# Unravelling the kinetics of time-resolved spectra by Matrix-Factorization without separability assumption and by Markov State Modeling with PCCA+ projection

**Renata Sechi**

Master thesis
Department of Physics
Freie Universität Berlin
July 10, 2021

Supervisors:
Prof. Dr. Karsten Heyne
PD Dr. Marcus Weber

# Unravelling the kinetics of time-resolved spectra by Matrix-Factorization without separability assumption and by Markov State Modeling with PCCA+ projection

Renata Sechi

July 10, 2021

## 1 Introduction

The dynamics of photo-activated molecular processes, such as bond breaking, bond formation, photocycles, can be resolved experimentally with time-resolved spectroscopic techniques. These techniques monitor the change of the optical signal (emission, absorption, scattering) in the excited sample as function of time. Today, these powerful methods have a large impact and broad application in the current research, since they can resolve processes in femto-picosecond time range. However, the explanation of the collected data and thus the interpretation of the detected process still requires a deeper understanding.
The datasets obtained with these spectroscopic techniques are very interesting, because they yield information about the processes happening in the reaction and the process dynamics. A process monitored by these spectroscopic techniques can be for example the excitation from the ground state to the second higher energy level.
This work studies the photoreaction of macromolecules (corroles), induced after excitation in the visible range and detected also in the visible range. These energies ranges allow to observe the photoreactions of the electronic energy levels of the molecules.
After these sentences it is easy to think of an energy landscape with a electron and jumping into the other energy levels (possibly, the whole drawn as filled circles and parabolas on the blackboard). That picture is good to understand the concepts of excitation and of energy levels, but it is hard to understand the results of the experiments with only that in mind. The samples in the experiments are molecular ensembles, which implies that, for whatever reason, electrons in a molecule will not react as the other ones. Hence, several processes will occur with different velocity or intensity, depending for example on the orientation of the molecule w. r. t. the excitation beam. Also the detection method and its accuracy play an important role in the interpretation.
The information obtained by the measurement of the reaction in time-resolved spectra is fascinating, because they detect very fast molecular processes and make them somehow "tangible". It is exciting to think of how to develop new methods to interpret the dynamics in the data. In particular, this study wants to answer to the questions:

1. which are the dominant processes in the reaction mechanism?
2. how to understand how dominant processes relate to each other, without making any previous assumption for that?
3. how probable is a reaction pathway?
4. how do dominant processes decay in time?
5. how much does the past of reaction influence the future steps?

In the language of physics and for mathematics, these questions translate in respectively, (1) to find a way to project the dynamics into a fewer subspaces that can still represent the main processes; (2) to compute a transition matrix, describing the transition probability between the dominant processes;(3) to read the transition pathways from the transition matrix without describing a dynamic model;(4) to compute a transition rate matrix; (5) to estimate the non-Markovian behavior (memory).

The main idea of the presented analysis methods is that the spectrum at time $t$ does not depend on the spectrum at time $t-1$. The following work analyzes time-resolved spectra by applying and developing the basis of two frameworks. The first method is called Matrix Factorization with PCCA+ (MF with PCCA+) and it is an application of the Non-Negative Matrix Factorization without Separability assumption [9, 29]. The second method consists on the computation of a Markov State Model and its projection with PCCA+ (MSM with PCCA+). Both Matrix Factorization and Markov State Modeling yield an estimation of the kinetic model of the studied system, without choosing a the model a priori for the kinetics.

This thesis introduces first the experimental method of transient absorption spectroscopy in the visible range; then, Global and Target Analysis, one of the most applied analysis-methods for time-resolved spectra is presented. The following sections will present the MF with PCCA+ (4.2), the MSM with PCCA+ (4.3), the application of these methods to artificial data. The last sections present heuristic methods for the estimation of the memory (5.2, 7), and the transition rate matrix (8). Finally, all the developed analysis tools are applied to the analysis of the Brominated Al-corrole and Sb-corrole transient absorption spectra.

# 2 Principles of pump-probe spectroscopy

The aim of this work is to develop a framework for the analysis of pump-probe spectra. These are time-resolved spectroscopic data. In the following, time-resolved spectroscopy is briefly introduced, as well as the principles of the pump-probe measurements.

## 2.1 Time resolved laser spectroscopy

Time-resolved measurements belong to spectroscopic techniques that enable to measure the time evolution of emission, absorption, scattering processes in a sample. Hence, it is possible to resolve dynamics of physical and biological systems and to analyse kinetics of chemical reactions [1]. Moreover, with ultrafast-pulsed lasers one can observe real-time processes in picoseconds and femtoseconds time-scale. However, the rapidity of laser pulses is continuously enhancing, and today the shortest generated laser pulse is in the attosecond ($10^{-18}$) range [10].

Datasets obtained with time-resolved spectroscopy are multiway data, more specifically two-ways data [27]. Each datapoint has two independent components: a spectral variable, such as the wavelength $\lambda$ or the frequency $\nu$; and as second component the delay time between two pulses or any other measure for the time-development of the spectral variable. These pairs of coordinates characterize, for example, the change of absorption in a sample after photoexcitation.

Broadly diffused time-resolved techniques are time-resolved fluorescence decay, Raman scattering and transient absorption (pump-probe) spectroscopy.

## 2.2 Pump-probe spectroscopy

Pump-probe spectroscopy is an experimental method that investigates femto and picosecond dynamics, overcoming the time-resolution limitation of the detectors ($10^{-10}$s-range).

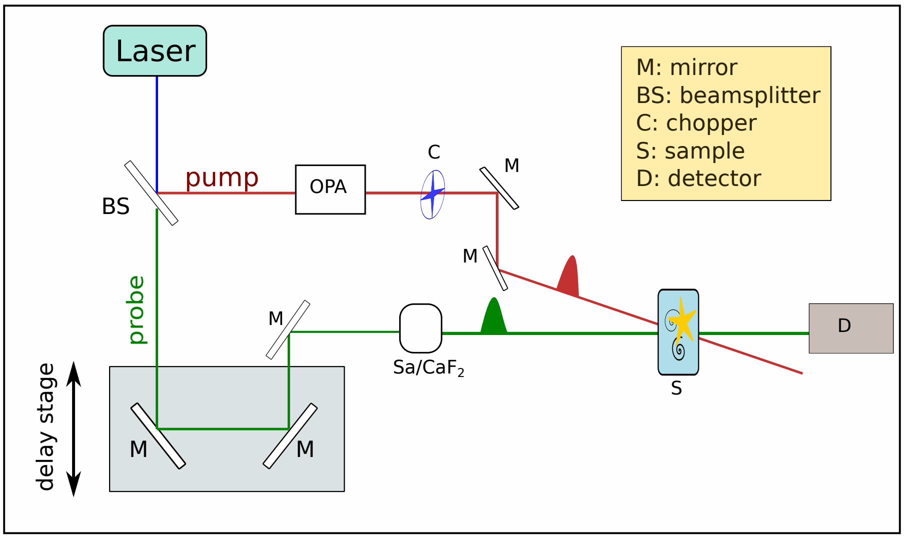


Figure 1: Schematic experimental setup of the pump-probe experiment.

A change in the sample is initiated by photoexcitation and the change is probed as function of the delay-time between the initiated change and the probe time. In the schematic experimental set up in figure (1), the laser beam is divided in pump beam and probe beam by a beamsplitter. The pump and the probe beam take different optical paths, so that the pump beam has a shorter path and reaches the sample first. The probe beam has a delay-time with respect to the pump beam of $t_{delay} = \Delta x/c$, with $\Delta x$ being the optical-path length difference and $c$ the velocity of light. The pump beam excites the sample; the photo-induced dynamics is monitored by measuring the signal with a probe beam at different delay-times. Different delay times are set by moving the delay stage, see fig. 1. One can pump and probe in different spectral ranges to excite the sample with a specific wavelength and to detect a broad range of wavelengths. Different excitation and probing ranges can monitor different processes. For example, electronic transitions are typical of the visible spectral range, whereas vibrations are in the infrared. Time resolution of the experiment depends on the pulse duration. A chopper rotates and stops the pump pulse with regular frequency. The measurements alternate two phases: 1. change of absorption of the non-excited sample (probe detection), 2. change of absorption by exciting the sample with the pump-pulse.
The pump light excites the sample with at a wavelength $\lambda$ and so the transmitted signal is

$$A_{pump} = -\log_{10} \frac{I_{pump}(\lambda)}{I_0}. \tag{1}$$

The probe light instead is a signal that is dependent on the delay-time $t$, so that

$$A_{probe} = -\log_{10} \frac{I_{probe}(t,\lambda)}{I_0}. \tag{2}$$

The detected change of absorption $\Delta A(t, \lambda)$ of a process with $r$ components is then given by:

$$\begin{aligned}\Delta A(t,\lambda) = A_{pump} - A_{probe} &= -\log_{10} \frac{I_{pump}}{I_0} + \log_{10} \frac{I_{probe}(t,\lambda)}{I_0} \\ &= d \sum_{j}^{r} \Delta\epsilon_j(\lambda) \Delta c_j(t), \end{aligned} \tag{3}$$

where $d$ is the thickness of the sample, $\Delta\epsilon_j(\lambda)$ is the difference of the extinction coefficient of the $j$ component of the sample with and without pumping, $\Delta c_j(t) = c_j(t) - c_j(t=0)$ is the difference of the concentration of the $j$ component at between time zero and delay-time $t$. In most cases, the $I_0$ is assumed to be the same for the pump and probe signal. Then, $I_0$ is canceled out in equation 3. The absorption is computed with Beer-Lambert-law [17].

Figure 2 illustrates the principal electronic processes occurring in the system during the experiment. The measured absorption change at the detector can be positive or negative. The main signal contributions are:

- Ground State Bleaching (GSB): negative signal; the pump pulse excite the molecules so that the ground state is depopulated. The absorption of the ground band decreases and more light arrives at the detector. The spectral shape of the GSB is the negative constant absorption spectrum of the sample. It appears instantaneously.

- Stimulated Emission (SE): negative signal; the first excited state is populated. The SE-signal decays with the depopulation of the excited state. Its signal has the shape of fluorescence, since it is typical of the emissive excited state (Kasha's rule). The emission of photons increases the pulse intensity and the detector measures more light intensity. It appears usually in the early stages of the photoreaction, but delayed fluorescence can occur as well.

- Excited State Absorption (ESA): positive signal; after population of the excited state, the incoming beam is absorbed and less light reaches the detector. It appears instantaneously.

- Product Absorption/Photoproduction (PP): positive signal; absorption increases and less light reaches the detector. The absorbed spectral range is new because of the product formation. It appears after decay of the SE but not total recovery of the GSB band. The presence of this signal can indicate a formation of a new molecular compound or a triplet state.

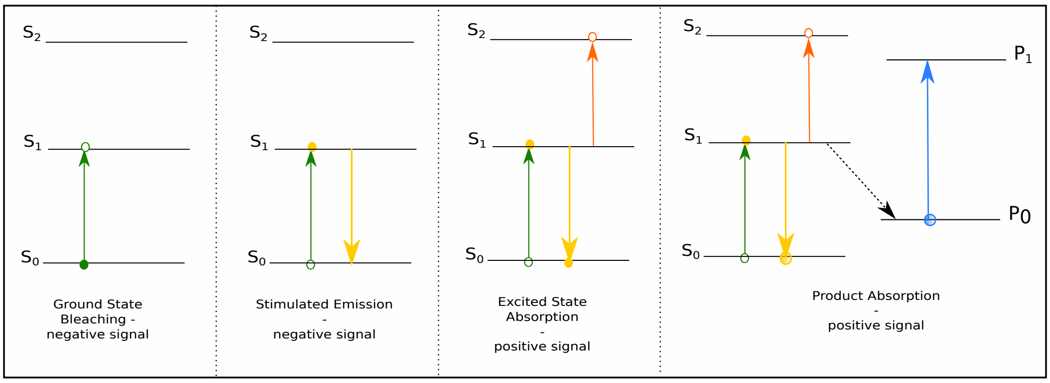

$S_2$
$S_1$
$S_0$
Ground State Bleaching - negative signal
$S_2$
$S_1$
$S_0$
Stimulated Emission - negative signal
$S_2$
$S_1$
$S_0$
Excited State Absorption - positive signal
$S_2$
$S_1$
$S_0$
$P_1$
P0
Product Absorption - positive signal

Figure 2: Main signal contributions in the pump-probe experiment.

## 3 Global and Target analysis: solving the mixture analysis problem

This section introduces Global and Target Analysis (short GloTarAn), a broadly diffused methodology to analyse time-resolved data. Time-resolved datasets collect the absorption change as a function of time. The overall change in the system is given by the contribution of a number $r$ of different components that variate under photo-excitation. The signal that is measured is a combination of the variation of the components in time. The task is then to analyse this "mixture", that is understanding what these components are and how much their change affects the system. Answer to these questions is the understanding the reaction of the system.

This problem of multivariate curve resolution can be modelled in different ways [4], Global and Target Analysis follows its simplest formulation and is a commonly routine method. Global Analysis is a tool that reduces the dimensionality of the mixture problem by identifying the main spectral components of the dynamics and their amplitude in time. The method assumes that spectroscopic data matrices can be fitted as a superposition of a number of decay exponential functions (time dependent) and their amplitudes, called Decay Associated Spectra [27, 32] (DAS). Global Analysis describes the data as a matrix that can be decomposed in

$$M = CE^T + D, \tag{4}$$

where $C$ describes the concentrations as function of time, $c_l(t) = exp(-t/\tau_i)$, with $\tau_i$ being the decay time of the $i$-component; $E$ describes the amplitudes spectral components $i$ as function of the wavelength, and $D$ describes the noise or other features such as offsets. The kinetics assumed is a first order kinetics. With fitting algorithms, such as Levenberg-Marquard or Alternate Least Squares, it is possible to identify the matrices and fit the decay times $\tau_i$.

Because of multi-exponential decays in the process, the global analysis is not enough to describe the data or to provide a sufficient understanding of them. This is because Global Analysis does not give information for the dynamics inter-components; as consequence, different dynamics are described in the same way. For example, Global Analysis is not sufficient to know if two spectral components decay sequentially ($A \to B$) or simultaneously ($A \to, , B \to$). The solution to this problem is to assume a *target model* for the dynamics. A target model is a particular model for the process, that describes parallel mechanisms, and sequential mechanisms and combinations of them for complex systems.

# 4 Two new perspectives for the data analysis

Matrix Factorization with PCCA+ and Markov State Modelling have been not yet applied in the context of pump-probe spectroscopy. Both Matrix Factorization Method with PCCA+ and Global and Target Analysis model the spectrum as a bilinear combination of spectral components and their contribution in time. In contrast to the standard assumption [27], MF with PCCA+ does not assume separability of the components and the outcome of the analysis is mainly determined by the optimization of an objective function. The objective function scores how much the known structural properties of the system have been satisfied by the approximation.

With respect to Global and Target Analysis, Markov State Modelling designs a stochastic dynamics from the dataset and does not assume *a priori* a kinetic scheme for the dynamic of the compounds.

## 4.1 Context for the concepts of configurations and compounds

When combining different methods from different scientific fields, it is useful to clarify the meaning of the adopted terms. The table 1 summarizes the considerations of this paragraph. In the section 3, the analysis is introduced talking about spectral components. Often the literature rather talk about *spectral species* or just *species* of the system. In the following, for MF with PCCA+ and MSM refer to *compounds*, since the outcome of these methods is not always a *chemical species* or a *spectral species*.

In the context of MF with PCCA+ and MSM, the *conformations* represent precise conditions of the system. The term microconformation or shortly conformation is used to identify a subset of the division of the conformation space in MSM. The so-called *dominant conformation* is a macroconformation that represents an important process for the system. The macroconformation is obtained by clustering of the conformations. The following text will explain further how to identify the dominant conformations (see 5.1). The analysis considers both microlevel and macrolevel processes; conformations live in microlevel and dominant conformations live in macrolevel dynamics.

| term | GloTarAn | MF with PCCA+ | MSM |
|---|---|---|---|
| compound | species/spectral compound | compound, component | compound |
| discretization | – | – | conformation/ microconformation , dominant conformations |

Table 1: Summary of the used terminology and the equivalent in the literature.

## 4.2 Matrix Factorization without separability and positivity assumption (MF with PCCA+)

This method is based on the *Non-Negative Matrix Factorization without the separability assumption*, described in [9]. The matrix is assumed to be factorizable in the multiplication of a matrix $W$, describing the dominant conformations and a matrix $H$, describing their relative concentration proportions in the system. Before going into details, it is important to point out that the *compounds* described by this method in the matrix $W$ do not have to be the *species* identified by Global and Target Analysis. Furthermore, the $H$ matrix do not represent the concentration of the compounds, but represents the relative contribution of a compound as function of time.

As previously mentioned, we assume that the spectrum $M$, $M \in \mathbb{R}^{n\times m}$ is given by

$$M = WH \tag{5}$$

with the matrix $W \in \mathbb{R}^{n\times r}$ and the matrix $H \in \mathbb{R}_{+}^{r\times m}$. The spectroscopic data are measured for $m$ time points and $n$ wavelength. The matrix $W$ represents $r$ component-fingerprints as function of the wavelength. Hereby the matrix $M$ does not have to be $r$-separable. A matrix $M$ is $r$-separable if there exists a factorization for which all $r$-columns of $W$ are equal to a column of $M$ [9]. If this were the case, a compound would be present at least once as 100% of the compounds in the process described by $M$. But this is not the case in experimental data this is not the case, since one measures a mixture of compounds, because in the sample not all the molecules are in the same condition. That is why the MF with PCCA+ is an algorithm particularly applicable to the analysis of experimental data. As a difference to the algorithm presented in [9], the entries of $W$ can be negative in this application, because the absorption change can be positive and negative. Each one of the $r$-rows of the matrix $H$ represents the proportion of the $r$-compounds as function of time $t$. Since $H$ represents the proportions, its entries are required to be positive and between $[0, 1]$ . The column sum of $H$ is a partition of unity (without optimization).

The authors of [9] model the evolution of the columns in $H$ as an autonomous, discrete-time Markov Process. They theorize that for the evolution of the compounds at time $t = i$ depends only on the conditions at time $t = i - 1$, and it is given by

$$H_i^T = H_{(i-1)}^T K. \tag{6}$$

$K \in \mathbb{R}^{r\times r}$ is a row-stochastic transition matrix[1]. The model also assumes an underlying autonomous Markov Process, so that $K$ is not time-dependent.
Note that unlike in the notation of [9], here the transition matrix is called $K$ for Koopman operator (matrix in discrete case). In facts, the Koopman operator describes the evolution of observables and this fits better in order to treat the concentration proportions in $H$. The factorization of the matrix $M$ as product of $W$ and $H$ has not a unique solution. One can find a set of solutions, but not all the solutions of this set will satisfy the necessary conditions to represent the system. To select the desired decomposition from the set of solutions, a penalty function $\Psi$ is defined. This function weights the required conditions for the elements of the factorization. The optimal solution for the decomposition of $M$ is obtained by minimizing the value of $\Psi^2$. The requirements the found matrices have to meet are

- the entries of $H$ are non-negative
- $H$ is column-stochastic
- the entries of $K$ are non-negative
- $K$ is row-stochastic

The penalty function $\Psi$ is defined as:

$$\Psi = \beta\left(\underset{i,j}{min} H_{ij}\right) + \gamma\left(\underset{j}{max} \mid 1 - \sum_{i}^{r} H_{ij} \mid\right) + \delta\left(\underset{i,j}{min} K_{ij}\right) + \\ \mu\left(\underset{i}{max} \mid 1 - \sum_{j}^{r} K_{ij} \mid\right), \tag{7}$$

where the coefficients $\beta, \gamma, \delta, \mu$ before each addend allow to design an objective function that fits the data characteristics. In comparison to [9], here $\Psi$ is slightly modified, since the spectral traces in $W$ can be both positive and negative. Thus, the penalty function $\Psi$ has only four requirements terms, instead of five.

The MF with PCCA+ algorithm is based on the one proposed by the authors of [9]. To the notation, for any matrix $Y$, $Y_+$ is the matrix $Y$ without the first row and $Y_-$ is the matrix $Y$ without the last row. A data matrix $M$, $M \in \mathbb{R}^{n\times m}$ and $M$ of rank $r$ is given.

- Singular Value Decomposition (SVD) of $M$ transposed: $M^T = U\Sigma V^T$.
- Define $\tilde{U}$, $\tilde{U} \in \mathbb{R}^{m\times r}$: the first column is the constant vector $(1, 1, ...1)^T$, the other columns are the first $(r-1)$ columns of $U$.
- Use PCCA+ to find $\tilde{H} = (\tilde{U}A)^T$.
- Use the Penrose-pseudoinverse to compute $\tilde{W} = M\tilde{H}^{-1}$ and $K$= $(\tilde{H}_-^{-1})^T\tilde{H}_+^T$
- minimize $\Psi$ for the requirements in order to find the optimal $A_{opt}$.
- reconstruct the proportions and compounds matrices with via $A_{opt}$: $H_{rec} = (\tilde{U}A_{opt})^T$, $W_{rec} = MH_{rec}^{-1}$, $K_{rec}$=$(H_{rec,-}^{-1})^T\ H_{rec,+}^T$.

The matrix $K$ is computed as the autocorrelation matrix between the $\tau$-time-shifted proportions $H_-, H_+$. With this relation, the matrix $K$ has the meaning of a Markovian transition matrix, since gives information on the $\tau$-step development of the concentration proportions.

[1] Only for readibility reasons, in the presentation of the algorithm, the dependence of $K$ on the lagtime $\tau$ is dropped. The lagtime $\tau$ is the time difference between $t = i$ and $t = i - 1$.

For simplicity, when presenting the results in the examples, the optimized quantities $H_{rec}$, $W_{rec}$,$K_{rec}$ will be referred to as $H$, $W$, $K_{MF}$.

This decomposition method allows to analyze experimental data with different structure, because the parameters of the objective function $\Psi$ can be adjusted to weigh more or less a feature rather than another. Therefore, setting the parameters in one or another way influences the final results of the decomposition, or better: one is going to find the best decomposition for those parameters, which can be very different to another one found with other parameters in $\Psi$.

## 4.3 Markov State Model (MSM)

In the following, time-resolved spectra are considered as trajectories evolving in time. The trajectory develops its dynamics in a $n$-dimensional space, where every wavelength $\lambda$ of the spectrum is considered to be a dimension of the space. A common framework to analyze systems dynamics is the Markov State Model(ing), short MSM. Two very good review on MSM are [12] and [16] .
Markov State Models are widely applied in the field of the molecular simulations, where they are almost the standard approach. At the base of the MSM is the Markov property: the system is memoryless and its time-development is determined only by the present conditions. MSM solves that the dynamics based on the master equation,

$$K(\tau) = \exp(\tau Q), \tag{8}$$

that thus describes a first order kinetics.

