STOCHASTIC MODELING OF BIOMOLECULAR SYSTEMS USING THE DATA-DRIVEN LANGEVIN EQUATION

Dissertation

vorgelegt von
Norbert Schaudinnus

zur Erlangung des Doktorgrades der
Fakultät für Mathematik und Physik der
Albert-Ludwigs-Universität, Freiburg im Breisgau

Betreut durch
Gerhard Stock

August 2015
<table>
<thead>
<tr>
<th>Dekan</th>
<th>Prof. Dr. Dietmar Kröner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betreuer</td>
<td>Prof. Dr. Gerhard Stock</td>
</tr>
<tr>
<td>Gutachter</td>
<td>Prof. Dr. Gerhard Stock</td>
</tr>
<tr>
<td></td>
<td>Apl. Prof. Dr. Heinz-Peter Breuer</td>
</tr>
<tr>
<td>Datum der mündlichen Prüfung</td>
<td>4.11.2015</td>
</tr>
</tbody>
</table>
List of Publications


List of Publications

1. Introduction 1

2. Theory and methods 7
   2.1. Molecular Dynamics (MD) Simulations 7
   2.2. Free Energy Landscapes (FEL) 10
   2.3. Principal Component Analysis (PCA) 11
   2.4. Autocorrelation 14
   2.5. States and Transition Rates 15
      2.5.1. Defining a state 16
      2.5.2. Relative clustering via relative coring 18
      2.5.3. Transition rates 18
      2.5.4. Markovianity 19
      2.5.5. Kramers’ theory 20

3. Model System and Hierarchical Dynamics 21
   3.1. AIB9 21
      3.1.1. Adequate collective coordinates 22
      3.1.2. Network of states 25
   3.2. Hierarchical Dynamics 28
      3.2.1. Hierarchical dependency of individual tiers 30
   3.3. Summary and conclusions 40

4. Data-driven Langevin modeling 41
   4.1. The Langevin equation 41
      4.1.1. Zwanzig’s approach 42
      4.1.2. Markovian Langevin equations 45
      4.1.3. Ito and Stratonovich calculus 46
      4.1.4. Noise-induced drift 48
   4.2. The Fokker-Planck Equation 48
      4.2.1. Solutions for small time steps 50
      4.2.2. Relation to Markovian Langevin equations 51
   4.3. The data-driven Langevin equation (dLE) 52
      4.3.1. The first order data-driven Langevin equation (dLE1) 52
      4.3.2. The second order data-driven Langevin equation (dLE2) 55
8.2.2. A generalizable fitting algorithm .................. 122
8.2.3. Examples ............................................. 126
8.2.4. Future perspectives .................................. 129

9. Conclusions .................................................. 131

A. Rates, transitions and first passage times ................. 135
   A.1. Transition and rate matrix, normalization ............... 135
   A.2. Continuity equation, stationarity and detailed balance .. 136
   A.3. Mean first passage times (MFPT) and transition times ... 136
      A.3.1. The MFPT ......................................... 136
      A.3.2. Two domains and relation to the transition time .... 137
      A.3.3. more than two domains ............................ 137
      A.3.4. Application to AIB₉ ................................ 139

List of Figures ................................................ 141

Bibliography .................................................. 143
### List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td>Molecular Dynamics</td>
</tr>
<tr>
<td>FEL</td>
<td>Free Energy Landscape</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Component</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>dPCA</td>
<td>dihedral Principal Component Analysis</td>
</tr>
<tr>
<td>LE</td>
<td>Langevin equation</td>
</tr>
<tr>
<td>mLE</td>
<td>model Langevin equation</td>
</tr>
<tr>
<td>mLE1</td>
<td>first order model Langevin equation</td>
</tr>
<tr>
<td>mLE2</td>
<td>second order model Langevin equation</td>
</tr>
<tr>
<td>dLE</td>
<td>data-driven Langevin Equation</td>
</tr>
<tr>
<td>dLE1</td>
<td>first order data-driven Langevin equation</td>
</tr>
<tr>
<td>dLE2</td>
<td>second order data-driven Langevin equation</td>
</tr>
</tbody>
</table>
Biomolecules are fundamental components of living organisms. Their variety ranges from small amino acids and peptide systems to large macromolecules such as proteins and are different in terms of biological functioning. While peptide systems may consist of only few amino acids, proteins form composites of long chains of amino acids or poly-peptides. Proteins perform a huge variety of regulatory functions like catalysis, DNA replication and transport, intra-cellular communication and are indispensable for many other biological processes. The specific appearance of proteins is regulated by biosynthesis inside cells. They adopt their functional shape by the process of protein folding. Within this particular procedure, they assume many intermediate metastable conformations which occasionally possess long lifetimes. The structure of proteins not only codifies their biological application but also the time scales of conformational changes. While protein folding takes place on a time scale of nanoseconds to seconds, transitions between intermediate states occur much more frequently. In fact, a large number of intermediate conformations of biomolecules that are crucial for their function exists which are primarily thermally populated [1, 2].

Experimentally, approved techniques like X-ray crystallography, magnetic resonance spectroscopy or force spectroscopy have been able to give insight into the composition of proteins [3–5]. However, detailed dynamical information on the folding process is not available in these experiments; the time-resolved accuracy is limited to averaged structures and specific events [6–8]. Therefore, events which are measured in experiments are usually only indirectly linked to the molecular structure.

Among other methods, classical molecular dynamics (MD) simulations have proven to recover experimental results, describing the structure and dynamics at atomistic detail [9–11]. Within this framework, the Newton equations of motion are propagated taking into account the interactions which dominate the folding process like hydrogen bonds, van der Waals interactions, electrostatic interactions and so on [12]. The interaction force fields are determined both empirically and by quantum chemical calculations. MD simulations have proven to be able to capture structure, dynamics and functioning of various
biomolecular systems. Although an increase of computational power over the last decade facilitates to access time scales of microseconds for comparably small and fast-folding proteins, relevant functional dynamics are often still beyond reach [13–16]. Improved schemes have been suggested to close that gap [17–22].

MD simulations generate huge amounts of data, resulting in a discretized time series of all atoms in Cartesian coordinates. However, one is often interested in few processes which occur on comparatively large time scales. In order to establish an appropriate subspace and restrict the perspective accordingly, several systematic approaches have been conducted to select a subset of system coordinates [23–27]. One of the most common methods is the principal component analysis [28–31]. To analyze and study dynamics on few designated collective coordinates, the Free Energy Landscape (FEL) provides an illustrative picture. It was proposed by Chan, Dill, Onuchic, Thirumalai, Wolynes and others [32–37] and describes conformational changes as sampling events along a rugged multidimensional energy surface. Dynamics along the FEL can be understood as propagations along a rugged surface driven by thermal fluctuations, as visualized in Figure 1.2. While local minima represent metastable states, different barrier heights denote the connectivity between states; high barriers correspond to sparse transition events while low barriers indicate frequent transition paths [35, 38–42].

Figure 1.1.: Time scales of different sampling schemes are also limited by the respective size of the system studied. Exact quantum calculations are limited to the shortest timescales, while MD simulations, working on atomistic detail are slower than coarse-graining methods. Monte-Carlo methods are able to even go beyond this time scale. Combined methods like QM/MM allow for an increased system size. Advanced sampling methods are designed to increase the accessible timescale for systems which show rare transition events. Illustration adapted from [14].
There are two different perspectives. On the one hand, the funnel picture relies on the internal free energy of a system. A designated state - the global minimum - exists which is referred to as the native state. It denotes the state in which the protein adopts if its properly folded conformation. A more recent picture defines a state as an ensemble of conformations which are structurally similar. Densely clustered regions therefore specify minima of the FEL while sparse regions represent transition states [36, 43]. This definition causes comparably larger local barriers or rather a more rugged FEL featuring a high diversity of conformational pathways [44]. Except for the property that neighboring states are kinetically associated with each other, the shape of the FEL depends on the collective coordinate along which it is evaluated.

Several models have been proposed to describe the dynamics of complex systems in a reduced space [45, 46]. Markov state models have become a famous approach to investigate the dynamics of protein folding [47–57]. They establish a network model which is governed by time discrete master equations. In this context, different conformations are considered as states which are interconnected by transition events. The object of Markov state models is to predict dynamics which occur on large time scales from short
MD samples\textsuperscript{1}. However, they rely on an a priori definition of discretized microstates and specific conditions, i.e. the Markov property. Besides, the definition of states is not that straightforward and can be ambiguous \[58\].

In contrast to Markov state models, this work will focus on an approach which is also based on MD sampling and does not rely on coarse-graining. It relies on the microscopic Langevin equation which was initially derived in order to describe Brownian dynamics \[59\]. Structurally, it resembles a Newton equation of motion but contains an additional stochastic force. It can be derived from first principles by introducing a system which is bilinearly coupled to the environment, the bath \[60–62\].

By further assuming a time scale separation between system and bath coordinates, one obtains a Markovian second order Langevin equation which reads

\[
\ddot{x} = f(x) - \Gamma(x) \dot{x} + F_s(x).
\]

It includes a Newton force \(f(x)\), a Stokes' friction term \(\Gamma(x) \dot{x}\) and the stochastic force \(F_s(x) = K(x) \xi(t)\) \[65–70\]. The latter in turn contains a diffusion field \(K\) and a stochastic noise variable \(\xi\) which is usually characterized by a Gaussian distribution with zero mean and unit variance, \(\langle \xi(t) \rangle = 0\) and \(\langle \xi_i(t) \xi_j(t') \rangle = \delta(t - t') \delta_{ij}\). The dynamics of this equation is fully determined by the three fields \(f\), \(\Gamma\) and \(K\).

In the overdamped limit, that is if \(\ddot{x} \ll \Gamma \dot{x}\), the Langevin equation reduces to

\[
\dot{x} = h(x) + D(x) \xi(t),
\]

which is described by a deterministic drift field \(h\) and a diffusion field \(D\). For this form of the Langevin equation, many efforts have been made recently to gain field estimates from a time series either for analytical aspects or in order to propagate dynamics of a system \[71–79\]. The data-driven Langevin equation (dLE) is a method to forecast the dynamics of a low-dimensional set of system coordinates \[80\]. These are specific collective coordinates. According to the requirements of the noise variable \(\xi\), the time scale on which the system coordinate is propagated has to exceed the memory kernel-decay of the bath which summarizes the remaining degrees of freedom. The basic idea of a data-driven description of a system is to correctly estimate the stationary fields from an underlying time series. The field information in turn is applied to generate an independent time series for the system coordinate. Hence, the algorithm operates in an iterative manner. With regard to the FEL, the dLE operates continuously on a suitable projection. This means that at each point of a free energy surface it is able to describe the dynamics and to deduce local information. On the contrary, Markov state models identify distinct states which

\textsuperscript{1}They can therefore be ranked among the rare event methods in Figure 1.1.

\textsuperscript{2}Independently, Caldeira and Leggett provided a similar approach for dissipative quantum systems \[63, 64\].
correspond to local minima of the FEL and summarize barrier information in transition rates. 

In this thesis, the data-driven Langevin algorithm and its application to multidimensional systems is studied, particularly concerning the application to biomolecular systems. It was recently shown to qualitatively recover the dynamics of a small peptide system [80]. This work also aims to further understand the adaptability of the data-driven Langevin equation and generalize its application spectrum. First of all, the robustness of the dLE concept is analyzed. Therefore, we for instance bias the information of the underlying data and study the influence onto the dLE performance. We also investigate the robustness of the field estimates in relation to the local sampling. An extended version of the dLE algorithm is provided in this work which allows for the propagation of systems that are governed by this second order Langevin equation. It will be shown, that we rely on this generalized form for a typical system coordinate which describes the dynamics of a biomolecular system. The specific system which we will use as a benchmark is the AIB$_9$ peptide. We elaborately studied the dynamical behavior of this small system and also discovered that it features a hierarchical mechanism that triggers conformational changes on the longest timescale, namely the transition from left-handed to right-handed helical structure. This specific system also serves as a reference for an advanced sampling tool. For this purpose we exploit the dLE feature that it does not rely on continuous input data. This application strongly resembles the concept of Markov state modeling, yet it does not request a clear separation of states.

