.4	0.1	0.1	0.3	42	170	0.9	4.0	1.0	0.6	1.0	0.6
Free mAT	0.3	0.3	0.7	0.7	43	140	1.9	3.6	0.6	0.7	0.6	0.7
Bound mAT	0.3	0.1	0.2	0.4	30	45	0.7	2.3	0.5	0.5	0.5	0.5

Table 1 Parameters from bi-exponential fitting shown in Fig. 8.10. Position of "f" and "b" symbols indicate (τ, A) results of each exponential. **β***-values closer to one indicate are narrower distribution of rise times around* **τ***.*

For free **mAA** without the ligand (circles at the top of **Fig. 8.10**), G and A ring-modes both reach their peak amplitude of one at the same time at 30 µs. This is also the case for free **mAT**, where G and A reach their peak amplitude at 20 µs (circles in the lower half). This signifies overall simultaneous melting of G-C and A-T base pairs rather than an *end-fraying* process for ligand-free DNA. These observations confirm results from a similar T-Jump experiment on self-complementary sequence 5'-GATATATATC-3' performed by Sanstead *et al.*, where two distinctive rise times and overall synchronous melting of G-C and A-T base pairs was observed[19,20].

At early rise times however, the G ring-mode shows a slightly sharper rise at early timescales ($\tau_1 \approx 40$ ns, black "f") than A ($\tau_1 \approx 120$ ns, red "f") in both sequences. This is unexpected, as G-C base pairs generally require more energy for dissociation due to their additional hydrogen bond over A-T base pairs. The G-C base pairs in mixed sequences **mAA** and **mAT** are located at the ends of the double helix and will therefore have a bigger exposure to the solvent, possibly leading to a slightly faster, initial melting response compared to A-T base pairs in the centre. A bi-exponential rise would suggest nanosecond-fraying of G-C base pairs and was also observed in the

experiments of Sanstead *et al.*, but it was not mentioned whether the nanosecond-rise of G was observed earlier than the nanosecond-rise of A. The ring-modes observed in **Fig. 8.10** are particularly sensitive to base stacking interactions along the strands and do not represent a direct measurement of Watson-Crick base pair distances. The nanosecond response may therefore characterise a destabilisation of the double helix before the main melting process occurs during τ_2.

The temporal traces for DNA in presence of Hoechst 33258 are shown as crosses in **Fig. 8.10** and were fitted using equation (8.1) as before (dashed lines). The data is taken at cell temperatures just below the now stabilised melting points, as before. The change in the fitting results from "free" to "bound" DNA is indicated as arrows in the plot. The most striking difference is a significant drop in amplitudes for almost all of the fitted exponentials upon binding, illustrated as downward pointing arrows in **Fig. 8.10**. The A ring-mode in **mAA** shows a drop in amplitude A_1 upon binding from 0.3 to 0.1 (see *Table* 1) for the first rise function on nanosecond timescales. The first rise of G in **mAA** on the other hand is able to maintain its amplitude upon binding (highlighted in *Table* 1). The same observations can also be made for the nanosecond rises in **mAT** and are reflected in the shape of the traces shown in **Fig. 8.10**: Within the first 100 ns, Hoechst 33258 binding results in a smaller response for A (red), but has little effect on the response of G (black). Structurally, it suggests that the G-C ends in the DNA-Ligand complex evolve during the melting process initially unperturbed from the minor groove binder. Melting of the A-T centre however is being affected right from the start of the measurement at 0 ns.

As time progresses towards microseconds, the second, exponential rise function continues to increase the intensity of both G and A ring-modes. In the free DNA, the second rise has the largest contribution to the overall T-Jump response with A_2 values between 0.6 and 0.7 ($A_1 + A_2 = 1$). In the bound case however, A_2 values drop significantly for G as well as A ring-modes so that the overall rise in amplitude $A_1 + A_2$ is below one. The A ring-mode in the bound case of **mAA** (red crosses in the top of **Fig. 8.10**) only reaches about 35% of the peak amplitude compared to the free DNA. The A ring-mode for the bound case of **mAT** (red crosses, bottom) rises slightly higher, up to 50% of the peak amplitude of free **mAT**. Observation of a smaller T-Jump response would imply that a smaller number of dsDNA strands is dissociating at the melting point. This can be evaluated by comparison to the results from FT-IR experiments. In the FT-IR difference spectra in **Fig. 8.5**, the rise of the A ring-mode shows a gradient of

3 mOD/°C for free **mAA** and 2.5 mOD/°C for free **mAT** at the corresponding melting point. When adding the minor groove binder to **mAA** and **mAT**, the observed gradients at the melting points are only slightly smaller, reaching 80% of the gradient in the free DNA cases. A slightly smaller gradient suggests that the melting of the DNA-Ligand complex occurs on a broader temperature range and the corresponding T-Jump response would therefore be expected to drop proportionally to 80% of the free DNA signal. The observed values in **Fig. 8.10** for **mAA** and **mAT** are with 35% and 50% however much smaller than the expected 80%. This discrepancy was reproducible when repeating the experiment with fresh prepared samples, ruling out inconsistencies in the T-Jump pulse, concentration or the sample thickness as explanations.

The blue trace in **Fig. 8.10** shows the T-Jump response of the previously mentioned DOH bend and indicates the evolution of the solvent temperature over time for these experiments. The trace shows that only a quarter of the initially induced temperature jump is left after 50 µs, so it seems possible that the unfolding process in the ligand-DNA complexes is prematurely interrupted by a depleted heat bath and will start to refold again. A square-shaped temperature gradient would be ideal so that the solvent temperature remains constant throughout the whole unfolding process, but its practical implementation is challenging when performing repeated measurements on a thin sample at a repetition rate of 500 Hz. By dividing the temporal DNA responses in **Fig. 8.10** (Black and red traces) by the temporal evolution of the solvent temperature (blue line), it is possible to compensate to some extent the loss of signal due to the drop in temperature. The corrected traces are plotted in **Fig. 8.11** and the rises up to 100 µs were fitted again using equation (8.1). The results of the fits are summarised in **Table 2**.

	A_1		A_2		τ_1 in ns		τ_2 in µs		β_1		β_2	
	G	**A**	**G**	**A**	**G**	**A**	**G**	**A**	**G**	**A**	**G**	**A**
Free mAA	0.20	0.24	0.80	0.76	110	700	**28**	**32**	0.6	0.5	0.9	1.1
Bound mAA	0.12	0.06	0.88	0.94	60	530	**280**	**240**	0.7	0.7	0.7	0.8
Free mAT	0.21	0.18	0.79	0.82	130	280	**19**	**20**	0.6	0.7	0.9	0.9
Bound mAT	0.16	0.09	0.84	0.91	50	210	**110**	**96**	0.8	0.6	0.7	0.8

Table 2 Parameters from bi-exponential fitting shown in Fig. 8.11. Position of "f" and "b" symbols indicate (τ, A) results of each exponential.