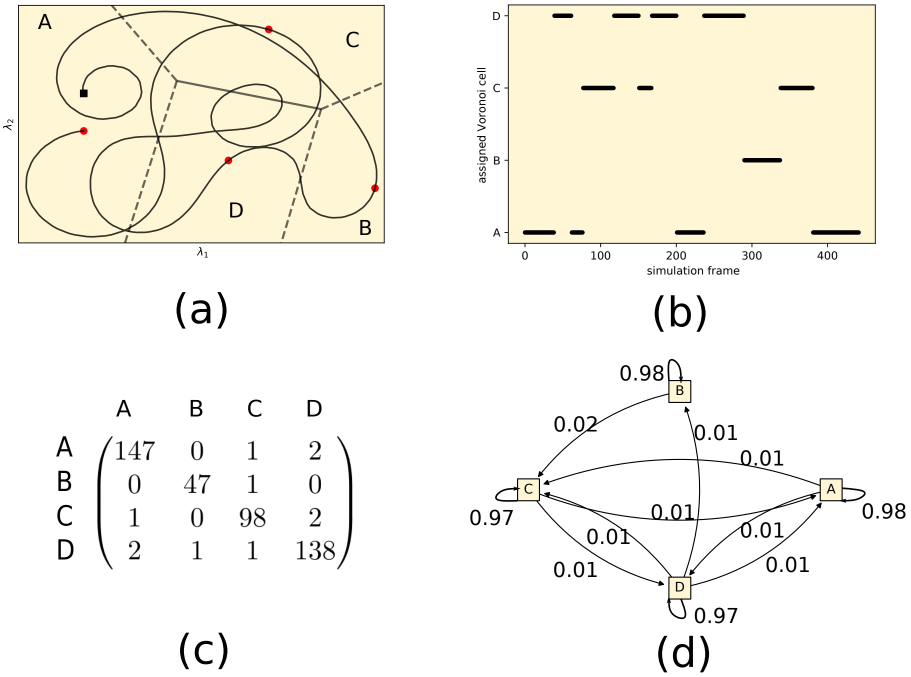
The idea of MSM (fig 4.3) is to divide the state-space in which the system (or simply a trajectory) lives into a $k$ number of micro-conformations. With this division, it is possible to measure how often the system evolves in time going from one of the microconformations to another. To count "how often" the system jumps between the microconformations, one also need to define a regular lag time $\tau$ for the counting. The requirement is that $\tau$ should be small enough to show the development of the process, but big enough so that any memory effect has decayed and the dynamic is Markovian. From the counting of the jumps between micro-states we can construct the so-called transition matrix $K(\tau)$. This matrix $K(\tau)$ is specifically the Koopman matrix and it is a $k \times k$-matrix whose rows describe the probability of the microconformation $i$ to go to all the other microconformations or to stay in the same state. So $K(\tau)$ is row-stochastic.
Take the spectrum $M \in \mathbf{R}^{m \times n}$, with $n$ wavelengths and $m$ time points.

$$f_{t+\tau}(a) = \left[\mathcal{K}(\tau) f_t\right](a) = \mathbb{E}[f_t(\tilde{X}_{t+\tau}) \mid \tilde{X}_t = a] = \int_\Omega p_\tau(b|a) f_t(b) d\mu(b). \tag{9}$$

In order to construct the transition matrix $K(\tau)$, discretization of the Koopman operator, one first discretize the conformation space into Voronoi cells. The picking algorithm was used to pick the centers of the Voronoi cells. So the transitions between cells are counted as "jumps". The Koopman matrix is then computed for a fixed lag time. In order to understand the system's dynamics, we further project the transition matrix with PCCA+.

The conformation space $\Omega$ in the MSM is discretized in Voronoi cells with tessellation $\Phi$. The Voronoi partition of the plane is based on a set of $n$ points $c_i \in \Omega$ and a distance measure, usually the Euclidean metric. Each cell $\Phi_i$ is the region of space in which all the points are closer to the center of $c_i$ of $\Phi_i$ than to the other centers $j, j \neq i$. A well-known case of a Voronoi tessellation in solid-state physics is the Wigner-Seitz cell, but this kind of tessellation is very common in natural science and not only[2]. In the analysis of the

[2] For example, the pastry chef Dinara Kasko uses the Voronoi tessellation to shape her cakes!



$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 147 & 0 & 1 & 2 \\ B & 0 & 47 & 1 & 0 \\ C & 1 & 0 & 98 & 2 \\ D & 2 & 1 & 1 & 138 \end{array}$$

Figure 3: Markov State Model. (a). The state space of the trajectory has been divided into 4 conformation, here Voronoi cells (A,B,C,D). (b). Each point of the trajectory is assigned to a conformation. (c) a count matrix for the transition between the Voronoi cells is computed. Finally, the matrix is row-normed to obtain the transition probability matrix, obtaining the MSM. The possible transition pathways described by the trajectory can be now understood by reading the transition probability matrix (d).

| | GloTarAn | MF with PCCA+ | MSM |
|---|---|---|---|
| Separability | yes | no | no |
| Kinetic Model | yes | no | no |
| Stochastic | no | no | yes |
| Fitting & Optimization | yes (Levenberg-Marquard, alternate least squares) | yes (objective function) | no |
| time-dependent ... | concentrations | concentration proportions | membership functions |
| Concentration of | chemical species | compounds | physical system's configurations |
| Advantages | -direct interpretation of the results<br>-broadly applied | -provides a stochastic model - direct interpretation of the results | -provides a stochastic model<br>-estimation of infinitesimal generator |

Table 2: Summary of the main features of the methods.

spectral datasets, the Euclidean metric can be weighted by the energy of the dimension. Considering each wavelength $\lambda_k$ in nm as a dimension for the distance metric, for each spectrum $s$ at delay-time $j$, one has

$$dist(c_i, s_j) = \sqrt{(c_i(\lambda_1) - s_j(\lambda_1))^2 + ... + (c_i(\lambda_n) - s_j(\lambda_n))^2}$$

in weighted case, each dimension is weighted such that $w_k = 10^7/\lambda_k$

$$dist_w(c_i, s_j) = \sqrt{w_1(c_i(\lambda_1) - s_j(\lambda_1))^2 + ... + w_n(c_i(\lambda_n) - s_j(\lambda_n))^2}$$

The conformation space $\Omega$ is finite, but high dimensional. As a result, the construction of a grid to discretize it is not possible or rather a very difficult task [19]. Furthermore, an analysis on the microconformations level does not provide information on the important processes of the dynamics, such as the creation of a new molecular product or the transition of the system to a spin-different electronic state. This change of conformation can be analyzed with clustering algorithms such PCCA+.

# 5 Projection of the process: PCCA+ and memory

Clustering (grouping objects) is necessary to make quantitative and also qualitative connections with experiments and experimental results. Clustering the conformations can be performed in different ways, here a spectral[3] clustering method, PCCA+, has been applied. The clustering process generates memory effects in the analysis. In the following, the spectral clustering algorithm PCCA+ is explained and the resulting memory effect is considered.

## 5.1 PCCA+

PCCA+ is a tool to connect the dynamics at microconformations-level to macroconformations -level [29], so that one can understand the overall process described by the spectra. But why doing that and how does it work in the context of spectral analysis? To answer to this question, one should look at the dataset from a different perspective.

Consider a time-resolved spectrum from different wavelengths and delay-times. The spectrum at time $t = 0$ shows positive signals and negative signals at certain wavelengths. For example, the spectrum at time zero is "+ - - +" signals. This form of spectrum at time $t = 0$ can be seen as a "conformation" for the process. For now, it does not matter what this conformation means. After a certain delay-time, a positive signal rises in place of a negative one, let's say "+ + - +". The spectrum has a new form, so a new conformation. One usually knows how to interpret this change, given a pre-existing knowledge of the system, other kinds of data etc. The two conformations are distinct and assume that they are relevant at macrolevel for the process. However, one has different spectra for many delay-times and it is difficult to assign them to one of these conformations. How much should a spectrum have a positive signal not to belong anymore to " + - - +" but to " + + -+"? To which conformation belongs a spectrum with signal "+ 0 - +"? Maybe it belongs to a third conformation, possibly to one of the first two conformations.

The described situation is very easy and simple; the problem is much more difficult. But basically what the algorithm PCCA+ does is to identify which macroconformations are the dominant ones and to assign all the others spectra with their microlevel form-changes to these dominant conformations.

[3] spectral: the clustering algorithm looks at eigenvalues and eigenvectors.

### 5.1.1 Principles of PCCA+

PCCA+ is a clustering algorithm that projects the microconformations with similar behaviour to fewer dominant or macro conformations[4]. This clustering algorithm relates microconformations, the conformation in which the spectra can be, and macroconformations, collections of microconformations grouped together by a similar feature.
From a mathematical point of view, macroconformations correspond to conformations $X$ that keep their structure upon application of the transition matrix $K(\tau)$:

$$X \approx K(\tau)X. \tag{10}$$

This means that $j$ macroconformations are eigenvectors to the eigenvalues $\ell_j \approx 1$, which are the dominant eigenvalues of $K(\tau)$. Note that transition matrices have eigenvalues $\ell_i \in [0,1], i \in [1,n]$ and that the first eigenvalue (Perron eigenvalue) $\ell_1 = 1$. Now PCCA+ uses $X$ is to score to which degree a microconformation belongs to/is a member of each one of the macroconformations. The vectors in $X$ are not membership functions yet, so the problem now is to compute the membership functions from these dominant eigenvectors.
Consider a transition matrix $K(\tau) \in \mathbb{R}^{n\times n}$ with eigenvalues $\ell_i, i \in [1,n]$ and eigenvectors $\tilde{X} \in \mathbb{R}^{n\times n}$. Solving the eigenvalue problem, i. e. $K(\tau)\tilde{X} = \tilde{\Lambda}\tilde{X}$, $\tilde{\Lambda} = diag(\ell_1, ...\ell_n)$, $r$ dominant eigenvalues are identified. Then the matrix of the dominant eigenvectors, $X \in \mathbb{R}^{n\times r}$, is the input of the PCCA+. To find the membership functions $\chi$, the algorithm has to project the matrix $X$ such that the entries of $\chi$ are not negative and form a partition of unity. That means finding a matrix $\mathcal{A} \in \mathbb{R}^{r\times r}$ such that with

$$\chi = X\mathcal{A} \tag{11}$$

the membership functions matrix $\chi$ satisfies

$$\chi_j(i) \in [0,1], i \in 1...n; \quad \sum_{j=1}^{r} \chi_j(i) = 1. \tag{12}$$

Equation 12 tells that $\chi_j(i)$ gives information about how much the $i$-th microconformation belongs to the $j$-th macroconformation [29, 19].
With the membership functions in $\chi$ the transition matrix of the macroconformations, $K^c(\tau)$ can be computed with:

$$K^c(\tau) = \langle \chi, \chi \rangle_\pi^{-1} \langle \chi, K(\tau)\chi \rangle_\pi, \tag{13}$$

with $\pi$ being the density distribution (e.g. uniform, stationary distribution). The PCCA+ projection applied in this work addresses both reversible and unreversible Markov processes, since the Schur vectors are used [20].

## 5.2 Memory-effect

In this article, the clustered Markov State Model of a spectrum is estimated via PCCA+. PCCA+ is a spectral clustering algorithm, an invariant subspace projection. This paragraph explains how the projection of the process into a finite state space and its clustering bring memory into the analysis, and how much memory affects the model. The memory effect is a consequence of the discretization of a Markov process. The process is memoryless (Markovian) in continuous space and time. When projecting the process into a finite number of conformation for the analysis, a dependence to the past is introduced. In computational molecular simulations this is called rebinding effect. When a ligand unbinds

[4] Instead of *macroconformations*, the term *metastable conformations* is broadly used in the literature of PCCA+. It refers to the long-time behaviour of the process. This is not the main topic of this work, wherefore the word *macroconformation* fits better.

from the receptor molecule, it is still near to the binding site. This spatial condition makes more likely that the molecule *rebinds* to the receptor in the next time step. So the past (the ligand unbinds) influences the future, introducing memory.
In time-resolved spectroscopic analysis, the introduction of memory is a phenomenon that occurs as well. Consider a system with two conformations $S_1$, $S_2$, see figure 5.2. Note that the "spatial condition" here is the change of absorption for each wavelength $\Delta A(\lambda)$. The conformation $S_1$ is characterized by a negative signal for lower values of the wavelengths $\lambda$, and positive signal for higher $\lambda$ values. In the conformation $S_2$, the signal is negative for every measured $\lambda$. The measured spectrum at delay-time $t = t_1$ (past) is assigned to conformation $S_1$, because the spectrum has clearly a negative and strong positive-signal range. At time $t = t_2$ the spectrum shows the decrease of the positive signal, but it can be still assigned to $S_1$. At $t = t_3$, the spectrum has still a positive signals, but mixed with negative ones. This spectrum is a mixture of both conformations $S_2\&S_1$. The spectrum at $t = t_4$ is assigned to $S_2$, since there is no positive signal. At time $t = t_5$, the spectrum is assigned again to $S_1$. Because of a short-time memory of the system, it is likely that the spectrum at $t = t_5$ shows again the features from the past (from $t_1, t_2, t_3$) than that it will become ocmpletely different in the next time-step. The $s$-conformation, the microconformation to which the system has been assigned depends on the past. These considerations are intuitive and clear by seeing the time-resolved datasets. On the macroconformations perspective, the designed model answer to the specific question:

> How probably will the system switch to $S_2$ in the next timestep, given that it is in $S_1$ now?

and not consider that the probability on the microlevel, i.e. $\Delta A(\lambda)$ for each $\lambda$, will be affected by the past position in the conformation space.

The application of PCCA+ and the definition of the membership functions $\chi$ allows to quantify and describe the memory effect. The question 5.2 can be slightly modified by defining membership functions $\chi$. This membership functions describe to which extent each microconformation belongs to all the macroconformations. In this way, the PCCA+ algorithm is applied and it will also provide a measure for the estimation of the memory effect.
Röblitz and Weber show in [19] that $K^c$, the $r \times r$- Koopman matrix projected via PCCA+, is given by

$$K^c = \mathcal{S}^{-1}\mathcal{T} \tag{14}$$

with

$$\mathcal{S} = \frac{\langle \chi, \chi \rangle_\pi}{\langle \chi, e_n \rangle_\pi} \quad T = \frac{\langle \chi, \mathcal{T}(\tau)\chi \rangle_\pi}{\langle \chi, e_n \rangle_\pi}, \tag{15}$$

and $\mathcal{S}, T, \in \mathbb{R}^{r\times r}$ with $r$ being the number of identified macroconformations. The transfer operator in continuous space is $\mathcal{T}$. The $\langle \cdot, \cdot \rangle_\pi$ is the $\pi$-weighted scalar product and $e_n = (1, 1, ...1)$ for $n$-microconformations. The weights in $\pi$ can be constant (uniform distribution) or $\pi$ can be the stationary distribution of $K(\tau)$. The stationary distribution is computed as the left eigenvector to the eigenvalue 1, i.e. $\pi = \pi K(\tau)$. The matrix $T$ represents the transition probability between macroconformations, i.e. $T_{ij}$ is the probability of going to conformation $i$ into conformation $j$ after a timestep $\tau$ [30]. The matrix $\mathcal{S}$ relates the PCCA+ projection $K^c$ to the "pure" transition matrix $T$. The matrix representation of 14 is [29, 8, 19]:

$$K^c = \mathcal{A}^{-1}\Lambda\mathcal{A}; \quad \mathcal{S} = \mathcal{A}^T\Pi\mathcal{A}; \quad T = \mathcal{A}^T\Pi\Lambda\mathcal{A}, \tag{16}$$

see sec. 5.1. Note that $\Pi$ are the weights of the scalar product (stationary distribution) in diagonal-matrix form, and that $\Lambda = diag(\ell_1, ..., l_r)$ is the diagonal matrix of the $r$ leading eigenvectors. Projecting the matrix with PCCA+, it holds [28, 5] that $K^c = \exp(Q^c\tau)$. That is, the projected Koopman matrix has an infinitesimal generator $Q^c$ (see sec. 8). The off-diagonal entries of $Q^c$ represent the rates between the dominant conformations (positive numbers) and the diagonal entries $Q^c_{ii}$ are given by

$$Q^c_{ii} = -\sum_{j\neq i}^{r} Q_{ij} \tag{17}$$

with $Q^c_{ii}$ being the diagonal entry of the matrix in the $i$-th row. $Q_{ij}$ represents the transition rates from dominant conformation $i$ to a conformation $j$ out of the $r$ dominant conformations of the process. Since the diagonal elements of the infinitesimal generator are computed as negative of the sum of the outgoing rates, the sum of all rates of the dominant conformations in the system is given by the trace of the $Q^c$. As in [29, 20, 30], one defines 14:

$$F = -trace(Q^c) = \tau^{-1}[(\ln\det(\mathcal{S})) - \ln\det(T))]. \tag{18}$$

Now, both $\mathcal{S}$ and $T$ are stochastic matrices and their determinant cannot be larger than one. If the metastable conformations are not very stable, the matrix $T$ has a lower-valued determinant, which means a high negative logarithm. The matrix $\mathcal{S}$ indicates the overlap between the membership functions of the dominant conformations ($\chi$). If the overlap is large, the determinant of $\mathcal{S}$ reduces the value of $F$. That means, the more the membership function overlap, the more the system is stable, or the lower $\det(\mathcal{S})$, the higher the memory effect. The determinant of $\mathcal{S}$, $det(\mathcal{S})$, is an indicator of the memory effect.

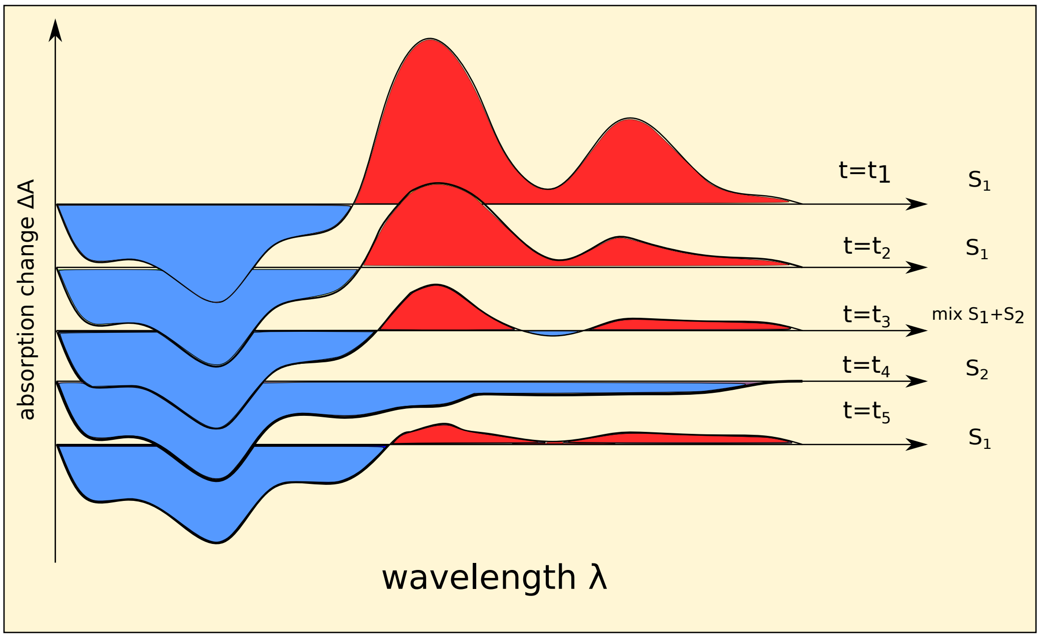


Figure 4: memory effect of high-dimensional conformation-space. Spectral signals at different delay-times show short-memory effect.

### 5.2.1 Memory effect in matrix form

In the previous paragraph, the memory effect in the context of Markovian models and of spectroscopic-data analysis has been introduced. From this follows the question of how determine by simply observation whether the analysis of the process carries some memory effect. Think of a clustered transition-matrix $K^c$ that shows some negative entries. Recalling eq. 14, eq. 15, the matrix $T$, the "pure" transitions, has by construction only positive entries. The entries are positive because the memberbership functions $\chi$ assume values between $[0, 1]$, the $e_n$ function is constantly 1. The entries of $T$ form row-wise a partition of the unity. Also the matrix $\mathcal{S}$ has by construction only positive entries because of the membership functions. However, its inverse, $\mathcal{S}^{-1}$, can have negative entries.

The multiplication $\mathcal{S}^{-1}T$ can yield a matrix with negative entries. If $\mathcal{S} = I$, the identity matrix, there is no overlap between the dominant conformations and so no short-memory effect. The $\mathcal{S}^{-1}$ is the identity matrix again and so $K^c$ has only positive entries. Hereby an example of a process that has been projected with PCCA+ in 4 dominant conformations. The clustered Koopman matrix is:

$$K^c = \begin{pmatrix} 0.919 & 0.080 & \underline{-0.017} & 0.018 \\ \underline{-0.005} & 0.969 & 0.058 & \underline{-0.022} \\ 0.003 & \underline{-0.004} & 0.982 & 0.020 \\ 0.000 & 0.000 & 0.001 & 0.998 \end{pmatrix}. \tag{19}$$

The matrix is stochastic (row-sum is 1), but there are some entries that are negative. The negative entries suggest that some short-memory effects are present in the projection. Assuming a uniform distribution $\Pi$, the overlap matrix $\mathcal{S}$ is then

$$\mathcal{S} = \begin{pmatrix} 0.688 & 0.215 & 0.042 & 0.056 \\ 0.223 & 0.401 & 0.319 & 0.057 \\ 0.024 & 0.177 & 0.503 & 0.295 \\ 0.026 & 0.025 & 0.238 & 0.711 \end{pmatrix};$$

$$(\mathcal{S})^{-1} = \begin{pmatrix} 1.890 & \underline{-1.393} & 0.924 & \underline{-0.421} \\ \underline{-1.446} & 4.692 & \underline{-3.395} & 1.149 \\ 0.534 & \underline{-1.889} & 3.842 & \underline{-1.487} \\ \underline{-0.196} & 0.516 & \underline{-1.199} & 1.88 \end{pmatrix}$$

The matrix $\mathcal{S}$ is not equal the identity matrix and its inverse shows negative values. The determinant is $det(\mathcal{S}) = 0.04$, so its close to zero. As aforementioned, a small-valued determinant indicates high memory effect.

Finally, consider that the value of $\mathcal{S}$ depends also on the experiment. The more often the spectrum is collected (so the smallest the delay-time), the bigger the discrepancy of $\mathcal{S}$ to the identity. The choice of the time-step ( how to choose a $\tau$ for computing $K(\tau)$) is a well-known problem in the analysis of simulated molecular trajectories. If $\tau$ is big, the process is Markovian, but possibly a lot of information is lost. If $\tau$ is too small, the process shows memory effects (the process is not Markovian).

# 6 Example: analysis of a computer-simulated sequential decay mechanism.

This work presents a different perspective in the analysis of time resolved spectra. The following example explains how to apply the introduced theory by analysing a sequential-dynamic decay ($A \rightarrow B \rightarrow$). Furthermore, the final paragraph considers some new feature resulting from the analysis. Appendix A discusses the application to a parallel decay dynamic, a reversible process and further consider the sequential decay. In the following, the dominant conformations are referred to with numbers.

### 6.1 Analysis of a sequential decay

The analysed process is a sequential decay of two compounds $A$ and $B$. For the MSM, the analysis starts with a discretization in 50 Voronoi cells. The combination of the analysis of the Koopman matrix and of the membership functions makes possible to understand the dynamics. The Koopman matrix for 2 dominant conformation is

$$K(\tau)_2 = \begin{pmatrix} 0.995 & 0.005 \\ 0.003 & 0.997 \end{pmatrix}. \tag{20}$$

The process goes from 1 to 2 more likely than from 2 to 1. Considering the membership to the conformations represented by the $\chi$ vectors, see fig 5(a), the process is at the beginning in conformation 2 for short time and it goes rapidly to 1. conformation 1 remains dominant during the development of the process and the system returns to its initial condition, conformation 2 ,at the end. Repeating the analysis with 3 dominant conformations, the Koopman matrix reads

$$K(\tau)_3 = \begin{pmatrix} 0.993 & 0.008 & -0.001 \\ 0.013 & 0.977 & 0.010 \\ -0.001 & 0.012 & 0.989 \end{pmatrix}. \tag{21}$$

The negative entries show memory effects, quantified as $det(\mathcal{S}) = 0.19$. The number is not almost zero, so that the memory effects are not substantial (and the modelling is appropriate). From the membership functions, see fig 6(a), the process starts equally being in conformation 1 and 2. The membership to these conformations drops abruptly and the process goes into conformation 3. From 3, the system goes back to conformation 2, which also drop and finishes back to conformation 1. Reading the matrix $K(\tau)_3$, one starts with conformation 1, from which it can only go to 2. From 2, the process can go almost equally to 1 and 3 again, but from 3 it can only go back to 2. Conformation 1 thus represent an empty conformation, conformation 2 the system mostly in compound A, then the membership to a new conformation rises (compound B is prevalent in the system), finally the system goes back to be empty (conformation 1 again).