This thesis is structured as follows: In Chapter 2, the central methods which we rely on in the remainder of this work will be discussed. A rough overview of molecular dynamics simulations will be given. Thereafter, the process to establish collective coordinates is delineated which in turn can be applied for further dLE. The remainder of this chapter will focus on quantifying dynamical properties of coordinates. A clustering scheme will be detailed which allows to identify states and number the interaction between pairs of states. Particularly, a scheme to identify states of a system in relation to an underlying data sample will be mentioned to evaluate the performance of the dLE.

Chapter 3 introduces and characterizes the small peptide system AIB$_9$ which is also central to our Langevin modeling, solvated both in polar water and non-polar chloroform. We observed a hierarchy of processes for this small peptide system in chloroform solution, relating microsecond conformational dynamics to sub-picosecond fluctuations of hydrogen bonds. Based on extensive MD simulations, we present a simple model to embed these multiscale dynamics and relate different tiers in the dynamic hierarchy of processes to one another. We also discuss the temperature dependency of the dynamics in this context.

In Chapter 4, we provide all fundamentals which are required for the concept of data-driven Langevin. This chapter is split in two parts: In the first part, the theoretical
fundamentals of the Langevin equation are introduced. Zwanzig’s approach to derive the formalism from first principles is elucidated, which in the limiting case results in a Markovian Langevin equation. I will briefly review the interrelation of the Langevin calculus with the Fokker-Planck equation, introducing the Kramers-Moyal coefficients which set the fields for the first order Langevin equation. The second part of Chapter 4 sets focus on the data-driven Langevin equation algorithm. Two different definitions for an estimator are introduced and discussed and field estimates are derived for the first order algorithm and extended to second order. In the remainder, general features of both algorithms are marked out.

Chapter 5 discusses different factors which influence the performance of the dLE on the basis of a simple one-dimensional double well model. Providing a model trajectory of a system for which the statistical and dynamical behavior is known, we apply simple modifications in order to analyze the effects on the dLE performance. Besides, two modifications of the input data are presented: ”Data pruning” is a preprocessing method that aims to reduce the amount of input data without harming the accuracy of the field estimates. It achieves a speed-up of the dLE algorithm. Delay embedding on the other hand is a method to preprocess the input data such that the first order dLE is able to imitate the performance of a system which is governed by a second order Langevin equation.

Chapter 6 concentrates on the application of the second order dLE to a biomolecular system, the AIB₉ peptide in chloroform solvent. On the basis of this system the necessity of the extended algorithm is exemplified. We will explicitly demonstrate and justify the capabilities of the second order algorithm as compared to the first order dLE. Beforehand, a detailed consideration of the second order dLE will be given. Among others, we compare different implementations and motivate our selection.

In Chapter 7, we introduce the application of the dLE for an enhanced sampling scheme. We set up a workflow which is based on many short MD samples that overlap in system space. Using the property of the dLE to not rely on continuous input data, we are able to predict long-term dynamics without extensive sampling. We will focus on the major challenges of this concept.

Chapter 8 gives an outlook on two more applications of the dLE: The concept of temperature rescaling is a different enhanced sampling approach. It is supposed to forecast the dynamics of a system at lower temperature provided that MD simulations at higher temperatures are available. A second application concerns the fields which determine the propagation of the Langevin equation. We introduce an algorithm which is able to model fields of a set of multidimensional system coordinates by coupled harmonic oscillators.

The last chapter will summarize and conclude the achievements of this work.
Many of the methods which are used in the remainder of this work are fundamental in the field of biomolecular dynamics. These “basics” will be discussed independently from the Langevin framework in this chapter, which is organized as follows. At the beginning, a short introduction to Molecular Dynamics simulations will be given. The basic ideas of the concept will be summarized in this part. I will further introduce methods to reduce the dimensionality of a biomolecular system and employ sets of system coordinates. This particular issue is used to develop a low-dimensional projection that maps all relevant features. It is also required for the concept of data-driven Langevin modeling which will be introduced later on. Static and dynamical measures to interpret and evaluate such a set of system coordinates will be listed and discussed later in this chapter. The remainder will focus on the identification of states and interaction of different states.

### 2.1. MD Simulations

MD simulations are an integral part of theoretical analysis of biomolecular dynamics [81]. Standard MD simulations are performed in atomistic detail. They propagate the Newtonian equations of motion in a classical manner. For $N$ interacting atoms the equations of motion read

$$m_i \frac{\partial^2 r_i}{\partial t^2} = F_i,$$

where $i \in 1..N$ denotes atom $i$ of mass $m_i$ at position $r_i$. 
Force field  The force $F_i = -\sum_j \frac{\partial V(r_{i,\ldots,r_N})}{\partial r_{i,j}}$ on each particle is derived from the potential energy. The latter typically has the form

$$V = V_{\text{bonds}} + V_{\text{angles}} + V_{\text{dihedrals}} + V_{\text{improper}} + V_{\text{non-bonded}} \quad (2.2)$$

$$V_{\text{bonds}} = \sum_{ij \in \text{bonds}} k_{b,ij} \frac{1}{2} (r_{ij} - r_{ij,0}) \quad (2.3)$$

$$V_{\text{angles}} = \sum_{ijk \in \text{angles}} \frac{k_{a,ijk}}{2} (\phi_{ijk} - \phi_{ijk,0}) \quad (2.4)$$

$$V_{\text{dihedrals}} = \sum_{ijkl \in \text{dihedrals}} k_{d,ijkl} (1 + \cos(n\theta_{ijkl} - \theta_{ijkl,0})) \quad (2.5)$$

$$V_{\text{improper \, dihedrals}} = \sum_{ijkl \in \text{imp.dih.}} \frac{k_{id,ijkl}}{2} (\xi_{ijkl} - \xi_{ijkl,0})^2 \quad (2.6)$$

$$V_{\text{non-bonded}} = V_{\text{L.J.}} + V_{\text{Coulomb}} = \sum_{i<j} \left( \frac{C_{ij,12}}{r_{ij}^{12}} + \frac{C_{ij,6}}{r_{ij}^6} + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right). \quad (2.7)$$

While bonds and angles are modeled harmonically, the periodicity of dihedral angles (see 2.2) motivates the cosine function. The improper dihedral components basically restrict torsion in dihedral space. The non-bonded interactions contain lowermost Coulomb interactions. Furthermore the Pauli repulsion preventing atoms from penetrating each other and dipole dipole interactions (van der Waals) are usually modeled by a Lennard-Jones potential for computational reasons, see (2.7). Equilibrium values ($r_{ij,0}$, $\phi_{ijk,0}$, $\theta_{ijkl,0}$, $\xi_{ijkl,0}$), all force constants ($k_j$), partial charges ($q_i$) and Lennard Jones coefficients ($C_6$, $C_{12}$) are determined via both ab initio calculations and fitting to experimental measurements.

Integration schemes  There are many popular integration schemes used to propagate the Newtonian equations (2.1). Besides numerical stability, they have to feature both consistency and convergence. One of the most prominent algorithms which has also been used in the simulations of our systems, is the Leapfrog-integrator. It iteratively calculates positions and velocities in the following manner:

$$v_i \left( t + \frac{\Delta t}{2} \right) = v_i \left( t - \frac{\Delta t}{2} \right) + \frac{F_i(t)}{m_i} \Delta t \quad (2.8)$$

$$r_i (t + \Delta t) = r_i(t) + v_i \left( t + \frac{\Delta t}{2} \right) \Delta t \quad (2.9)$$

As the name of the algorithm suggests, the propagation of velocities so to speak leapfrogs position propagation, being shifted by a time step $\Delta t/2$. The algorithm is third order convergent in $r_i$. Being ab-initio, prior to the simulation a starting configuration is set.
Chapter 2. Theory and methods

The Velocities are therefore typically sampled Maxwell-Boltzmann distributed, while the positions are usually determined from experimental findings, e.g. crystal structure may be accessed by X-ray scattering experiments. An energy minimization prior to the actual simulation is applied to equilibrate the system and determine initial positions $r_i$.

MD simulations can be performed considering different statistical ensembles. While the micro-canonical ensemble (NVE) simply integrates eqs. (2.8),(2.9), other ensembles involving temperature or pressure coupling require additional constraints that are realized in the coupling to a bath.

**Temperature coupling** The temperature of the system can be derived from the total kinetic energy of the system,

$$E_{\text{kin,tot}} = \sum_i \frac{m_i}{2} v_i^2 = \frac{N_{df}}{2} k_B T,$$

(2.10)

where $N_{df}$ denotes the degrees of freedom. When simulating a canonical ensemble (NVT), the temperature has to be kept constant. This is realized by schemes that smoothly correct deviations of the system temperature $T_0$ from the actual temperature $T$ by rescaling the velocities in eq. (2.10). The simplest thermostat for example, the Berendsen-thermostat, adjusts $T$ following the differential equation $\frac{dT}{dt} = \frac{T - T_0}{\tau}$, where $\tau$ sets the convergence speed. A more tricky temperature coupling is achieved with the “Velocity-Rescaling”-thermostat [82]: in addition to the Berendsen-thermostat, it includes a stochastic correction. The kinetic energy of the system is then adjusted according to

$$\frac{dE_{\text{kin}}}{dt} = \frac{E_{\text{kin},0} - E_{\text{kin}}}{\tau} + 2 \sqrt{\frac{E_{\text{kin}}E_{\text{kin},0}}{N_{df}\tau}} dW,$$

(2.11)

where second term denotes a Wiener process (which is discussed in the context of Langevin Modeling in Ch. 4 in more detail). For systems far away from equilibrium, the deterministic part which represents the Berendsen-thermostat dominates. The stochastic part essentially enhances the quality of the temperature distribution.

Pressure coupling and more elaborate modifications of the above described scheme are carried out following similar ideas. They are not central to this work and detailed for example in the Gromacs manual [83].
2.2. Free Energy Landscapes (FEL)

The FEL is a static quantity that characterizes a system and is often used to investigate properties of biomolecular systems [84–87]. It is defined via

\[ \Delta G(r) = -k_B T \ln P(r), \]  

(2.12)

where \( P \) denotes the probability distribution associated with an in general multidimensional reaction coordinate \( r \), \( T \) the temperature and \( k_B \) the Boltzmann constant. Conventionally, the global minimum of the free energy landscape is set \( G = 0 \) via \( G(\max_r P(r)) = 0 \). The shape of the free energy landscape provides information about the stability of states, which are represented by local minima, their connectivity and transition paths or barriers separating them. Notably, in contrast to the full coordinate space, the choice of a reaction coordinate already yields a projection of the FEL. Assume, the full picture is monitored along the coordinate \( r^T = (x^T, r_{n+1}, ..., r_N) \), we get the projection

\[ \Delta G(x) \sim -k_B T \ln \int dr_{n+1}... \int dr_N P(x, r_{n+1}, ..., r_N). \]  

(2.13)

Thus, since the structure of the FEL depends on the choice of the reaction coordinate in a crucial manner, its significance relies on the selected collective coordinate.

MD simulations provide a high dimensional picture of Cartesian coordinates that include overall translation and rotation. Apart from that, minor important fluctuations of side chains are enclosed in this picture. The latter are irrelevant to understand the basic conformational dynamics of the overall system.

In order to monitor internal conformational changes and restrain to the variations that substantially effect structural changes, the dihedral angles \( \phi, \psi \) along the backbone are often applied. The joint distribution of a pair of dihedral angles \( P(\phi_i, \psi_i) \) that are assigned to the same residue \( i \), the so-called Ramachandran plot or simply Ramachandran, visualizes and distinguishes conformational states of that residue.

**Alanine peptides** Throughout this thesis, methods and concepts will often be exemplified on the basis of two small alanine peptides. Particularly, we will restrain to dialanine and hepta-alanine. Both systems have been subject to research in many applications. For instance, hepta-alanine was introduced related to a specific coordinate transformation [31], see page 12. Dialanine has been used as a benchmark system to test several methodologies. Lastly, it has been simulated in the group at different temperatures, to study the temperature-dependent behavior in the framework of a bachelor thesis\(^1\).