Sanstead *et al.* reported average rise times of ≈200 ns for τ_2 and ≈30 µs for GATATATATC when probing the centre of the transition curve in a T-Jump experiment with 20 Hz repetition rate and a longer cell[20]. This agrees well with the rise times observed in **Fig. 8.11** for free **mAT** ($\tau_1 \approx 200$ ns, $\tau_2 \approx 20$ µs) and confirms that the previously plotted traces in **Fig. 8.10** were affected by depletion of the heat bath over time, predominantly at late time delays.

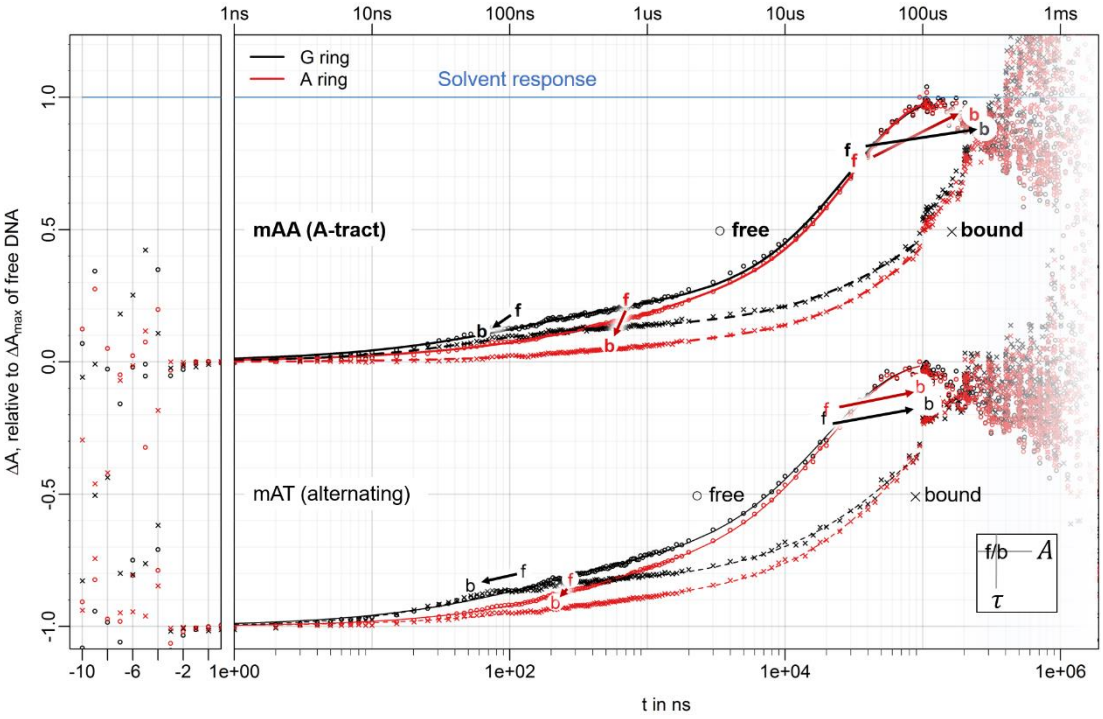


Fig. 8.11 Same plot as in Fig. 8.10 but with all traces divided by the temporal trace of the temperature gradient. The blue trace for the DOH bend is therefore always one. The fitted parameters of each single exponential function (τ, A) is illustrated as the (x,y)-location of points "f" and "b" throughout the plot. "f" indicates results from free DNA, "b" indicates results from DNA in presence of Hoechst 333258. Colours indicate association to G (black) and A (red) ring-modes.

Fitting of the corrected traces shows that amplitudes A_2 observed for the second rise are much higher than for A_1 (*Table* 2 suggests 0.2 for A_1 and 0.8 for A_2). The second rise therefore characterises the main part of the unfolding process that will eventually give rise to the spectroscopic changes seen in steady-state measurements. The microsecond rise times τ_2 in *Table* 2 show little variation between G and A ring-modes and again indicate synchronised unfolding of G-C ends and A-T centre. The τ_2 values signify that minor groove binding of Hoechst 33258 delays the main unfolding process by a factor of 7-10 for A-tract sequence **mAA** (30 µs to 260 µs) and by a factor of 5 for alternating sequence **mAT** (20 µs to 100 µs). This indicates a more effective stabilisation of the double helix in the case of **mAA**. The actual rise times may differ slightly, as fitting was

only carried out until 100 µs but the qualitative difference between bound **mAA** and **mAT** responses is also visible in **Fig. 8.11**, as bound **mAT** reaches a value of 0.5 earlier than bound **mAA**.

These observations suggest an initial destabilisation of the duplex structure on nanoseconds that leads to a main unfolding process on microseconds. Minor groove binding of Hoechst 33258 is affecting each of these processes to a different extent. The A ring-modes in centre of the mixed sequences are in direct contact with the ligand and are affected by the minor groove binder right from the outset of the measurement. The G ring modes at ends of the sequences remain initially unperturbed from the ligand and show the same destabilisation on nanosecond timescales as in the ligand-free case. Stabilisation of the A/T centre relative to the initially unperturbed G/C-ends therefore suggests end-fraying of the dsDNA structure on nanosecond timescales. As time progresses, unfolding of the G ring-modes appears increasingly restrained. It is possible that neighbouring A-T base pairs, which are stabilised by the ligand, prevent G-C structures from unfolding any further. The data suggests that A and G ring-modes rise together during τ_2 and reach their maximum at the same time, which indicates a concerted unfolding process of all base pairs. The region in between τ_1 and τ_2 would then mark the timeframe when stabilising effects from the minor groove binder become less localised and interactions along the strand become relevant.

The rise times in **Table 2** further indicate that the delay of τ_2 upon binding is affected by the ligands ability to stabilise the DNA, as the optimum binding sequence **mAA** returned a longer delay than suboptimal sequence **mAT**. With the main part of the unfolding process taking longer, the surrounding solvent will have had more time to dissipate the kinetic energy out of the sample and into the CaF_2 windows, which act as a heat sink due to their high thermal conductivity (9.71 W/(m K) for CaF_2[36], 0.59 W/(m K) for Water[37]). The experimental results obtained with this technique therefore accentuate nanosecond responses and suppress microsecond dynamics as seen in **Fig. 8.10**.

8.4.4 T-Jump DNA response at pre-melting conditions

Analysing the dynamics observed below the melting transition can be informative to the minor groove binding process as well. FT-IR data below the melting transition shows no significant difference between the steady-state responses of G/C and A/T structures. T-Jump measurements at these temperatures however give a little more

insight. **Fig. 8.12** shows data for **mAA** and **mAT** of free and bound case at a cell temperature of 35°C below the individual melting points. With the T-Jump magnitude being about 9°C, the heat introduced to the system is too small to reach the inflection point of the melting transition and the majority of the double strands should remain hybridised throughout this experiment.