MF with PCCA+ is applied with parameters $\beta = 100, \gamma = 10, \delta = 1, \mu = 10$ for the objective function (eq.7). The analysis has been performed for two compounds and three dominant conformations. In order to compare the two methods, the membership functions $\chi$ are considered instead of $H_{rec}$. The reconstructed transition matrix with two dominant conformations is

$$K_{MF2} = \begin{pmatrix} 9.9985e-01 & 1.4811e-04 \\ 9.8185e-05 & 9.9990e-01 \end{pmatrix}. \tag{22}$$

The transition between the conformations is very seldom, however it is more likely to go from 1 to 2. The analysis of the membership functions suggests that the system starts and ends in conformation 1, conformation 2 develops in the middle, see figure 5(b). Thus, conformation 2 can be seen as the empty-system conformation. Now the decomposition with 3 domaninat conformations yields a reconstructe transition matrix:

$$K_{MF3} = \begin{pmatrix} 1.023 & 0.026 & -0.046 \\ -0.073 & 0.916 & 0.149 \\ 0.003 & 0.004 & 0.994 \end{pmatrix} \tag{23}$$

The negative entries hereby show some memory effects. The matrix shows a process that from conformation 3 goes to conformation 1 and conformation 2; from 1 it goes to 2, from 2 to 3. The analysis of the membership functions shows that the profiles of conformation 1 and 2 are similar to the one found with the standard analysis methods, however, conformation 3 has the role of a conformation that characterizes the system when is empty.

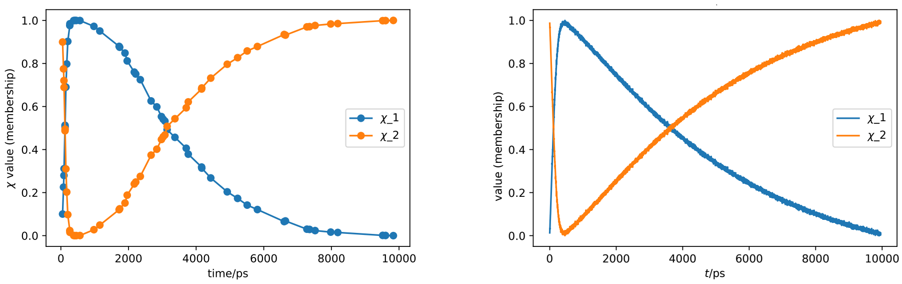
As difference to the MSM, hereby the conformation 2 is never a dominant conformation during the process, see figure 6(b). The interpretation of electronic dynamics with MF with PCCA+ and MSM is different to the interpretation with canonical Global and Target Analysis, because the analysis determines system conformations representing small processes, rather than at chemical compounds.

## 6.2 Start conformation, sink conformation

In figure 7, the analysis shows a new conformation. This conformation dominates almost entirely in process at the beginning and drops very quickly after few analysis steps. For the simple sequential-decay process with 3 dominant conformations, (sec. 6), this conformation rises again at the end, when the process has finished its cycle. This conformation is called *start conformation*.

The standard analysis methods do not identify the start conformation. In the process described in sec. 6, only two species $A$ and $B$ would be identified by the standard analysis methods. PCCA+ identifies an additional conformation because from a mathematical point of view, the so-called *start conformation* corresponds to a dominant conformation (see sec. 5.1). The start conformation has this meaning for the dynamics in relation to the other membership conformations and their time developments. In easy dynamics or in processes that are cyclic and returns to the start situation, the start conformation rises at the beginning, decays when new compounds characterize the system, and rises again at the end of the process. This means, the start conformation rises again when the process does not belong anymore to the compounds conformations, that is when the other compounds decay. The start conformation represents an empty system, in the case of the sequential process in section 6.

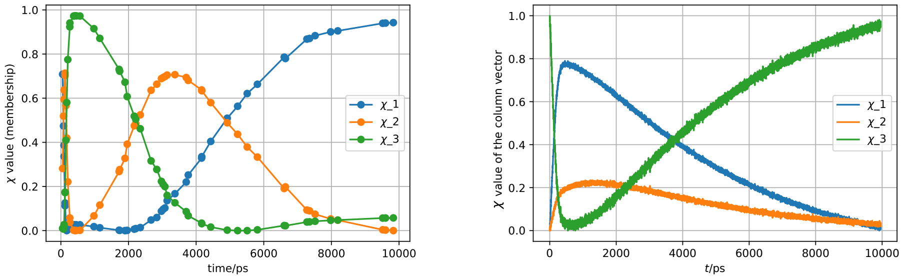
If the process returns again to its initial condition, the start conformation rises again one the reaction has happened. But in case the system ends up in a product state after the reaction, the start conformation can also decay without arising anymore. In this case, another conformation rises only at the end of the dynamics. The system does not change conformation anymore once reached this last conformation. This conformation is called *sink state* and once reached cannot be left. In other words, the sink state has only an incoming flow. The sink state can be a new-product state that is very stable and that the system does not leave. In the case of the presented example, the *start state* has the role of *sink* state.



(a) $\chi$-vectors obtained with MSM

(b) $\chi$-vectors obtained with MF with PCCA+

Figure 5: Comparison of the $\chi$ vectors computed with MSM and MF with PCCA+ (2 dominant-conformations-system).



(a) $\chi$-vectors obtained with MSM

(b) $\chi$-vectors obtained with MF with PCCA+

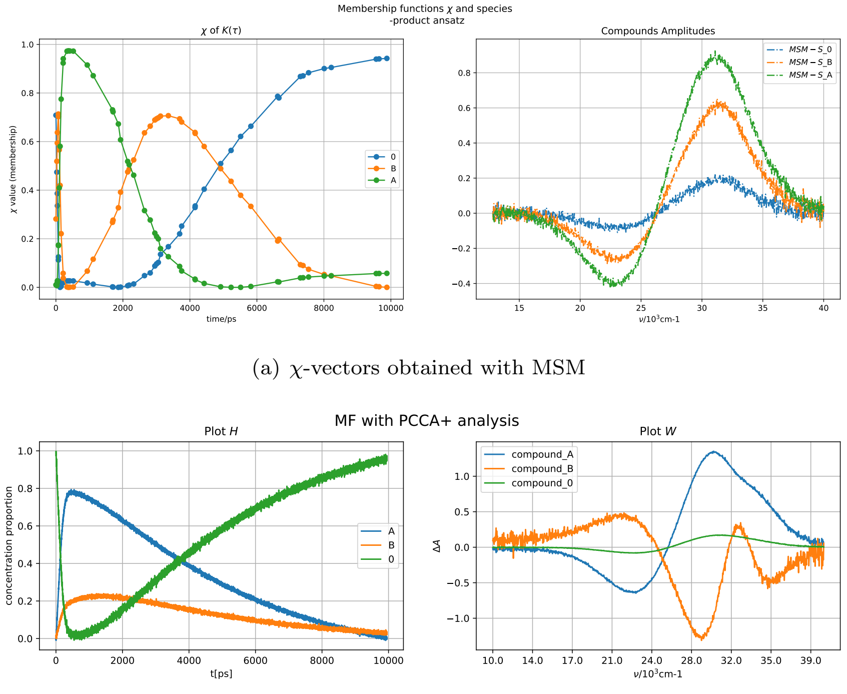
Figure 6: Comparison of the $\chi$ vectors computed with MSM and MF with PCCA+( 3-dominant-conformations-system).

# 7 Comparison and equivalence of projections: PCCA+, EDMD, DMD

In the last paragraphs, the Koopman operator $\mathcal{K}$ and then the Koopman matrix $K$ have been introduced as mathematical objects to describe the time evolution of spectral compounds (more general, of system's observables). This chapter will establish the connection between two methods for the projection of the Koopman operator, a special Galerkin projection (PCCA+) and the extended dynamic mode decomposition (EDMD). As explained before, projection of the Koopman operator are needed to treat measurements or real-world data, in general.
Establishing connections between methods allows to understand better the meaning of the computed objects in every step of the analysis.
An observable is "something that can be measured", so it is a function on the state space. Spectra are optical signals and so they are observables themselves. In the case of the pump-probe spectra, one can think about them as very complex functions on the state space, depending on wavelength and delay-time.
The concentration proportions $H_i$, $H_i \in \mathbb{R}^m$ is a row-vector which contains $m$ delay-time evaluations of the $i$th observable. $H_i$ only depends on the delay-time component, so the rows of $H$, $H \in \mathbb{R}^{r\times m}$ are the time-development of a collection of $r$ observables.
Since $H$ depends only on the delay-time and not on the wavelengths, it can be used to quantify the memory effect of the system in a similar manner as in section 5.2, by the overlap of the $r$ observables $H_i$. The equivalence of EDMD and MF with PCCA+ can gives the foundation to introduce how to compute the minimal memory by MF with PCCA+. The last subsection establish the connection between the Dynamic Mode Decomposition for the MF with PCCA+.

## 7.1 Extended Dynamic Mode Decomposition

Extended dynamic mode decomposition (EDMD) is a method to approximate the Koopman operator. The following text explains first how the EDMD can be applied to datasets analysis, and second it constructs a relationship between EDMD and physical observables in experiments.
Consider an autonomous, discrete-time dynamical system specified by the set $(\Omega, t, \mathbb{F})$, where $\Omega \in \mathbb{R}$ is the state space, $t \in \mathbb{R}^+$ is the (here) discrete time and $\mathbb{F} : \Omega \to \Omega$ is the evolution operator. The evolution operator $\mathbb{F}$ acts on states $x \in \Omega$. The Koopman operator $\mathcal{K}$ does not act directly on $x$, but is acts on functions of state space, $f \in \mathcal{F}, f : \Omega \to \mathbb{C}$ [2]. Functions on the state space as $f$ are called *observables* and those can be measured in the



(a) $\chi$-vectors obtained with MSM

(b) $\chi$-vectors obtained with MF with PCCA+

Figure 7: Three dominant conformations system. Comparison of the $\chi$ vectors computed with MSM and of the $H_{rec}$ vectors computed with MF with PCCA+(left side). Compound amplitudes computed with a product ansatz for MF and MSM (right side). Compounds name are named accordingly to the interpretation in the text.

experiments, so $\mathcal{F}$ denotes the space of the scalar observables. The Koopman operator is defined as $\mathcal{K} : \mathcal{F} \to \mathcal{F}$, that is it maps the (here) discrete time-evolution of functions on the state space to functions on the state space. The action of the Koopman operator is then

$$\mathcal{K}f = f \circ \mathbb{F}, \tag{24}$$

so with the Koopman operator one considers a different dynamical system $(\mathcal{F}, t, \mathcal{K})$.
Recalling from the previous sections, $m$ is the number of delay-times functions evaluations, $r$ is the number of observables. The EDMD framework requires [21, 33]:

1. a set of $m$ states of the system, $\mathbb{X} \in \Omega$, $\mathbb{X} = \{x_i\}_{i=1}^{m}$ and for each state in $\mathbb{X}$, its corresponding evolution, $\mathbb{Y} = \{y_i\}_{i=1}^{m}$ with $y_i = \mathbb{F}(x_i), i = 1, ..., m$;

2. a dictionary of basis functions $\mathbb{D} = (\phi_1, \phi_2, ..., \phi_r)$, where $\phi_i \in \mathcal{F}$ are our observables. Furthermore, we define a function $\Phi : \Omega \to \mathbb{C}^{1\times r}$

$$\Phi(x) := [\phi_1(x) \quad \phi_2(x) \quad .. \quad \phi_r(x)]^T \quad \forall x \in \Omega, \tag{25}$$

which maps a state $x$ to a vectors containing the evaluation of the observables in $\mathbb{D}$.

Instead of considering $\mathcal{F}$, the whole space of observables, it is possibile to consider a subset of observables $\mathcal{F}_r$, $\mathcal{F}_r \subset \mathcal{F}$, from which we choose the elements of $\mathbb{D}$.

### 7.1.1 Projection of $\mathcal{K}$

To work with a framework of real or experimental measurements,the Koopman operator $\mathcal{K}$ is projected with respect to the basis set $(\phi_1, ..., \phi_r)$ and this projection yields a finite-

dimension Koopman matrix $K \in \mathbb{R}^{r \times r}$.
Hence, the Koopman matrix governs the evolution over $\mathcal{F}_r$:

$$\Phi_{\mathbb{Y}}^T = K \Phi_{\mathbb{X}}^T \tag{26}$$

where $\Phi_{\mathbb{X}} := [\Phi(x_1)|...|\Phi(x_m)]$, $\Phi_{\mathbb{Y}} := [\Phi(y_1)|...|\Phi(y_m)]$.
In order to approximate the Koopman operator $K$, we consider a function $f, f \in span(\mathcal{F}_r)$ such that

$$f(x) := \sum_{i=1}^{r} a_i \phi_i = \Phi(x)\mathbf{a}, \tag{27}$$

with $\mathbf{a} \in \mathbb{R}^r$ being some scalar coefficients and $r$ is finite . The action of the operator $\mathcal{K}$ on $f$ can be reduced to the following [33]

$$(\mathcal{K}f)(x) = (f \circ \mathbb{F})(x) = (\Phi \circ \mathbb{F})(x)\mathbf{a} = \Phi(x)(K\mathbf{a}) + \xi(x) \tag{28}$$

where $\xi$ is a residual term. A way to explain why this residual $\xi$ rises is to say that $\mathcal{F}_r$ is not *a priori* invariant subspace of the Koopman operator $\mathcal{K}$ [33]. The best approximation of the Koopman operator for the sets $\mathbb{X}, \mathbb{Y}$ is given by minimizing

$$\begin{aligned} J &= \frac{1}{2}\sum_{j=1}^{m} |\xi(x_j)|^2 \\ &= \frac{1}{2}\sum_{j=1}^{m} |((\Phi \circ \mathbb{F})(x_j) - (\Phi(x_j)K)\mathbf{a}|^2 \\ &= \frac{1}{2}\sum_{j=1}^{m} |(\Phi(y_j) - (\Phi(x_j)K)\mathbf{a}|^2 \end{aligned} \tag{29}$$

Note again that $y_j = \mathbb{F}(x_j)$. By minimizing the expression in equation 29, one obtains

$$K = \mathcal{S}^\dagger T \quad \text{with} \tag{30}$$

$$\mathcal{S} = \frac{1}{m}\sum_{j=1}^{m} \Phi(x_j)^* \Phi(x_j); \tag{31}$$

$$T = \frac{1}{m}\sum_{j=1}^{m} \Phi(x_j)^* \Phi(y_j). \tag{32}$$

The expression for $K$ given in equation 30 can be reconducted to 26. Writing $\mathcal{S} = \Phi_{\mathbb{X}}^\dagger \Phi_{\mathbb{X}}$ and $T = \Phi_{\mathbb{X}}^\dagger \Phi_{\mathbb{Y}}$ we have

$$K = \mathcal{S}^\dagger T = (\Phi_{\mathbb{X}}^* \Phi_{\mathbb{X}})^\dagger \Phi_{\mathbb{X}}^* \Phi_{\mathbb{Y}} = \Phi_{\mathbb{X}}^\dagger (\Phi_{\mathbb{X}}^*)^\dagger \Phi_{\mathbb{X}}^* \Phi_{\mathbb{Y}} = \Phi_{\mathbb{X}}^\dagger \Phi_{\mathbb{Y}} \tag{33}$$

In this section, the approximation of the Koopman operator has been derived for a finite number of observables, $r$.
The relationship between EDMD and the data analysis of time-resolved spectra is direct. In particular, it is straightforward to relate MF with PCCA+ (sec. 4.2) and EDMD. In spectroscopy, one measures observables evolving in time and the states $x_i$ are not accessible from the measurements. The Koopman transition matrix $K$ that one wants to find is an object that describes the evolution in time of the concentration proportions of the compounds in the system. This means, one can compute the Koopman matrix using the matrices $H$, so equation 6.
In the context of EDMD, $H$ is the vector of observables (eq. 25); so $H \in \mathbb{R}^{r \times m}$ is a matrix

representing the time evolution of a dictionary of $r$- observables for a data set of $m$ time steps. We define the lagtime $\tau$ as the delay-time between the dataset points, so between the observable at time $t = (i-1)$ and $t = i$.
The description of MF with PCCA+, the dictionary has been chosen by means of the SVD so that $r$ are the leading eigenfunctions that can approximate well the Koopman operator. In this way, the formulation in equation 6 is the EDMD-formulation for the projection of the Koopman operator. The EDMD-projection has a meaning for the matrix factorization; the equivalent formulation obtained by the minimization of the residuals (eq. 29) is equivalent to 6, with the columns of $H[:, i]$ being the set of basis functions.

## 7.2 Galerkin projection

A Galerkin projection of the Koopman operator $\mathcal{K} \in \Omega$ is a projection onto a finite dimensional state-space $\mathcal{V} \in \Omega$ with $r$ basis functions $(\phi_1, ..., \phi_r)$. The projection has to satisfy:

$$\langle \phi_i, \mathcal{K}\phi_j \rangle_\pi = \langle \phi_i, K\phi_j \rangle_\pi \tag{34}$$

with inner product defined as $\langle h, g \rangle_\pi = \int h(x)^* g(x) d\pi(x)$ and $\pi$ being the distribution of $x$. As for the EDMD, one minimizes the residual so that the finite Galerkin-approximation of the Koopman operator $\mathcal{K}$ is given by a matrix $K$ with entries

$$K_{ij} = (\langle \phi_i, \phi_j \rangle_\pi)^{-1} (\langle \phi_i, \mathcal{K}\phi_j \rangle_\pi). \tag{35}$$

For the Galerkin method, two matrices $\tilde{\mathcal{S}}$ and $\tilde{T}$ are defined as

$$\tilde{\mathcal{S}}_{ij} = \langle \phi_i, \phi_j \rangle_\pi \tag{36}$$

$$\tilde{T}_{ij} = \langle \phi_i, \mathcal{K}\phi_j \rangle_\pi \tag{37}$$

and so $K = \tilde{\mathcal{S}}^{-1}\tilde{T}$.
In order to relate the EDMD and the Galerkin projection, the Monte Carlo approximation must hold. In facts, the EDMD approximation of the Koopman operator converges to the formulation of the Galerkin projection ( equation 35) for large $m \to \infty$ and if $x \sim \pi$

$$\mathcal{S}_{ij} = \lim_{m\to\infty} \frac{1}{m} \sum_{l=1}^{m} \phi_i^*(x_l)\phi_j(x_l) \qquad \approx \int_\Omega \phi_i(x)^* \phi_j(x) d\pi(x) = \langle \phi_i, \phi_j \rangle_\pi \qquad = \tilde{\mathcal{S}}_{ij} \tag{38}$$

$$T_{ij} = \lim_{m\to\infty} \frac{1}{m} \sum_{l}^{m} \phi_i^*(x_l)\phi_j(y_l) \quad \approx \int_\Omega \phi_i(x)^* \phi_j(F(x)) d\pi(x) = \langle \phi_i, \mathcal{K}\phi_j \rangle_\pi \quad = \tilde{T}_{ij} \tag{39}$$

Assuming that the Monte Carlo approximation holds, the resulting transition matrix $K_{MF}$ from the matrix factorization method (sec. 4.2) has the meaning of the EDMD-projcted Koopman matrix, but because of PCCA+ it can be treated as a Galerkin-projected matrix.If for the EDMD one chooses the observables in the dictionary to be a set of $\chi_j, j \in 1, ..., r$, then they correspond to an invariant subspace of the Koopman operator/matrix for the first discretization without clustering. The PCCA+ (sec. 5.1) is a special Galerkin-projection for which one chooses $\chi = (\chi_1, ..., \chi_r)$ as basis functions for the projection. The set of basis functions $\chi$ is computed from the Koopman matrix as explained in 5.1. The $\chi$-projected Koopman matrix has the characteristics that its projection and its propagation commute (see for example [18]). Therefore the commutation relation can be achieved with a finite number of basis functions, $r$, and no complete (infinite) $L^2$-basis is needed.
Usually the requirements of PCCA+ is to find the distribution $\pi$. Doing it is not possible, but one can approximate $\pi$by sampling. In this work, the configurations do not have only a

spatial component (the amplitude at a wavelength $\lambda_i$), but are characterized also by a time value. The association to the time component means that a certain spectral shape occurs only after a certain delay time. The configurational space has also a time component and the time component is also the reason why the analysis allows to estimate the membership to a certain dominant configuration (a certain $\chi_i$) as function of the delay time. As a consequence, the number of dominant conformations in the distribution $\pi$ will change as function of time. This is why in the examples the number of dominant conformations and their meaning vary if one analyzes early delay-times or the whole spectrum is included.

## 7.3 Implications of the equivalence of PCCA+ and EDMD

In section 7.1, 7.2, PCCA+ and EDMD has been compared and their equivalence has been showed. This relatinship between the two approaches can be used to estimate the minimal memory effect arising from the analysis by MF with PCCA+.

Being able to estimate the minimal memory effect by matrix decomposition is an interesting tool. Estimating the memory effect allows to notice if the rank of the decomposition is correct, since it is likely that if the number of dominant conformations is too high, then these will strongly overlap (small determinant of $\mathcal{S}$). Furthermore, one has a measure to understand how strongly the time-evolution of the (electronic) states in the systems is carrying information from the past. Finally as other advantage, one can vary the parameters in the objective function, obtain $H$ and observe how the memory effect increases or diminishes. For a good decomposition, the memory effect is small (so $det(\mathcal{S}) \approx 1$).

Given this motivation, one obtains the observable $H$ and compute, following the EDMD framework, the Koopman transition matrix $K_{MF}(\tau)$, see eq. 6. As EDMD and PCCA+ are equivalent, the Koopman transition matrix $K_{MF}$ can be given by a Galerkin projection of the "true" transition matrix $T_{MF}$ and the overlap matrix $\mathcal{S}_{MF}$

$$K_{MF} = \mathcal{S}_{MF}^{-1} T_{MF}. \tag{40}$$

The basis functions used to for the projection are the rows of $H$. With this decomposition, the EDMD projected transition matrix has an infinitesimal generator; this the steps as in sec. 5.2, the

$$\mathcal{S}_{MF} = \frac{\langle H^T, H^T \rangle_\pi}{\langle H^T, e_n \rangle_\pi} \tag{41}$$

estimates the minimal memory effect in the dataset analysis.

The result in eq. 41 is only possible because for the PCCA+-projected $K_{MF}$ *there exists* an infinitesimal generator $Q$. This step is very important, since with the only EDMD framework one cannot claim $Q$ exists.

## 7.4 Dynamic mode decomposition

EDMD is an extension of dynamic mode decomposition (DMD) [33]. For DMD, given two data sets $\mathbb{X} \in \Omega$, $\mathbb{Y} \in \Omega$, $\mathbb{X} = \{x_i\}_{i=1}^m$ and $\mathbb{Y} = \{y_i\}_{i=1}^m$ with $y_i = \mathbb{F}(x_i), i = 1, ..., m$, the Koopman matrix is given by

$$\tilde{K} = \mathbb{Y}\mathbb{X}^\dagger \tag{42}$$

So, as difference to the EDMD, there is no choice of a dictionary of observables. The Koopman matrix is estimated without a functions that maps $\mathbb{Y}$ and $\mathbb{X}$. Another way to see it is that DMD is an EDMD in which one chooses a dictionary of $r$ identity functions $\mathbb{D} = (e_1, ..., e_r)$.

