Gareth Moore
Sulgeneckstrasse 50
3007 Bern
+41 78 744 1187

DATA SCIENCE PROJECT

Transient Absorption Spectroscopy
Conceptual Design Report

27st September 2019

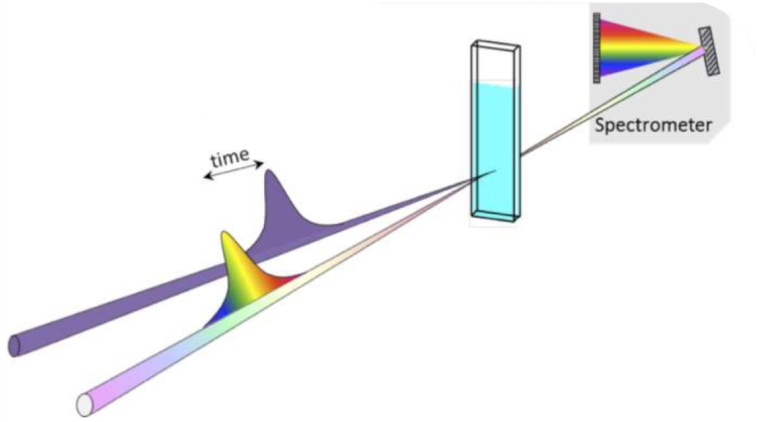
ABSTRACT

Transient Absorption Spectroscopy (TAS) is an experimental method which allows us to spectroscopically follow physical processes on a femtosecond timescale. The processing, cleaning, presenting and analysis of TAS data is something well documented but not at all standardized. In this report we formulate standard data flow and procedural methods for signal processing, data cleaning, pre-processing as well as data representation. We also discuss the analysis methods and the way in which these too can be implemented on a standardized way that produces both reliable and efficient analyses.

TABLE OF CONTENTS

INTRODUCTION

Transient Absorption Spectroscopy (TAS):
TAS is a method in which we can spectrally follow processes that result from electromagnetic excitation in a sample (in this case Organic Photo-voltaic). Using a 1 kHz pulse laser a pump (excitation) beam is made at a specific wavelength/energy, as well as a broad white light beam. The pump pulse excites the sample and at some time (time delay) later the probe beam measures the absorption spectra. The difference between the spectra of the pumped and the unpumped sample is then recorded. Spectral signatures of S1 states, excitons, free electrons and holes, amongst others, can be seen and followed in time with femtosecond resolution.



OBJECTIVE

The analysis of TAS data is not something new, however the procedures and methods are far from optimal and therefore do not always produce usable results in a timely or rigorous way.

The objective of this project is to develop a sound data flow and procedure for:

- Signal processing
- Data cleaning and preprocessing (including automation)
- Spectral Analysis

A signal processing method will be implemented in order to pre-clean data in improve measurement efficiency.

The data cleaning and processing will be migrated from a combination of Labview, MATLAB and Igor Pro to a single Python code.

The standardization of Spectral analysis will be discussed in terms of atomization and removing as much of the human element from the analysis in order to produce consistent and reliable analysis.

METHODS

The methods used will be broken up into data acquisition and data analysis.

Data Acquisition:
Data acquisition is done by generating a spectra with a pulsed laser system. The light pulse that contains the spectral data is dispersed through a prism and onto a pixelated CCD camera where the intensities of small wavelength bins are measured at a 1 kHz repetition rate.

The signal from each pixel is treated using a noise reduction method developed be Anderson et al.[1] which excludes outliers based on the mode of binned signal (described later).

Data acquisition and signal processing is done and controlled using LabView on a dedicated lab PC. Data is then saved onto a local server.

Data Analysis:
Data is handled as a 2D numpy array (using numpy and glob packages) in Python (2017 Mac 3.6 GHz Intel Core i7 Processor, 16 GB 2400 MHz DDR4 Ram).

- Python packages used:
- Numpy for array handling
- Glob for directory navigation
- SciPy's stats, special and optimize (for fitting)
- Math are used for data cleaning.
- Matplotlib for plotting

Data cleaning involves background subtraction and chirp correction.

A Global fitting procedure is preformed (described below) using Igor Pro software.
Spectral decomposition is done using either MATLAB or Python.

All plotting is done either using Pythons Matplotlib package of in Igor Pro.

DATA

Signal from CCD

All data was collected by me at the Department of Chemistry.