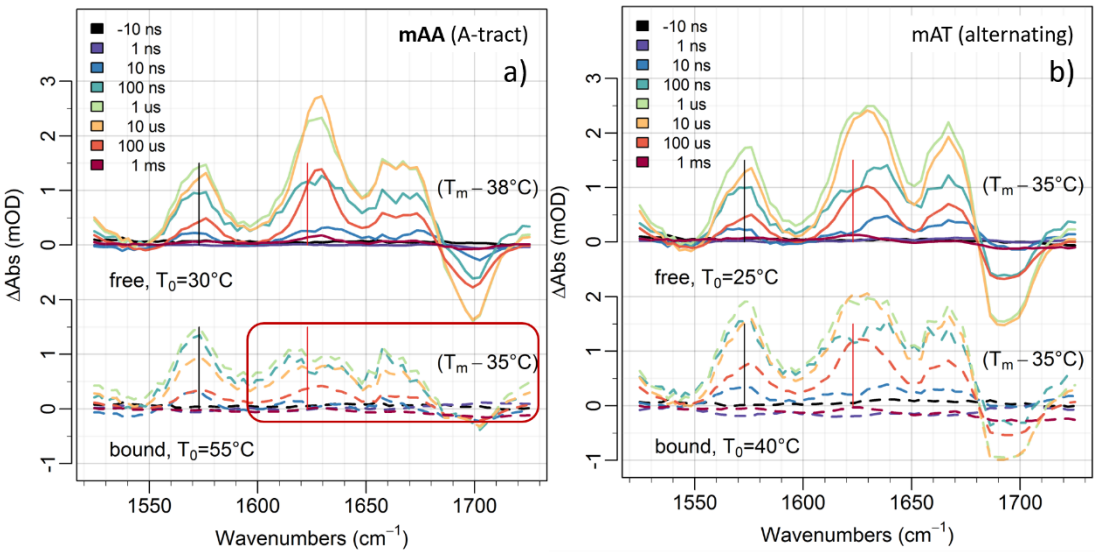


Fig. 8.12 T-Jump response of mAA in a) and mAT in b) in presence/absence of Hoechst 33258 at pre-melting conditions, 35°C below the melting point. Colours indicate time.

The spectral features of the free DNA sequences (top, solid lines) observed at that temperature are in good agreement with responses obtained at the melting point (see **Fig. 8.9**, rise of G and A ring-modes), albeit with a much smaller, overall change in absorbance (≈2 mOD peak amplitudes instead of 10 mOD) as a result of less unfolding. The spectral features of alternating sequence **mAT** in **Fig. 8.12b** on the right do not seem to be affected by the addition of Hoechst 33258, as the response of free (top) and bound (bottom) sequence looks fairly similar. The bound sequence **mAA** in **Fig. 8.12a** on the other hand shows a distinctively smaller rise of the A ring-mode and a few other features above 1650 cm^{-1} (highlighted area in red) compared to the free DNA response. The G ring-mode at 1575 cm^{-1} (black vertical line) however still rises to the same height as in the ligand-free case of **mAA**, suggesting that the drop in signal for A is not simply due to a smaller path length or concentration. This again suggests an increased base-pair stability of the A-T centre over G-C ends, which is in accordance with the results discussed at the melting point in the previous section: A/T structures in the minor groove are stabilised by the ligand and show a smaller response upon binding in the T-Jump experiment, while G/C structures dissociate unperturbed from the binder.

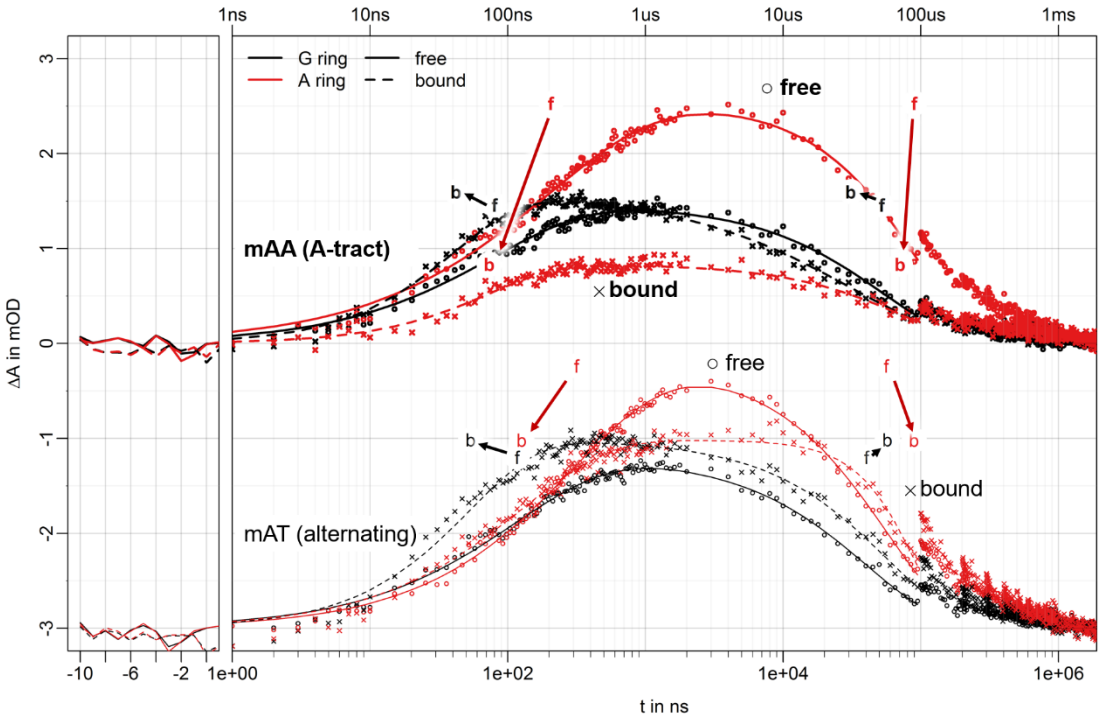


Fig. 8.13 Temporal traces of T-Jump data for 1575 cm^{-1} (G ring-mode, black) and 1623 cm^{-1} (A ring-mode, red) at pre-melting conditions. Cell temperatures are 35°C below the corresponding melting point. The fitted parameters of each single exponential function (τ, A) is illustrated as the (x,y)-location of points "f" and "b" throughout the plot. "f" indicates results from free DNA, "b" indicates results from DNA in presence of Hoechst 333258. Colours indicate association to G (black) and A (red) ring-modes.

The temporal traces of both ring-modes at pre-melting conditions seem to follow the same narrative. The traces shown in **Fig. 8.13** (traces not normalised) lack the previously observed rise on microsecond timescales and instead only contain a small rise on nanoseconds, followed by a decay on microseconds. Each trace was therefore fitted from 0 to 100 µs to the sum of a rise and a decay outlined in equation (8.2):

$$\Delta A(t) = A\left(-e^{\left(-\frac{t}{\tau_{\text{rise}}}\right)^{\beta_{\text{rise}}}} + e^{\left(-\frac{t}{\tau_{\text{decay}}}\right)^{\beta_{\text{decay}}}}\right), \tag{8.2}$$

where A denotes the total peak amplitude. The results of the fits are summarised in **Table 3**. Rise times τ_{rise} on the order of 40-400 ns and low peak amplitudes of a few mOD indicate that the T-Jump pulse at pre-melting conditions is only destabilising the double helix without inducing significant unfolding.