In the steps of the NMF without separability assumption, we have that the PCCA+ modifies the left singular vectors $\tilde{U}$ so that the Koopman matrix is given by

$$\begin{aligned} K &= (H^{-\dagger}_{rec,-})^T H^T_{rec,+} \\ &= (\tilde{U}_- \mathcal{A}_{opt})^\dagger \tilde{U}_+ \mathcal{A}_{opt} \\ &= \mathcal{A}^{-1}_{opt} - (\tilde{U}^\dagger_- \tilde{U}_+) \mathcal{A}_{opt} \\ &= \mathcal{A}^{-1}_{opt} \tilde{K} \mathcal{A}_{opt}. \end{aligned} \tag{43}$$

Since we used a linear transformation of the $\tilde{U}$, the EDMD-projected matrix $K$ is a linear transformation of the DMD-projected matrix $\tilde{K}$.

# 8 Estimation of kinetic rates and decay times: computation of the infinitesimal generator

The previous sections 4.2 and 4.3 explain how those methods are applied to compute the transition matrix $K(\tau)$ from the datasets. The following text develops the concept of infinitesimal generator and illustrates different possibilities for the data-driven estimation. The infinitesimal generator is an important object to understand the systems, since it describes the kinetic rates between the (dominant) conformations. The kinetic rates in the diagonal of $Q$ are the inverse of the decay times.

For an autonomous[5], discrete-time Markovian process, the instant time-evolution of an observable $f$ in the system is given by

$$\begin{aligned} \frac{d}{dt} f_t|_{t=0} &= \lim_{t\downarrow 0} \frac{\mathcal{K}(t) - \mathcal{K}(0)}{t} f_0 \\ &=: \mathcal{Q} f_0, \end{aligned} \tag{44}$$

where $\mathcal{Q}$ is called infinitesimal generator and the time is discretized in steps $\tau$, so that $t = i\tau$, $\tau > 0$, $i \in \mathbb{N}^+$. Because of this time-discretization, in the following the time variable will be indicated by $\tau$. In that equation 44 holds for every value of $t = i\tau$, the Koopman operator will be given by

$$\mathcal{K}(\tau) = \exp(\mathcal{Q}\tau) \tag{45}$$

with matrix exponential

$$\exp(\mathcal{Q}\tau) = \sum_{i=0}^{\infty} \frac{(\tau \mathcal{Q})^i}{i!} \tag{46}$$

The discrete infinitesimal generator $Q$ has row-sum zero and it holds

$$Q_{ii} = -\sum_{j \neq i}^{n} Q_{ij} \tag{47}$$

with $Q_{ii}$ is the diagonal entry of the matrix in the $i$-th row, $Q_{ij}$ is the off-diagonal element representing the transition rates from conformation $i$ to a conformation $j$ out of the $n$ conformations of the process. Assuming the first-order kinetics reaction introduced before, the diagonal entries of the infinitesimal generator will represent the life times of the conformation $i$. In other words, the incoming flux of the conformation $i$ is given by the sum of the outgoing flux from all the other conformations $j \neq i$.

[5] If the Markov process it not autonomous, then the infinitesimal generator change as function of time. See, for instance [24].

The approximation of the infinitesimal generator $Q$ can be obtained from the Koopman transition matrix $K(\tau)$ obtained with the MSM. In discrete case, the relation between Koopman matrix and its infinitesimal generator is defined as

$$Q = \frac{d}{d\tau}K(\tau), \quad K(\tau) = \exp(\tau Q). \tag{48}$$

Because of the relation 48, the Koopman operator has semigroup properties ($K(t+s) = K(t)K(s)$). However, when dealing with datasets, the discretization in space and time of the process introduces an error, so that the transition matrix may not form a semigroup. As consequence of the loss of semigroup property, also the exact infinitesimal generator for this matrix may not exist. However, it is possible to estimate it by different approaches from the datasets .
The following paragraphs will discuss two methods that approximate the matrix exponential, the matrix logarithm (sec. 8.1.1), and the Newton extrapolation ( 8.1.3). Further, section 8.1.2 explains a method that exploits the limit-definition of the infinitesimal generator. Hereby only these methods for the inversion of the matrix exponential (and so the approximation of the infinitesimal generator) are considered; however, there are many (more than 19!) methods to approximate the matrix exponential [15] and so many other ways for extract $Q$ are still to be implemented.

## 8.1 Methods for the approximation of the infinitesimal generator

This subsection describes three methods for the computation of the infinitesimal generator, given a dataset. For each method, advantages and disadvantages are explained. Table 3 summarizes the text. In case of finite state-space processes, the infinitesimal generator is a matrix. The discretized infinitesimal generator has to be interpreted as transition rates between the conformations of the process, since for the dataset is modelled as a first-order kinetics mechanism.

### 8.1.1 Matrix logarithm

Assuming that the exponential relation between Koopman matrix and rate matrix holds, $Q$ can be computed straightforward as matrix logarithm of the Koopman matrix at time $\tau$ [14, 13]

$$Q_{logm} = \frac{log(K(\tau))}{\tau}. \tag{49}$$

and for multiple of the lag-time $\tau$, the rate matrix will be $Q^i_{logm} = \frac{log(K(i\tau))}{i\tau}$, $i \in \mathbb{N}$, and $log(\cdot)$ is the matrix logarithm in natural basis. Since the logarithm of a matrix appears in eq. 49, the required solution for the estimation of the rate matrix has to exist, be real and unique.
Given $K \in \mathbb{R}$, the logarithm of a real matrix only exists (a) if $K$ is non-singular, and (b) if each Jordan block of $K$ belonging to a negative eigenvalue occurs an even number of times [3]. In numerical computation, the result for $K$ is not always an invertible matrix. In general, a matrix has not just one logarithm; if $Q = log(K)$, then $Q + 2k\pi$ is also a logarithm of $K$. If one put the constraint of $Q$ to be real, then the condition of the Jordan blocks must be satisfied[11, 3], else $K$ has a matrix logarithm but it is not real.
The matrix logarithm of a real, non-singular matrix is a real matrix if each Jordan block belonging to a negative eigenvalues occurs an even number of times [3]. Furthermore, for the solution $Q$ to be unique, all the eigenvalues of $K$ have to be positive and real and the Jordan blocks do not have to repeat. The conditions that the matrix $K$ have to meet in order to yield a real, unique solution for its logarithm are many and it is difficult to guarantee that these conditions are always satisfied. Because of these requirements, the

use of the logarithm can lead to good results for very easy and small systems, but tends to be less reliable for complex cases, see the example of the computation of the rate matrix for the sequential decay in sec. 8.2. Another problem of the matrix logarithm is that if the system has reached an equilibrium condition (for example by going back to the ground state), the logarithm is hard to compute. For example, see the sink conformations in the Br-Al-corrole 9 and the Sb-corrole 10 analysis.
The rate-matrix resulting from the logarithm-method has a strong dependency on the presence of negative entries in the Koopman matrix and on the choice of the lag-time $\tau$ for the computation of the Koopman matrix. Numerical experiments on small reaction processes showed that the matrix-logarithm approximation leads to best results for finest lagtimes [22]; however, for very short lagtimes a strong memory effect has to be considered and slowly-developing processes are not taken into account.

#### 8.1.2 Finite differences

The finite-difference method is maybe the most diffused approach for the computation of the rate matrix. The idea is to treat the rate matrix $Q$ as infinitesimal object and to consider a discretized version of eq. 44. The infinitesimal generator computed by linear finite-difference approximation is given by

$$Q_{fd} = \frac{K(\tau) - K(0)}{\tau}. \tag{50}$$

Note that it numerical cases is possible that $K(0) \neq I_n$. This happens mostly if we compute $K^c(0)$, so the PCCA+ projected transition matrix. In facts, if the membership functions of the dominant overlap, then the $K^c(0)$ will not be the identity matrix. $K(t=0) \neq I_n$ happens also without any clustering projection, when using a radial basis function (e.g. a Gaussian) as membership function for the points in the data. The approximation in equation 50 is exact in the limiting case of $\tau \to 0$. Note that the approximation of 50 is the first-order polynomial approximation of the matrix-exponential function.
The finite-differences approximation depends only on the choice of the lag-time $\tau$ for the computation of the transition matrix. Therefore, the transition matrix will contain only information regarding events happening on the specific time-scale of $\tau$.
The choice of $\tau$ is not an easy task. The data are modelled as a Markovian process, but if $\tau \to 0$ strong memory effects rise. On the other hand, if $\tau \gg 0$, the resolution of the experiment gets lost completely together with information about fast-timescales events.

#### 8.1.3 Newton polynomial extrapolation

The Newton's polynomial extrapolation [22] is a multistep approach for the approximation of the infinitesimal generator. The method uses a set of Koopman matrices computed for different lag-times for the approximation of the matrix-exponential function. The outcome is a matrix that incorporates the information from different-timescales events.
Assuming that equation 48 holds and given a set of Koopman matrices $K_i = K(i\tau)$ for $i = 0, ..., n$ corresponding to $n+1$ lag-times, the $n$-th order approximation to $K$ is given by the Newton polynomial

$$\Gamma(\tau) = \sum_{i=0}^{n} \binom{\tau}{i} \Delta^i_{i/2}, \tag{51}$$

where the $\Delta^i_{i/2}$ are the divided differences of the Koopman matrices $K_i$. The finite differences for equidistant-knots are given by

$$\begin{cases} \Delta^i_0 = K_i & \\ \Delta^{k+1}_{i+1/2} = \Delta^k_{i+1} - \Delta^k_i & k \text{ even} \\ \Delta^{k+1}_i = \Delta^k_{i+1/2} - \Delta^k_{i-1/2} & k \text{ odd} \end{cases}. \tag{52}$$

| Method | Formula | Advantages | Disadvantages |
|---|---|---|---|
| Matrix logarithm | $Q_{i\tau}^{log} = \frac{log(K(i\tau))}{i\tau}$ | good results for short lagtimes | - not unique, non always numerically feasible |
| Finite Differences | $Q_{fd} = \frac{K(\tau)-K(0)}{\tau}$ | - easy to compute<br>- always numericall feasible | result depends on the choice of $\tau$ |
| Newton extrapolation | $\Gamma'(\tau) = \sum_{i=0}^{n} \frac{1}{i!} \sum_{k=0}^{i-1} \frac{\prod_{j=0}^{i-1}(\tau-j)}{\tau-k} \Delta_{i/2}^{i}$ | information from different timescales<br>- always numerically feasible | Runge's phenomenon |

Table 3: Summary of the methods applied for the estimation of the infinitesimal generator.

The $\tau$-derivative of the Newton polynomial for equidistant interpolation-knots (so in $\tau$-steps) [6, 26] is then given as

$$\Gamma'(\tau) = \sum_{i=0}^{n} \frac{1}{i!} \sum_{k=0}^{i-1} \frac{\prod_{j=0}^{i-1}(\tau - j)}{\tau - k} \Delta_{i/2}^{i}. \tag{53}$$

The estimation of (53) for $\tau = 0$ yields to obtaining $Q$. In facts, the time derivative of the Koopman matrix at time $t = 0$ is equal to the infinitesimal generator. Note that the finite-differences approach (sec. 8.1.2) is a first-order polynomial interpolation, so the Newton's method can be seen as an extension of it.

The advantages of the Newton extrapolation method is that it is not connected to the choice of a single lag-time $\tau$, as for the matrix logarithm and for the finite-differences. This enables to incorporate the information from multiple time points of the experimental data and not to bias only fast or only slow processes. Furthermore, the Newton's extrapolation method is always computationally feasible given a set a Koopman matrices, since it works with addition and scalar multiplication of the Koopman matrices.

The downside of this method is that Newton's polynomials show oscillatory behavior at the boundaries, the so-called Runge's phenomenon [22, 7]. This can decrease considerably the quality of the approximation. However, this phenomenon tends to appear for high-degree polynomials; in the analysis of time-resolved spectra, the degree of the polynomial is rarely higher than 10, so it is unlikely to expect the Newton polynomial to shows Runge's phenomenon.

## 8.2 Estimation of rates from example 6.

In the following we consider again the example (6) of the synthetic dataset describing a sequential decay of two species $[A \rightarrow B \rightarrow 0]$. The rate matrix has been computed for the three methods described in the precedent paragraphs of this section 8, each time applying the two methods of matrix factorization (4.2) and MSM (4.3). The analysis will involve three dominant conformations. For this kind of systems the discretized infinitesimal generator has the meaning of the rate matrix for the first-order kinetic process, that is why here the terms are interchangeable.

For these examples the rate matrices have been computed by first projecting the transition matrices via PCCA + and then using one of the methods for the estimation in section 8. However, one could also first approximate the rate matrix and then project it via PCCA+. With PCCA+ the order of propagation and projection commute. You can convince yourself of it with this Jupyter notebook.

#### 8.2.1 Estimation of the rate matrix with MF with PCCA+

The matrix has been factorized using the parameters of section 6 for the objective function $\Psi$.
The quality of the rate matrices can decrease because the strong negative values of the transition matrix. This means, the rate matrices will not have the standard form (rowsum zero, diagonal entry negative sum of the other entries). As a result, it is hard to interpret the meaning of these rate matrices.

**Matrix logarithm** In the example of the sequential decay, the presence of negative entries in $K_{MF}$ (eq. 23) has a large impact on the estimation of the rate matrix with matrix logarithm:

$$Q_{logm} = \begin{pmatrix} 0.02 & 0.03 & -0.05 \\ -0.08 & -0.09 & 0.15 \\ 0.00 & 0.00 & -0.01 \end{pmatrix}. \tag{54}$$

It is very difficult to interpret this matrix. One could have different guesses on some rows, but the dynamic itself is very hard to determine uniquely.

**Finite Differences** When computing the transition matrix as in 6, the transition matrix for $\tau = 0$, $K_{MF}(0)$ is not always the identity matrix. In facts, the transition matrix will encounter also overlap effects of the columns in $H$. Therefore, one can call $K_{MF}(0)$ is called also *mass matrix*.

When using the mass matrix, the infinitesimal generator for the finite differences method is

$$Q_{fd} = (-)\frac{K_{MF}(\tau) - K_{MF}(0)}{\tau} = \begin{pmatrix} 0.51 & -0.15 & -0.36 \\ -0.53 & 0.76 & -0.23 \\ -0.19 & -0.09 & 0.28 \end{pmatrix} \tag{55}$$

Naming the first dominant conformation A, the second B and the last 0, the first row shows a process that goes very fast from A to 0, then from B the fastest process goes to A and from 0 the fastest process goes to A. So from this matrix the slowest processes show again a sequential decay $[A \to B \to 0]$, but there are also others processes happening with faster rates (twice as fast, circa).

**Newton's extrapolation** In order to apply the Newton's extrapolation formula, one has first to compute the $K_{MF}$ for different lagtimes $\tau$. It is sufficient to consider that for the $K_{MF}$ it must hold $H^T(t) = K(\tau)H^T(t+\tau)$. Now, the time is discrete and for simplicity each column of $H$ is $\tau$-shifted w.r.t. the next one. Then, the Koopman matrix for a different timestep $i\tau, i \in \mathbb{N}^+$ can be obtained by taking out the *last* $i$ columns to $H$ for $H_-$ and taking out the *first* $i$ columns to $H$ for $H_+$. In this way, a set of transition matrices can be directly obtained and used as input for the Newton's extrapolation method.

$$Q_{nt} = \begin{pmatrix} 1.014 & -0.295 & -0.715 \\ -1.065 & 1.511 & -0.462 \\ -0.381 & -0.172 & 0.554 \end{pmatrix} \tag{56}$$

The obtained result is a matrix whose entries are very similar to the one of the finite-differences method. It is to notice that by adding interpolation points to the polynomial approximation, the resulting rate matrix has a constant scaling to the finite difference one. So the computation of the rate matrix with this method does not yields to further

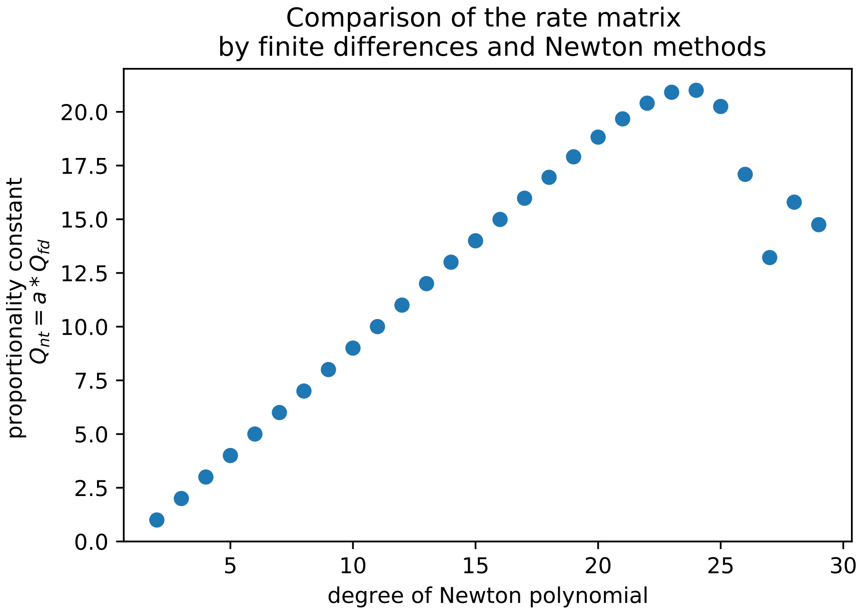


Figure 8: The rate matrix obtained with the Newton's approximation is proportional to the one computed via finite-differences method for the first 25-degree polynomials.

information about the system. The relation between the species is the same that was computed with the finite-differences method, $[A \to B \to 0]$. As shown in figure 8, this scaling relation of the Newton rates matrix w.r.t. the one of the finite difference method stays until the 25-th degree Newton polynomial. After, the rate matrix obtained with Newton is very different to the finite-differences one, maybe due to the Runge's oscillatory phenomenon.

### 8.2.2 Rate matrix with MSM and PCCA+

For the finite-difference (8.1.2) and Newton(8.1.3) method, it is necessary to compute the transition matrix for a lagtime $\tau = 0$, which is the identity matrix in this case, because the MSM is constructed as explained in sec 4.3 from the count matrix.
From the interpretation of the transition matrix and its membership functions in the MSM method in sec. 6, the first dominant conformation is called 0, the second is $B$ and the third is $A$. The resulting rate matrices show also negative values in the off-diagonal elements that make the interpretation of the dynamics harder.

$$Q_{logm} = \begin{pmatrix} -0.007 & 0.008 & -0.001 \\ 0.013 & -0.023 & 0.01 \\ -0.001 & 0.012 & -0.011 \end{pmatrix};$$

$$Q_{fd} = \begin{pmatrix} -0.007 & 0.008 & -0.001 \\ 0.013 & -0.023 & 0.011 \\ -0.001 & 0.012 & -0.011 \end{pmatrix};$$

For finite-differences and matrix logarithm, the rates are the very similar. The improvement of the matrix-logarithm approximation is given by the fact that the transition matrix in eq. 21 shows very light-weighted negative entries. Although their interpretation cannot be so clear, the information that we have is that the flux from $A$ goes to $B$ and there is no outgoing flux $A \to 0$; from $B$, there is outgoing flux to $A$ and to 0; from 0, the outgoing flux is to $B$. Moreover, the outgoing flux to $B$ has always a contribution of both $A$ and 0. Summing up the information from these matrices, $[A+0] \to B$ and $(B \to A,\ B \to 0)$ are not clear information about the dynamics in the system.

$$Q_{nt} = \begin{pmatrix} -0.008 & 0.009 & -0.001 \\ 0.02 & -0.034 & 0.014 \\ -0.002 & 0.02 & -0.018 \end{pmatrix}$$

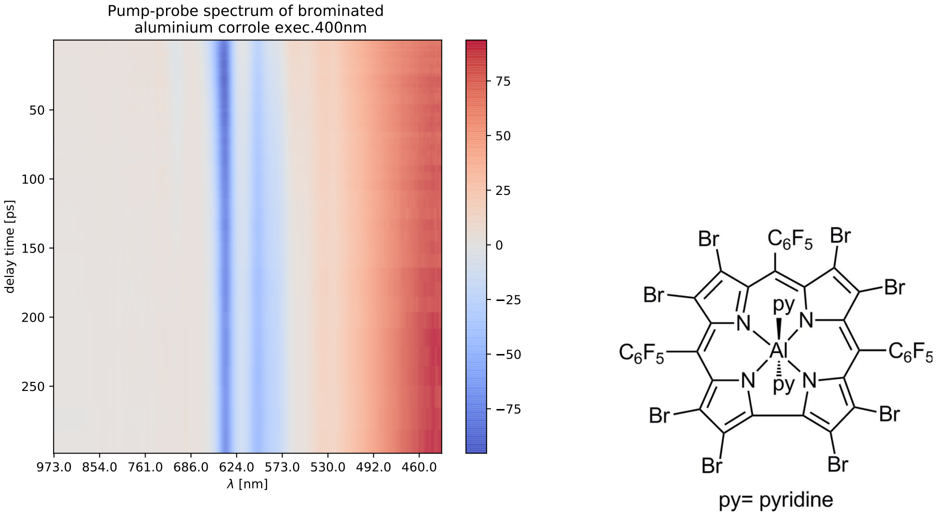
with Newton's extrapolation, $Q_{nt}$ a second order polynomial has been used. By incorporating information about the transition matrix with bigger lagtimes, the sequential character of the dynamics is become more clear. The magnitude of the rates is different and from the third row is clear that the main contribution of the flux to $B$ is coming from $A$, form the second row is clear that $B \to 0$ is the slowest process. So from the Newtons's matrix is easier to interprete the sequential character of the reaction.

# 9 Analysis of Brominated Al-Corrole

The analyzed dataset analyzed is a fs VIS pump–supercontinuum probe spectrum of the photoreaction of the hexacoordinated Al(tpfc-Br8)(py)$_2$, studied in [25]. The sample is excited at 400 nm with fs-pulse and the change of absorption is monitored as function of the delay-time. The spectrum is displayed in figure 9(a) as heatmap as well as the structure of the molecule 9(b). Global and Target Analysis where performed in [25]. Summing up the outcome of the analysis in [25], the data describe that the system leaves the ground state (GSB) $S_0$ and jumps to higher excited states $S_x$. From $S_x$, it cools down to energetic-lower excited states $S_1$ and $S_2$ with a time constant of 250fs. The energy is then redistributed via two cooling processes of 2ps and 20 ps time constants. Finally the system in $S_1$ reaches the triplet state $T_1$ with a time constant of 95 ps. The energetically higher singlet excited-state that the excited molecule reach is called *Soret band* (absorption maximum at 455nm). The singlet excited states $S_1$ (637nm) and $S_2$ (600nm) are called *Q bands*. Theory and further explanations about Soret and Q band can be found in [17]. The DAS were modelled with a sequential model and the figure 25 schematizes their interpretation.

The analysis is carried out from 300 fs until 300 ps and until 70 ps, so that for both cases the non-linear optical phenomenon are not considered.

The spectrum is measured at different pump-probe delay-times, which are exponentially distributed. Since the Koopman matrix $K(\tau)$ for an autonomous Markovian process is given for a fixed lag-time $\tau$, so the function `stroboscopic_index`, see sec. 4.3, is applied to the data.