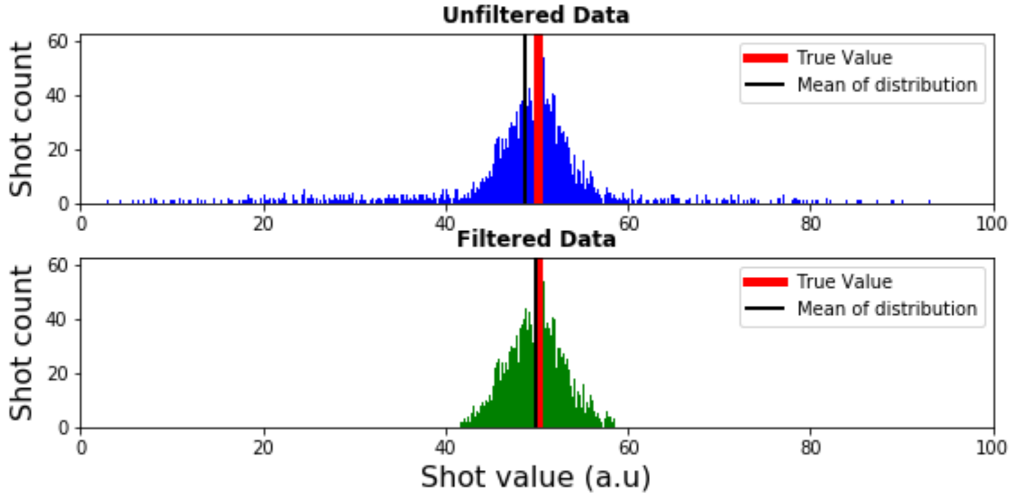


Figure 1: *The signal received from one pixel of the CCD camera for one time delay before filtering (top, blue) and after filtering (bottom, green).*

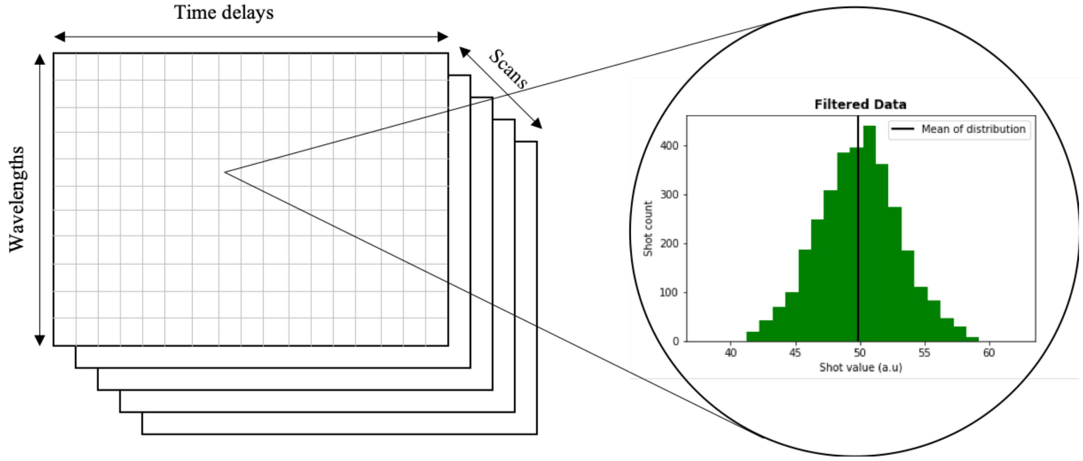


Figure 2: *Illustration of how the processed signal is averaged with the mean value becoming one point in the data, along with an illustration of the shape of the data coming from the signal processing step.*