	A in mOD		τ_{rise} in ns		τ_{decay} in µs		β_{rise}		β_{decay}	
	G	A	G	A	G	A	G	A	G	A
Free mAA	**1.5**	**2.5**	80	206	52	90	0.7	0.6	0.8	1.0
Bound mAA	**1.6**	**0.8**	42	73	31	71	0.9	0.9	0.6	0.9
Free mAT	**1.8**	**2.8**	118	317	41	56	0.7	0.7	0.8	0.9
Bound mAT	**2.0**	**2.0**	54	126	57	89	0.9	0.7	0.8	1.7

Table 3 Parameters from bi-exponential fitting shown in Fig. 8.13. Position of "f" and "b" symbols indicate (τ, A) results of each exponential.

Peak amplitudes for G ring-modes remain almost constant upon binding (from 1.5 to 1.6 mOD for **mAA** and 1.8 to 2.0 mOD for **mAT**), while A ring-modes again drop in intensity (for **mAA** from 2.5 to 0.8 mOD and 2.8 to 2.0 mOD for **mAT**). This is clearly visible for **mAA** in the top of **Fig. 8.13** (red traces) and from the fit-results illustrated as downward pointing arrows. The drop in peak amplitude of A is more distinctive for **mAA** (30% of the ligand free signal) than for **mAT** (70%), signifying a qualitative difference between an A-tract and an alternating minor groove at pre-melting conditions. The τ_{rise} value seems to decrease upon binding for both, G as well as A ring-modes in both sequences. The reason for this is unclear, as it could be an effect due to binding as well as due to a higher cell temperature in the bound cases. Sanstead *et al.* has shown that both, the nanosecond destabilisation and microsecond unfolding dynamics follow Arrhenius behaviour in the GATATATATC sequence[20], where the unfolding rate increases with temperature and therefore reduces the rise time. This could be a possible explanation for the observed reduction of τ_{rise}. It does not however, explain the delayed unfolding process observed for bound DNA in the previous chapter. This is still assumed to be a consequence of binding, as higher temperatures are expected to lead to faster rise times.

These results show that the minor groove binder affects the nanosecond dynamics of the dsDNA already at pre-melting conditions. Both sequences **mAA** and **mAT** show suppression of the A ring-modes upon binding with the "A-tract" target sequence **mAA** being affected to a larger extend. This indicates stabilised A/T structures in the minor groove due to presence of Hoechst 33258. G/C ends seem to remain unaffected by the ligand and respond with similar amplitudes as in the free DNA.

8.4.5 Comparison to pure A/T sequences pAA and pAT

Pure A/T sequences **pAA** (5'-TTAAATTTTT-3') and **pAT** (5'-TTATATATTT-3') were analysed similar to **mAA** and **mAT** at cell temperatures of 5-8°C below their corresponding melting point to probe the main transition of the melting curve and maximise the T-Jump response. **Fig. 8.14** shows the spectral responses obtained for **pAA** and **pAT** in absence/presence of Hoechst 33258. Due to limited instrument access at the Rutherford Appleton Laboratories, measurements for **mAA** and **mAT** were prioritised and **pAA** and **pAT** were only measured once. Sequence **pAA** in presence of Hoechst 33258 was only measured up to a cell temperature of 65°C, as this sample already reached a maximum T-Jump response at 40°C. The FT-IR melting response discussed earlier indicated two transition points, but the FT-IR data was only collected after the T-Jump experiment had been carried out, so that a second maximum in the T-Jump response was not anticipated at the time. In the following analysis, the first transition point at 45°C in the FT-IR data is used as melting point for the "bound" **pAA** – case and the maximum T-Jump response at $T_c = 40$°C is being analysed.

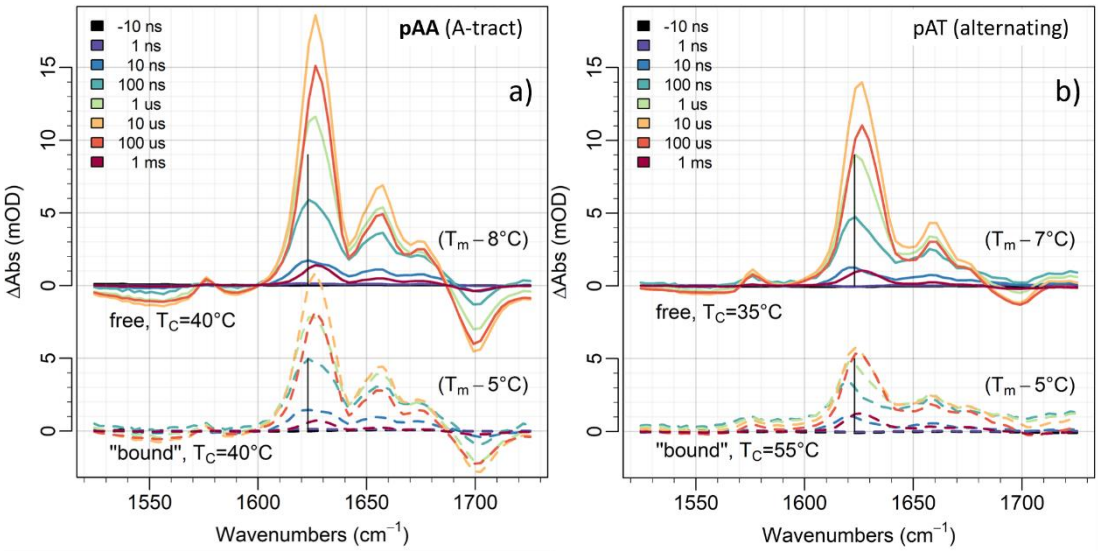


Fig. 8.14 T-Jump response of pAA in a) and pAT in b) in presence/absence of Hoechst 33258 probing the corresponding melting transition.

The spectral features for the dsDNA strands without ligand (solid lines) again compare well with the corresponding FT-IR data in **Fig. 8.5b**. Just as in the steady-state measurements, there are no significant spectral features present at 1575 cm^{-1} due to the lack of guanine nucleobases in these sequences. If the previously discussed findings for the effect of minor groove binding to **mAA** and **mAT** are valid, then the following characteristics would be expected for **pAA** and **pAT** when interacting with Hoechst

33258: Minor groove binding should increase the melting point and slightly shift the negative T_{2s} carbonyl feature at around 1700 cm^{-1} to higher wavenumbers. The ring-mode of adenine at 1623 cm^{-1} should reach its maximum later in time and complete unfolding is likely to be obstructed due to fast dissipation of heat out of the sample. This would affect the overall magnitude of the T-Jump response and result in smaller signals. The reporter mode of adenine in case of **pAA** and **pAT** should show the response of base pairs at the ends of the double strands and within the minor groove, so that the temporal response might be less well defined than in **mAA** and **mAT**.