\(^1\)We will return to this aspect later, in Sec. 8.1.
All poly-alanines have a characteristic Ramachandran in common which is shown in Figure 2.1. E.g. the locations \((\phi_i, \psi_i) \approx (-50^\circ, -50^\circ)\) account for the right-handed and \((\phi_i, \psi_i) \approx (50^\circ, 50^\circ)\) for the left-handed \(\alpha\)-helical structure of the single residue, while \((\phi_i, \psi_i) \approx (-135, 135)\) is assigned to the \(\beta\)-sheet structure of this residue. The free energy picture of the distribution \(-k_B T \ln P(\phi_i, \psi_i)\) provides transition paths connecting these metastable conformations and the height of potential barriers to overcome in the reduced picture. A major task is to interpret the FEL of a high-dimensional system. Apart from resolving and distinguishing states of interest, projections also disguise connectivity and distort effective barriers. This feature will be discussed later. In order to overcome this difficulty, schemes that provide collective coordinates by reduction of the dimensionality are applied.

![Ramachandran plot and corresponding FEL](image)

**Figure 2.1.:** Ramachandran plot and corresponding FEL: The left picture schematically shows, how different conformations are aligned in \(\phi-\psi\)-space. The central picture shows the probability distribution \(P(\phi, \psi)\) for a typical alanine residue. This representation is referred to as Ramachandran. The corresponding FEL projection of this representation is shown on the right plot: The logarithmic representation reveals metastable states (minima, dark color coding) and barriers connecting the different stable states.

### 2.3. Principal Component Analysis (PCA)

There are many brands of dimensionality reduction techniques which are also applied in other scientific fields like pattern recognition, image processing and neural computation. Apart from an intuitive selection of collective coordinates which is motivated by knowledge about the process of interest, there are many projection schemes that work in a systematic manner. They all aim to elaborate a sub-manifold of the high-dimensional system which is capable to cover all dynamical features of interest. Therefore, besides cov-
2.3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical method for reducing the dimensionality of data while retaining as much of the variation in the data as possible. It does this by identifying the directions (principal components) in which the data vary the most. The data is then projected onto these principal components.

Mathematically, the covariance matrix of a coordinate set \( \mathbf{r} = \{ r_i \} \) is given by

\[
\text{Cov}_{ij} := \langle (r_i - \langle r_i \rangle) (r_j - \langle r_j \rangle) \rangle,
\]

(2.14)

where \( \langle \cdot \rangle \) denotes the average. The method then finds the directions (principal components) that maximize the variance of the projection of \( \mathbf{r} \) onto these directions. The variance of the principal components is equal to the corresponding eigenvalues,

\[
\text{Var}(V_i) = \lambda_i.
\]

(2.16)

The cumulative fluctuations

\[
F(n) := \sum_{i=1}^{n} \lambda_i
\]

(2.17)

are often used to estimate the amount of relevant PCs. In this case, a correlation between the size of the variance and the structure of PCs is presumed (see Figure 2.3). Since \( \text{Cov}(V; V) = \text{diag}(\lambda_1, \ldots, \lambda_n) \), the PCs are globally linearly uncorrelated.

The PCA is illustrated in Figure 2.2.

Dihedral Principal Component Analysis (dPCA) A PCA can be applied to any set of coordinates. For example, Cartesian coordinates defining positions of atoms can be used (cPCA), which directly result from MD simulations. However, they still include overall translation and rotation of the system, which has to be removed in first place. This is often accomplished by fitting the structure to a reference. On the other hand, the dihedral angles already represent internal coordinates. The dihedral Principal Component Analysis (dPCA) features a modified version of PCA that considers the periodicity of these angular coordinates [31, 93, 94]. In fact, cyclic coordinates provide a manifold which complicates the assignation of collective coordinates since they are only locally Euclidean [95, 96]. Due to the circularity of a dihedral angle \( \Theta \in \phi, \psi \), it is not straightforward to
define a metric which is required to calculate means and variances that are contained in the covariance matrix. In order to facilitate the use of the Euclidean metric, each $\Theta$ is mapped onto the Cartesian coordinates of the angle on the unit circle,

$$\Theta \mapsto \begin{cases} 
  x = \sin \Theta \\
  y = \cos \Theta
\end{cases} \quad (2.18)$$

This bijective transformation doubles the number of coordinates, but facilitates to compute averages and variances that map monotonously to the correct non-Euclidean averages along the circular variable. Subsequently to this transformation, a common PCA is applied.

It was shown recently, that the dPCA is even more robust than a Cartesian PCA since it avoids artifacts caused by fitting to a reference [97]. The dPCA is used to establish a set of system coordinates that will be used in the context of Langevin modeling. In contrast to the bath coordinates, system coordinates highlight structure (see 4.1.2). One-dimensional projections of the FEL in Figure 2.3 show that this property separates the first PCs 1-5 from the remaining using the small peptide hepta-alanine as example. PCs 1-5 cover more than 60% of the cumulative fluctuations.
2.4. Autocorrelation

The FEL of a reduced system of coordinates does not necessarily imply dynamical properties of the system for several reasons: On the one hand, the reduced picture induces a bias to the force which relates to the Newtonian equations of motion. On the other hand, the projection has the ability to induce spacial varying diffusion. However, information about the time scale of conformational changes in biomolecular systems are extremely relevant to understand the system dynamics. Since MD simulations produce continuous trajectories, raw data that is used to establish a set of collective coordinates already possesses temporal information, \( r = r(t) \) which in turn yields a time dependent trajectory in PC space. A simple quantity to measure the time scale on which a single collective coordinate \( x \) acts is the (time-lagged) autocorrelation function,

\[
C_x(\tau) = \frac{\langle [x(t) - \langle x(t) \rangle] [x(t + \tau) - \langle x(t + \tau) \rangle] \rangle_t}{\langle x^2(t) \rangle_t} = \frac{\text{Cov}(x(t), x(t + \tau))}{\text{Cov}(x(t), x(t))},
\]

where \( \tau \) denotes a lag time and the average \( \langle . \rangle \) is taken over all \( t \in \{0, \delta t, 2\delta t, ..., (T - \tau)\} \) of the discrete time series of length \( T \). Assuming a wide-sense stationary process (mean and variance exist and are time-independent), eq. (2.19) is just a function of the time difference \( \tau \). By definition, eq. (2.19) is restricted to \([-1 : 1]\) and \( C_x(0) = 1 \). Typically, a time series loses memory about its past for long lag times \( \tau \). This implies that points are statistically independent and the autocorrelation vanishes, \( \lim_{\tau \to \infty} C(\tau) = 0 \). The decay of the autocorrelation function specifies the time scale on which the associated coordinate alters. For a stationary process, the Wiener-Khinchin theorem allows for a comparable representation of the time scales by computing the power spectral density as the Fourier-transform of the autocorrelation function.

Numerous processes along a structured coordinate are reflected in various decay times since dynamics evolve on several timescales. A multi-exponential ansatz can be used to respect several processes at once [98], \( C(x, \tau) = \sum_i^n c_i \exp^{-\tau/\tau_i} \), where \( \tau_i \) monitors the time scale of the \( i \)-th process. For the limit of a time scale separation, we can recover the
slowest process dominating the associated autocorrelation:

\[ \tau_j \gg \tau_i \quad \forall i \neq j \Rightarrow C(\tau) \overset{\tau \to \tau_j}{\approx} c_j \exp^{-\tau/\tau_j} \tag{2.20} \]

Although the autocorrelation maps the dynamical information of a coordinate to a single highly-averaged quantity, it contains dynamic information about the system and information about time scales. The latter property is also used in the modified linear projection scheme “time-lagged independent component analysis” (TICA) [99–101]. This scheme performs a non-unitary linear transformation by simultaneously diagonalizing the covariance matrix and a time lagged covariance matrix \( \text{Cov}(r(t), r(t + \tau)) \).

### 2.5. States and Transition Rates

Since the autocorrelation function is an averaged quantity, it generally does not allow for tracking and resolving the dynamics of single conformational changes. In fact, it contains information about a multitude of processes along cuts of the coordinate space. Yet, in order to describe the dynamical behavior of a system in more detail, one first needs to characterize different processes. For this purpose, a clear definition of states featuring different conformations is required. A static definition is rather simple: For example, the local minima of the free energy landscape define metastable states and the spread of each local minima gives reasonable information about the importance of this state. The Ramachandran of a particular residue contains characteristics about the possible stable configurations of this residue (see Figure 2.1). Once states are identified in a single Ramachandran, the setting of product states [102] yields a way to register each possible configuration and address it to a state.

However, it is not straightforward to separate different states from each other. Fluctuations within a state may for example overlap with other states in a purely geometric picture. Therefore, dynamical information has to be considered additionally. The connectivity between different states is also not straightforward. It can be masked by the projection on the reaction coordinate selected [31]. This is visualized in Figure 2.4: It shows a FEL that consists of 6 states. The full connectivity shall be resolved along the reaction coordinates \( r_1, r_2 \). A simple projection solely onto the coordinate \( r_1 \), as indicated above in the illustration, gives the impression of a wrong connectivity. For instance, the states 2 and 4 which are not connected to each other in the “full” picture appear to be adjacent. The connectivity between state 1 and 2 on the other hand gets lost in the reduced picture.

Finding suitable reaction coordinates that distinguish all relevant metastable states is a major task in general. Once a set of states is established, one can track transitions that
2.5. States and Transition Rates

Figure 2.4.: Scheme to illustrate loss of connectivity information by projection: While the full connectivity is apparent along the reaction coordinates \( r_1, r_2 \), the free energy picture projected onto \( r_1 \) solely displays a misleading connectivity, see text. Adapted from [31].

occur and characterize the overall system dynamics. A network of states can be built to summarize interconnectivity and understand the time scales of biological functioning. These transition networks are a pre-stage for the construction of Markov state models [48, 58]. Transition state theory addresses the interaction of pairs of states that are located on a potential energy surface and separated by a transition state. The latter is characterized as a saddle-point along the barrier connecting a pair of states. We only refer to two excerpts in this section: Theoretically, Kramer’s reaction rate theory provides analytical expressions that relate effective barrier heights to transition rates [103]. Since we designed a simple model system which is mentioned later on, the particular application on this system will be mentioned. Besides, we will concentrate on the identification and definition of metastable states characterizing a system and briefly on the definition of transition rates.

2.5.1. Defining a state

In order to define a state, we first of all delimit it geometrically by defining core regions. This definition is then dynamically extended to obtain a time series of clusters.
Geometrical definition of cores

The core region is defined in a static manner. Since the local extrema of Ramachandran plots (see page 24 for the actual application) clearly delimit stable conformations, they are for example an appropriate coordinate space to start with. To define states, cores are first determined by confined regions. The extent of the core region is restricted to the local environment of the extrema. The geometry crucially depends on the coordinate to which the coring is applied. Adjusting the size of a core is not straightforward: on the one hand, the more delimited a core region is chosen, the lower is the probability that different states will be mixed. On the other hand, the possibility arises that a state affiliation is not recorded at all if the core region is too confined. In a self-consistent manner, it is reliable to vary the core size and finally utilize a stable core size concerning the stability of the observables of interest that result from it. Particularly, these observables are rates describing important transitions. They are supposed to remain constant for small variations of the core size.

Dynamical definition of clusters

Once the definition of cores is completed, the state of each frame of the time series that describes the propagation of a system lying within a core can be associated with this specific core. Still there are “neutral” regions that have not been assigned so far. The definition of clusters assigns them, following the time series in chronological order (see Figure 2.5): Once a core region of a state is entered, the cluster affiliation is assigned to the according core. It remains in this cluster until another core region is entered and so on. Except for initial time frames in which the system did not enter any core region yet, each time frame within a time series is now related to a cluster. This simple dynamical definition of states as clusters allows for the further processing: it yields a time series of states/clusters which can be finally considered to compute transition rates and thus quantitatively analyze the connectivity.