Data after Analysis

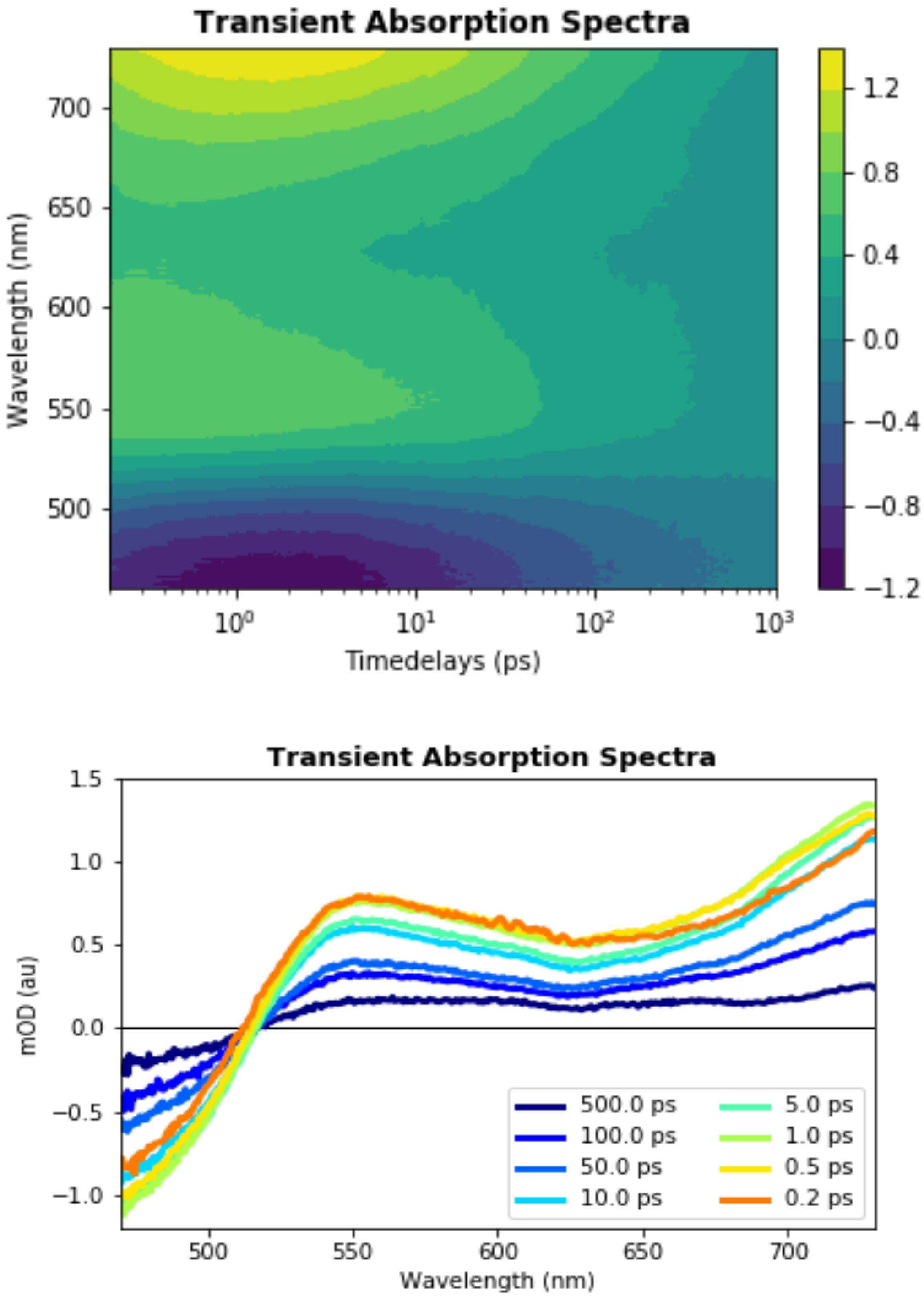


Figure 2: *Figure showing the contour plot of the TAS spectra after cleaning and chirp correction (top) and spectra at different time delays (bottom) .*