The spectral features shown in **Fig. 8.14a** are not very conclusive to confirm minor groove binding to **pAA**. The negative T_{2s} band at 1700 cm^{-1} does not appear to shift in frequency when adding Hoechst 33258 to the sample. The largest T-Jump response for bound **pAA** was found at the same cell temperature as the ligand-free DNA at 40°C. The loss in magnitude of the T-Jump response is present and could be the result of a delayed unfolding process, although excitation of a smaller fraction of DNA modes is plausible as well. So far, there is no sign of Hoechst 33258 binding to **pAA**.

The response of sequence **pAT** in **Fig. 8.14b** is ambiguous as well. A shift of the T_{2s} at 1700 cm^{-1} is possible, but the signal intensity is quite low. The largest T-Jump response was found at a cell temperature of 55°C when Hoechst was added to the DNA, 20°C higher than in the ligand-free case. This is the only clear evidence that minor groove binding is occurring for **pAT**.

The temporal traces of the A ring-mode might give further insight into the unfolding process and is plotted in **Fig. 8.15**. The traces of the free DNA were normalised to one and the responses for the "bound" cases were scaled accordingly by the same factor to retain the relative ratio from free to bound responses, as before. Each trace was fitted from 0 ns to 30 µs to a sum of two exponential rise functions as defined in equation (8.1), which is similar to the previous analysis of data from the mixed sequences at the melting point.

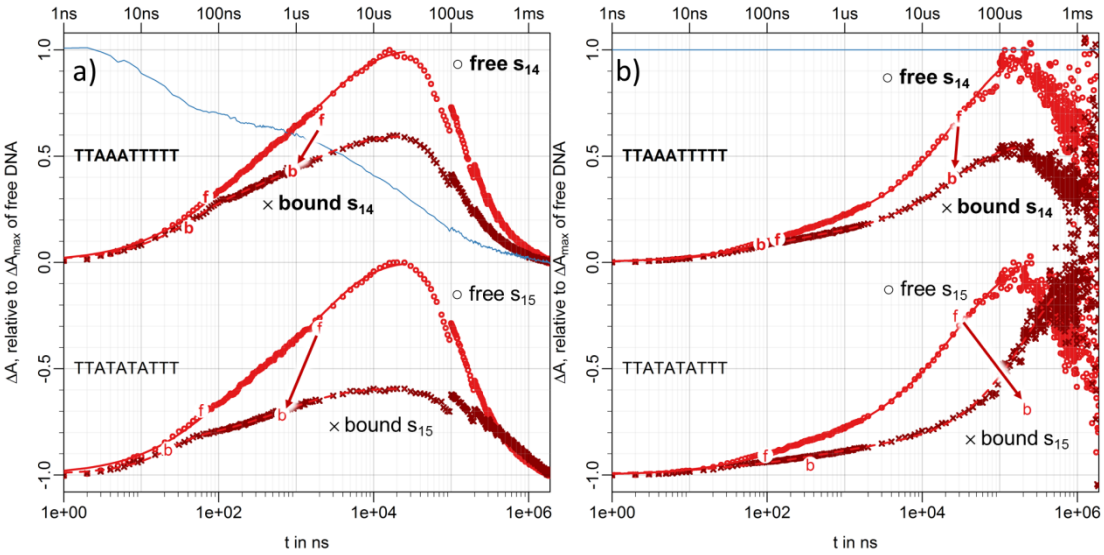


Fig. 8.15 a) Temporal traces for the band at 1623 cm^{-1} (A ring-mode, red), normalised to the maximum signal of the corresponding free DNA (similar to Fig. 8.10). Blue trace indicates the temporal evolution of the DOH bend. b) Similar plot as in a) but each trace was divided by the response of the solvent before performing the normalisation step. Traces were fitted to the sum of two rising exponentials as defined in equation (8.1). Position of "f" and "b" symbols indicate (τ, A) results of each exponential.

Fig. 8.15a summarises the results of the bi-exponential fitting. The rise of the A ring-mode in free **pAA** and free **pAT** is almost identical (traces in lighter red). Addition of the minor groove binder lowers the peak amplitude of the A ring-mode in both sequences (traces in darker red), but definitive proof of minor groove binding still remains elusive. The temporal response of the solvent is shown as a trace from one in blue and again indicates significant cooling of the solvent throughout the measurement window. Only when accounting for the gradual loss in temperature, as discussed earlier, by dividing each trace by the response of the solvent, the results become more indicative. The corrected traces in **Fig. 8.15b** show similar unfolding dynamics for ligand-free sequences **pAA** and **pAT** (light red), as before. Addition of Hoechst 33258 to **pAT** (darker red, bottom) shows a delayed rise on microsecond timescales, indicating a delayed unfolding process. Fitting of these traces to equation (8.1) reveals that the microsecond rise for **pAT** upon binding is delayed by a factor of 8 from 27 µs to 226 µs (red arrow in the bottom of **Fig. 8.15b**) and suggests that some of the A-T base pairs in the alternating sequence **pAT** are stabilised due to interaction with the Ligand. The addition of Hoechst 33258 to **pAA** (top of **Fig. 8.15b**) has a different effect on the T-Jump response. The microsecond rise time remains in **pAA** when adding the ligand (from 29 µs to 24 µs) and the overall response only drops in amplitude, indicated by the downward pointing arrow in red. The T-Jump response therefore seems to describe the same unfolding process as in the ligand-free case, only on a smaller fraction of

dsDNA. This seems plausible from the observed melting curve for "bound" **pAA**, shown in **Fig. 8.6**, where the first transition point at 45°C only accounts for about half of the overall response of the DNA during melting. The remaining fraction only starts to melt at a second transition point, closer to 80°C. This seems to confirm the theory that the transition point at 45°C originated from a significant fraction of free DNA in the sample and that unfolding of stabilised DNA-ligand complexes may only be observable at cell temperatures closer to 80°C. A possible explanation for this could be an insufficient concentration of Hoechst 33258 available in solution. The structure of this binder is relatively hydrophobic, which can cause aggregation and precipitation over time. This measurement usually would have been repeated under normal circumstances but limited access to the T-Jump instrument at the Rutherford Appleton prevented further experiments.

8.5 Conclusions and Future Work

Drug discovery relies on the ability to assess and differentiate binding interactions of potential new therapeutic molecules not only in great molecular detail but also in a reasonable amount of time. The present study was intended as a proof-of-principle to investigate minor groove binding using time-resolved, temperature-jump experiments in the infrared region. The newly developed instrument used in this project focuses on minimising the acquisition time to allow for complex experiments on a large number of samples in the future; as would be needed for screening experiments of novel minor groove binders. The interaction of Hoechst 33258 to 5'-AAATTT-3' and 5'-ATATAT-3' minor groove motifs provided a well-established reference system involving subtle structural changes that remain challenging to characterise directly *in situ*.