(a) Transient absorption measurement of Al(tpfc-Br8)(py)$_2$ excited at 400 nm.

(b) Molecular structure of the Al(tpfc-Br8)(py)$_2$, adapted form [25]

Figure 9: Experimental vis-pump supercontinuum-probe spectrum of the Al(tpfc-Br8)(py)$_2$ and its molecular structure.

## 9.1 MF with PCCA+

**300ps analysis** The analysis of the dataset taken as a matrix shows 4 dominant conformations, basing the choice of the number of leading singular values. The used parameters for the optimization of the objective function $\Psi$ are [100, 10, 10, 1]. The results of the analysis are displayed in figure 10.

Considering the concentration proportions plot on the left, the dominant conformation $A_1$ is the start conformation; the process belongs mainly to $A_2$ in range 10-60 ps, then it moves to $A_4$ until 170 ps and finally the conformation $A_3$ rises until the end of the measurement. The relation between the dominant conformations can be read from the transition matrix $K_{MF}(\tau)$, with $\tau = 1ps$. If not specified otherwise, the analysis will be constructed always for 1ps as lagtime.

The transition matrix for this process,

$$K_{MF} = \begin{pmatrix} 0.882 & 0.171 & -0.013 & -0.042 \\ 0.020 & 0.953 & -0.012 & 0.040 \\ -0.002 & 0.001 & 0.977 & 0.024 \\ 0.008 & -0.008 & 0.033 & 0.967 \end{pmatrix}, \tag{57}$$

shows metastable dominant-conformations (the transition probability to other conformations is very low), except of $A_1$, which has a considerable transition probability to $A_2$. By reading the transition matrix, one observes that the strongest transition pathway will be $A_1 \to A_2 \to A_4 \to A_3$. However, $A_3$ is not a sink state, and has a small probability of going back to $A_4$ and $A_2$.

The sequential pattern $A_1, A_2, A_4, A_3$ can be observed also from the 710nm range; the stimulated emission signal is very small for $A_1$, it increases in $A_2$, it decreases to zero in $A_4$ and it positive in $A_3$. The guess is then that $A_1$ and $A_2$ must belong to the first stage of the reaction, when the system is in singlet, so before of the intersystem crossing. In particular, since the negative signal at 650nm is increased for $A_2$, then $A_2$ could include stimulated emission processes.

The interpretation of $A_4$ is not straightforward, because this state has triplet features, but from the transition matrix there is also a modest probability to go back to the ground state $A_1$. For these reasons, (a) $A_4$ can represent the dynamic of triplet-state arising, because the signal at 710nm is next to zero, so the triplet is not completely there; or (b) $A_4$ represents the system in a triplet state $T_2$, which has very similar energy to the $S_1$ state.

Finally, $A_3$ can be assigned to represent a stable triplet-state, $T_1$ since the signal for this dominant conformation is clearly positive around 710 nm. This qualitative assignment matches the order of the time scales for the photoreaction described in [25], $A_1 + A_2$ are processes of singlet states, so they are in the first 100ps-range. $A_4 + A_3$ are associated to triplets and they rise after the 100ps, which fits the 95ps time-constant of [25].

The transition matrix for this analysis in eq. 57 shows some negative entries, given by the memory effect. With the formula in eq. 41, one can estimate the memory effect to be very high ($det(\mathcal{S}_{MF}) = 0.002$). Since the memory effect is connected to the overlap between the concentration proportions, one can guess that two or more dominant-conformations describe very similar or strongly connected processes. Here, the guess would be that $A_4$ and $A_3$ both describe two energetically near triplet states, respectively $T_2$ and $T_1$. These triplet states overlap a lot and the analysis model has a strong memory effect.

The overlap of the concentration proportions is also more likely to rise for big delay times. If less molecules are excited, then the signal becomes weaker and the separation between the features of the different excited states is harder to realize. This is why it can occur that the values of some concentration proportions oscillate for bigger delay times. Figure 12(a) schematizes the interpretation of the analysis with some squares, since the character of $A_4$ is not clear.

Because of the several differences between the amplitudes of $A_1$ and $A_2$ and also guided by the information of [25], it is interesting to study and observe the character of the dominant conformations in the early states of the photoreactions. The following paragraph analyzes the portion of the spectrum for which $A_1$ and $A_2$ are the dominant conformations and the conformations with triplet features do not yet prevail.

The interpretation of the most probable path for the MF with PCCA+ analysis is schematized in figure 12(a).

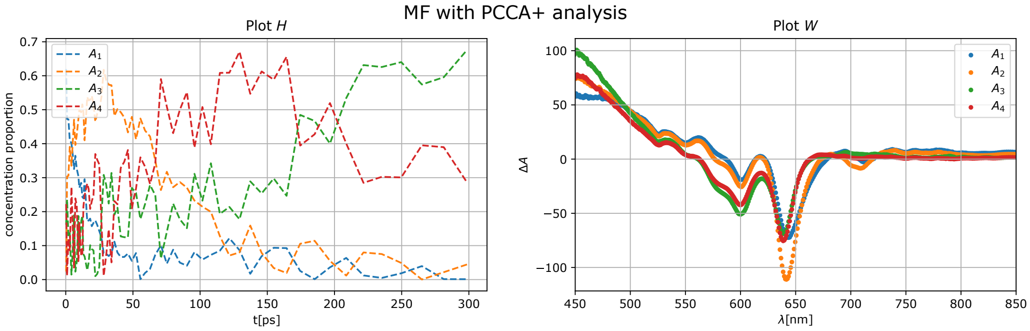


Figure 10: MF with PCCA+ analysis of the Brominated Al-corrole spectrum until 300ps as delay-time. On the right, the amplitude of the curves show a photo-product formation around 710 nm.

**300fs-70ps analysis** Since the triplet character is too strong in comparison to the fast processes in singlet-states, the analysis can be carried until 70ps. This enables to study also the fast-decaying processes, because they will have a similar weight in the portion of analyzed dataset. In particular, here the aim is to study if it is possible to distinguish the fast processes represented by $A_1$ and $A_2$ in the previous paragraph.

The result of the analysis is represented in 11 and the dominant conformations are called $a_1, a_2, a_3, a_4$. On the left plot in 11, one can notice the sequence $a_1 \to a_2 \to a_3 \to a_4$. Because of $det(S_{MF}) = 0.001$, the high memory effect for this decomposition indicates that the overlap between the conformations is big, as one can see from the concentration-proportions curves. This fact can implies that the processes described are very similar and so they are difficult to separate; alternatively, it means that the processes have similar weight and so similar decay-times. You can see it for example by looking at the curves $a_3$ and $a_4$ in range 40-60ps of figure 11.

The transition matrix for this analysis is

$$K_{MF} = \begin{pmatrix} 0.952 & 0.051 & 0.004 & -0.007 \\ 0.004 & 0.988 & 0.008 & -0.001 \\ 0.000 & 0.014 & 0.979 & 0.007 \\ 0.003 & -0.014 & 0.015 & 0.996 \end{pmatrix}. \tag{58}$$

This matrix shows that the most probable reaction dynamic are $a_1 \to a_2 \to a_3 \to a_4$ and $a_1 \to a_2 \to a_3 \to a_2$; in particular, $a_2 \leftrightarrow a_3$, meaning possibly that this two conformations represents similar processes in which the system oscillates its membership and then transits to $a_4$. From $a_2$, it is possible to go back to $a_1$ as well. This interpretation is represented graphically in 12(b).

Considering further figure 11, Conformations $a_1$ shows moderate ESA in the region 450-550 nm, and small GSB signals in the 550-650nm region. Around 710nm, $a_1$ shows very moderate negative signals. So maybe this conformation only shows a state in which the system left the ground state, but the SE signals (also negative) are not, or only partially present.

Conformation $a_2$ shows a similar profile as $a_1$ in the 450-630nm region, but it shows an increased negative signal at the wavelength of the fluorescence (650 and 710 nm). This means that the negative signals in the curve in $a_2$ show GSB+SE. Moreover, the curve $a_2$ is a bit blue-shifted w.r.t. $a_1$.

In the curve $a_3$ the ESA is higher in the region 450-500nm and in the region $>$ 750nm and the signal around 710nm is slightly blue-shifted and less-negative than in $a_2$. This can mean that a ESA signal is added to the SE signal at 710nm, showing a very early-stage of intersystem crossing, in which some molecules is already in triplet state. However, since the transition matrix show a possible transition to $a_1 \to a_3$, it is not clear it $a_3$ is a triplet state. Since $a_3 \to a_1$ is not possible, $a_3$ is not $S_1$. Perhaps $a_3$ is a singlet state with energy very similar to the triplet.

The curve of conformation $a_4$ is generally blue-shifted and the ESA signal in the region 450-550nm and shows no SE signal in the 710nm region, which means that the system is migrating to the triplet state. There is less ESA in the 450-550 nm region, because less molecules absorbs in that region, since some molecules are in triplet. Also the negative signal at 650nm is decreased for $a_4$.

Following this information and comparing qualitatively them to the interpretation of [25], one can say that $a_1$ is similar to the DAS at 0.2ps, because it only shows the excitation of the sample; $a_2$ and $a_3$ then show other singlet excited states (similarly to DAS 2ps and DAS 20 ps) and they both show SE and ESA of the singlet systems. However, $a_3$ incorporates already some triplet characteristics. Finally, $a_4$ shows in between of $S_1$ and $T_1$ is figure 25. The time-scales in which the dominant conformations prevail are roughly similar to the decay times estimated in [25]. The interpretation of the results is schematized in figure 12(b).

**Analysis of decay times via MF, 300ps dataset.** From the previous results, it is possible to compute the transition rates for this datasets. Our analysis will focus here only on the 300ps dataset. For the different approximation of the infinitesimal generator, the coarse grained infinitesimal generator $Q$ is obtained with the application of the PCCA+ algorithm. From the transition rates matrices, the rates are computed by the inverse the diagonal entries. The life times( timescales) $\tau$ are the inverted diagonal entries of the projected infinitesimal generators onto 4 dominant conformations:

$$\begin{aligned}\tau^{logm} &= (-7.829, -20.094, -42.475, -29.859)\\ \tau^{fd} &= (-8.453, -21.318, -43.766, -30.389)\\ \tau^{nt} &= (-6.641, -17.098, -40.645, -28.781)\end{aligned}$$

The order of the $\tau$s for each method is the same of the dominant conformations, so $\tau^{fd}(1) = -8.453$ is the decay time of $A_1$ computed with the finite-differences method. The decay times estimated with matrix logarithm and finite difference methods are more similar to each other than to the rates of the Newton method. The one obtained with the Newton's extrapolation of fourth order are different. However, the ratio between the inverted diagonal rates constants for each method is roughly the same. The fact that the relative decay times of the systems conformations are stable hints that velocity with which the system mutates its conformations is well represented by the different infinitesimal-generator computations. Comparing the information of the time constants to the decomposition, the slowest process is the third one, $A_3$,with a constant of $\approx 40$. However, since the unit of the decay times is not clear, whereas it is possible to say that the first process is 5 times faster than the third one and so on.

## 9.2 MSM

The Markov State Modeling method is applied to the dataset. To do it, each wavelength $\lambda$ is taken as a dimension and the conformation space is discretized in Voronoi cells. The centers or the Voronoi cells are picked with the picking algorithm [31] from the trajectory data. This enables to assign a delay time to each a Voronoi center $c_i$ and afterwards to identify the time ranges in which the dominant conformations develop.

**300fs-300ps analysis** By building the MSM with the whole dataset with 50 Voronoi cells, the Schur values show a small, but still significant gap after the first 5 Schur values, so that the transition matrix is projected into 5 dominant conformations via PCCA+. The

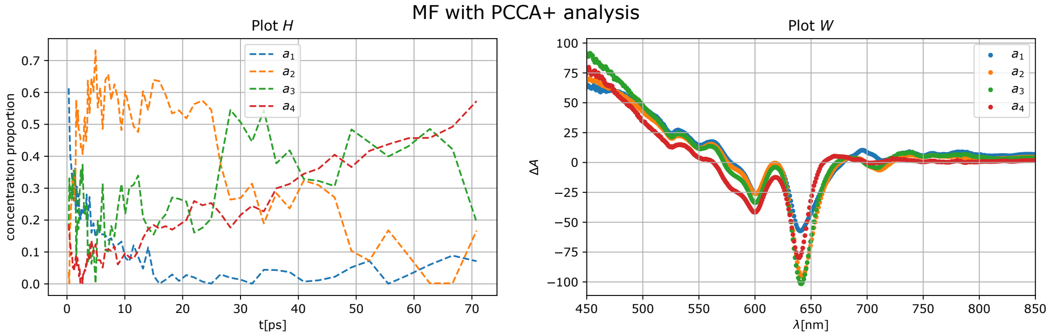
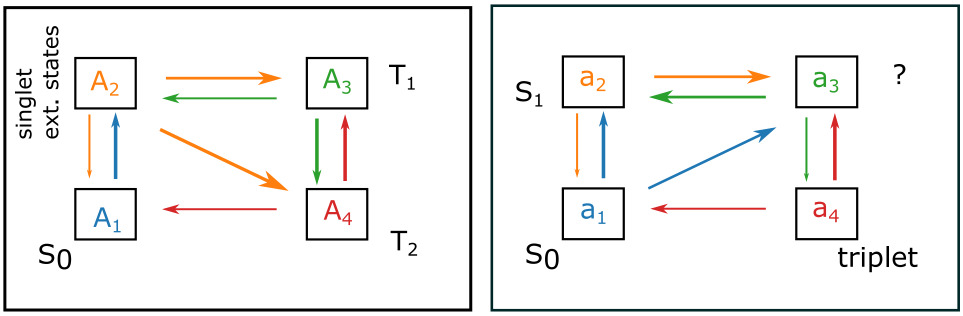


Figure 11: MF with PCCA+ analysis of the Brominated Al-corrole spectrum until 70ps delay-time.



(a) Brominated Al-corrole, MF with PCCA+, 300 ps

(b) Brominated Al-corrole, MF with PCCA+, 70 ps

Figure 12: The dominant conformations computed with the MF with PCCA+ analysis represent different processes of the reaction. Here the most probable path is schematized. The color coding is the same used in the figure 10 and 11.

membership vectors of each dominant conformation, the columns of $\chi$, are represented in figure 13 (left). Using a *product ansatz*, it is possible to obtain also the amplitudes of the dominant conformations (fig. 13, right). Note that, since one is using a product ansatz, it is hard to discern the amplitudes in the 450-500 nm region, because the positive excited state absorption signal is very strong. However, the interpretation of the amplitudes is still possible by considering other spectral regions in which the changes are clearer.
From figure 13, the dominant conformations alternates in sequential way $B_1 \rightarrow B_2 \rightarrow B_4 \rightarrow B_3 \rightarrow B_5$. $B_1$, $B_2$, $B_4$ rise and full decay in the first 70 ps. The first state, $B_1$ is the start conformation and represents the excitation of the system, since it rises only at the beginning and decays completely after 25ps. In its amplitude curve, $B_1$ presents negative signals at the place of GSB, moderately positive signal at the place of ESA (450-500 nm). $B_2$ shows an increment of the negative signal at 650 nm and a negative signal around 710 nm. Its amplitude curve is slightly red-shifted w.r.t. the other dominant conformations. This red-shift is hard to interpret: if $B_1$ is the ground state, then the system should reach a state with higher energy ($\propto$ blue). Most likely, the red-shift is apparent because of a broader negative signal. In fact, this small red-shift around 650 and 710 nm could be GSB+SE signals together. Since the interpretation of this conformation is not clear, the following paragraph will investigate only the first 35ps, in which the system is mainly in $B_2$.
The dominant conformation $B_4$ shows decreased negative signal at 710nm and increased positive signal (ESA) in the $>$ 710nm region. $B_4$ is probably a transition state between

some of the molecules in the sample moving to the $T_1$ state and others still in the singlet-system. $B_3$ has a similar shape as in $B_4$, but shows a loss in the ESA-signal in the $>750$ nm region. Finally, $B_5$ shows a positive signal in the 700-710 nm region, a decreased negative bleaching signal and an increased bleaching signal in the 570 nm region, due to the loss of ESA. This suggests that the population in $B_5$ is mainly in triplet state and it is starting to transfer back to the $S_0$, since the bleaching is starting to recover. However, what one observe is not the recovery of the bleaching signal, but the recovery of the SE signal. This piece of information can be gained by observing the transition matrix. $B_5$ is a sink state for the dynamics and the transition probability to $B_1$ is zero.
The coarse grained transition matrix helps in understanding the relation between the dominant conformations, especially for $B_2$, $B_4$ and $B_5$

$$K_{MSM}(\tau) = \begin{pmatrix} 0.880 & 0.206 & 0.083 & -0.124 & -0.045 \\ -0.008 & 0.917 & -0.051 & 0.120 & 0.022 \\ 0.003 & 0.023 & 1.001 & -0.041 & 0.014 \\ 0.002 & -0.009 & 0.024 & 0.991 & -0.008 \\ -0.000 & -0.001 & -0.007 & 0.007 & 1.001 \end{pmatrix}. \tag{59}$$

The coarse-grained Koopman matrix describes a process that from $B_1$ goes to $B_2$ and less likely to $B_3$. From $B_2$ the process goes mainly to $B_4$ and partially to $B_5$. From $B_4$ the process goes to to $B_3$. $B_3$ is an sink conformation. In case the process ends up in $B_5$, coming from the dominant conformations $B_2$ or $B_3$, $B_5$ is an sink conformation as well. Overall the dynamics shows a character that is sequential, but the sink states can be reached also with other less probable pathways. Since $B_2 \rightarrow B_1$ is not possible, we exclude that $B_2$ represents the $S_1$ state. It is likely that $B_4$ represents $S_1$, because of the transition probability. However, $B_4$ has also some triplet character, so the meaning of $B_4$ is $S_1 + T_2$. The high transition probability $B_4 \rightarrow B_3$ suggest that they have very similar character. From the $B_3$ amplitude, we know that $B_3$ is a triplet state $T_2$. This triplet state can go back to the singlet state, because of $B_3 \rightarrow B_1$; and it can also relax to another triplet state $T_1$ represented by $B_5$.
The transition probability $B_5 \rightarrow B_4$ can be explained by the fact that there is a small part of $T_2$ is $B_4$, so we see the system that oscillates between two triplet states $T_1$ ($B_5$) and $T_2$ ($B_3$). The fact that $B_3$ and $B_5$ have a sink- character confirms that they represent to triplet states with very similar energy, because the probability of changing spin again is very low. The memory effect for this $MSM$ quite high ($det(\mathcal{S}) = 0.005$, because of the strong overlap of the membership functions (see the $>$ 100ps time-range in 13, left). Because of the red-shift, assigning this state to cooling processes (time constants 2ps and 18 ps) is hard. If there is a red-shift, then the energy should be higher. Nevertheless, the time-range in which the system is mainly in $B_2$ (2ps to 25ps) matches with those cooling processes. The following paragraph will study deeply the time-range of this dominant conformation.

**300fs-35ps analysis** In the previous section, the analysis has been applied to the dataset in its whole time-extension. However, the interpretation of $B_2$ was not so clear, because the presence of slow processes does not permit to resolve faster ones. A way to circumvent the problem of biasing slow processes is to stop the analysis to earlier delay times. In this way, the fast processes are not so fast compared to the other ones; as a result, one can work only with similarly lasting processes.
To build the MSM here 20 Voronoi cells were used and the first 4 Schur values were identified for the PCCA+ projection. Figure 15 shows the outcome of the analysis. The analysis shows a start process $b_1$ in sub-picsecond time range, a process $b_4$ in the first 10 ps, a process $b_2$ in 12-16ps and it ends in $b_3$ for the rest of the time. The dominant conformation $b_1$ only shows bleaching signature, but almost not stimulated emission characters around 710nm, so it can be assigned to represent the system right after photoexcitation. The negative

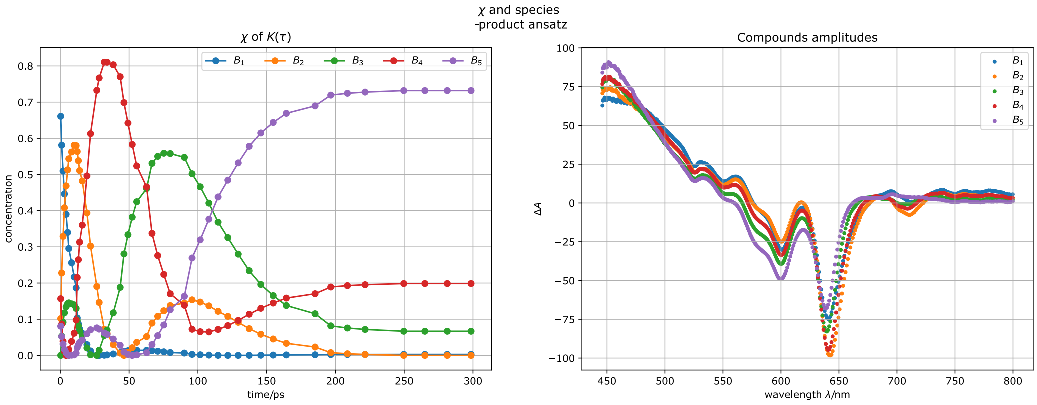


Figure 13: MSM analysis of the brominated Al-corrole dataset with PCCA+ projection into 5 dominant conformations.

peaks in the curves of $b_4$ and $b_2$ show a small blue-shift, especially in the 710nm region (corresponding to SE). The dominant conformation $b_3$ shows an increase of the negative signal at 650 and 600 nm, a small increment of ESA at 450 nm, but still the same SE signal at 710 nm. Therefore, $b_3$ describes a conformation in which the population is still in singlet, but it is already relaxed to the first excited state $S_1$. This analysis suggests the possibility that the excited-singlet system must have more than 2 excited singlet states, because of the small blue-shift of the $b_4$ and $b_2$ (they have different energies). The transition matrix for this MSM is very interesting

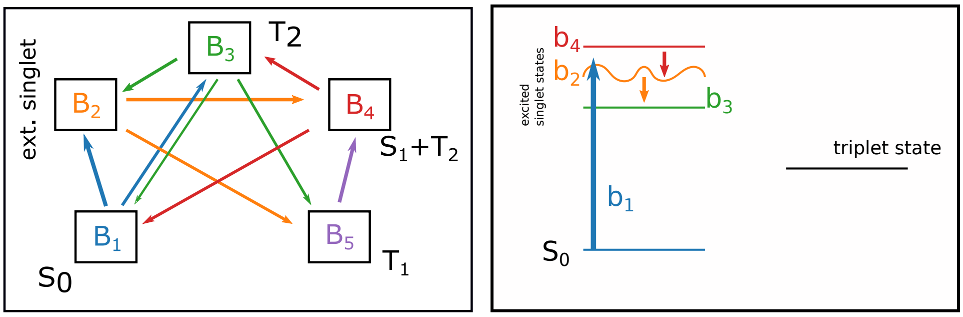
$$K_{MSM}(\tau) = \begin{pmatrix} 0.509 & -0.156 & 0.027 & 0.620 \\ 0.112 & 0.908 & 0.145 & -0.164 \\ -0.002 & -0.018 & 0.996 & 0.024 \\ -0.027 & 0.103 & -0.032 & 0.956 \end{pmatrix} \tag{60}$$

starting from $b_1$, one can have here several pathways, the most probable is $b_1 \rightarrow b_4 \rightarrow b_2 \rightarrow b_3$, however, from $b_1$ the system can evolve in $b_4$ or $b_3$, from $b_2$, it goes to $b_3$ but it can also reach $b_1$ again; the conformation $b_2$ is only accessible from $b_4$. This non-reversibility $b_4 \rightarrow b_2$ indicates that $b_4$ must represent a fast vibrational cooling process. If the excitation of the system to $b_2$ cannot occur from $b_1$, ic can mean that $b_1$ and $b_2$ have different symmetry in the wavefunction. Also, if the system takes the path $b_1 \rightarrow b_3$, it is very unlikely to reach $b_2$ and $b_4$. Summing up, $b_2$ and $b_4$ are transition states, $b_1$ describes the photoexcitation and $b_3$ is a stable excited state (probably most of the population in $S_1$). The memory effect in this MSM is lower than in the whole-spectrum case (0.058), because the differences between the processes that translates in less overlap of the membership vectors.