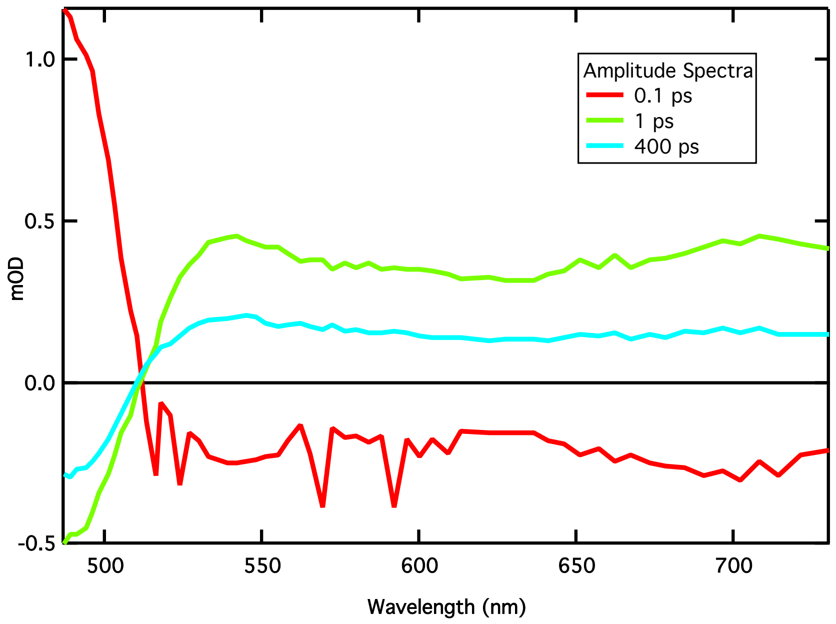


***Figure 3:** Figure showing the Amplitude spectra with associated time constants resulting from the Global Analysis.*

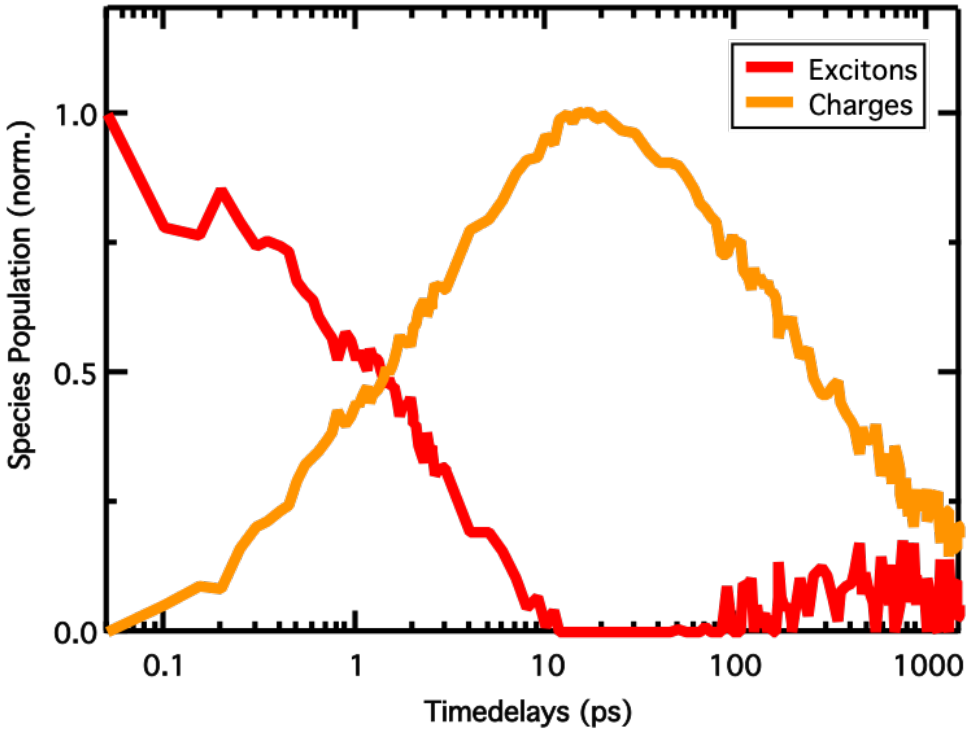


***Figure 4:** Figure showing the results of spectral decomposition where we see the splitting of excitons into free charges and then some recombination at the end.*

METADATA

All files are stored on internal server (NAS DS916+ from Synology, 10 TB capacity), and is password protected for members of research group.

Data successful experiments, which is used in scientific publications, will be deposited for re-use (public access, CC0 license) on the BORIS Publication Repository of the University of Bern. Each data set will be identified with a DOI (Digital Object Identifier). The Supporting Information of the corresponding scientific publication will contain a link to the repository and data DOI.

Folder structure for raw data:

- Year -> Researcher_name -> Date -> Filename.

Filenames:

- Experiment_sample_conditions_trial.
- For example for a transient absorption experiment done for the second time on Perylene with excitation at 400 nm and a fluence of 100 nJ: TA_Per_400nm_100nJ_2.

Readme file is added to each raw data file containing:

- Date
- Name of measurement set
- Name of researcher
- Sample details and measuring conditions
- Abbreviations in the file naming

Folder structure for processed data:

- Sub-project -> Technique -> Measurement_set -> Individual_Measurement --> Treated_files.

Readme file is added to each raw data file containing:

- Name of sample and measurement set
- Name of researcher
- Details of analysis
    - Methods
    - Software

DATA QUALITY

Data is deemed to be of high enough quality if:

- All spectral features are visible and distinguishable
- Dynamics of features are able to be fit with a reasonable chi-square goodness of fit

As data is measured on CCD in wavelength bins, the data quality is limited by the quality at the edge of the wavelength range. The quality needs to remain good enough over a large enough range that the two requirements can be met.