Steady-state, FT-IR melting experiments were used as guidance and validation measurements for time-resolved experiments by providing melting curves and expected spectral responses for each sample. The magnitude of the jump in sample temperature induced by the laser pulse was estimated to about +9°C by performing preliminary T-Jump measurements on TFA in D_2O. The high repetition rate of 500 Hz required complete re-equilibration of the sample temperature within 2 ms, which made it necessary to reduce the sample thickness to 6 μm. The CaF_2 windows of the cell acted effectively as heat sinks for this purpose, leading to a stretched exponential decay with a lifetime of ≈50 μs. This is considerably faster than another, previously published methodology with a re-equilibration of several milliseconds[38]. The method allows for

short acquisition times within minutes but also comes with a drawback for obtaining absolute kinetic traces, as observed in the DNA melting responses.

The simultaneous measurement of the solvent response within the spectral window proved to be extremely useful as its temporal trace provided a relative reference for the momentary temperature at each point in time. It was shown that this trace can be used to separate the sub-ns heating response of the DNA from spectral changes due to conformational rearrangement during melting. From this basis, a pre-processing step was developed and applied to the acquired T-Jump data.

Two ring-modes of guanine and adenine at 1575 cm^{-1} and 1623 cm^{-1} were identified as temperature sensitive reporter modes that indicated disruption to the double stranded structure. These were treated as marker bands for nucleobases at the ends of the double helix (G ring-mode) and in the centre (A ring-mode). DNA unfolding was observed to contain a small, initial destabilisation process at nanosecond timescales followed by the main unfolding process on microsecond timescales. Unfolding of the G-C ends and A-T centres in the ligand-free DNA occurred simultaneously, which is in accordance with previously published results[19,20].

Binding of Hoechst 33258 was shown to suppress the initial destabilisation of A-T base pairs in the minor groove at nanosecond timescales. This effect was particularly visible at pre-melting conditions, where the microsecond response from unfolding was minimal. The nanosecond response of G-C base pairs at the ends however remained unaffected by the ligand and followed the response obtained from free DNA until about 100 ns. The subsequent unfolding processes of G-C ends and A-T centres were both affected by the ligand and took at least 5 times longer than in the free DNA, leading to a delayed, but still simultaneous de-hybridisation of the double helical structure. It is suspected that after about 100 ns, unfolding of G-C end structures becomes restrained by the progress of neighbouring A-T base pairs, leading to an initial end-fraying process on nanosecond timescales that evolves into a correlated unfolding process in case of the DNA-ligand complex.

The T-Jump response of sequence **mAA** with the optimal binding site for Hoechst 33258 was affected to a larger extent by the presence of the ligand than the response of **mAT** with a suboptimal minor groove. Unfolding of **mAA** was delayed by a factor of 7-10, whereas unfolding of **mAT** was delayed by a factor 5 when Hoechst 33258 was

added to the sample. Nanosecond destabilisation at pre-melting conditions was clearly suppressed for **mAA** (drop to 30% of free DNA peak amplitude upon binding) but less obvious for **mAT** (drop to 70%). This shows that the selectivity of the ligand to a specific minor-groove is reflected in the kinetic profiles observed in T-Jump measurements.

Replacement of the G-C ends in **mAT** with A-T ends in **pAT** did confirm observations of a delayed unfolding process for minor groove binding. The replacement of G-C ends from **mAA** to **pAA** led to an unexpected melting behaviour, possibly due to a non-equimolar DNA-ligand-ratio for this particular sequence. T-Jump measurements were only performed on the first of two melting transitions for this DNA-Ligand complex which was assigned to a significant residual concentration of free DNA, characterised by a melting process similar to that of the free DNA. The second transition is assumed to be the expected response of bound DNA but was not measured due to limited instrument access.

The response obtained in T-Jump experiments intimately depends on the kinetic energy within the sample at any given moment in time. The small sample volume used in this study allowed for a fast acquisition of data but at the same time limited the information available from millisecond timeframes. The perturbed sample lost its thermal energy gradually throughout the measurement and therefore immediately affected the melting response of the DNA. This affected measurements at late time points in particular and has led to premature refolding when probing stabilised DNA sequences, as the absolute magnitude of these responses was significantly smaller than expected. It was shown that the effect of the gradual loss in temperature could be accounted for in the results to some extent and a comparison of obtained rise times to previously published results seemed to confirm this methodology. Screening experiments in particular rely on relative differences between ligands and a large number of repeated measurements so that a short acquisition time may be favourable for this type of application.

These results give a promising outlook for the utilisation of time-resolved infrared spectroscopy methods to assess binding interactions of novel DNA ligands. The method presented is completely label-free, capable of observing structurally sensitive information and can cover a large temporal window that is critical when studying molecular processes. The methodology in its present form sheds further light onto the

cause of melting point stabilisation while leaving enough overhead in terms of measurement time for more demanding spectroscopy methods. An extension to 2D-IR spectroscopy will combine the large temporal window of T-Jump experiments with insights into the molecular connectivity of the reporter modes and could further the understanding of melting dynamics in DNA-Ligand complexes. The comparably high repetition rate used in this study will allow to take a large, three-dimensional datasets (two frequency axes and one time axis) within one or two hours for a single sample, leaving enough time for repeats and other experiments needed for screening experiments.

8.6 References

1. Cai, X., Gray, P. J. & Von Hoff, D. D. DNA minor groove binders: Back in the groove. *Cancer Treat. Rev.* **35**, 437–450 (2009).

2. Dasari, M. *et al.* H-Gemcitabine: A New Gemcitabine Prodrug for Treating Cancer. *Bioconjug. Chem.* **24**, 4–8 (2013).

3. Saini, K. S. *et al.* New orally active DNA minor groove binding small molecule CT-1 acts against breast cancer by targeting tumor DNA damage leading to p53-dependent apoptosis. *Mol. Carcinog.* **56**, 1266–1280 (2017).

4. Nimesh, H. *et al.* Synthesis and Biological Evaluation of Novel Bisbenzimidazoles as Escherichia coli Topoisomerase IA Inhibitors and Potential Antibacterial Agents. *J. Med. Chem.* **57**, 5238–5257 (2014).

5. Dale, A. G., Hinds, J., Mann, J., Taylor, P. W. & Neidle, S. Symmetric Bis-benzimidazoles Are Potent Anti-Staphylococcal Agents with Dual Inhibitory Mechanisms against DNA Gyrase. *Biochemistry* **51**, 5860–5871 (2012).

6. Ranjan, N. *et al.* Selective Inhibition of Escherichia coli RNA and DNA Topoisomerase I by Hoechst 33258 Derived Mono- and Bisbenzimidazoles. *J. Med. Chem.* **60**, 4904–4922 (2017).

7. Neidle, S. DNA minor-groove recognition by small molecules (up to 2000). *Nat. Prod. Rep.* **18**, 291–309 (2001).

8. Bazhulina, N. P. *et al.* Binding of Hoechst 33258 and its Derivatives to DNA. *J. Biomol. Struct. Dyn.* **26**, 701–718 (2009).