From this analysis, it is possible to compare $b_2$ and $b_4$ to the cooling processes found with global analysis in [25] with time scales 2ps and 18ps. With this model of analysis the amplitudes of $b_2$ and $b_4$ are very similar, so it is not clear to which process assign them exactly. However, one knows in which time-ranges they occur and how they are related to the other conformations $b_1$ and $b_3$, so that it is possible to know that they are fast transient states with higher energy w.r.t. $b_1$. The diagram shows the resulting reaction scheme.

As general remark, if the interpretation of the dominant conformation is clear, then an energy-like schema is used in the diagrams. Else, a graph represents the relation between the dominant conformations that is extrapolated by the transition matrix and it assigns a meaning to some of components, when clear.

**Analysis of decay times via MSM, both 300ps dataset.** Here we compute the time decay times of the dominant conformations of the MSM of the Br-corrole. So these decay



(a) Brominated Al-corrole, MSM with PCCA+, 300 ps (b) Brominated Al-corrole, MSM with PCCA+, 35 ps

Figure 14: The dominant conformations computed with the MSM with PCCA+ analysis represent different processes of the reaction. Here the most probable path is schematized. The color coding is the same used in the figure 13 and 15. Note that the 35ps analysis only shows a system in the singlet-symmetry states.

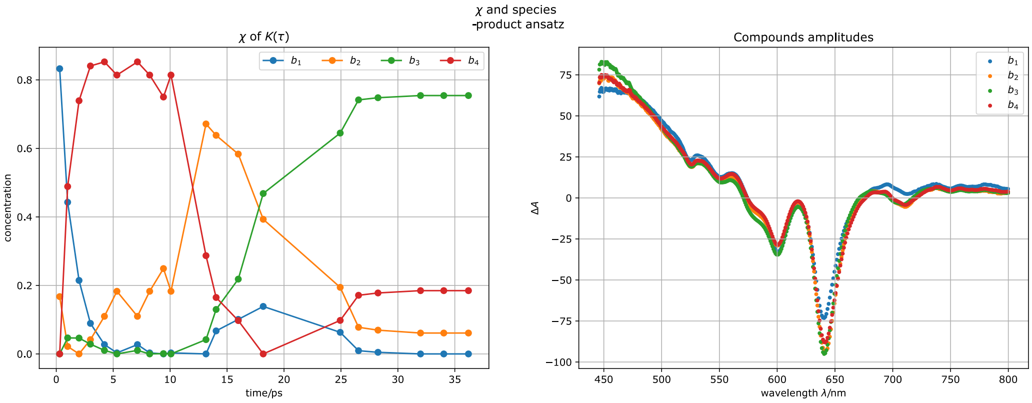


Figure 15: MSM analysis of the brominated Al-corrole dataset until 35 ps, with PCCA+ projection into 4 dominant conformations.

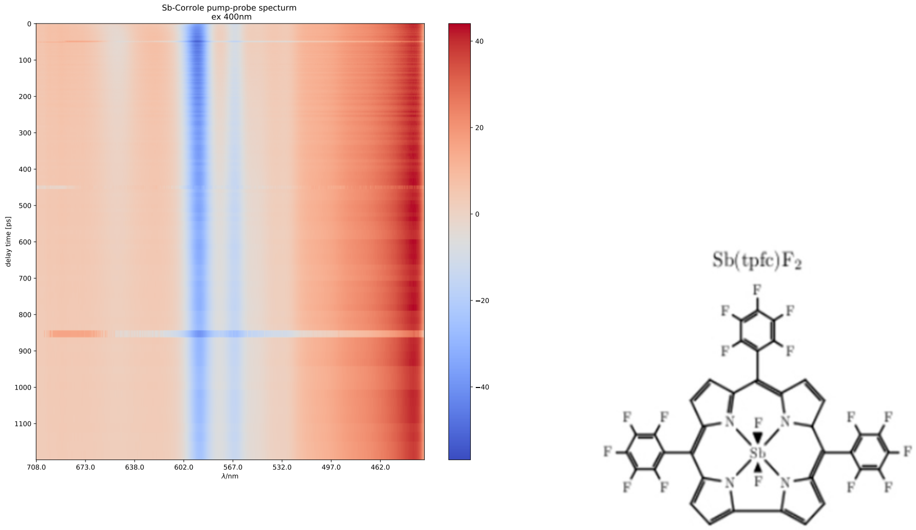
times corresponds to the $B_i$ scheme (diagram 14(a)). The decay times are different then for the MF analysis, because the meaning of the processes represented by the dominant conformations in the MSM is not the same.

$$\tau^{logm} = (-7.889, -11.806, 478.860, -135.618, 869.856)$$
$$\tau^{fd} = (-8.338, -12.006, 957.722, -117.035, 928.281)$$
$$\tau^{nt} = (-12.143, -7.273, -990.673, -36.293, 25.807)$$

The Newton's method yields very different results from the other ones. The fifth and the third conformations are sinking and this is conformed by the long decay times found with the finite-differences and matrix-logarithm methods. The positive decay time found with the Newton method does not suggest a sink character of the fifth dominant conformation. Moreover, this decay time is positive and that the results of the Newton method differ considerably from the other ones, so in this case the outcome obtained with 4-grade polynomials is not indicative for the analysis.

# 10 Analysis of Sb-Corrole

In this section, MF with PCCA+ and MSM with PCCA+ methods are applied to the analysis of the fs broadband vis-pump vis-probe spectrum of 5,10,15-tris-pentafluorophenyl-corrolato-antimony(V)-trans-difluoride (Sb-tpfc-F2). The photoreaction described by this dataset has been analyzed in [34]. The experiments show that after excitation at 400 nm the system is excited to high singlet excited states, then on a short time scale (0.5-20 ps) the system relaxes to lower-energy singlet states due to internal conversion (0.5ps) and cooling processes on time scales of 10 ps and 20 ps [34]. Because of the presence of the Sb-atoms, intersystem crossing from the first excited singlet state ($S_1$) to the first triplet state ($T_1$) was observed on the time scale of 400 ps. The dataset in this whole time-extension (from fs to ns) and wavelength extension [430-700] nm, as well as the structure of the molecule are displayed in figure 16. The Soret band peak for this molecules is at at 410 nm, whereas two Q band peaks are measured at at 565 nm and 590 nm [34].
The main issue of these analysis methods is that they are very dependent on the predominance of a process in time . It is hard to identify fast processes while analyzing the dataset, because they are not "dominant conformations". This "weighting-problem" does not undermine the possibility to apply the MF with PCCA+ and the MSM to the data. In facts, one can study the dataset for different time ranges, here in ns and $\sim$ 10 ps ranges.



(a) Transient absorption measurement of Sb-tpfc-F2 excited at 400 nm.

(b) Molecular structure of the Sb-corrole, adapted form [34]

Figure 16: From the spectrum the rise of the triplet state is evident as when the stimulated emission signal is not recovered.

## 10.1 MF with PCCA+

**ns-range analysis** The analysis is carried out for a ns-range (whole dataset). The results of the decomposition in 5 dominant conformations are displayed in figure 17. From the plot of the concentrations proportions $H$, it is clear that the dynamics is mainly modelled by two conformations, $C_4$ and $C_1$, whereas the other conformations have minor role. The spectral amplitudes in $W$ (right plot) are very noisy for the conformations $C_2$, $C_3$, $C_5$, especially between $620-660nm$. The fingerprint for $C_4$ is almost a constant line with a small negative peak in the GSB region (590 nm); it can be interpreted as a *start conformation* because its concentration decays with progressing delay times. The dominant conformation $C_1$ has two small peaks in the 560-590 region and its amplitude "steps" as almost a constant line in the 640-700 nm region. The conformation $C_5$ owns, especially for large delay-time ranges, a still significant fraction in the concentration profiles. From the transition matrix obtained by the MF, one can read different pathways for the dynamics. The transition matrix $K_{MF}$ for this system is

$$K_{MF}(\tau) = \begin{pmatrix} 1.053 & -0.014 & 0.011 & -0.160 & 0.042 \\ 0.381 & 0.770 & -0.059 & -0.437 & 0.099 \\ -0.001 & 0.025 & 0.810 & 0.220 & -0.035 \\ -0.007 & 0.004 & 0.001 & 1.015 & -0.004 \\ -0.339 & 0.100 & -0.020 & 0.898 & 0.760 \end{pmatrix}. \tag{61}$$

Starting from the fourth line ($C_4$), the process can go to $C_2$ or $C_3$. If it goes to $C_2$, it ends up in $C_1$ (sink state) or $C_5$. If from $C_4$ one goes to $C_3$, the process ends up in $C_4$ or $C_2$. For this factorization, there is almost no Markovian character. Using the measure for the memory introduced in the previous chapters, $det(\mathcal{S}_{MF}) = 0$. The analysis evidences that the process has very weak Markovian character, for example there are sink states in the transition matrix. Furthermore, the absolute value of the negative entries in the transition matrix suggests a significant overlap of the dominant conformations.

The interpretation one can give to the conformations is that $C_4$ is the start conformation; $C_2$ and $C_3$ represent likely some transition excitation processes (maybe relaxations from the Soret band), but their concentration proportion is almost zero (very few population), so their amplitude is most likely noise. $C_5$, with an offset, shows some positive signals (ESA) in the 450 nm range and clear GSB signals, that can be interpreted as bleaching-recovery, but it also has positive amplitudes in the triplet ESA-region. $C_5$ represents a late state in the dynamic, in which the population is partially in triplet but it is going back to the ground-state. Finally, $C_1$ is of difficult interpretation, but it is maybe a state in which no change take place, so it can be seen together with $C_5$ as ending state of the reaction.

If one considers the amplitudes in $W$ of the conformations $C_1$ and $C_5$ together, one could compare them to the constant DAS in [34, Fig. 7(b)]. The constant DAS presents the triplet decay and the decay of the GSB [34]. Because of the "transitional character" of $C_2$ and $C_3$ and the noisy reconstructed spectrum, it is hard to assign them to a process or to compare them to a DAS.

**300fs-35ps range analysis** Keeping the same parameters for the optimization of the objective function, now the analysis covers the first 35 ps. Several changes happen in the spectrum in the first picoseconds; since the following analysis until 35ps is not biased by the long-time processes, fast decaying processes can be distinguished.

The figure 18 shows the results of the analysis. Observing the concentration proportions (left in 18), the tendency is that only a single dominant conformation at the time is the main component of the system. One can see the sequence $c_1$, $c_2$, $c_3$, and then $c_4$ and $c_5$ alternates and have an opposite line-profile. Together, $c_4$ and $c_5$ constitute $\approx 80\%$ of the concentration proportions values form 10ps and since they alternate, it is likely that they represent the same process.

$c_1$ and $c_2$ have positive amplitude peaks at 570 and 590 nm, so at the wavelengths of the bleaching absorption. $c_2$ also shows positive amplitudes in the ESA range ($< 500nm$). Since $c_1$, $c_2$ are very fast-decaying, and their amplitude is similar, one can only say that they represent fast early-stage processes, but it is hard to identify more about their character.

The dominant conformation $c_3$ is also only present in the first 5ps delay times. In $c_3$ there is clear bleaching signature at 570-590nm, as well as ESA in the 440-510nm range and ESA at 620 nm. This information hints that $c_3$ could be the configuration that is described in [34] as Soret-band population.

Considering their amplitudes, $c_4$ and $c_5$ have nearly the same spectrum, $c_5$ is apparently red-shifted w. r. t. $c_4$, the apparent red-shift can be assigned to stronger SE signal which broadens the curves. Nonetheless they represent probably two energetically near electronic states or the same state. We think that $c_4$ and $c_5$ are two Q-band states, because of the increment of the negative signals, resulting from the rising SE of the Q-band at 590 and 650 nm. As a consequence of the increment of the SE, the bleaching peak at 590 results broader. The population of the Q-band also justify why there is a positive signal at big wavelengths [34]. There is no clear sign of triplet formation in the amplitude of the dominant conformations, but it is reasonable to think that a very small part of the process is already changing spin-symmetry.

It is just a case that, comparing the amplitudes to the DAS in [34, Fig. 7(b)], one can see similarities of $c_4$ and $c_5$ with the constant DAS, and also with the 380-ps DAS because of the negative signal at 650nm. This represents so a triplet formation, but it does not match with the interpretation of the analysis.

The dominant conformations $c_1$ and $c_2$ are start conformations, but they show in the 550-650 nm range a similar behaviour to the $DAS_3$ and $DAS_4$. Since the curves resulting in this analysis tends to have an offset, it is hard to assign them, because it is difficult to distinguish fluctuations. However, one can say that $c_1$ and $c_2$ represent maybe fast cooling processes or they are just numerical artefacts. It is not evident to which DAS in [34] belongs the conformation $c_3$ , because the amplitude between 650-700 nm decays to zero. If $c_3$ were the same process of the constant DAS ( decay of the triplet state), then it should have different time-development and it should be a sink state, so at the end of the reaction and with zero transition probability to other conformations in the transition matrix. Here it is not the case as expected, also because only the first 35ps of the photoreaction are analyzed. What one can say is that $c_3$ shows the amplitudes of the system in the Soret band. The transition matrix shows a tendency to the sequential decay between the dominant conformations

$$K_{MF} = \begin{pmatrix} 0.003 & 0.946 & 0.040 & 0.054 & -0.133 \\ -0.001 & 0.032 & 1.014 & 0.041 & -0.117 \\ 0.012 & 0.075 & 0.316 & -0.201 & 0.807 \\ 0.008 & 0.005 & -0.050 & 0.596 & 0.443 \\ 0.004 & 0.046 & 0.043 & 0.514 & 0.394 \end{pmatrix}. \tag{62}$$

Mainly one dominant conformation has one high-transition-probability entry, so that the overall process would read $c_1 \to c_2 \to c_3 \to [c_4 + c_5]$. $c_4$ and $c_5$ form a sink state for the process. $c_1$ and $c_2$ are transient states.

For this factorization, the minimal memory effect is very high, because $det(\mathcal{S}_{MF}) = 0.008$. But this is also given by the high overlap of $c_4$ and $c_5$, which are clearly depend on each other. In the diagram 22(b), the $c_5$ is represented as energetically higher than $c_4$ because of the transition probabilities in the transition matrix. However, it is to bear in bind that $c_4$ and $c_5$ are very close.

**Analysis of decay times via MF, 1.2ns dataset.** As done for the Brominated Alcorrole, the rate matrix is approximated with three different methods, matrix logarithm ($logm$), finite differences ($fd$), Newton's ($nt$) approximation, respectively. Here the degree of the Newton's polynomial is 4.

$$\begin{aligned} \tau^{logm} &= (16.054, -3.812, -4.765, 58.048, -3.711), \\ \tau^{fd} &= (19.027, -4.345, -5.265, 66.943, -4.168), \\ \tau^{nt} &= (-11.916, -4.395, -3.899, 243.177, -5.4) \end{aligned}$$

Due to the high memory of the MF Koopman matrix, the all the rate matrices approximations do not yield good results, especially for the sink conformations $C_1$ and $C_4$. These two

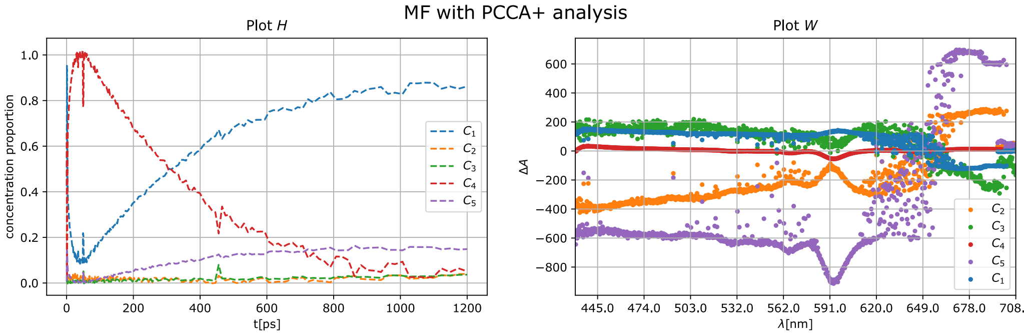


Figure 17: MF with PCCA+ in ns time-range. Parameters of the $\Psi$ [100, 10, 10, 1]. The concentration-proportion plot on the left shows that only two processes, $D$ and $A$, have an important role in the dynamics.

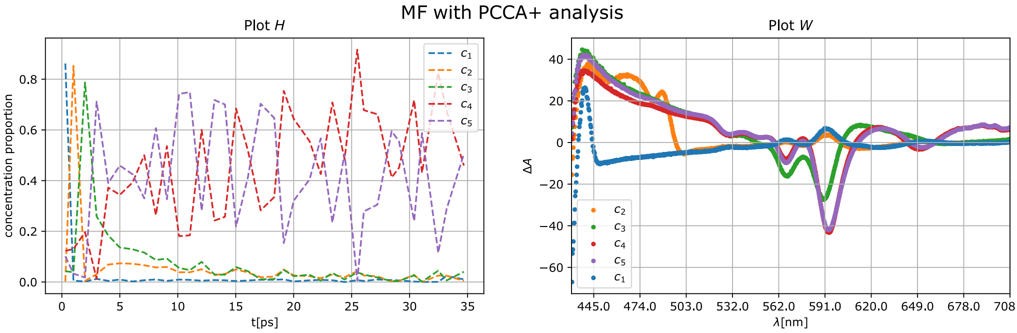


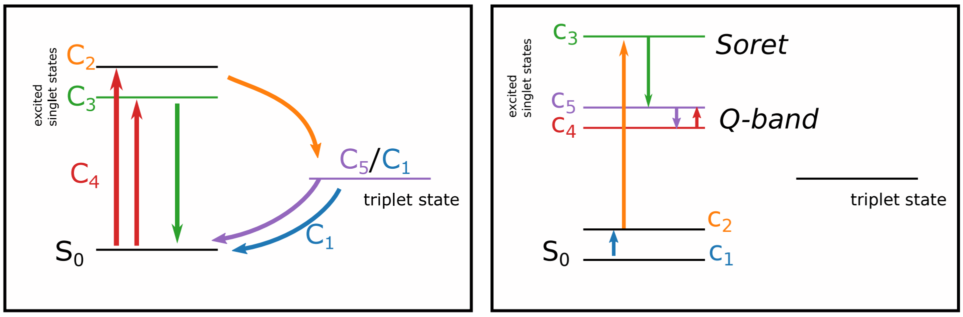
Figure 18: MF with PCCA+ in 35ps time-range. Parameters of the $\Psi$ [100, 10, 10, 1]. The concentration-proportion plot on the right shows that $c_4$ and $c_5$ are similar singlet state, possibly two Q-band states. $c_1$, $c_2$, $c_3$ show a strong transitional character.

dominant conformations have probability higher than 1, which makes the approximations of matrix logarithm and finite differences not suitable. However, one can interpret the positive values as the fact that there is not outgoing flux from a sink conformation.

## 10.2 MSM

The procedure of building a Markov State Model is the same used for the example of the Brominated Al-Corrole (9). For the Markov State Model the configuration space has been discretized into Voronoi cells. The centers of the Voronoi cells have been selected with a picking algorithm and a energy-weighted Frobenius norm for the assignment of the Voronoi cells.

**ns analysis** The conformation space has been discretized into 60 Voronoi cells. The analysis of the Schur values of the resulting transition matrix suggests that there are 4 dominant conformations. Using the *product ansatz* with the $\chi$ vectors, it is possible to reconstruct some spectral amplitudes as well. The figure 20 shows the results. From the left plot, the membership in the $\chi$ vectors suggests a a sequential process ($D_4 \rightarrow D_3 \rightarrow D_1 \rightarrow D_2$) that evolves slowly with $\approx$ 100-ps time steps. For the plot of the compounds amplitudes, one observe GSB signals in the 550-600nm regions for the dominant conformations $D_4, D_3, D_1$. This conformations also have decreasing negative signals at 650 nm. The gradually decreasing negative signal at 650nm suggests that $D_4, D_3, D_1$ are



(a) Sb-corrole, MF with PCCA+, 300 ps

(b) Sb-corrole, MF with PCCA+, 35 ps

Figure 19: The dominant conformations computed with the MF with PCCA+ analysis represent different processes of the reaction. Here the most probable path is schematized. The color coding is the same used in the figure 18 and 17. Note that the 35ps analysis only shows a system in the singlet-symmetry states.

comparable to the triplet formation. In particular, $D_3$ shows increased negative signal at 590 nm and increased ESA at 650-700 nm. The system absorbs more in that energy-range, which suggest that $D_3$ represents part of the system's population in the $Q$ band, which has a different ESA range.The dominant conformation $D_2$ represents the triplet decay and bleaching recovery, because the bleaching signal at 591 nm of $D_2$ is less than in the other conformations and that the amplitude of $D_2$ is positive at 650 nm. The interpretation of the spectrum in the range of 430- 500 nm is not clear, however, for $D_1$ one can see that the peak at 440 nm is higher, which means that the ESA increases. The transition matrix shows a sequential decay with $D_2$ as sink state

$$K_{MSM} = \begin{pmatrix} 0.996 & 0.004 & -0.001 & 0.000 \\ -0.001 & 1.000 & 0.001 & 0.000 \\ 0.009 & -0.002 & 0.993 & -0.000 \\ -0.000 & 0.001 & 0.010 & 0.989 \end{pmatrix}. \tag{63}$$

Starting reading the information in the transition matrix from $D_4$, it is ($D_4 \to D_3 \to D_1 \to D_2$). The resulting scheme from this analysis is very clear, but it is not precise regarding the processes under the 100-ps-time-scale. It is possible to compare the $D_4, D_3, D_1$ to the 380-ps DAS and the $D_2$ to the constant DAS in [34].
For this Markovian Model, the minimal memory effect is $det(\mathcal{S}_{MSM}) = 0.08$ which means that the membership functions in the columns of $\chi$ overlap not to strongly[6]. The reaction scheme is summarized in figure 22(a). Note that for $D_1$, $D_3$, $D_4$ , part of the population is already in triplet state. Note also that $D_4$ represents 1. the excitation to the Soret band, 2. the de-excitation to the Q-band, since $D_3$ has a different ESA w.r.t. $D_4$.

**300fs-50ps range analysis** This MSM analyzes the first 50ps of the dataset. The Schur-values show a clear gap after the first 4 values, so 4 dominant conformations are chosen. The results of the analysis after the PCCA+ projection are displayed in 21. As in the MF with PCCA+ case, the first dominant conformation, $d_1$, decays abruptly in sub-ps timescale. The the dominant conformation $d_2$ rises for the first 5 ps. The second dominant conformation to rise for the first 5 ps is $d_2$; the system is mainly in $d_4$ until 40ps delay-time and in the last 10ps $d_3$ has the biggest membership value.
The amplitude of $d_1$ (right plot in 21) is hard to assign to a specific process; it has only a

[6] In general, for this kind of data is it hard to have membership functions that do not overlap at all, since then one would have completely different spectral shapes.