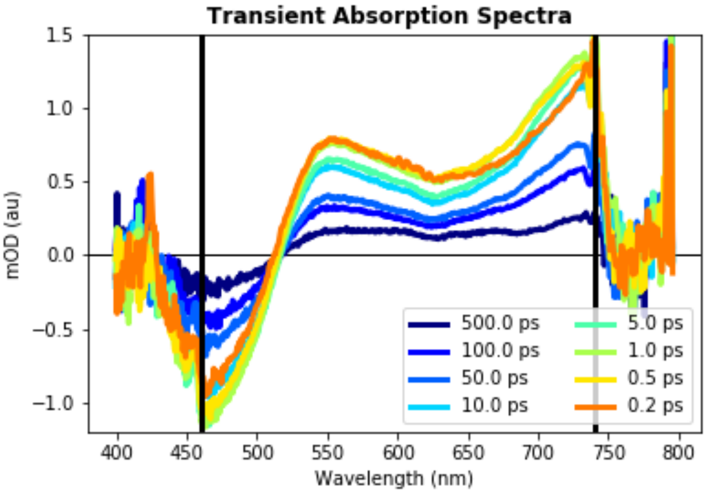
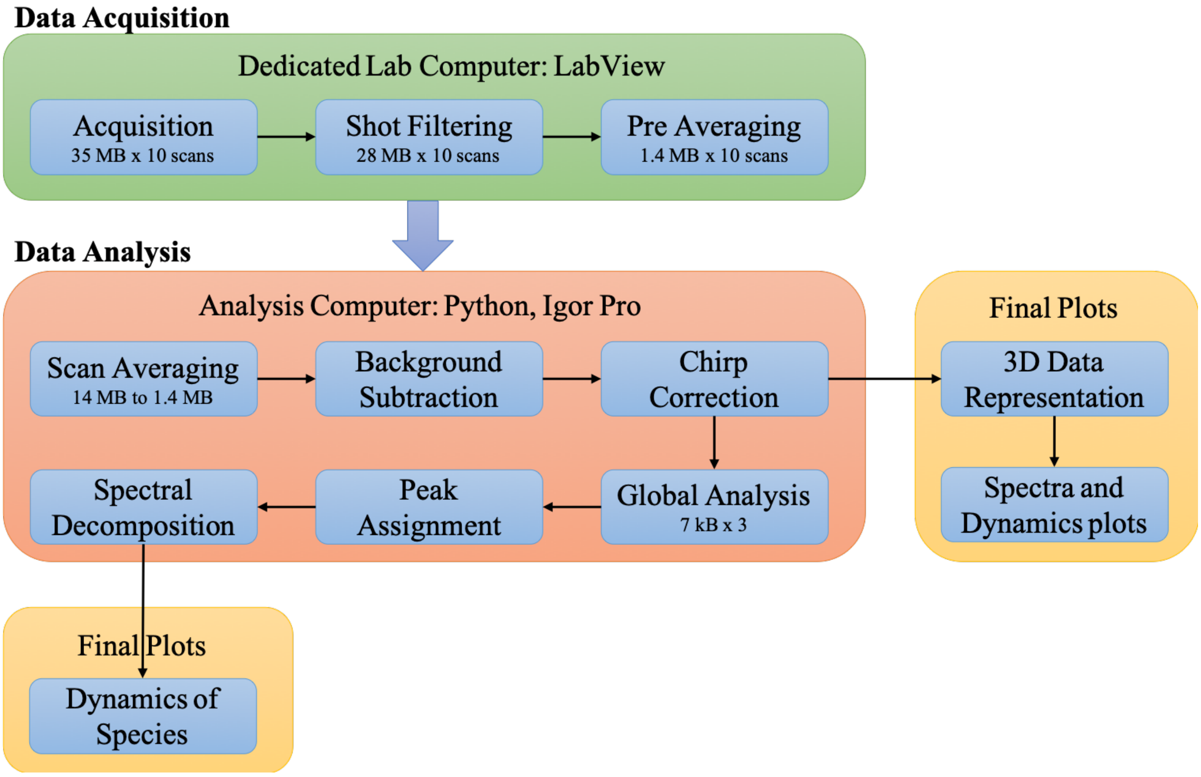


Figure 4: *Figure showing TA Spectra at different time delays (colors) with black vertical lines indicating the spectral region that contains acceptable data quality.*

In this case the data quality criteria are met as all relevant spectral features are at least partially represented in a way that can be well fit. If this were not the case then the results of the Global Analysis (fitting) may have skewed time constants and the spectral decomposition would not represent all species present in the process.

If data quality is not high enough solutions could include troubleshooting the experimental setup, but mainly increasing the number of scans preformed so that the averaging results in a mean that is closer to the real value.

DATA FLOW



Data Acquisition:

Data is acquired using a pixelated Charge-coupled Device (CCD) camera which records light intensities for wavelength bins. Each bin records around 4000 shots (4 seconds from 1 kHz pulse Laser), these shots are filtered for extraordinary outliers and the mean value is taken as the value for that pixel. This is done for each pixel of the CCD and then for each time delay. The range of time delays are scanned through on average 10 times and each scan is saved separately as tab delimited .txt files of around 1.4 MB each. Each scan is saved separately in order to more easily identify and remove anomalies. Data is saved on dedicated University of Bern servers (NAS DS916+ from Synology).