9. Fornander, L. H., Wu, L., Billeter, M., Lincoln, P. & Norde, B. Minor-Groove Binding Drugs: Where Is the Second Hoechst 33258 Molecule? *J Phys Chem B* **117**, 5820–5830 (2013).

10. Lankas, F., Spackova, N., Moakher, M., Enkhbayar, P. & Sponer, J. A measure of bending in nucleic acids structures applied to A-tract DNA. *Nucleic Acids Res.* **38**, 3414–3422 (2010).

11. Ramakers, L. A. I. *et al.* 2D-IR Spectroscopy Shows that Optimized DNA Minor Groove Binding of Hoechst33258 Follows an Induced Fit Model. *J. Phys. Chem. B* **121**, 1295–1303 (2017).

12. Bostock-Smith, C. E., Harris, S. A., Laughton, C. A. & Searle, M. S. Induced fit DNA

recognition by a minor groove binding analogue of Hoechst 33258 : fluctuations in DNA A tract structure investigated by NMR and molecular dynamics simulations. *Nucleic Acids Res.* **29**, 693–702 (2001).

13. Wilson, W. D., Tanious, F. A., Fernandez-Saiz, M. & Rigl, C. T. Evaluation of Drug–Nucleic Acid Interactions by Thermal Melting Curves. in *Drug-DNA Interaction Protocols* 219–240 (Humana Press). doi:10.1385/0-89603-447-X:219

14. Bueren-Calabuig, J. A., Giraudon, C., Galmarini, C. M., Egly, J. M. & Gago, F. Temperature-induced melting of double-stranded DNA in the absence and presence of covalently bonded antitumour drugs: Insight from molecular dynamics simulations. *Nucleic Acids Res.* **39**, 8248–8257 (2011).

15. Beitz, J. V., Flynn, G. W., Turner, D. H. & Sutin, N. Stimulated Raman effect. A new source of laser temperature-jump Heating. *J. Am. Chem. Soc.* **92**, 4130–4132 (1970).

16. Kubelka, J. Time-resolved methods in biophysics. 9. Laser temperature-jump methods for investigating biomolecular dynamics. *Photochem. Photobiol. Sci.* **8**, 499 (2009).

17. Ma, H., Wan, C., Wu, A. & Zewail, A. H. DNA folding and melting observed in real time redefine the energy landscape. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 712–6 (2007).

18. Zhang, X. *et al.* Studying Protein–Protein Binding through T-Jump Induced Dissociation : Transient 2D IR Spectroscopy of Insulin Dimer. *J. Phys. Chem. B* **120**, 5134–5145 (2016).

19. Sanstead, P. J., Stevenson, P. & Tokmakoff, A. Sequence-Dependent Mechanism of DNA Oligonucleotide Dehybridization Resolved through Infrared Spectroscopy. *J. Am. Chem. Soc.* **138**, 11792–11801 (2016).

20. Sanstead, P. J. & Tokmakoff, A. Direct Observation of Activated Kinetics and Downhill Dynamics in DNA Dehybridization. *J. Phys. Chem. B* **122**, 3088–3100 (2018).

21. Steinel, T., Asbury, J. B., Zheng, J. & Fayer, M. D. Watching hydrogen bonds break: A transient absorption study of water. *J. Phys. Chem. A* **108**, 10957–10964 (2004).

22. Chung, H. S., Khalil, M., Smith, A. W. & Tokmakoff, A. Transient two-dimensional IR spectrometer for probing nanosecond temperature-jump kinetics. *Rev. Sci. Instrum.* **78**, 063101 (2007).

23. Fritzsch, R. *et al.* Rapid Screening of DNA–Ligand Complexes via 2D-IR Spectroscopy and ANOVA–PCA. *Anal. Chem.* **90**, 2732–2740 (2018).

24. Banyay, M., Sarkar, M. & Graslund, A. A library of IR bands of nucleic acids in solution. *Biophys. Chem.* **104**, 477–488 (2003).

25. Lee, C. & Cho, M. Vibrational dynamics of DNA. II. Deuterium exchange effects and simulated IR absorption spectra. *J. Chem. Phys.* **125**, 114509 (2006).

26. Doorley, G. W. *et al.* Tracking DNA Excited States by Picosecond-Time-Resolved Infrared Spectroscopy: Signature Band for a Charge-Transfer Excited State in Stacked Adenine–Thymine Systems. *J. Phys. Chem. Lett.* **4**, 2739–2744 (2013).

27. Hithell, G. *et al.* Long-Range Vibrational Dynamics Are Directed by Watson–Crick Base Pairing in Duplex DNA. *J. Phys. Chem. B* **120**, 4009–4018 (2016).

28. Peng, C. S., Jones, K. C. & Tokmakoff, A. Anharmonic vibrational modes of nucleic acid bases revealed by 2D IR spectroscopy. *J. Am. Chem. Soc.* **133**, 15650–15660 (2011).

29. Alniss, H. Y. Thermodynamics of DNA Minor Groove Binders. *J. Med. Chem.*

acs.jmedchem.8b00233 (2018). doi:10.1021/acs.jmedchem.8b00233

30. Liepinsh, E., Otting, G. & Wüthrich, K. NMR observation of individual molecules of hydration water bound to DNA duplexes: direct evidence for a spine of hydration water present in aqueous solution. *Nucleic Acids Res.* **20**, 6549–6553 (1992).

31. Lankas, F. *et al.* Critical Effect of the N2 Amino Group on Structure , Dynamics , and Elasticity of DNA Polypurine Tracts. *Biophys. Chem.* **82**, 2592–2609 (2002).

32. Breslauer, K. J., Frank, R., Blocker, H. & Marky, L. A. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci.* **83**, 3746–3750 (1986).

33. Szyc, Ł., Yang, M., Nibbering, E. T. J. & Elsaesser, T. Ultrafast Vibrational Dynamics and Local Interactions of Hydrated DNA. *Angew. Chemie Int. Ed.* **49**, 3598–3610 (2010).

34. Liu, Y., Guchhait, B., Siebert, T., Fingerhut, B. P. & Elsaesser, T. Molecular couplings and energy exchange between DNA and water mapped by femtosecond infrared spectroscopy of backbone vibrations. *Struct. Dyn.* **4**, 044015 (2017).

35. Tros, M. *et al.* Picosecond orientational dynamics of water in living cells. *Nat. Commun.* **8**, 904 (2017).

36. Ballard, S. S., McCarthy, K. A. & Davis, W. C. A Method for Measuring the Thermal Conductivity of Small Samples of Poorly Conducting Materials such as Optical Crystals. *Rev. Sci. Instrum.* **21**, 905–907 (1950).

37. Touloukian, Y. S., Liley, P. E. & Saxena, S. C. Thermophysical Properties of Matter - The TPRC Data Series. Volume 3. Thermal Conductivity - Nonmetallic Liquids and Gases. *Def. Tech. Inf. Cent.* 1–708 (1970).