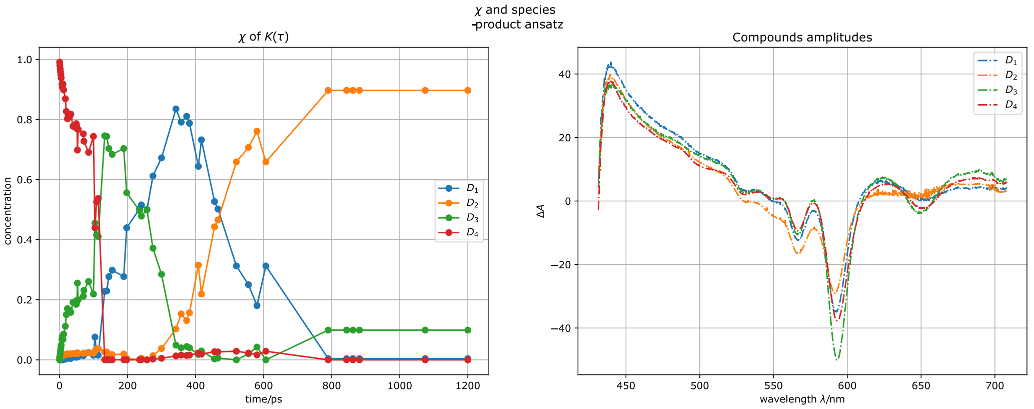


Figure 20: MSM of Sb-corrole with 4 dominant conformations, 60 Voronoi cells for the discretization of the conformation space. The analysis covers to the whole dataset, until 1.2ns.

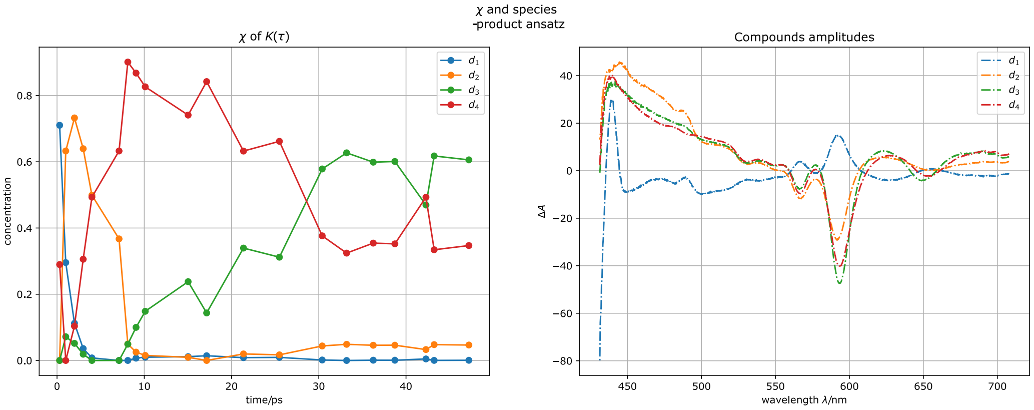


Figure 21: MSM of Sb-corrole with 4 dominant conformations, 30 Voronoi cells for the discretization of the conformation space. The MSM analysis only the first 50 ps of the spectrum.

small GSB signal at 590 nm, small ESA signal in the 700 nm range and an ESA peak at 440 nm. However we can say that $d_1$ shows that a part of the system has left the ground state, but for its amplitude there are no stimulated emission signals. Another way to interpret $d_1$ is that this conformation is a numerical artifact.

$d_2$ shows increased negative signals at 590 and 570 nm and ESA signals at 450nm and at 700 nm. It can be interpreted as the system that has left the ground state and it is excited to the Soret band, since the ESA ranges correspond also to the analysis of the fluorescence done in [34]. The spectral shapes of $d_3$ and $d_4$ are very similar, however the amplitude of $d_4$ is red-shifted w.r.t, $d_3$. The dominant conformation $d_4$ is a long-lasting process (30ps) and it shows SE signals in the 590 nm region, negative signal in the 650nm region and more positive signals between 500-570nm and between 650-700nm. It can be assigned to another band in the singlet system, the Q-band, because of the ESA positive signals added to the GSB ones and because of the SE. Finally, the conformation $d_3$ shows less ESA in the 450-500nm range, increased negative signal in the 500-600nm range, increased positive signal in the 650-700nm range. Since the shape of this dominant conformation shows strong SE signals, one can assign this conformation to a system which populates mostly the first excited state. Nonetheless, $d_3$ and $d_4$ are similar conformations because of their overlap in the $\chi$ profiles and because of their noticeably similar spectral amplitudes. Another option could be that $d_4$ is another excited state in the Q-band from which the system rapidly decays to $S_1$ (in $d_3$). From the comparison to [8, Fig.7(b), 7(c)] or here fig. 26, the amplitude of the dominant conformation $d_3$ has a similar shape of the 380-ps component, however the conformations $d_3$ has a different meaning and this is likely a coincidence. Comparing $d_2$ and $d_1$ to the other DAS is not direct. The analysis of the rate matrix can given more information about these two dominant conformation and their relationship to the results of global analysis.

The transition matrix shows here also a sequential scheme:

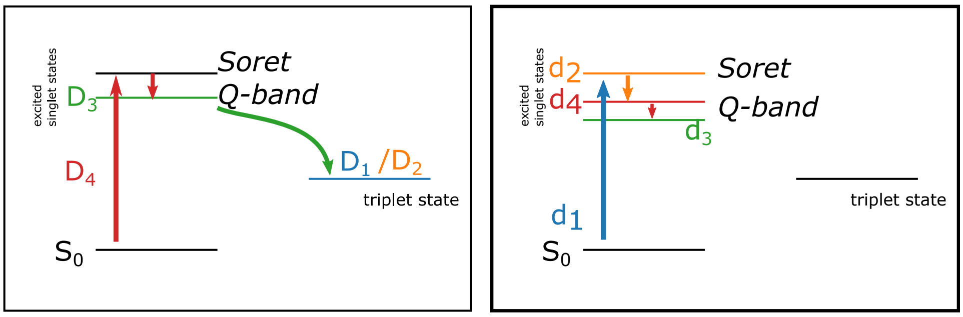
$$K_{MSM} = \begin{pmatrix} 0.001 & 0.899 & 0.097 & 0.004 \\ 0.000 & 0.829 & -0.047 & 0.218 \\ 0.000 & 0.023 & 1.006 & -0.029 \\ -0.000 & -0.001 & 0.039 & 0.962 \end{pmatrix} \tag{64}$$

From the matrix, it is clear that $d_1$ is a very fast start-process, whereas $d_3$ is a sink state. The triplet formation is not visible within this analysis until 50ps. Still it is to see that the process from the early stages conveys to $d_3$, which also makes it likely to be the $S_1$ state. From $S_1$, the system can go in the triplet state or back to the ground state.

The most probable dynamic-scheme from the transition matrix is $d_1 \to d_2 \to d_4 \to d_3$. There is also a consistent probability of going to $d_3$ directly from $d_1$, which means that some molecules go very fast from the ground state to the Q-band directly from the ground state. The overlap of the membership functions $d_3$ and $d_4$ for this MSM is moderate and so the memory effect $(det(\mathcal{S}_{MSM}) = 0.091$. This means that the Markovianity assumption here represent a good approximation. The meaning of the dynamic is schematized in figure 14, in which $d_4$ and $d_3$ are interpreted as Q band states.

**Analysis of decay times via MSM, both 1.2 and 50ps datasets.** Finally we consider the decay times computed for the MSM method for 300ps of the Sb-corrole. The resulting decay times are very similar.

$$\begin{aligned} \tau^{logm} &= (-595.333, -194.321, -193.769, -169.747), \\ \tau^{fd} &= (-597.368, -194.985, -194.269, -170.248), \\ \tau^{nt} &= (-597.368, -194.985, -194.269, -170.248). \end{aligned}$$



(a) Sb-corrole, MSM with PCCA+, 300 ps. $D_1$, $D_3$, $D_4$, part of the population is in triplet state. $D_4$ represents 1. the excitation to the Soret band, 2. the de-excitation to the Q-band, since $D_3$ has a different ESA w.r.t. $D_4$.

(b) Sb-corrole, MSM with PCCA+, 300fs-50 ps analysis. The analysis shows the excitation of the Q-band.

Figure 22: The dominant conformations computed with the MSM with PCCA+ analysis represent different processes of the reaction. Here the most probable path is schematized. The color coding is the same used in the figure 21 and 20. The early delay-times analysis shows a system in the singlet-symmetry states.

By analyzing the decay times resulting from the analysis of the first 50 ps with matrix logarithmus and finite differences

$$\tau^{logm} = (2.983e-02, 2.848e+04, 2.439e+01, 8.373),$$
$$\tau^{fd} = (1.000, 7.856e+04, 2.490e+01, 8.888)$$

which are very similar to the one found to in [34]. So only from the analysis of the first 50ps it is possible to resolved the fast-decaying processes. For the analysis with 30 Voronoi cells of the first 50 ps it was not possible to construct the Koopman operator for different delay-times, so only the rate matrices with these two methods are presented.

# 11 Discussion

Beyond the interpretation of the results, the analysis of the experimental datasets of the brominated Al-corrole (9) and Sb-corrole (10), together with the theory, infer the advantages and limitations of each method.

MF with PCCA+, based on [9], uses the singular value decomposition to factorize the matrix $M$. As many left singular vectors as many leading as many leading singular values are taken as input for the PCCA+ algorithm. The input of the clustering algorithm is an affine transformation, because the set of left singular vectors is shifted so that the first vector is constant. After the application of PCCA+, a penalty function $\Psi$ further optimizes the outcome for different parameters. Throughout this algorithm, the matrix is factorized so that $M = WH$, with the row of $H$ being the concentration proportions of the compounds as function of time, and $W$ being the amplitudes of the compounds, as well as a transition probability matrix $K$.

In general, MF with PCCA+ gives the best decomposition for the chosen set of parameters of the objective function $\Psi$. There is not a rigorous method to choose those parameters, which can be advantageous, because it allows to simply adapt the parameters to a data type, or it allows to give more weight to a feature of interest. Weighting a feature more than another is a trade-off for the quality of the decomposition. Not having a rigorous way to choose the parameters set is an issue when one wants to distinguish the quality of a decomposition from the results obtained with another parameter set. A way to score the quality of the decomposition is to multiply the respective $H$ and $W$ and see how the multiplication reconstructs the dataset.

The objective function $\Psi$ operates only a small modification of the concentration-proportion vectors in $H$. The previous application of PCCA+ separates the concentration proportions of the dominant conformations already. Instead, the amplitude matrix $W$ is mostly influenced by the choice of parameters. The entries of $H$ can be only between [0,1], so also a small variation of these entries causes a considerable change in $W$.

The change in the amplitude shape of a compound in $W$ implies that the interpretation of the corresponding phenomenon is harder. This happens especially when the curve has an amplitude next to zero in a spectral range of interest, because then only the comparison to the amplitudes of the other curves can help in interpreting the meaning of the process.

The MF with PCCA+ is a method that leads to concentration proportions (in $H$) and so it cannot give a quantitative measure of the concentration (as Global Analysis does), but only a qualitative one. This could be maybe a problem when comparing the results with the experiments.

When the concentration proportions, the rows of $H$, have a parallel or mirroring curve development, the analysis yields similar amplitudes in the respective species (in matrix $W$). This behaviour of co-dependence of two compounds implies an high memory effect, because the concentration proportion vectors overlap (see eq. 41). The reason of the co-dependent curve development is that the compounds cannot be fully separated by PCCA+ and so they end up having a mirroring profile development. Its implications are that the two compounds represent similar processes (resulting in comparable amplitudes in $W$), or that the two compounds are the same process, but they cannot be properly discerned to each other via the only application of PCCA+ because of the noise.

As seen in the results on the Corrole analysis, sometimes the concentration-proportions curves can oscillate, that is their profile rises, goes down and rises again multiple times. Even if this up and down of the concentration is a possible behaviour of the compounds, what one expects is that the concentration of a conformation rises and then decay, The systems should then go to another conformation, not going back and forth between them or rapidly changing the equilibrium between the compounds. The oscillatory behaviour is mostly given by the noise. In general, the curves of the concentration proportions has to be interpreted so the global behaviour of the curve profile, rather than for the differences

between two points.
Experimental data are affected by noise. Noise can mask a signal, so that a dominant conformation might not be found; or noise might add to a signal, so that two dominant conformations actually represent the same process. The interpretation of the transition probabilities in the Koopman matrix helps in distinguishing the effect of the noise: if two dominant conformation form a "sink" together, then they are probably the same process. In real data, the concentrations can become negative because of noise. Here, the concentration proportions in MF with PCCA+ or the membership functions of the MSM are, by construction, always positive. This does not contradict that there is noise in the system. The decompositions include noise as element of the system, and the concentrations are not concentrations of chemical species, but *concentration proportions* (MF) and *membership* (MSM) of dominant conformations. Going back to the problem of MF with PCCA+, the concentration-proportions curves zig-zaging, this behaviour is a result of the noise taken as part of the compounds.
Summing up, the MF with PCCA+ is a very advantageous method for analyzing different kinds of time-resolved spectra, since one can modify the parameters of the objective functions. It is particularly suitable for experimental data, because it does not require the separability assumption for the decomposition. However, some aspects such as the treatment of the noise, the separation of the compounds and the choice of the parameters still need further developments.

In the MSM with PCCA+ method, the main idea is to look at the time-resolved data as they were a trajectory in as many dimensions as wavelengths and the absorption difference is the value assumed by each coordinate, wavelength at a given time-step, the delay time. From this assumption, a Markov State model is built: first the configuration space is divided into Voronoi cells, after each point (so each spectrum at a given delay-time) is assigned to a Voronoi cells, then the transitions of the trajectory between the Voronoi cells are counted. A count matrix is computed from the transitions, the row-normed count matrix is called Koopman matrix. By PCCA+ projection using the leading eigenvectors of the Koopman matrix as basis functions, the matrix is clustered so that it represents only the transition probabilities between the dominant conformations.
As a difference to the MF with PCCA+, MSM with PCCA+ method does not require an objective function, but the results of the modeling and specifically of the clustering strongly depend on the first Voronoi tessellation used to compute the transition matrix. Many or too few Voronoi cells accentuate the non-Markovian character of the trajectory and so affect the quality of the MSM. Too few Voronoi cells might not be enough to represent the configuration space, so that some processes will be missed. Also the allocation method of the centres the Voronoi cells has a large impact to the result: if for example one uses a regular grid to place them, then the initial changes are not represented at all.
The distance from each spectrum at a given delay-time to the centers of the Voronoi cells is computed by weighted Euclidean-norm, in which the weights are given by the energies. However, one could also weight some wavelengths-ranges of interest more than others, so that the change in these ranges has more influence to the assignment to a Voronoi cell. This different weighting system allows to characterize more the parts of the spectrum in which one sees a small signal change which is very important for the dynamic. For example, by weighting more a range of the triplet signal, one could analyze better its rise and decay. At the moment, the MSM with PCCA+ does not have this kind of *ad hoc* specifications, but only the energy weighting has been used for the assignment of the Voronoi cells.
Adopting a product ansatz for the MSM decomposition is a big assumption. For the Global Analysis, the spectrum is given by a linear combination of concentrations (exponential decay functions) and amplitudes. For the MSM method, the structure of the membership function $\chi$ is not a sum of decay functions *a priori*, but each vector $\chi$ represents the degree of association of the system to that observable as function of time, so the curve has

a meaning as a whole. This is why the compounds amplitudes obtained with MSM with PCCA+ can be very different from the results of the MF with PCCA+.
In general, the characteristic meaning of the membership functions $\chi$ does not assure to identify the "species" of the reaction. Rather, one determines some processes that have a strong importance, regardless if they are dynamic (movement of the system between two states) or static (the system populates a state). This thesis illustrates a way to interpret the amplitudes obtained by product ansatz. In the MSM method the noise is part of the dominant configurations profiles, as for the MF with PCCA+. The presence of the noise can, as aforementioned, influence the number of dominant configurations.
In experiments, the signal becomes weaker with time, because fewer and fewer molecules are excited. The concentration curve decreases. In other words, the number of particle $N$ changes. The membership function vectors, the columns of $\chi$, are always between 0 and 1 and they always sum up to one at each timestep. For example, one vector of $\chi_j$ has value 0.8 at time $t = t_i$, and it is 0.8 at time $t = t_{i+n}$, $n > 0$. The value 0.8 tells in both cases that 80% of the system belongs to the dominant conformation $\chi_j$; but it does not mean that the amount of molecules in the dominant conformations is the same for both $t_i$ and $t_{i+n}$. The $\chi$ vectors do not represent *quantitatively* how many molecules are excited. Rather, the $\chi$ represent *qualitatively* the excitation state of the system at each time.
The MSM with PCCA+ is based on the condition of partition of the unity, which means that the sum of the membership functions is always one. This could make more difficult the treatment of experimental data. Still, this method have a very simple set-up (discretization of the configuration space and then invariant projection) and it is helpful to study systems for which there is not much knowledge.
The results from the analysis with MF with PCCA+ and with MSM with PCCA+ are (usually) not the Decay Associated Spectra. Sometimes the curves can be similar to them, but a priori one cannot say that they are the same. The amplitudes of the compounds show features that characterize the system's conditions, and the amplitudes curves can be interpreted only by comparing each curve to the other ones. The relative change between the curves enables to assign a meaning to what the amplitudes profiles represent.
For both methods, the meaning and the interpretation of a dominant conformation is found by considering three elements: the amplitudes, the kinetics (concentration proportions or membership functions) and the Koopman transition matrix. The transition matrix is an indicator of the transition pathways in the photoreaction and gives information about the kinetics. In the other methods, the transition matrix is not used for the interpretation.
In general, for both methods the Markovian process assumed is autonomous, which means that the transition matrix does not depend on time, but only on a fixed lat-time $\tau$. So when the measurements do not scale linearly the delay time, but exponentially, the data set has to be pre-rocessed so that each time-step gets the same weight. This is done with the application of the function *stroboscopic-index*[23] to the raw-data. The curve for each wavelength then looks like a stepping function. An aspect that would change noticeably change the quality of the analysis is to interpolate the spectra by using a spline ( or another interpolation function) for each wavelength . This is something that can be done in future work.

Both the methods are useful for processes that have similar timescales. Short processes cannot be detected by only analyzing eigenvalues or singular values, because those would just describe the slowest processes. In this cases, the analysis of time-windows of the dataset can help in investigating fast-decaying processes. The identification of the 2 Q-band conformations in the first reaction stages of the Sb-corrole 10 shows how the analysis of time-windows can be helpful. In facts, the analysis of the 1ns dataset did not allow to identify these singlet states.
Both the methods are very strong for the computation of the transition probability between the conformations, that together with the time-development and the amplitudes completes a framework to interpret the data.

The approximation of the rate matrix $Q$ for both methods depends strongly on the memory introduced in the computations of the transition matrix. Heuristically, the tendency is to have poor estimation of the transition rates if the Koopman matrix has multiple negative values, which also implies overlap between the conformations. The computation of the rate matrix for time-resolved spectra is still a topic to study. The biggest challenge is to understand the meaning of the values of the decay times (1/rate) of the compounds and to associate them with a time unit or a proper scaling. At the moment, the computation of the decay times is only important to understand the *relative* decay velocity between the dominant conformations.

The Python-scripts for the presented algorithms are online here.

# 12 Conclusion

This thesis aimed to unravel the kinetic of photo-activated processes from time-resolved spectral data, without assuming any decay model to explain the reaction. To puzzle out the kinetic model, two Koopman-operator based methods have been developed and applied to analyse the data. These procedures are called MF with PCCA+ (sec. 4.2) and MSM with PCCA+ (sec. 4.3). The PCCA+ part indicates the projection of the dynamics into dominant conformations, which can represent chemical and physical species, but also noise or reaction processes. A dominant conformation is everything whose "shape" is important for the considered data set. The reaction pathway or kinetic scheme is read from the transition probabilities between the dominant conformations, that is by studying the Koopman transition matrix $K(\tau)$. The transition matrix provides not only a transition pathway, but describe the transition probability between all the dominant conformations. Hence, less probable transition pathways can be found as well.

The analysis of the experimental transient absorption spectra of the corroles molecules with MF and MSM infer a sequential pathway for the reaction. A sequential model was suggested also by the authors of [34, 25], which investigated the reactions by global analysis.

The methods developed in this thesis find conformations which have similar dominance degree, which implies that the processes have similar timescales that are all similarly slow for the process. This does not guarantee to identify fast, but chemically important reaction-steps by analyzing the whole dataset. However, considering only small time-windows for the data allows to study also fast processes.

This thesis shows that sequential processes can be studied by the MF with PCCA+ and MSM with PCCA+. Both methods still need refinements for the treatment of the noise and the spectral amplitudes, which are conditioned by the positivity of the membership functions and by the product ansatz. Further studies could focus on alternative ways to obtain the spectral amplitudes, to the computation of the noise levels, and the physical and quantitative meaning of the transition rates.

Likewise, future studies could optimize the modeling by estimations of the memory effect. The less memory, the less overlap between the conformations, so the more distinct the meaning of the processes that the conformations represent.

# References

[1] Halina Abramczyk. "8 - Selected Methods of Time-Resolved Laser Spectroscopy". In: *Introduction to Laser Spectroscopy*. Ed. by Halina Abramczyk. Amsterdam: Elsevier Science, 2005, pp. 175–217. ISBN: 978-0-444-51662-6. DOI: https://doi.org/10.1016/B978-044451662-6/50009-5. URL: https://www.sciencedirect.com/science/article/pii/B9780444516626500095.

[2] Marko Budišić, Ryan Mohr, and Igor Mezić. "Applied Koopmanism". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 22.4 (2012), p. 047510. DOI: 10.1063/1.4772195. eprint: https://doi.org/10.1063/1.4772195. URL: https://doi.org/10.1063/1.4772195.

[3] Walter J. Culver. "On the Existence and Uniqueness of the Real Logarithm of a Matrix". eng. In: *Proceedings of the American Mathematical Society* 17.5 (1966), pp. 1146–1151. ISSN: 0002-9939.

[4] Anna de Juan and Romà Tauler. "Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review". In: *Analytica Chimica Acta* 1145 (2021), pp. 59–78. ISSN: 0003-2670. DOI: https://doi.org/10.1016/j.aca.2020.10.051. URL: https://www.sciencedirect.com/science/article/pii/S0003267020310771.

[5] Luca Donati et al. "Estimation of the infinitesimal generator by square-root approximation". In: *Journal of Physics: Condensed Matter* 30.42 (Oct. 2018), p. 425201. ISSN: 1361-648X. DOI: 10.1088/1361-648x/aadfc8. URL: http://dx.doi.org/10.1088/1361-648X/aadfc8.

[6] G. Engeln-Müllges and F. Reutter. *Formelsammlung zur numerischen Mathemathik mit C-Programmen*. 2. Auflage. BI-Wiss. Verlag, 1990.

[7] James F Epperson. "On the Runge example". In: *The American Mathematical Monthly* 94.4 (1987), pp. 329–341.

[8] Franziska Erlekam et al. "Modeling of Multivalent Ligand-Receptor Binding Measured by kinITC". In: *Computation* 7.3 (2019). ISSN: 2079-3197. DOI: 10.3390/computation7030046. URL: https://www.mdpi.com/2079-3197/7/3/46.

[9] Konstantin Fackeldey et al. *Analyzing Raman Spectral Data without Separability Assumption*. 2020. arXiv: 2007.06428 [math.NA].

[10] Thomas Gaumnitz et al. "Streaking of 43-attosecond soft-X-ray pulses generated by a passively CEP-stable mid-infrared driver". In: *Opt. Express* 25.22 (Oct. 2017), pp. 27506–27518. DOI: 10.1364/OE.25.027506. URL: http://www.opticsexpress.org/abstract.cfm?URI=oe-25-22-27506.