Data Analysis:

- Data is retrieved from the server
- Loaded into a numpy array in Python
- All scans are averaged into one 2D matrix **M[Wavelength, Timedelays]**
- Background calculated and subtracted
- Chirp correction is applied. The cleaned data is then saved again as a tab delimited .txt file.
- The data is then plotted as either a 3D representation using Matplotlibs contour function or in slices along wavelength or timedelay axis.

- Data is then loaded into Igor Pro for Global analysis fitting, this produces 3 or 4 amplitude spectra associated to each decay constant of size no more than 7 kB.

- The amplitude spectra and contour plots are used for peak assignment, either manually or potentially using a machine learnt algorithm.

- Spectral components either from the amplitude or normal spectra are used to perform a spectral decomposition from which the dynamics of spectral species can be plotted.

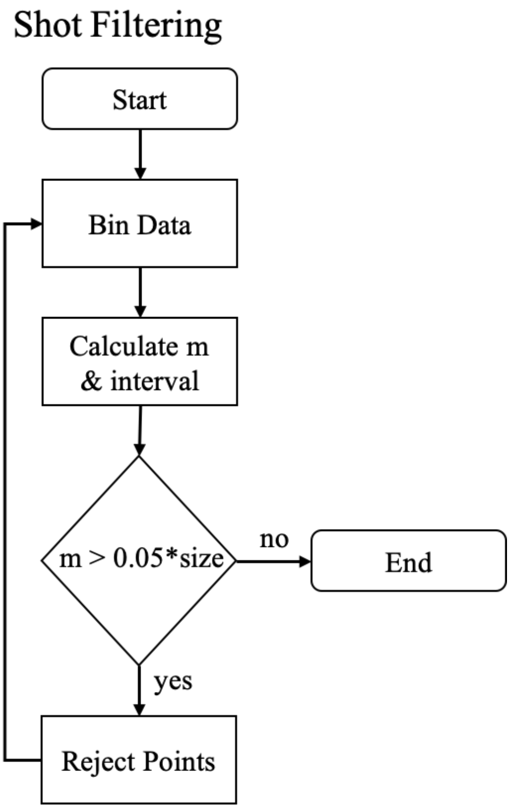
The data flow (analysis) can be seen in the accompanying Jupyter Notebook along with explanations.

DATA MODELS

Models for signal processing (shot filtering):
The model developed by Anderson et al.[1] takes the signal and puts it into a predefined number of bins. The mode of the bins is then determined with the population of that mode being denoted as m. A selection criteria is then determined by:

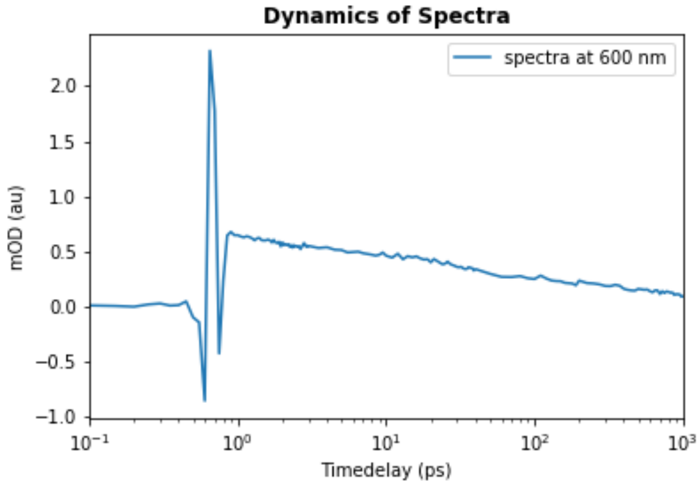
$$n = \sqrt{2}\, erf^{-1}\left(\frac{x-m}{x}\right)$$
$$M - n\sigma < M < \ \ M - n\sigma$$

Where erf^{-1} is the inverse error function, x is the number of shots and n is a scaling factor. M is the mode value and σ being the standard deviation of the distribution. If the m value is still larger than 5% of the original size then the points outside of the interval are rejected and the data is re-binned. If the binning of the data (with constant number of bins) results in a m value that is smaller than the 5% of the original size of data then the signal filtering stops.