38. Jones, K. C., Peng, C. S. & Tokmakoff, A. Folding of a heterogeneous -hairpin peptide from temperature-jump 2D IR spectroscopy. *Proc. Natl. Acad. Sci.* **110**, 2828–2833 (2013).

9 Conclusions and Future Work

The previous chapters have outlined different approaches to study DNA-Ligand interactions with time-resolved spectroscopy methods. 2D-IR spectroscopy has been successfully applied as a screening technique in chapter 6 to study minor groove binding of Hoechst 33258. This chapter has introduced ANOVA-PCA as a scalable analysis method in order to cope with an ever-increasing abundance of spectroscopic data. ANOVA-PCA has enabled the differentiation and categorisation of twelve different minor groove interactions by analysing their impact on the 2D-IR spectrum. Minor groove binding of Hoechst 33258 to its target sequence 5'-AAATTT-3' has shown a subtle, but distinctive 2D-IR response that could be distinguished from responses of incompatible minor grooves. The structural information attainable from the 2D-IR method has confirmed the sensitivity of the DNA marker band T_{2S} to minor groove binding, which identified replacement of the spine of hydration in the minor groove by the ligand upon binding. Hoechst 33258 minor groove binding was shown to have little impact on the principal, vibrational relaxation of carbonyl and ring-modes of the nucleobases, although more subtle dynamic changes could not be ruled out.

After evaluating 2D-IR spectroscopy as a screening technique in chapter 6, the method was extended to a range of four hairpin polyamide ligands screened across a narrow range of dsDNA sequences with 5'-TGNNCA-3' minor groove (N = G, C, A or T). 2D-IR difference spectra were analysed via ANOVA-PCA and identified an increased selectivity for 5'-TGTACA-3' sequences in cases for polyamide ligands with isopropyl side group substitution. While all DNA-ligand combinations indicated a perturbation to G/C-structures upon binding, combinations with 5'-TGTACA-3' in particular returned an additional set of cross-peaks, possibly indicating an interaction between vibrational modes of the DNA with the ligand. Analysis of the time-resolved 2D-IR responses of individual DNA-ligand combinations identified a possible perturbation to vibrational dynamics in the sample upon binding, suggesting a competing energy transfer process in complexes for ligands with isopropyl residue modification.

The strength of the ANOVA-PCA approach utilised is clearly its scalability to large datasets. The rate determining step of the experiments in both chapters 6 and 7 was the sample preparation. High-throughput sampling techniques such as HPLC-style autosamplers with sample loop, or the implementation of multiwell plates could not

only increase sample throughput for 2D-IR spectroscopy, but may improve reproducibility of the measurements due to standardisation of the sampling conditions. ANOVA-PCA in its current implementation is however limited in its ability to extract the dynamic information available from 2D-IR spectroscopy, as observed when analysing the principal relaxation dynamics in chapter 6. Chapter 7 has demonstrated how a global fitting approach leads to a more detailed representation of the dynamic features collected and identifies a possible route to extract more subtle dynamic responses in the future. The analytical approach used in these chapters allow for a quick survey of a large spectroscopic dataset to identify key DNA-ligand combinations in a first step, which then provides detailed structural and temporal information from individual 2D-IR responses in a second step.

The spectroscopic methods used in this thesis provide a complementary approach to existing analysis methods in order to assess the selectivity of new ligand candidates. 2D-IR spectroscopy is fast, label-free, works for solutions and can provide temporal information to characterise ultrafast structural fluctuations on femtosecond timescales. The 2D-IR responses observed for the two types of ligands studied in this thesis were quite distinct from each other. Future projects could focus on a selection of different minor groove binders to gain a more comprehensive understanding of the underlying structural perturbations across different binding situations. Different ligand concentrations and the effect of salt and pH-value could provide further insight, how reporter modes at the DNA-ligand complex are affected by various binding conditions. It is further possible to study the same interactions with different reporter modes, such as NH/NH_2-stretching modes at the nucleobases, phosphate stretching modes at the backbone, or using isotope labelling to identify specific interactions in the spectrum.

While the initial two chapters studied DNA-ligand interactions at equilibrium conditions, chapter 8 examined Hoechst 33258 binding again via temperature-jump, time-resolved, infrared spectroscopy. The stabilising effect of minor groove binding to the DNA duplex structure was studied by investigating the unfolding dynamics of bound DNA from nanosecond to millisecond timescales. Ligand binding initially stabilised the duplex structure at the minor groove on nanosecond timescales, relative to the unperturbed DNA-end structures, which caused initial end-fraying of the bound DNA. Minor groove binding further increased the time required for synchronised unfolding along the strand by a factor of 5-10, depending on the minor groove interaction studied. The results indicated that the stabilising effect of the ligand is only

initially localised to the binding site and will affect unfolding dynamics of adjacent structures as well, indicating an allosteric effect of the ligand. The instrumental setup in its current implementation suffers from nonuniform heating across the sample. This effect can be mitigated by simultaneous excitation of the sample from both sides of the cell and possibly by applying additional coatings to the cell windows to adjust the thermal conductivity at the interface to the sample. A separate publication will further discuss benefits and drawbacks of the employed temperature-jump implementation in more detail. Future work on this instrument will also include the development of a 2D-IR spectrometer that captures non-equilibrium processes from nanosecond to millisecond timescales, while resolving changes to the vibrational coupling between modes. This may provide further insight how ligand binding is affecting the structural dynamics of the DNA.

10 Acknowledgements

First of all, I need to thank my supervisor Neil Hunt for his continuous support throughout the last four years. From advice on experiments over in-depth discussions of experimental results to guidance on scientific writing; his insights and feedback were invaluable to me and indispensable to the outcome of this thesis. Thank you for all the quick email replies, your open-door policy and the seemingly never-ending stream of comments on my drafts. It has always been a pleasure working in this group.

I would like to thank my colleagues Lucy, Sam and Gordon, as well as Niall and Lennart for helping me through the long evenings at RAL, long days at Strathclyde and the far too long train rides in between. Thank you for the warm welcome in Scotland and showing me the culinary heritage that you call square sausage. And thank you Glasgow for your curry houses.

Further I would like to thank everyone at RAL, including Tony Parker, Greg Greetham, Paul Donaldson, Ian Clark, Mike Towrie and Emma Gozzard for their ongoing support during measurements and their invaluable expertise regarding instrumentation. Thank you, Tony, for asking all the right questions at the wrong time to keep us on our toes. I will miss them.

I also want to thank Matthew Baker for his advice regarding all things analytical together with his group for the many social events in the last years, including organisation of the Clirspec Summer School and Spec 2018.

I would also like to thank our collaborators in chemistry, Glenn Burley, Giacomo Padroni and Khalid Aman for providing polyamide ligands and the many discussions regarding minor groove binding. Thank you, Khalid, for working through countless number of samples together at RAL.

Finally, I want to thank my partner Lisa for her patience, her positivity and her relentless support in the last eight years of our adventure so far, as well as in our new journey as a family together. Thank you, Mats, for giving me a hard deadline to finish this thesis once and for all.