Figure 2.5.: Scheme to illustrate dynamical clustering: Once the system enters the core $A$, it is associated with cluster $A$ (red). Once core $B$ is entered, its association switches to cluster $B$ (black).
2.5. States and Transition Rates

2.5.2. Relative clustering via relative coring

As pointed out in Sec. 2.3, dPCA is a powerful tool to establish a set of system coordinates from a given set of dihedral angles. Although the coordinate transformation is bijective by construction, there is no functional relation in general that maps the definition of cores in the Ramachandran representation onto the PCs of the dPCA. This is induced by the dimensionality reduction of the image space from \( D \) to \( d < D \). Thereby, the mapping is no longer necessarily injective. However, the method that is primarily used in this work is directly operating on a reduced space of few PCs (see Sec. 4.3). In order to establish a cluster trajectory from a given time series, we use a relative coring procedure which works as follows. Let \( y_n \) be the collective coordinate that describes frame \( n \) of the discrete time series in a reduced space of dimension \( d \). We now want to establish a cluster trajectory along this coordinate. A discrete time series \( \{x_1, \ldots, x_N\} \), for which the associated core trajectory \( \{c_1(x_1), \ldots, c_N(x_N)\} \) is known, shall be our reference. In the applications mentioned in this work, \( x(t) \) represents one or several MD trajectories on which we apply coring prior to a coordinate transformation. The cluster associated with \( y_n \) is now inherited from the local cluster affiliation of the reference trajectory: one searches for the \( k \) nearest neighbors \( \{x_i\}, i \in \{1, \ldots, k\} \), where a \( d \)-dimensional Euclidean norm is applied to measure the distance \( \Delta_{in} = \|y_n - x_i\| \). The time frame \( y_n \) is then affiliated to the core to which the majority of reference frames is related to. In case of a tie, the farthest reference frames are discarded until a majority of core affiliations is obtained. The cluster trajectory is then established conventionally as described in Sec. 2.5.1. Relative Clustering is applied explicitly in Sec. 6.2.

2.5.3. Transition rates

The existence of a cluster trajectory allows for the calculation of transition rates \( k_{ij} \) between a pair of states \( i \) and \( j \). Let \( C(\tau) \) be a matrix that counts the total number of transitions between all pairs of states that occur for a fixed lag time \( \tau \), \( C(\tau) = \{C_{ij}(\tau)\} \). The so-called transition matrix \( T \) then results from \( C \) by row-normalization,

\[
T_{ij}(\tau) = \frac{C_{ij}(\tau)}{\sum_j C_{ij}(\tau)}.
\]  

Each element accounts for the probability of the system to switch from state \( i \) to \( j \) within \( n = \tau/\delta t \) time frames (where \( \delta t \) denotes the time step of the discrete trajectory). The transition rate then reads

\[
k_{ij}(\tau) = \frac{T_{ij}(\tau)}{\tau}.
\]
Chapter 2. Theory and methods

Notably, \( k_{ij} \) crucially depends on the lag time. For small lag times, high-frequent interstate fluctuations are still monitored; these can arise both from improper coring or clustering. Moreover, they could just monitor time scales which are not of interest, e.g. since the comparative data record is measured at another time scale. Large lag times on the other hand omit a detailed dynamical picture since states that possess a comparably small metastability are skipped. The transition rate matrix is closely related to the rate matrix of the associated rate master equation, see A.1. Different notations also introduce a column-normalization.

2.5.4. Markovianity

Markov processes form a sub-category of dynamical processes [54]. Markov State Models (MSM) for instance completely rely on systems that feature Markovianity [104]. A time discrete process \( x(t) = x_n \) of length \( T \) is called Markovian if it fulfills the following conditions:

- the instantaneous change \( x_n \rightarrow x_{n+1} \) relies only on the time step \( n \) and not on previous history
- \( x_n \) is ergodic if all states are visited infinitely often for an infinite length of the time series, \( T \rightarrow \infty \). The ensemble average of a quantity then equals the time average.

In MD, at constant temperature there exists a stationary state density distribution \( P(x) \) which is Boltzmann distributed,

\[
P(x) = Z^{-1} \exp(-H(x)/k_B T),
\]

where \( H \) denotes the Hamilton function and \( Z \) the respective partition function.

- \( x_n \) is reversible, meaning that detailed balance\(^2\) is fulfilled: if \( p_i \) denotes the population of state \( i \), we demand

\[
p_i k_{ij} = p_j k_{ji}.
\]

The Chapman-Kolmogorov equation has to be fulfilled by a Markov process. If \( T_{ij}(\tau) \) denotes the stationary probability to switch from state \( i \) to \( j \) in one time step \( \tau \), it reads

\[
p_j(n+1) = \sum_i T_{ij}(\tau)p_j(n).
\]

Iteratively, it follows for \( T = \{T_{ij}\}, P = \{p_i\} \)

\[
P(n + m) = T(\tau)P(n + m - 1) = T^m(\tau)P(n) \equiv T(m\tau)P(n).
\]

\(^2\)the continuity equation is consequently fulfilled, see A.2
2.5. States and Transition Rates

2.5.5. Kramers’ theory

As mentioned above, Kramers provided an analytical solution for the transition rate of a special problem [103] which will be utilized in Ch. 5. The mean first passage time (MFPT) \( \tau_{MFPT} \) can be calculated analytically in this case. The definition of this quantity arises from the solution of a one-dimensional Smoluchowski equation (see page 51). The MFPT denotes the average time of a system spent in a confined region until it reaches a certain boundary. For a one-dimensional double well potential (see page 69) which is characterized by two absorbing boundaries, \( \lim_{x \to \pm \infty} U(x) \to \infty \) and a single local maximum at \( x_{max} \), \( \tau_{MFPT} \) yields a simple expression. If one locates a boundary at the single local maximum in order to calculate the MFPT needed to leave the region to the left, one obtains

\[
\tau_{MFPT} = \frac{\gamma}{k_B T} \int_{-\infty}^{x_{max}} dy \exp \left( \frac{U(y)}{k_B T} \right) \int_{-\infty}^{y} dz \exp \left( -\frac{U(z)}{k_B T} \right),
\]

(2.27)

where \( \gamma \) denotes the friction constant. Kramers’ approach is based on the following assumptions that simplify the expression above: Assuming the thermal energy \( k_B T \) to be sufficiently small, one can express the potential in its harmonic approximation around the left minimum at \( x_{min} \) for the inner integral. In the same manner, we can argue for the outer integral since it is dominated by the potential:

\[
U(x) \approx U(x_i) + \frac{1}{2} U''(x)|_{x=x_i} (x - x_i)^2 \quad i \in \{0, max\}
\]

(2.28)

If this approximation is applied, both integrals in (2.27) can be evaluated separately:

\[
\int_{-\infty}^{y} dz \exp \left( -\frac{U(z)}{k_B T} \right) = \sqrt{\frac{2\pi k_B T}{U''(x_{min})}} \exp \left( -\frac{U_{min}}{k_B T} \right)
\]

\[
\int_{-\infty}^{x_{max}} dy \exp \left( \frac{U(y)}{k_B T} \right) = \sqrt{\frac{\pi k_B T}{-2U''(x_{max})}} \exp \left( \frac{U_{max}}{k_B T} \right)
\]

The barrier crossing-time is now approximately twice the MFPT. Therefore, the transition rate from left to right state reads

\[
k_{trans}(l \to r) = \frac{1}{2\tau_{MFPT}} = \frac{-U''(x_{max})}{2\pi\gamma} \exp \left( \frac{\Delta U}{k_B T} \right),
\]

(2.29)

where \( \Delta U = U(x_{max}) - U(x_{min}) \) denotes the potential energy barrier. Analogously, the transition rate in the other direction also follows eq. (2.29), substituting \( x_{min} \).
MD simulations are performed in order to understand biomolecular systems concerning their biological functioning in the first place. Although relevant systems are mostly proteins which possess a minimum size, many functionally important effects can already be illustrated by comparatively small systems, often on the basis of FELs [2, 35, 39, 105]. This chapter will focus on a small peptide system. It assembles only 9 residues which is about the minimum length required to develop metastable conformations like \( \alpha \)-helices. Our particular system features some remarkable dynamical temperature dependent properties and acts on several timescales [106]. This will be discussed later in this chapter [107].

### 3.1. AIB\(_9\)

The \( \alpha \)-aminoisobutyric acid (AIB) is an amino acid with the structural formula \( \text{H}_2\text{N}-\text{C}((\text{CH}_3)_2-\text{COOH} \). It is non-proteinogenic and typically synthesized. Our investigations of the small peptide chain AIB\(_9\) is mainly motivated by an experimental application studying energy transport [7]. In this context, the property of the peptide chain to form 3\(_{10}\) helices was used to study energy transport. An AIB\(_9\) peptide was provided with a photo-switch at the N terminus and dissolved in chloroform. Vibrational modes were excited and the propagation of the signal along the peptide backbone was further studied using 2D IR spectroscopy [108]. The experiments were supported by non-equilibrium simulations [7] and energy transport in AIB\(_9\) further studied utilizing for example a semi-classical system bath ansatz [109]. In order to entirely understand the AIB\(_9\) system, we continued investigating the peptide studying its conformational dynamics. Not including the photo-switch, we established a slightly different AIB\(_9\) system which is completely symmetric to the central residue. Regarding various aspects, we analyzed the dynamical behavior both solvated in chloroform and water. The system in aqueous solvent served as a benchmark for an enhanced sampling method (see Ch. 7) [110]. In chloroform, we observed a hierarchical scheme where separate processes on fast timescales trigger and
determine dynamics on larger timescales [107]. This aspect will be further elucidated below.

### 3.1. Adequate collective coordinates

A first step to analyze the dynamical and statistical properties of our small peptide system is the finding of a proper set of reaction coordinates. Therefore, the backbone dihedral angles $\phi_i, \psi_i$ were used as they have the property to neglect overall rotation and translation of the molecule but monitor only internal conformational changes which affect the peptide backbone. We excluded the termini, since they exhibit high-frequent uncorrelated fluctuations and restrain to the central residues $i \in \{3, 4, 5, 6, 7\}$. Figure 3.1 illustrates that this selection is reasonable: concerning the decay of autocorrelation and the cross correlation, the $\phi_i$ are clearly separated for the inner residues 3 to 8, which is a misleading observation. As discussed in detail below, these dihedral angles account for the slowest dynamics, disclosing the transition from left-handed to right-handed helical conformation (see Sec. 3.2). The residues $\psi_i$ on the other hand account for the faster dynamics, illustrating an excitation on a faster timescale (see 3.2.1). Taking them into account, the residues 3 to 7 are clearly separated from the termini. In order to fully describe these inner residues, we restrict our definition of the internal coordinates to this subset, having in mind that we neglect resolving the fluctuations close to the termini. This still yields a reduced phase space of $2 \times 5 = 10$ dihedral angles. To further reduce this set of reaction coordinates, a dPCA is applied subsequently (see Sec. 2.3). In this way, cumulative coordinates are established, sorted according to their overall fluctuations. The conformational change from completely right-handed to left-handed helical structure is resolved along the first PC. Individual transition paths are distinguishable along higher PCs. For instance, the two major pathways are resolved along the edges of the free energy landscape $F$ in its projection onto the first two PCs while minor populated conformations are located in the central region of this projection.

The dPCA also yields a clear timescale separation (see Figure 3.3). While the dynamics of PC 1–5 take place on a nanosecond (ns) timescale, the autocorrelation of PC 6 already decays an order of magnitude faster. Since the slowest process of the system is resolved along PC 1, the autocorrelation decay is already separated from the following PCs by an order of magnitude. We rely on the property of time scale separation in the context of Langevin modeling, see Sec. 4.3.

---

1Self-consistency also justifies this choice of coordinates. Our model approach which is central to the remainder of this work, is able to recover the system properties only for this selection.
Figure 3.1.: Uncorrelated fast fluctuations of terminal residues: shown are autocorrelations (left column) and cross correlation coefficients (right column) for backbone dihedrals $\phi$ (first row) and $\psi$ (second row) of all residues.