Chirp Correction:
In order to maintain temporal resolution, the spectra are taken in a single shot by a broad 'white light' pulse ranging from 450 nm to 730 nm. This white light goes through several optics (mirrors, filters, lenses etc.) and because the refractive index of these optics is dependent on the wavelength of light passing through, the pules is lengthened and becomes chirped. When a pulse is said to be chirped this means that the wavelengths are not equally distributed throughout the pulse and so the higher frequency parts (lower wavelength) of the pulse arrive at the beginning and the lower frequency (longer wavelengths) at the end of the pulse. When a spectrum is measured, with a chirped white light pulse, the temporal onset of the spectra is different for each wavelength. The time of the onset at each wavelength is measured and then fitted to an exponential calibration curve. This calibration curve is then used to make a ‘true time’ axis that is different for each wavelength. The data is then interpolated back onto the native time in a way that the spectral onset at each wavelength is the same.

Global Analysis:



As the bands in the TA spectra can be broad, there is often an overlap of signals. A way to distinguish between processes is to fit the dynamics (decay) in time of the spectra and link the features that have the same decay time constants. As seen in the figure, the dynamics of the signal consist of the Gaussian (pump pulse) convoluted with a sum of exponential decays (normal decay processes). In the Global Analysis, the dynamics of each wavelength is simultaneously fitted with a convolution of a Gaussian and the sum of 1-4 exponential decays. During the simultaneous fitting the width of the Gaussian and the time constants of the exponential decays are kept linked. The amplitudes of the exponentials with the linked time constants are then plotted back into an amplitude spectrum potentially showing spectral signatures of the different species.

Spectral Decomposition:

The goal of spectral decomposition is to take the predetermined spectra of individual species (found from Global analysis or previously determined) and to fit a weighted sum of individual spectra in a way that the full spectra is reconstructed using a non-negative linear least-square fit at each time delay. The result of this are dynamics of each individual spectra and therefore the dynamics of each species being observed (electron, hole, exciton etc.). From these dynamics we can determine lifetimes of species or conversion of one species to another (exciton to charge, or exciton to triplet, or recombination of charges or excitons), this is the ultimate goal of TAS.

Standardization:

Global Analysis and Spectral Decomposition still require a large amount of personal experience with the data. With Global Analysis, the more exponentials that are put into the fit the easier the fit is, however this is often not linked to real physical processes. Similarly the components that are put into the Spectral decomposition also require a lot of experience to see what should be put in as a component and what should not in order to make physical sense. To develop a machine learnt algorithm for setting up the conditions for these two steps would be the ultimate goal of the project. This would create a way for spectra to be rigorously analyzed so that consistent and reliable results can be produced.

RISKS

Experimental:
Experimental errors are always a risk in data collection, especially in calibration of spectrometers which can lead to errors that are not easily caught by eye. The impact of experimental errors would be wrong data and ultimately wrong or misleading scientific results. This can only be mitigated by being as experimentally rigorous as possible in terms of following scientific methods and protocols.

Data loss:
Raw or processed data can be lost due to system crash or improper saving. This would result in having to redo experiments. To protect against data loss it is automatically synchronized (continuously), to an internal data storage server (NAS DS916+ from Synology, 10 TB capacity), which is managed by the IT Service of the Department. On the NAS, the raw data is read-only. Data is copied onto analysis computer were all folders are also synchronized to the NAS and can be accessed there with a password. The entire NAS is automatically backed up daily to a second University building (3 backups are kept at the time). The server is also protected against digital threats (firewall).

Analysis/Automation errors:
Errors in analysis can result in misleading scientific results. Similarly to experimental errors, these can be reduced by following methods rigorously and mainly by frequent data visualizations that help catch data processing that is not going as expected.

PRELIMINARY STUDIES

Preliminary studies are shown in a Jupyter notebook that is submitted along with this report. Module 2 was done on a different topic and so there is no poster to present.

CONCLUSIONS

Although no scientific conclusions were drawn from this project, we can see that a standardized procedure for signal processing, data cleaning and presentation can be done in a rigorous and efficient way. The standardization of the data processing leads to more reliable and consistent results.

The more complex analysis was described with the potential for future standardization of this analysis discussed in order to further improve both the efficiency of the measurements as well as their scientific rigor.