[11] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Third. The Johns Hopkins University Press, 1996.

[12] Brooke E. Husic and Vijay S. Pande. "Markov State Models: From an Art to a Science". In: *Journal of the American Chemical Society* 140.7 (2018). PMID: 29323881, pp. 2386–2396. DOI: 10.1021/jacs.7b12191. eprint: https://doi.org/10.1021/jacs.7b12191. URL: https://doi.org/10.1021/jacs.7b12191.

[13] Stefan Klus et al. *Data-driven approximation of the Koopman generator: Model reduction, system identification, and control*. 2019. arXiv: 1909.10638 [math.DS].

[14] A. Mauroy and J. Goncalves. "Linear identification of nonlinear systems: A lifting technique based on the Koopman operator". In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. Dec. 2016, pp. 6500–6505. DOI: 10.1109/CDC.2016.7799269.

[15] Cleve Moler and Charles Van Loan. “Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later”. In: *SIAM Review* 45.1 (2003), pp. 3–49. DOI: 10.1137/S00361445024180. eprint: https://doi.org/10.1137/S00361445024180. URL: https://doi.org/10.1137/S00361445024180.

[16] Vijay S. Pande, Kyle Beauchamp, and Gregory R. Bowman. “Everything you wanted to know about Markov State Models but were afraid to ask”. In: *Methods* 52.1 (2010). Protein Folding, pp. 99–105. ISSN: 1046-2023. DOI: https://doi.org/10.1016/j.ymeth.2010.06.002. URL: https://www.sciencedirect.com/science/article/pii/S1046202310001568.

[17] William W Parson. *Modern optical spectroscopy*. Vol. 2. Springer, 2007.

[18] Bernhard Reuter et al. “Generalized Markov State Modeling Method for Nonequilibrium Biomolecular Dynamics: Exemplified on Amyloid $\beta$ Conformational Dynamics Driven by an Oscillating Electric Field”. In: *Journal of Chemical Theory and Computation* 14.7 (2018). PMID: 29812922, pp. 3579–3594. DOI: 10.1021/acs.jctc.8b00079. eprint: https://doi.org/10.1021/acs.jctc.8b00079. URL: https://doi.org/10.1021/acs.jctc.8b00079.

[19] Susanna Röblitz and Marcus Weber. “Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification”. In: *Advances in Data Analysis and Classification* 7.2 (2013), pp. 147–179. DOI: 10.1007/s11634-013-0134-6.

[20] Susanne Röhl, Marcus Weber, and Konstantin Fackeldey. *Computing the minimal rebinding effect for non-reversible processes*. 2020. arXiv: 2007.08403 [math.NA].

[21] Christof Schütte, Péter Koltai, and Stefan Klus. “On the numerical approximation of the Perron-Frobenius and Koopman operator”. In: *Journal of Computational Dynamics* 3.1 (Sept. 2016), pp. 1–12. ISSN: 2158-2491. DOI: 10.3934/jcd.2016003. URL: http://dx.doi.org/10.3934/jcd.2016003.

[22] Renata Sechi, Alexander Sikorski, and Marcus Weber. “Estimation of the Koopman Generator by Newton’s Extrapolation”. In: *Multiscale Modeling & Simulation* 19.2 (2021), pp. 758–774. DOI: 10.1137/20M1333006. eprint: https://doi.org/10.1137/20M1333006. URL: https://doi.org/10.1137/20M1333006.

[23] Alexander Sikorski, Renata Sechi, and Luzie Helfmann. *cmdtools*. 2021. DOI: 10.5281/zenodo.4749330. URL: https://github.com/zib-cmd/cmdtools.

[24] Alexander Sikorski, Marcus Weber, and Christof Schütte. “The Augmented Jump Chain”. In: *Advanced Theory and Simulations* 4.4 (2021), p. 2000274. DOI: https://doi.org/10.1002/adts.202000274. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adts.202000274. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/adts.202000274.

[25] T. Stensitzki et al. “Ultrafast electronic and vibrational dynamics in brominated aluminum corroles: Energy relaxation and triplet formation”. In: *Structural Dynamics* 3.4 (2016), p. 043210. DOI: 10.1063/1.4949363. eprint: https://doi.org/10.1063/1.4949363. URL: https://doi.org/10.1063/1.4949363.

[26] Petre Teodorescu, Nicolae-Doru Stanescu, and Nicolae Pandrea. *Numerical analysis with applications in mechanics and engineering*. John Wiley & Sons, 2013.

[27] Ivo H.M. van Stokkum, Delmar S. Larsen, and Rienk van Grondelle. “Global and target analysis of time-resolved spectra”. In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1657.2 (2004), pp. 82–104. ISSN: 0005-2728. DOI: https://doi.org/10.1016/j.bbabio.2004.04.011. URL: http://www.sciencedirect.com/science/article/pii/S0005272804001094.

[28] Marcus Weber. *A Subspace Approach to Molecular Markov State Models via a New Infinitesimal Generator*. 2011.

[29] Marcus Weber. “Implications of PCCA+ in Molecular Simulation”. In: *Computation* 6.1 (2018). ISSN: 2079-3197. DOI: 10.3390/computation6010020. URL: https://www.mdpi.com/2079-3197/6/1/20.

[30] Marcus Weber and Konstantin Fackeldey. “Computing the Minimal Rebinding Effect Included in a Given Kinetics”. In: *Multiscale Model. Simul.* 12.1 (2014), pp. 318–334. DOI: 10.1137/13091124X.

[31] Marcus Weber, Konstantin Fackeldey, and Christof Schütte. “Set-free Markov state model building”. In: *The Journal of Chemical Physics* 146.12 (2017), p. 124133. DOI: 10.1063/1.4978501. eprint: https://doi.org/10.1063/1.4978501. URL: https://doi.org/10.1063/1.4978501.

[32] Luuk J. G. W. van Wilderen, Craig N. Lincoln, and Jasper J. van Thor. “Modelling Multi-Pulse Population Dynamics from Ultrafast Spectroscopy”. In: *PLOS ONE* 6.3 (Mar. 2011), pp. 1–14. DOI: 10.1371/journal.pone.0017373. URL: https://doi.org/10.1371/journal.pone.0017373.

[33] Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. “A Data–Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition”. In: *Journal of Nonlinear Science* 25.6 (June 2015), pp. 1307–1346. ISSN: 1432-1467. DOI: 10.1007/s00332-015-9258-5. URL: http://dx.doi.org/10.1007/s00332-015-9258-5.

[34] Clark Zahn et al. “Ultrafast Dynamics of Sb-Corroles: A Combined Vis-Pump Supercontinuum Probe and Broadband Fluorescence Up-Conversion Study”. In: *Molecules* 22.7 (2017). ISSN: 1420-3049. DOI: 10.3390/molecules22071174. URL: https://www.mdpi.com/1420-3049/22/7/1174.

## 13 Acknowledgements

I would like to thank Dr. Jose Luis Pérez Lustres, without whose insight and patience this thesis would not have been possible . I am also grateful for the support and friendship of my colleagues at AG Heyne (FUB) and the Computational Molecular Design Group (ZIB).

# A Illustrative examples: transition matrix

Hereby 3 processes with different underlying dynamics are analyzed with the two aforementioned methods. At beginning of each subsection, there is a short summary of the results and their interpretation. Then each transition matrix is analyzed in detail.
The represented dynamics are: Process 1: $A$ decays, $B$ decays. The two species do not communicate
Process 2: $A \to B \to$. It is a directed reaction.
Process 3: $A \leftrightarrow B$ and the system reaches equilibrium fast.
The parameters of the penalty function in the MF used are: $\beta = 100, \gamma = 10, \delta = 1, \mu = 10$. For all the data the analysis starts from the row 50. In the following description of the results, the system's conformations, are referred to as $1, 2, 3$.
Note that adding more Voronoi cells for the discretization of the state space improves the characterization of the dynamics (the sink state is more absorbing/sinking, the transitions are clearer). See comparison for matrix 2 in its section.

| Process | Dynamics | MF with PCCA+ | MSM |
|---|---|---|---|
| #1 Parallel Decay | $[A+B] \to$ | 2 dominant-conformations: 1=$[A+B] \to$ and 2= the system is empty | directed dynamics from state 1 to state 2. |
| #2 Sequential Decay | $A \to B \to$ | from 1 go to 2; 2 is sink state. | from 1 go to 2; 2 is sink state. state 1: the system is populated, state 2: the system is empty |
| #3Reversible Process | $A \leftrightarrow B$ | 2 communicating dominant-conformations but almost sink state. 1: the system is equilibrating, 2: the system is equilibrated | from state 1 to 2, 2 is sink conformation (from analysis of the first 1000 times) |

Table 4: Comparison of the interpretation of the dynamics with MF with PCCA+ and MSM. The processes have been analyzed also with 3 clustering conformations in order to have a better understanding of the meaning of the two clusters.

## A.1 Process 1

### A.1.1 MSM

(A+B decay), 50 voronoi,

$$K(\tau) = \begin{pmatrix} 0.99048 & 0.0118779 & -0.00235808 \\ 0.00236807 & 0.992563 & 0.00506864 \\ 3.89444e-05 & 0.000694883 & 0.999266 \end{pmatrix} \tag{65}$$

shows a sequential dynamics, directed from conformation $1- > 2- > 3$. analysis with 2 conformations, since the representation of the $\chi$ shows a vector that is always smaller than the other 2, suggesting that the number of clusters used is too high.

$$K(\tau) = \begin{pmatrix} 0.99666 & 0.00333953 \\ 0.000352406 & 0.999648 \end{pmatrix} \tag{66}$$

shows 2 species, one decays and the other rises, better representation of the dynamics.

### A.1.2 MF with PCCA+

- 3 clusters:

$$P_{rec3} = \begin{pmatrix} 0.997836 & 0.00583599 & -0.00378184 \\ 0.00135933 & 0.995968 & 0.00274842 \\ -0.00112329 & 0.00343267 & 0.997626 \end{pmatrix} \tag{67}$$

  Interpretation: from 1 go to 2, from 2 likely to 3 and less likely to 1, from 3 go to 2 back.

- 2 clusters:

$$P_{rec2} = \begin{pmatrix} 0.999921 & 7.88226e-05 \\ 1.59912e-05 & 0.999984 \end{pmatrix} \tag{68}$$

  Interpretation: two sink conformations. From this analysis one sees that the transition between the conformations is not allowed. This can suggest that: (i) the system has two independent conformations that do not communicate A, B; (ii) a conformation (A+B) decays to form a product, a conformation (?)rise form the decay of (A+B)

## A.2 Process 2

### A.2.1 MSM

matrix2,(A zu B) 25 voronoi cells,

$$K(\tau) = \begin{pmatrix} 0.992791 & 0.00838098 & -0.00117236 \\ 0.0140016 & 0.973656 & 0.0123425 \\ -0.0016633 & 0.0143459 & 0.987317 \end{pmatrix} \tag{69}$$

2 clusters, 25 Voronoi cells:

$$K(\tau) = \begin{pmatrix} 0.994477 & 0.00552344 \\ 0.00291951 & 0.99708 \end{pmatrix} \tag{70}$$

Now with 50 Voronoi cells:

$$K(\tau) = \begin{pmatrix} 0.995033 & 0.0049667 \\ 0.00258549 & 0.997415 \end{pmatrix} \tag{71}$$

Interpretation: not clear, goes from conformation 1 to conformation 2 and less likely to come back to 1. 3 clusters: 25 voronoi,

$$K(\tau) = \begin{pmatrix} 0.992791 & 0.00838098 & -0.00117236 \\ 0.0140016 & 0.973656 & 0.0123425 \\ -0.0016633 & 0.0143459 & 0.987317 \end{pmatrix} \tag{72}$$

50 voronoi:

$$K(\tau) = \begin{pmatrix} 0.99318 & 0.00795601 & -0.00113582 \\ 0.0125521 & 0.976994 & 0.0104538 \\ -0.00139596 & 0.0121704 & 0.989226 \end{pmatrix} \tag{73}$$

Interpretation:both matrices have negative entries, but they describe the same process, only with different precision in the decimal numbers. From 1 goesn to 2, from 2 goes almost with the same probability to 3 and 1. From 3 only go back to 2. This indicates probably that conformation 2 is an intermediate conformation and "switches over" the direction of the reaction to conformation 3. Probably conformation 2 and 3 can be clustered.

### A.2.2 MF with PCCA+

- 3 clusters:

$$P_{rec3} = \begin{pmatrix} 1.02252 & 0.0258619 & -0.0459583 \\ -0.0733033 & 0.916256 & 0.149195 \\ 0.00319144 & 0.00356556 & 0.993576 \end{pmatrix} \tag{74}$$

  Interpretation: from 1 go to 2; from 2 go to three; from 3 go equally probably to 1 and 2. The fact that the system has the same probability to go back to 1 and 2 suggests that 1 and 2 are "similar", or better that they can be further clustered into the same conformation. Moreover, in this sense the clustered 1-2 conformation is also sink.

- 2 clusters:

$$P_{rec2} = \begin{pmatrix} 0.999852 & 0.000148112 \\ 9.81847e-05 & 0.999902 \end{pmatrix} \tag{75}$$

  Interpretation: from 1 go to 2; 2 is sink conformation.

## A.3 Process 3

### A.3.1 MSM

For the following analysis, 50 Voronoi cells have been used to discretize the conformation space. All data, 2 clusters:

$$K(\tau) = \begin{pmatrix} 0.986476 & 0.0135239 \\ 3.67097e-05 & 0.999963 \end{pmatrix} \tag{76}$$

Interpretation: from 1 go to sink conformation 2. Try to analyse again with 3 clusters:

$$K(\tau) = \begin{pmatrix} 0.974192 & 0.032637 & -0.00682919 \\ 0.00863059 & 0.963833 & 0.0275368 \\ 9.38688e-05 & -1.80855e-05 & 0.999924 \end{pmatrix} \tag{77}$$

$1->2$ from 2 goes to 3 and 1. 3 is sink conformation, again.
SInce the described process is very fast, I think there is an oversampling of the bigger times scales. So I analyse only the early times data:
- from 50 to 2000 time points in the matrix:

$$K(\tau) = \begin{pmatrix} 0.976333 & 0.0330198 & -0.00935323 \\ 0.00461976 & 0.970113 & 0.025267 \\ 0.000119269 & 0.000198472 & 0.999682 \end{pmatrix} \tag{78}$$

from 1 to 2, from 2 to 3 and from 3 equally going from 1 and 2, so 1 and 2 are the same conformation? cluster with 2:

$$K(\tau) = \begin{pmatrix} 0.985723 & 0.0142769 \\ 0.000103479 & 0.999897 \end{pmatrix} \tag{79}$$

Since the dynamics is still not clear enough, the analysis is limited to the first 1000 time points of the data:

$$K(\tau) = \begin{pmatrix} 0.978013 & 0.0290124 & -0.00702536 \\ 0.00356202 & 0.973437 & 0.023001 \\ 7.70176e-05 & 0.000634617 & 0.999288 \end{pmatrix} \tag{80}$$

from 3 i go only to 2, from to i likely come back to 3 and i go a bit to 1, from 1 i can go back to 2. I would cluster conformation 2 with 3. now same but with 2 clusters:

$$K(\tau) = \begin{pmatrix} 0.985837 & 0.014163 \\ 0.000197261 & 0.999803 \end{pmatrix} \tag{81}$$

The sink conformations, so not clear. Nonetheless, the $\chi$ vectors show the correct dynamics. The sink conformation represents the system in equilibrium and the other conformation represents the equilibration process of species A and B.

### A.3.2 MF with PCCA+

- 3 clusters:

$$P_{rec3} = \begin{pmatrix} 0.99178 & 0.0182374 & -0.0101492 \\ -0.00321412 & 0.957116 & 0.046416 \\ 0.000919371 & 0.00896158 & 0.990053 \end{pmatrix} \tag{82}$$

  Interpretation: cluster conformation 1 with conformation 2 because communicating. conformation 3 is the second conformation.

- 2 clusters:

$$P_{rec2} = \begin{pmatrix} 0.990641 & 0.00958259 \\ 0.000512424 & 0.999475 \end{pmatrix} \tag{83}$$

  Interpretation: the conformation 1 and 2 are communicating bust almost sink conformation.

# B Illustrative examples: rate matrix for 3 dominant conformations

In section

$$K^c(\tau) = \begin{pmatrix} 1. & -0. & 0. & 0. & -0. \\ 0.004 & 0.996 & 0. & 0. & -0. \\ -0.002 & 0.006 & 0.996 & -0. & 0. \\ 0.001 & -0.004 & 0.006 & 0.995 & 0.001 \\ -0. & 0.002 & -0. & 0.007 & 0.992 \end{pmatrix} \tag{84}$$

# C Figures for comparisons

To ease the comparison to the analysis of the Brominated Al-corrole [25] and the Sb-corrole, the figures needed for the comparison are reprinted in the following.

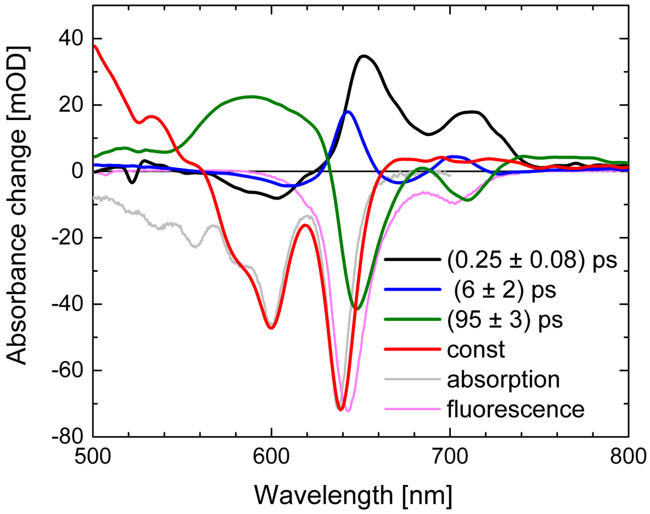


Figure 23: Global analysis of brominated Al-corrole, fit with 4 DAS. Reprint from [25].

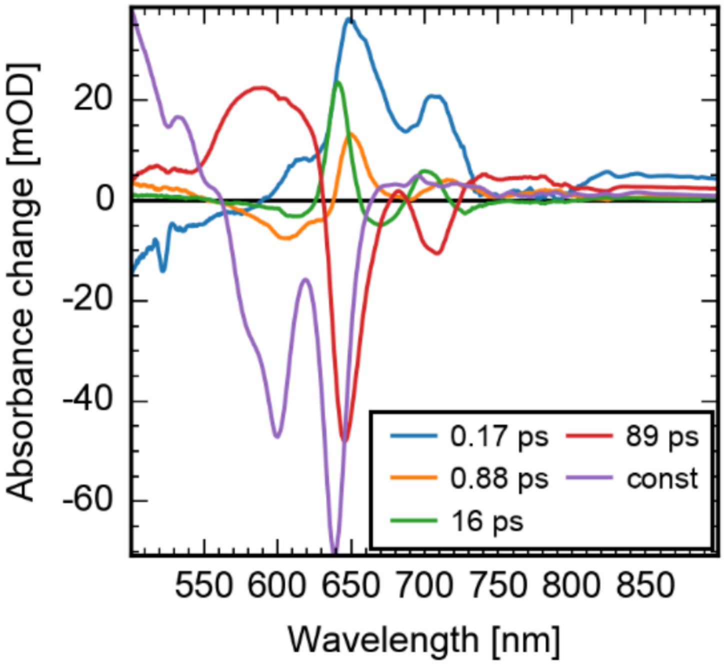


Figure 24: Global analysis of brominated Al-corrole, fit with 5 DAS. Reprint from [25].

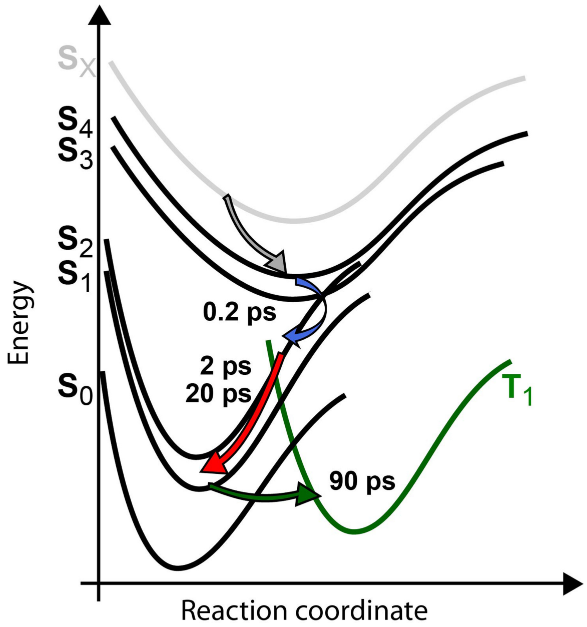


Figure 25: Energy level diagram resulting from the global analysis of the Brominated Al-corrole dataset. Reprint from [25].

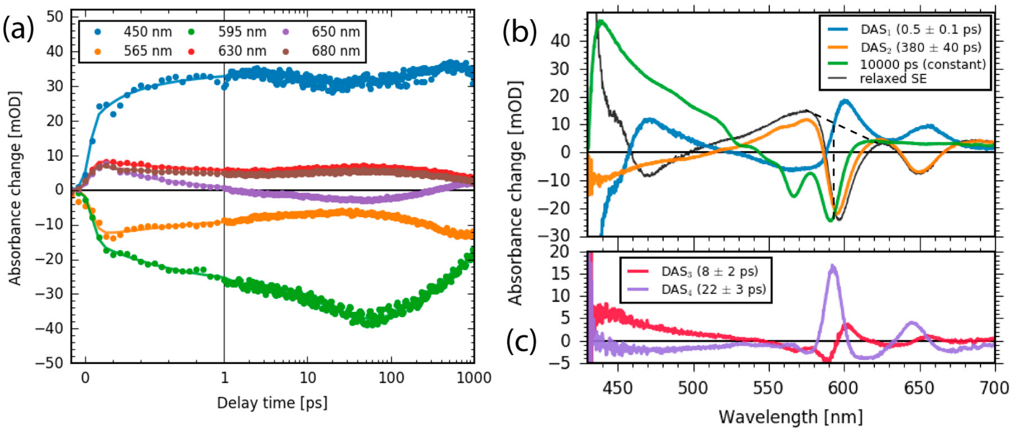


Figure 26: Global analysis of Sb- corrole with 5 DAS. Reprint from [34]

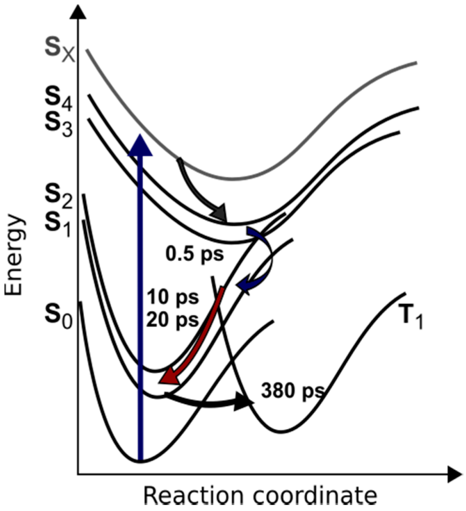


Figure 27: Energy level diagram resulting from the global analysis of the Sb-corrole dataset. Reprint from [34].

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende

Masterarbeit selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe.

Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch bei keiner anderen Universität als Prüfungsleistung eingereicht.

10.07.2021

Renata Sechi