Figure 3.2.: dPCA of AIB$_9$ solvated in water: the free energy landscape already clearly distinguishes left-handed ($L$) and right-handed ($R$) helical conformation along $V_1$ and further resolves the major transition paths along the edges of $V_2$. 
3.1. AIB$_9$

Figure 3.3.: Autocorrelations of AIB$_9$ solvated in chloroform: A clear time scale separation is observed both at $T = 300$ K (left) and $T = 320$ K (right) and allows to establish a system bath classification.

Figure 3.4.: State definition in the AIB$_9$ peptide system: Each residue 3 – 7 reveals two distinct states $(l, 0)$ and $(r, 0)$ corresponding to left- and right-handed helical conformation. The existence of these states is independent from the solute. In contrast to water solute (right side), we observe the existence of additional less metastable “excited” states $(l, 1)$, $(r, 1)$ for the peptide system solvated in chloroform. The latter occur twice for each conformation. They account for two possible transition paths in each Ramachandran and are distinguished by the superscript *.
3.1.2. Network of states

In order to further investigate the dynamical behaviour of our peptide system, we identified the metastable states using a dynamical clustering method. The Ramachandran plot of the AIB peptide in Figure 3.4 reveals two distinct local minima at $(\phi_i, \psi_i) \approx (50^\circ, 45^\circ)$ and $(-50^\circ, -45^\circ)$ corresponding to the left- and righthanded helical conformation for each residue $i$. We refer to these regions as cores in the following. They occur both in the water and the chloroform solute. The geometry of the core was set circular with a radius $r_0 = 50^\circ$. This size turned out to yield stable results for the transition rates, meaning that small variations did not affect the transition rates. A straightforward definition of product states as 5-tuples is accomplished: while the entirely left- and right-handed conformation correspond to $L = (lllll)$ and $R = (rrrrr)$, the state $(rllll)$ for instance corresponds to the constellation of only residue 3 being in a right-handed, and the other residues 4-7 in the left-handed helical conformation. Figure 3.5 and Figure 3.6 highlight a networks of these states and point out their connectivity. In order to construct these networks, we applied the dynamic clustering method described in Sec. 2.5, using the core definitions as shown in Figure 3.4. The network is aligned such that the location of the states corresponds to their placement in the free energy landscape along the first two principal components (see Figure 3.2). Major transition paths are labeled with the total number of transitions that were counted for 8 MD trajectories at 300 K$^2$. It must be pointed out that the length of the MD runs for the water solvent is twice the amount of the runs for chloroform. As indicated by the arrow labels in Figure 3.5, the most frequent transitions are single flips of the edge residues 3, 7 from the entirely helical conformation. The symmetry of the node locations with respect to the center of each network shows that the dPCA qualitatively reflects the symmetry of the system. Local vicinity of clusters in this projection does not imply a high connectivity. In fact, the least populated states are located in the central area of the FEL representation where the actual volume of the phase space is increased. Some states are accessed only sparsely but show a rather high metastability, e.g. $rrllr$ in CHCl$_3$ is accessed 5 times and has a life time of $\approx 1$ ns. Populations of our ensemble are increased in the left-handed conformation for both solvents, e.g. the fraction $L/R$ is $\approx 1.98$ for chloroform and $\approx 3.06$ for water. The network model fulfills the required condition to be Markovian (see Sec. 2.5.4). To verify this, we prepared an equally populated 32-dimensional state $P$, $P_i(t = 0) = 1/32$, $i \in \{1..32\}$ and propagated the rate Master equation, $P(t = n \cdot \tau) = M(n \cdot \tau)P(t = 0) \equiv M(\tau)^nP(t = 0)$, where the latter relation

$^2$see Sec. 3.2 for a detailed description of the simulation setup with chloroform solvent and Sec. 7.2.2 for the simulations in water solvent
Figure 3.5.: Networks for AlB$_9$ at 300 K solvated in H$_2$O (top) and CHCl$_3$ (bottom). Node centers are placed according to the position of the cores in the FEL along PCs $V_1$, $V_2$, respectively. The labels denote the configuration of the inner residues 3-7. The node size reflects the overall population of the related state while arrow opacity and size correspond to the magnitude of the transition rate. Major transition events are labeled with the number of total transitions observed. For the system in water, a time step of $\delta t = 10$ ps was used while the time step $\delta t = 1$ ps was used for chloroform.
only holds if the system is Markovian. The transition matrix

\[ M_{ij}(\tau) = \begin{cases} -k_{ii}, & i = j \\ k_{ij}, & i \neq j \end{cases} \]  

(3.1)

contains the transition rates \(k_{ij}\) that are measured for different lag times \(\tau = n\delta t\). \(\tau = 10\text{ps}\) was detected to be the minimum time scale to construct Markov model. To illustrate the Markov property, Figure 3.7 shows a propagation of \(P\) for \(t = 1\) ns, a common multiple of 4 different lag times. Although they are varied by one order of magnitude, the distributions \(P(t = 1\text{ns})\) coincide.
3.2. Hierarchical Dynamics

In many biomolecular systems, dynamical properties are crucial for their functioning [111]. The dynamical behavior of the AIB$_9$ peptide can be characterized by a hierarchy of different processes. This denotation shall indicate, that the system contains different processes which are mutually dependent. The hierarchy which we stated for our peptide system possesses three different levels [107]. While the breaking and opening of hydrogen bonds takes place on a picosecond timescale, another excitation from helical conformation to an intermediate state (which will further be referred to as 01 transition) takes place on a nanosecond timescale. The latter transition is a condition for the overall conformational change from left- to right-handed helical structure. This dynamical hierarchy is illustrated in Figure 3.8 by the solid line: The minima marking the edges of a fictive reaction coordinate correspond to the most stable configuration of a single residue, the ground state 0 with a hydrogen bond formed (abbreviated by $b$). I will use the tuple $(l, 0, b)$ to denote this joint state in the following, where $l$ specifies the a left-handed helical conformation. Analogously, I will write e.g. $(r, 1, f)$ to name the excited state 1 without a hydrogen bond formed ($f$ for free) in the right-handed helical conformation. The abbreviated form $(1, f)$ does not distinguish a state with respect to its helical conformation. $(0, f)$ corresponds to the states adjacent to the ground states in Figure 3.8. $(0, b)$ and $(0, f)$ are dynamically connected on tier $C$. In order to trigger the overall transition on tier $A$, the system has to

![Figure 3.7: Markovianity test for AIB$_9$ in CHCl$_3$ at 300 K. Shown are the probabilities that states are occupied at $t = 1$ ns. The system was initiated equipartitioned, $P_i = 1/32$. Populations are measured for four different lag times: 10, 20, 50 and 100 ps.](image)
pass the excited state \((1, f)\) which is directly connected with \((0, f)\). The latter transition is described by tier \(B\).

The dashed line in the schematic visualization of Figure 3.8 corresponds to a related model based on a free energy landscape picture [1, 112]: while intra-state transitions occur frequently, inter-state transitions are less common. The main difference in this picture is, that there is no designated excited state that has to be overcome but different levels of energy barriers.

Hierarchical properties can be characterized in more detail utilizing an applicable model.

![Figure 3.8: Illustration of the Hierarchical model describing AIB$_9$](image-url)

**Figure 3.8.:** Illustration of the Hierarchical model describing AIB$_9$: The ground states marking the edges of a fictive reaction coordinate are left on a ps time scale. Solid line: This process is a prerequisite to an excitation that occurs on a ns time scale. The latter process itself is required to allow for an overall transition from the left to the right ground state which we observe on a $\mu$s time scale. The dashed line corresponds to a different picture describing a similar hierarchical model approach (see text).

Related to our small peptide system, this idea will be further discussed in the remainder of this chapter. Apart from being small enough to generate sufficient statistics, this system proved to be an ideal “toy model”. This is because the dynamics associated with the different tiers are already visible on the level of the time series, if a representation that uses proper collective coordinates is chosen. In order to establish a temperature dependent analysis of the hierarchical character of this system, we performed extensive MD simulations. For each of the temperatures \(T = 240, 260, 280, 300, 310, 320, 330\) and \(350\) K, eight MD runs of 2 $\mu$s length were conducted, simulating a NVT ensemble. The MD setup used a leapfrog integrator with a time step of 2 fs. The trajectory was recorded using a time step of 0.2 ps. This proved to be sufficient to still account for the fastest time scale in detail, namely the forming and breaking of hydrogen bonds.

As discussed above, the symmetry of the system is reflected along the first PC. The slow-
est process, the switching of residues from left- to right-handed conformation is resolved in the time series of this PC. In terms of internal coordinates, this process is visualized identically by the sum of inner dihedrals \( \phi_i, \phi_{\text{sum}} = \sum_{i=3}^7 \phi_i \), see Figure 3.9 (a). A sequential flipping of inner residues results in the overall transition \( L \leftrightarrow R \). At \( T = 300K \), on average 8 transitions are counted per trajectory, which gives an estimate of 0.25 \( \mu s \) for the time scale of the \( L \leftrightarrow R \) transition.

The preceding \( 0 \leftrightarrow 1 \) process is monitored along the inner dihedrals \( \psi_i \). Since the Ramachandran of AIB\(_9\) (see page 24) is symmetric to the origin, mirroring facilitates a separation of both states:

\[
f(\psi_i) = \begin{cases} 
\psi_i & \text{for } \phi_i < 0 \\
-\psi_i & \text{else.}
\end{cases}
\]

Figure 3.9 shows a representative cutout of the time series of this observable by means of the backbone dihedral \( \psi_5 \) in a time window, where \( \phi < 0 \): the excitations \( 0 \rightarrow 1 \) occur rather irregularly and the excited state \( 1 \) is quite unstable. We record a transition rate of about 1 ns\(^{-1}\) at 300 K for this excitation, while the relaxation rate is two orders of magnitude higher.

The primary process of hydrogen bond breaking is represented by the coordinate \( \theta_i \), the hydrogen bond angle. \( \theta_{i-2,i+1} \) denotes the angle which is enclosed between the fictive connecting lines from oxygen atom \( O_{i-2} \) to the atoms \( H_{i+1} \) and \( N_{i+1} \). The length of the fictive line connecting \( O_{i-2} \) and \( H_{i+1} \) is referred to as hydrogen bond length. It is visualized for \( i = 5, 6 \) in Figure 3.10. The timescale of hydrogen bond breaking is about 50 ps, while the formation of a bond happens five times more frequently.

### 3.2.1. Hierarchical dependency of individual tiers

In the following, simple concepts are established to model tiers individually in more detail. These are linking dynamically adjacent pairs of tiers: Both forming and breaking of hydrogen bonds are associated with the local \( 0 \leftrightarrow 1 \) transition and the latter in turn with the overall helical transformation \( l \leftrightarrow r \), which are again prerequisite to the overall helical transition \( L \leftrightarrow R \). Different pairs of tiers will be independently connected by their dynamical properties as a function of the temperature. Concluding, the developed chain of interlinked tiers will be interpreted in terms of a hierarchical model.

The easiest conceptional approach that we suggest to model the transition rates obtained directly from MD simulations in the following is a simple Arrhenius-ansatz,

\[
k_\alpha = g_\alpha \exp^{-\Delta U_\alpha/k_BT},
\]

where \( \alpha \) denotes the specific transition process, e.g. \( 0 \rightarrow 1 \).
Figure 3.9.: Collective coordinates illustrating tiers of hierarchical dynamics in \textit{AIB}_9 at 300 K: (a) The first PC represents the slowest time scale (red) of the \( l \leftrightarrow r \) and \( L \leftrightarrow R \) transition similarly as the observable \( \phi_{\text{sum}} = \sum_{i=3}^{7} \phi_i \). (b) The \( 0 \leftrightarrow 1 \) transition is monitored along the backbone dihedral \( \psi_i \) (red) (c) The hydrogen bond angle \( \theta_{i-2,i+1} \) basically illustrates hydrogen bond forming and opening (red). The insert zooms to sub-ps fluctuations of this trajectory. Green lines reflect the cluster affiliation for the last two plots.
3.2. Hierarchical Dynamics

The energy difference $\Delta U$ denotes the corresponding potential energy barrier that has to be trespassed in order to perform a transition. The constant prefactor $g$ can be interpreted as a temperature independent waiting time. It inherits the time scale of the next lower tier. Both $\Delta U$ and $g$ characterize a simplified picture. The results of the ansatz (3.2) will be discussed in the following, where it is adapted to measured data. The basic transitions characterizing the different tiers are consolidated on page 37.