REFERENCES

[1] Rev. Sci. Instrum. 78, 073101 2007

[2] J. Phys. Chem. Lett. 2018, 9, 1885−1892

Appendix

Data management plan (DMP)

1 Data collection and documentation
1.1 What data will you collect, observe, generate or reuse?

Spectroscopic data in .txt or .csv files will be produced (about 1 Terabyte per year)

- Transient absorption spectroscopy data

1.2 How will the data be collected, observed or generated?

- The other experimental data will be collected from home-built spectroscopy setups running on Labview acquisition programs. We have tested the validity of the experiments and established protocols for data acquisition in years of experience in the field.
- Experiments are typically repeated twice to confirm repeatability.
- Folders containing raw data on each experimental setup are organized as Year -> Researcher_name -> Date -> Filename.
- Filenames of raw data have the form: Experiment_sample_conditions_trial. For example for a transient absorption experiment done for the second time on Perylene with excitation at 400 nm and a fluence of 100 nJ: TA_Per_400nm_100nJ_2.
- Data is treated (baseline subtraction, chirp correction...) and analyzed (least- square fitting of dynamics) using Matlab, Python and IgorPro software.
- Data will be treated and analyzed on the individual computers of the researchers in a folder structure: Sub-project -> Technique -> Measurement_set -> Individual_Measurement --> Treated_files.

1.3 What documentation and metadata will you provide with the data?
For each set of measurements (typically collected over the duration of 1 day), there will be three types of documentation:

- The lab book of the researcher (paper format) containing all information about the sample preparation, the detailed optical alignment of the spectroscopic setup and the precise measuring conditions. This is cross-referenced to the raw data via the date and file names.
- A Readme file that contains the date, name of the measurement set, name of the researcher, sample details and measuring conditions is added to each folder containing the raw data collected during a set of measurements. Abbreviations in the file naming are defined here, a

reference to the lab book (volume, page) is given and conditions to access the data are indicated.

- A Readme file containing the details of data treatment and analysis (procedures, software used) is added to the folder containing the analysis on the individual computers of the researchers.

Moreover, for published data, the details of the experimental techniques and data analysis are described in the experimental part of the publication.

2 Ethics, legal and security issues

2.1 How will ethical issues be addressed and handled?

There are no ethical, legal or security issues. The project does not involve any personal data, sensitive data or data subjected to a confidentiality agreement.

2.2 How will data access and security be managed?

All electronic data is deposited on a group-internal data storage server (NAS from Synology). This can only be accessed with a password and if permission is granted from the administrator (Research Group Leader) for access to specific folders. All computers of the group (individual ones and for data acquisition) are also password-protected.

2.3 How will you handle copyright and Intellectual Property Rights issues?

The research is not expected to lead to patents. IPR issues will be dealt with in line with the institutional policy of the University of Bern. As the data is not subjected to a contract and will not be patented, data related to scientific publications will be released as open data under Creative Commons CC0 license (as described in 4.1).

3 Data storage and preservation

3.1 How will your data be stored and backed-up during the research?

The raw data is initially collected and stored on desktop computers linked to the individual spectroscopic experiments (about 500 GB storage capacity per system). Data is automatically synchronized (continuously), to an internal data storage server (NAS DS916+ from Synology, 10 TB capacity), which is managed by the IT Service of the Department. On the NAS, the raw data is read-only. Researchers copy their data to their individual computers, where they do the data analysis. All their folders are also synchronized to the NAS and can be accessed there with a password/permission by the concerned group member and the administrator. The entire NAS is automatically backed up daily to a second University building (3 backups are kept at the time). The server is also protected against digital threats (firewall). Lab books are stored in the laboratory and remain in the laboratory when researchers leave the group.

3.2 What is your data preservation plan?

- Since the volume of our data remains in a reasonable range, there is no necessity to delete any data, and there is no legal obligation to destroy data.
- All data from the completed project will be preserved/ archived on our NAS system (where storage capacity can be added if necessary) for a period of 20 years.
- Raw data will be archived in .txt format with corresponding .txt Readme files documenting the data (see 1.3), which should permit to reanalyze/reinterpret the experiments also in a far future.
- Data of highest relevance and potential value for re-use (data from successful experiments on which publications are based) will be transferred to a repository as described under 4.1.

4 Data sharing and reuse
4.1 How and where will the data be shared?

Data of highest relevance from successful experiments, which is used in scientific publications, will be deposited for re-use (public access, CC0 license) on the BORIS Publication Repository of the University of Bern (and the BORIS Research Data Repository). Each data set will be identified with a DOI and will contain the raw as well as the treated data with the corresponding Readme files giving information on the data set, measurement details and data treatment (see 1.3). The Supporting Information of the corresponding scientific publication will contain a link to the repository and data DOI.

4.2 Are there any necessary limitations to protect sensitive data?

For large spectroscopic data sets, we sometimes analyze different aspects of the data over a longer period of time and therefore re-use the raw data for several scientific publications. Therefore, if the data set has not been completely exploited from our side, we might wait up to one year before making it publicly available on BORIS under CC0 license.

4.3 All digital repositories I will choose are conform to the FAIR Data Principles.

Yes

4.4 I will choose digital repositories maintained by a non-profit organization.

Yes