**How bond breaking triggers the $0 \leftrightarrow 1$ (01) transition**

The most frequent transition is the forming and breaking of hydrogen bonds. Since hydrogen bonds are known to be helix stabilizing, the opening is a prerequisite for a variation of the dihedral angles. A snapshot of the helical structure of AIB$_9$ is shown in Figure 3.10. In order to define states, we used a geometrical coring considering the hydrogen bond angle $\theta_{i-2,i+1}$ to identify the formation of a 3$_{10}$ helix and $\theta_{i-3,i+1}$ to identify the formation of an $\alpha$ helix for each residue $i$. This relation between dihedral and hydrogen angles is plausible since it correlates most with the $0 \leftrightarrow 1$ transition: the population of the ground state 0 and the bonded configuration is maximized for this relation, as shown in Table 3.1. We consider a hydrogen bond to be formed if the bond length is below 3.6 Å and the angle $\theta$ is less than 30° for both the $\alpha$- and the 3$_{10}$-helical constellation. On the other hand, the non-bonded / free configuration is occupied if the bond criteria applies to neither of both helical constellations [113, 114]. Since trajectories that monitor these two quantities exhibit rapid and large fluctuations, we additionally demand that a change of state only occurs if the system remains in one configuration at least for 1.2 ps which corresponds to 6 time steps. Both 3$_{10}$ and $\alpha$ helical conformations are formed, with a ratio of about 30% to 70%. They will not be distinguished in the following. The resulting transition rates that we measured for different temperatures are

<table>
<thead>
<tr>
<th>$\Delta P [%]$</th>
<th>O1</th>
<th>O2</th>
<th>O3</th>
<th>O4</th>
<th>O5</th>
<th>N4</th>
<th>N5</th>
<th>N6</th>
<th>N7</th>
<th>N8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi_3$</td>
<td>-16</td>
<td>-75</td>
<td>-2</td>
<td>2</td>
<td>7</td>
<td>-76</td>
<td>-12</td>
<td>-6</td>
<td>-7</td>
<td>-4</td>
</tr>
<tr>
<td>$\psi_4$</td>
<td>-28</td>
<td>-45</td>
<td>-62</td>
<td>-5</td>
<td>8</td>
<td>-2</td>
<td>-75</td>
<td>-26</td>
<td>-6</td>
<td>-6</td>
</tr>
<tr>
<td>$\psi_5$</td>
<td>-14</td>
<td>-19</td>
<td>-59</td>
<td>-65</td>
<td>5</td>
<td>-20</td>
<td>-7</td>
<td>-80</td>
<td>-42</td>
<td>-16</td>
</tr>
<tr>
<td>$\psi_6$</td>
<td>-12</td>
<td>-18</td>
<td>-12</td>
<td>-51</td>
<td>-57</td>
<td>-16</td>
<td>-11</td>
<td>-10</td>
<td>-82</td>
<td>-47</td>
</tr>
<tr>
<td>$\psi_7$</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
<td>-30</td>
<td>-53</td>
<td>-9</td>
<td>0</td>
<td>-4</td>
<td>-16</td>
<td>-80</td>
</tr>
</tbody>
</table>
shown in Figure 3.14 (a), compared with the simple exponential Arrhenius ansatz. The error bars correspond to the standard deviation due to averaging over the inner residues 4, 5, 6. For the simple exponential model, a waiting time $1/g_{bf} \approx 3.7$ ps and an effective barrier $\Delta U_{bf} \approx 2.5k_B T_0$ ($T_0 = 300$ K) were determined. Thus the resulting time scale of a transition is about 45 ps for $T = T_0$.

The projection of the free energy landscape along the dihedral $\psi_5$ in Figure 3.10 (c) illustrates the idea that the loosening of hydrogen bonds facilitates the transition $0 \to 1$: while the green curve is restricted to the contribution of the bond state, the red curve considers only the non-bonded configuration. The angular range $\psi_5 < 0$ corresponds to the state 0 while $\psi_5 > 0$ matches the configuration 1. The result shows, that $(1, b)$ is not occupied at all while the configurations $(0, f)$ and $(1, f)$ are separated by a reasonable free energy barrier.

The $0 \to 1$ transition thus can be easily described by the reaction equation

$$
(0, b) \xrightarrow{k_{bb}} (0, f) \xrightarrow{k_{bf}} (1, f),
$$

(3.3)
3.2. Hierarchical Dynamics

Figure 3.11.: Model scheme for 01 transition and population of states: The scheme on the left illustrates the $0 \rightarrow 1$ transition: hydrogen bond breaking induces an “excitation” in the ground state (left dashed box) $(0, b) \rightarrow (0, f)$ which allows for the $0 \rightarrow 1$ transition. The right plot shows the temperature-dependent population of the different joint states.

where the individual $k$ denote the transition rates in abbreviated notation. The model is illustrated in Figure 3.11, where the overall population of the three states contained in our model is shown as a function of temperature in the same color coding. The population diagram demonstrates, that state $(1, b)$ is negligible and that the population of the excited state $(1, f)$ is increased for higher temperatures, mostly at the cost of the population of the ground state $(0, b)$. The reaction equation (3.3) can be modeled using the macroscopic rate equation ansatz

\[
\frac{d}{dt}(0, b) = -k_{bf}(0, b) + k_{fb}(0, f),
\]

\[
\frac{d}{dt}(0, f) = -[k_{fb} + k_{f01}] (0, f) + k_{bf}(0, b),
\]

\[
\frac{d}{dt}(1, f) = k_{f01}(0, f).
\]

The process $(0, b) \rightarrow (1, f)$ that provides the $0 \rightarrow 1$ transition can be extracted from this equation, assuming a time scale separation of the forming and opening of hydrogen bonds and 01 transition, namely

\[
k_{mf}, k_{fb} \gg k_{f01}.
\]

This assumption proves to be reasonable: Figure 3.12 shows that $k_{mf} > 2k_{f01}$, the bond-forming rate $k_{fb}$ is separated nearly by an order of magnitude from $k_{f01}$. The interpretation of this observation is, that the 01 transition is a sequential process with a high back-rate $k_{fb}$, compared to the forward-rate $k_{f01}$. Using assumption 3.5, the first two lines of (3.4) are stationary on a time scale that describes the process $(0, b) \rightarrow (1, f)$ and one directly
obtains
\[
\frac{d}{dt}(1, f) = \frac{k_{R1} k_{bf}}{k_{fb}} (0, b) \equiv k_{01, \text{model}}(0, b),
\]  
which defines a model transition rate \( k_{01, \text{model}} \). The numerical solution of rate equation (3.4), compared with the model eq. (3.6) is shown in Figure 3.13 for a system initiated in state \((0, b)\) at time \( t = 0 \) for \( T = 300 \text{ K} \). In order to compare the performance of this model, we measured the \( 0 \rightarrow 1 \) transition autonomously. Therefore, a refined definition of cores on the Ramachandran that resolves the exited states is used, see page 24. We introduced the state 1 at \((\phi_i, \psi_i) \approx (-68^\circ, 45^\circ) \) and \( \approx (68^\circ, -45^\circ) \) with a radius of \( r_1 = 18^\circ \) each. As illustrated in the Ramachandran picture, there are several transition pathways.

Figure 3.12.: Rates describing competing processes concerning \( 0 \rightarrow 1 \) transition in \( \text{ns}^{-1} \): (a) transition \((0, f) \rightarrow (1, f)\) and (b) \((0, f) \rightarrow (0, b)\). The black lines account for the measured rates while the green ones reflect the simple exponential ansatz (3.2).

Figure 3.13.: Time resolved \( 0 \rightarrow 1 \) transition for \( T = 300 \text{K} \): The rate equations (3.4) are solved numerically, initially preparing the system in the state \((0, b)\). The solid lines show the evolution of the three states following eq. (3.4) while the dashed line indicates the evolution of state \((1, f)\) using the model (3.6).
In this representation in particular, two transition pathways are revealed: transitions can happen both clockwise and counter-clockwise along the Ramachandran. While the first case is accounted for by the above definition of state 1, the definition of the excited state is extended to record counter-clockwise $0 \rightarrow 1$ events (and clockwise $1 \rightarrow 0$ events). Additional local extrema are located at $(\phi_i, \psi_i) \approx (45^\circ, -100^\circ)$ and $\approx (-45^\circ, 100^\circ)$. We assigned them to additional cores $(r, 1)^*, (l, 1)^*$, each with a radius $r^*_1 \approx 17.3^\circ$. To determine the transition $0 \rightarrow 1$ directly, we again set a minimum residence time of $1.2$ ps in a core as a precondition to switch the cluster. For each of the inner residues 4, 5, 6, we average over the four different possible $0 \rightarrow 1$ transitions in each residue and subsequently average over these residues. Figure 3.14 (b) shows that the model 3.6 (red curve) qualitatively matches the measured rates for this process (black curve). The exponential model eq. 3.2 yields the parameters $1/g_0 = 0.7$ ps and $\Delta U_{01} = 7.1 k_B T_0$, which results in a time scale of $1$ ns at $T = T_0$ for the $0 \rightarrow 1$ transition.

**How the process $0 \leftrightarrow 1$ triggers $lr$ transitions**

The $0 \rightarrow 1$ transition itself is a prerequisite to the $l \leftrightarrow r$ transition. Due to the symmetry of the Ramachandran, we can consider both processes, $l \rightarrow r$ and $r \rightarrow l$, at once. Similar to the previous case, we also observe competing processes: the high-frequent $1 \rightarrow 0$ back-transition is opposing the sparse event $(1, f, l) \leftrightarrow (1, f, r)$. Rates describing both processes are shown in Figure 3.15 as a function of temperature. The time scale separation of both processes is larger than an order of magnitude. We measured $k_{1l \rightarrow r}^{-1} \approx 200$ ps for the transition between excited states of left- and right-handed helical conformation, while we found a relaxation rate $k_{10}^{-1} \approx 10$ ps. This time scale separation could in principle motivate a similar model like the $0 \rightarrow 1$ transition. However, this approach is not possible at this tier since the transition $l \rightarrow r$ of a single residue is not independent of the conformation of other primarily neighboring residues. This relative dependency of the system is demonstrated in Figure 3.16, where the population of states $(l, 0)$, $(l, 1)$, $(r, 0)$ and $(r, 1)$ is shown for each residue. For comparison, populations restricted to the events that either the neighboring residue $i + 1$ towards the C-terminus performs the transition from core $(l, 0)$ to core $(r, 0)$ or vice versa, are shown. We observe, that the probability of a residue to be in one of the excited states is generally increased in this case. Particularly, for the transition $(l, 0) \rightarrow (r, 0)$, the probability that the residue already occupies state $(r, 0)$ is increased. This observation fortifies the assumption, that the overall conformational change from left- to right-handed helical conformation $L \leftrightarrow R$ is cooperative. Since cooperative behavior does not allow for the construction of a simplified Markov model\(^3\), we can only motivate

\(^3\)This observation does not affect the construction of a Markovian network model in 3.1.2, since we consider product states of all inner residues in that context while we do not distinguish different
Figure 3.14.: Transition rates of hierarchical model in ns$^{-1}$: (a) hydrogen bond breaking, (b) the 0 $\rightarrow$ 1 transition, (c) the l $\leftrightarrow$ r transition and d) the mean overall transition from left-handed to right-handed helical conformation and vice versa, L $\leftrightarrow$ R. The black lines account for the measured rates while the green curves reflect the simple exponential ansatz (3.2). Red lines show the results of the hierarchical model. The dashed lines in d) are linear fits indicating the biphasic behavior of the system for $T \lesssim 300$ K. Error bars account for averaging over residues 4, 5, 6 except for d) where they report the standard deviation of mean first passage time.
a hierarchical model ansatz which implicitly considers the average dynamics of the $0 \rightarrow 1$ transition. For this purpose, we allow for a temperature dependence of the factor $g_{lr}$ in eq. (3.2) which is assumed proportional to $k_{01}$, namely

$$k_{lr}(T) = c_{lr}k_{01}(T) \exp\left(-\Delta U_{lr}/k_BT\right).$$

(3.7)

The red line in Figure 3.14 (c) indicates the quality of this assumption. It is slightly improved compared to the simple exponential ansatz (3.2). A hierarchical dependence in this sense manifests the interpretation, that the biphasic behavior which can be observed in the $l \leftrightarrow r$ transition at $T = 300$ K, is already inherited from the $0 \leftrightarrow 1$ transition.

Figure 3.16.: Non-Markovianity of the $l \leftrightarrow r$ transition: The left figure shows the overall population of states $(l,0)$, $(l,1)$, $(r,0)$ and $(r,1)$ for the central residues $i = 3$ to 7. The other figures show the population restrained to the event, that the neighboring residue $i + 1$ is performing a transition from one ground state to the other.

permutations here.
Ir and LR transitions

The biphasic behavior of the AIB$_9$ system is indicated by the dashed lines in Figure 3.14 (d). The effect is comparably distinct both for the $l \leftrightarrow r$ transition and the $L \leftrightarrow R$ transition. A comparison of the transition rates measured for these processes shows, that the time scale of both is similar. For $T = T_0$, for example, we record a time scale $1/k_{lr} \approx 140$ ns and $1/k_{LR} \approx 262$ ns. This relation is physically intuitive: The bottleneck that facilitates the transition $L \leftrightarrow R$ is the switching of the central residue 5. The event may therefore be regarded as the bottleneck along the barrier separating the two most metastable conformations of the molecule. The temperature-dependent flexibility of residue 5 mainly dictates the time scale of the transition, while the residues which are located closer to the termini are only able to slightly decelerate the transition $L \leftrightarrow R$. The network on page 26 supports this picture: transitions that occur predominantly along two pathways are numerically close to the transitions that are recorded for the central residue. The non-Markovian character of the $l \leftrightarrow r$ transition, illustrated in Figure 3.16 confirms this behavior: once a neighboring residue switched its conformational orientation, a bias is introduced to the transition probability of the neighboring residues which compels the other residues to align similarly as the central one.

The hierarchical character

To summarize and connect the achievements of this section, I will start on top of what we called a hierarchical model: We located the biphasic behavior of the $L \leftrightarrow R$ transition to be inherited by the $l \leftrightarrow r$ transition which we observe for the inner residues. The same feature appears less distinct in the $0 \rightarrow 1$ transition, which indicates that this process is related to the $l \leftrightarrow r$ transition in turn, see Figure 3.14. Eq. (3.7) assumes a modified Arrhenius ansatz that relates the attempt frequency to the subordinated process. Consequently, a rise in the $l \rightarrow r$ transition rate occurs at the same temperature where the frequency of $0 \rightarrow 1$ excitations exhibit a sudden increase. This observation seems intuitive if one considers the representation of the Ramachandran for a single residue (see page 24), since the excited state 1 is a necessary prerequisite for the $l \leftrightarrow r$ transition. In a different representation, we quantitatively stated that the breaking of hydrogen bonds is required to facilitate the $0 \rightarrow 1$ excitation. For this special relation, we could utilize the simplified chemical master equation (3.6) to recover the dynamics of the $0 \rightarrow 1$ transition. To summarize, while we figured a simple physical model connecting the $l \leftrightarrow r$ transition to the process $0 \leftrightarrow 1$, the latter event itself is connected to the process of hydrogen bond forming and closing with high accuracy. Overall, these relationships inherently connect the process which acts on the longest time scale to the most frequent event which is separated by more than four orders of magnitude.
3.3. Summary and conclusions

I introduced the AIB₉ peptide as a benchmark system that exhibits dynamics on different time scales. Using different reaction coordinates and constructing conformational states for proper analysis, the system was shown to exhibit a hierarchical structure featuring dynamics on different tiers. We could trace back the transition of the most elaborate transition - from overall left-handed to right-handed helical conformation and vice versa - to the forming and breaking of hydrogen bonds, which occurs about $10^4$ times more frequently. An important conclusion which results from the hierarchical character of the AIB₉ system concerns the application of coarse-graining methods. The limitations of this approach are obvious by means of the AIB₉ system: hydrogen bond interactions already have to be resolved in detail in order to contain and preserve the hierarchical character of the system.
4

Data-driven Langevin modeling

4.1. The Langevin equation

Langevin originally proposed a stochastic differential equation that was named after him to describe Brownian motion of a particle in viscous environment [59]. This application of his formalism is the most simplistic example for a system that cannot be described deterministically, if the particle mass $m$ is so small that its velocity is governed by thermal fluctuations. The Langevin equation (LE) in its original form can be interpreted as an extended form of the deterministic Newton equation considering thermal energy. Extending the equation of motion for the particle in 1D by an additive stochastic term yields the LE [67]

$$F(t) = F_d(t) + F_s(t).$$

(4.1)

While the deterministic force on the right hand side of this equation is simply Newtonian, $F_d(t) = -\gamma mv(t)$, the stochastic force $F_s(t)$ is only defined via its moments. Explicitly, the differential equation reads

$$\dot{v} + \gamma v = F_s(t)/m.$$  

(4.2)

Setting the identity $\langle F_s \rangle = 0$ for the first moment, we recover the Newton equation in the averaged quantities. In order to describe the second moment, the thermal energy is considered. In thermal equilibrium, according to the equipartition theorem the system possesses the mean energy

$$\frac{m}{2} \langle v^2(t) \rangle = \frac{1}{2} k_B T.$$  

(4.3)

The stochastic force is assumed to be not correlated on “sufficiently” long time scales, $\langle F_s(t) F_s(t') \rangle = c \delta(t-t')$. “Sufficiently” in this context means that the system looses memory with respect to the propagation time step. Using eq. (4.2), the second momentum of the velocity distribution yields

$$\langle v(t_1)v(t_2) \rangle = v_0^2 e^{-\gamma(t_1+t_2)} + \int_0^{t_1} dt_1' \int_0^{t_2} dt_2' e^{-\gamma(t_1+t_2-t_1'-t_2')} c \delta(t_1'-t_2').$$
4.1. The Langevin equation

where the rightmost expression reduces\(^1\), \(\frac{\gamma}{2} e^{-\gamma |t_1 - t_2|} - e^{-\gamma (t_1 + t_2)}\). We get an expression for \(c\), using the equipartition relation (4.3) for large \(t = t_1 \approx t_2\),

\[
\langle F_s(t) F_s(t') \rangle = 2\gamma k_B T m \delta(t - t').
\] (4.4)

This relation, a fluctuation-dissipation theorem (FDT) [62, 66], contains the thermal information and highlights a stochastic uncertainty. Integrating eq. (4.2) using the above FDT yields the famous Einstein relation

\[
\langle x^2 \rangle \sim t.
\] (4.5)

The basic concept of the LE and stochastic differential equations is already explained in this brief example. The principle of adding a stochastic force with zero mean is more generalizable and is explained in the following in more detail. The comparison of Einstein’s and Langevin’s approach denotes an equivalence between the stochastic calculus and a deterministic differential equation, the Fokker-Planck equation. I will consider this issue and the connection between both perspectives - the Kramers-Moyal expansion - in Sec. 4.2.

4.1.1. Zwanzig’s approach

The particle of the above example can be considered as a system and the surrounding fluid which exhibits arbitrary fluctuations can be understood as a bath coupled to this system. This picture already covers the whole idea of a Langevin equation. It basically describes the dynamics of a system coupled to a bath. While the dynamics of the system coordinates are explicitly described, the bath characteristics are summarized in the stochastic force term. A famous approach to derive the Langevin equation from first principles was introduced by Zwanzig [62]: Let \(H_{tot}\) be the Hamiltonian that describes both system and bath for a - without loss of generality- one-dimensional system coupled bilinearly to a harmonic bath. Then

\[
H_{tot} = H_{sys} + H_{bath} = \frac{p^2}{2m} + U(x) + \sum_{i=1}^{N} \left( \frac{p_i^2}{2m_i} + \frac{\omega_i}{2} \left( q_i - \frac{c_i}{\omega_i^2} x \right)^2 \right),
\] (4.6)

where the system is described via \(x, p\) and the i-th bath coordinate via \(q_i, p_i\). The frequency of individual bath oscillators is denoted by \(\omega_i\), the coupling strength is contained in \(c_i\),

\(^1\)since \(\int_0^{t_1} dt_1' \int_0^{t_2} dt_2' f(t_1', t_2') \delta(t_1' - t_2') = \int_0^{\min(t_1, t_2)} dt_1' f(t_1', t_2' = t_1')\)
We get \( N+2 \) equations of motion from the canonical equations,

\[
\begin{align*}
    \frac{dx}{dt} &= \frac{p}{m} \\
    \frac{dp}{dt} &= -U'(x) + \sum_i c_i \left( q_i - \frac{c_i}{\omega_i^2} x \right) \\
    \frac{dq_i}{dt} &= p_i \\
    \frac{dp_i}{dt} &= -\omega_i^2 q_i + c_i x,
\end{align*}
\]

where the masses of the harmonic oscillators are set to \( m_i = 1 \) for simplicity. If the time dependence of the system is known, the equations of motion of the bath coordinates can be solved. Setting \( x_i = (q_i, \tilde{p}_i)^T \) where \( \tilde{p} = p/m, b = (0, c_i)^T \) and

\[
A = \begin{pmatrix} 0 & \omega_i^2 \\ -\omega_i^2 & 0 \end{pmatrix},
\]

they can be written in the simplified form

\[
\frac{d}{dt} x_i = A x_i + b.
\]

This equation can be solved analytically,

\[
x_i(t) = \exp(At)x_i(0) + \int_0^t \exp(A(t-s)b(s)) \, ds,
\]

with

\[
\exp(At) = \sum_{j=0}^{\infty} \frac{(At)^j}{j!} = \begin{pmatrix} \cos(\omega_i t) & \sin(\omega_i t)/\omega_i \\ -\omega_i \sin(\omega_i t) & \cos(\omega_i t) \end{pmatrix}.
\]

Partial integration of (4.12) then yields an expression which allows to recast eq. (4.8),

\[
\frac{dp}{dt} = -U'(x) - \int_0^t p(t-s)K(s) \, ds + F_s(t),
\]

where

\[
K(s) = \sum_{i=1}^{N} \frac{c_i^2}{\omega_i^2} \cos(\omega_i s),
\]

\[
F_s(t) = \sum_{i=1}^{N} c_i \left( p_i(t=0) \frac{\sin(\omega_i t)}{t} + \left( q_i(t=0) - \frac{c_i}{\omega_i^2} \right) \cos(\omega_i t) \right).
\]
The second part on the right hand side of eq. (4.13) obviously relates the derivative of the friction with its past. Therefore, $K$ is called memory-kernel. For a dense bath, meaning a continuous spectrum $\{\omega_i\} \rightarrow \omega$, we can introduce a functional mapping $c = c(\omega)$. Eq. (4.14) then adopts the integral form

$$K(t) = \int_0^\infty g(\omega) \frac{c^2(\omega)}{\omega^2} \cos(\omega t) d\omega,$$

(4.16)

where $g(\omega)$ specifies the density of frequencies. The special Debye case $g(\omega) \sim \omega^2 \forall \omega < \omega_D$ gives

$$K_D(t) \sim \frac{\sin(\omega_D t)}{\omega_D^2 t}.$$

(4.17)

This is a representation of the delta distribution for high cut-off frequencies $\omega_D$, since $\lim_{t \rightarrow 0} K_D(t) \sim 1/\omega_D$. Thus, the memory dependence of the second term in eq. (4.13) vanishes and the expression becomes Markovian.

$F_s$ indicates at that point already, that the stochastic character is contained in (4.15). Since it is defined by the initial distribution of discrete points $p_j(t = 0), q_j(t = 0)$, it is a known function of time in general. However, in the continuous limit $\{\omega_i\} \rightarrow \omega$, the number of degrees of freedom becomes infinite. $F_s$ can then be replaced by continuous Fourier series defining a distribution which is determined by the initial conditions of the bath. If the initial configuration is sampled from the Boltzmann distribution $f(\{p_i\}, \{q_i\}) \sim -\exp(H_{\text{bath}}/k_B T)$, the first and second moments are

$$\langle p_i(t = 0) \rangle = 0 \quad \langle q_i(t = 0) - \frac{c_i}{\omega_i^2} \rangle = 0$$
$$\langle p_i(t = 0)^2 \rangle = k_B T \left\langle \left( q_i(t = 0) - \frac{c_i}{\omega_i^2} \right)^2 \right\rangle = \frac{k_B T}{\omega_i^2}.$$

These relations directly characterize the stochastic force:

$$\langle F_s(t) \rangle = 0$$
$$\langle F_s(t) F_s(t') \rangle = k_B T K(t - t').$$

(4.18)

(4.19)

Obviously, this fluctuation-dissipation theorem resembles (4.4). It links the fluctuations of the stochastic force with the memory Kernel which in turn reflects the dissipation of the system.
4.1.2. Markovian Langevin equations

While we restrained to the one-dimensional system for introductory purposes, a general extension is just straightforward. Let \( x \) be an \( d \)-dimensional mass-weighted system coordinate which is twice differentiable, \( U(x) \) be a potential function, \( \Gamma(x) \) a friction- and \( K(x) \) a diffusion-matrix which are all time-independent. A generalized Markovian LE can then be written in the form

\[
\ddot{x} = f(x) - \Gamma(x) \dot{x} + K(x) \xi(t),
\]

(4.20)

where \( f(x) = -\nabla U(x) \) denotes a drift field induced by a potential gradient. Since this equation is second order, it will be referred to as second order model Langevin equation (mLE2).\(^3\) The stochastic force \( F_s(x, t) = K(x) \xi(t) \) is aligned such that \( \xi \) is sampled from a normal distribution that obeys

\[
\langle \xi(t) \rangle = 0
\]

(4.21)

\[
\langle \xi(t) \xi^T(t') \rangle = \mathbb{I} \delta(t - t').
\]

(4.22)

Apart from zero mean, the stochastic force is uncorrelated pairwise for system components and in time in the Markovian limit. Since the Fourier transform of (4.22) is a constant function of the frequency, this is often referred to as the white-noise case. The solution of eq. (4.20) is completely determined by the three fields \( f, \Gamma \) and \( K \). The mLE2 eq. (4.20) adopts a simplified form in the so-called “strong friction” limit or overdamped limit: If the external force acting on the system is negligible, compared with the frictional force (\(|\ddot{x}| \ll |\Gamma(x) \dot{x}|\)), eq. (4.20) reduces to a first order LE (first order model Langevin equation (mLE1)),

\[
\dot{x} = h(x) + D(x) \xi(t),
\]

(4.23)

where

\[
h(x) = -\Gamma^{-1}(x) \nabla U(x) = \Gamma^{-1}(x) f(x)
\]

(4.24)

\[
D(x) = \Gamma^{-1}(x) K(x)
\]

(4.25)

denote drift and diffusion fields. These two fields summarize the dynamical behavior governed by eq. (4.23). The presence of the fluctuation-dissipation theorem [62, 65, 68]

\(^2\)We implicitly demand a diagonal form of the mass tensor at this point.

\(^3\)The preposition “model” is used in order to discriminate it from the dLE, which is introduced in Sec. 4.3.
4.1. The Langevin equation

further interconnects the fields of the model Langevin equations (mLEs),

\[ \langle K(x)K^T(x) \rangle = 2\Gamma(x)k_BT, \]  
\[ \langle D(x)D^T(x) \rangle = 2\Gamma^{-1}(x)k_BT. \]

4.1.3. Ito and Stratonovich calculus

Due to its stochastic character, the mLEs eqs. (4.23) and (4.20) are not completely defined or - more strictly speaking - do not exist mathematically [66]. In an integral formulation, eq. (4.23) adopts the form

\[ x(t+\delta t) = x(t) + \int_t^{t+\delta t} h(x(t'))dt' + \int_t^{t+\delta t} D(x(t'))dW(t'), \quad dW(t') = \xi(t')dt'. \]

(4.28)

While the first integral is a Riemann integral since \( h \) is continuous, the second integral is a stochastic Stieltjes integral. For the special case of constant \( D \), it reduces to a Wiener integral, \( dW \) being the so-called Wiener increment. It is formally solved by partitioning the time interval \([t, t+\delta t]\). Introducing partial sums

\[ S_n = \sum_{i=1}^{n} D(r(\tau_i)) (W(t_i) - W(t_{i-1})) , \]

(4.29)

the stochastic integral is defined as the limiting case

\[ \int_t^{t+\delta t} D(r(t'))dW = ms - \lim_{n \to \infty} S_n, \]

(4.30)

where the mean square limit on the right hand side denotes that the standard deviation of \( S_n \) converges in the mean square against a fixed value \( S \),

\[ \lim_{n \to \infty} \langle (S_n - S)^2 \rangle = 0. \]

Still the definition eq. (4.30) allows for alternative choices of intermediate points

\[ \tau_i(\alpha) = \alpha t_i + (1-\alpha)t_{i-1}. \]

(4.31)

At this point, different stochastic interpretations diverge, the most prominent of these are the Ito calculus using \( \alpha = 0 \) and the Stratonovich calculus using \( \alpha = 1/2 \).

The Ito interpretation turns out to be most advantageous for numerical purposes, since it facilitates the propagation of the Langevin equation in an Euler manner. If we presume that a function \( f \) is describing an adapted process, that is if \( f(t) \) it is statistically independent of \( W(s) - W(t) \) for \( s > t \), only the Ito interpretation of a stochastic differential
Chapter 4. Data-driven Langevin modeling

equation fulfills the identity
\[ \langle \int_t^{t+\delta t} f(t')dW(t') \rangle = 0 \]  
(4.32)
for the mean of the integral over the stochastic variable. Further, the identity
\[ \int_t^{t+\delta t} f(t') \left( dW(t') \right)^{2+N} = \begin{cases} \int_t^{t+\delta t} f(t')dt' & N = 0 \\ 0 & N > 0. \end{cases} \]  
(4.33)
also only holds in the Ito interpretation [66]. It facilitates a numerical propagation of the Ito Langevin equation, utilizing the Wiener increment \( dW \equiv \delta t^{1/2} \). The numerical implementation of the Langevin equation eq. (4.23) in this case reads
\[ x(t + \delta t) = x(t) + h(x)\delta t + D(x)\xi(t)\sqrt{\delta t}, \]  
(4.34)
where \( \xi(t) \) is a Gaussian-distributed random number. The above discretization is also referred to as Euler-Maruyama method and central to the data-driven Langevin equation, Sec. 4.3.

Using the above identities which also hold for the multidimensional case, one can derive how the Ito interpretation affects any function \( f \) of a stochastic variable \( x \). For the Langevin equation eq. (4.23), this yields
\[ df(x(t)) = \sum_i h_i(x(t)) \frac{\partial}{\partial x_i} f(x(t))dt + \frac{1}{2} \sum_{i,j} \{ D(x(t))D^T(x(t)) \} \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} f(x(t))dt \]
\[ + \sum_{ij} D_{ij}(x(t)) \frac{\partial}{\partial x_i} f(x) dW_j(t). \]  
(4.35)

Despite the fact, that the alternative Stratonovich interpretation is not straightforward to integrate, functions of the stochastic variable transform similarly as in ordinary differential equations (i.e. \( df(x(t)) = dx(t)f'(x(t)) \) in the one-dimensional case). Both pictures are related as follows [66]:
\[ h_i(x, t) = h_{S,i}(x, t) + \frac{1}{2} \sum_{j,k} D_{k,j}(x, t) \frac{\partial}{\partial x_k} D_{i,j}(x, t), \]  
(4.36)
\[ D(x, t) = D_S(x, t) \]  
(4.37)
where the index \( S \) on the right hand side of both expressions denotes the use of the Stratonovich interpretation.
4.1.4. Noise-induced drift

The discrepancy of Ito and Stratonovich interpretation is reflected in the second term of the transformation formula eq. (4.36). It is denominated noise-induced or spurious drift since it only appears in case of a multiplicative noise which means for space dependent diffusion fields. We can circumvent this problem in the one-dimensional case, applying a coordinate transformation. Setting \( \tilde{x}(t) = x(t)/D(x) \), eq. (4.23) becomes

\[
\tilde{x}(t) = \frac{\tilde{h}(\tilde{x}(t))}{D(x)} + \xi(t) \equiv \tilde{\tilde{h}}(\tilde{x}) + \xi(t),
\]

where \( \tilde{\tilde{h}} \) presumes that \( \tilde{x} = \int dx/D(x) \) exists and is invertible. However, for the multi-dimensional case the ambiguity remains. It arises from the noise property to be delta-correlated.

As the limit of a process that shows finite noise correlation, the Markovian case can be shown to reduce to a stochastic differential equation with the same coefficients as if interpreted in the Stratonovich sense [66, 69]. For this reason, it is also often referred to as the more physical interpretation. Following this idea, we have to consider a the noise-induced shift\(^4\) in the physical interpretation of a LE interpreted in the Ito-sense. The properties and consequences of this effect as well as various interpretations are widely discussed in the literature [115–119].

4.2. The Fokker-Planck Equation

A different approach to describe the dynamics of a stochastic system is directly from the perspective of ensemble averages [66, 67]. For example, the local average of the displacement (or drift) of the system coordinate \( x \) at \( x' \) which is governed by a Langevin process can be written as

\[
\langle x(t + \delta t) - x(t) \rangle \big|_{x(t)=x'} = \int dx(x - x')p(x, t + \delta t | x', t),
\]

where the integral covers all possible realizations of the system coordinate. The quantity \( p \) denotes the conditional probability of the system to be at \( x \) after a time interval \( \delta t \), being at \( x' \) beforehand. The probability distribution \( P \) of the system can be expressed via this conditional probability, following the idea of the Chapman-Kolmogorov equation eq. (2.25). In integral form, it reads

\[
P(x, t + \delta t) = \int p(x, t + \delta t | x', t)P(x', t)dx'.
\]

\(^4\)which is therefore at times also referred to as Ito-shift
Chapter 4. Data-driven Langevin modeling

The integrand can be evolved in a Taylor series around \( x = x' \),

\[
p(x, t + \delta t|x', t)P(x', t) = p(x' + (x - x'), t + \delta t|x - (x - x'), t)P(x - (x - x'), t)
\]

\[
= \sum_{n=0}^{\infty} \frac{(x - x')^n}{n!} \left(-\frac{\partial}{\partial x}\right)^n p(x', t + \delta t|x, t)P(x, t)
\]  

(4.41)

Finally, a time evolution of the probability distribution can be derived in the limit \( \delta t \to 0 \) using the above evaluation,

\[
\frac{\partial}{\partial t} P(x, t) = \lim_{\delta t \to 0} \frac{P(x, t + \delta t) - P(x, t)}{\delta t}
\]

\[
= \lim_{\delta t \to 0} \sum_{n=1}^{\infty} D^{(n)} P(x, t),
\]  

(4.42)

where

\[
D^{(n)} = \int \frac{(x - x')^n}{n!} p(x', t + \delta t|x, t)dx'
\]  

(4.43)

are the so-called Kramers coefficients. Incidentally, for \( n = 1 \), we recover the drift eq. (4.39).

**Pawula Theorem**  The expansion eq. (4.42) collapses for special cases. This can be easily shown using Hoelder’s inequality which is valid for positive transition rates \( p(x') \equiv p(x', t + \delta t|x, t) \),

\[
\left( \int (x - x')^{2n+m} p(x')dx' \right)^2 \leq \int (x - x')^{2n} p(x')dx' \int (x - \tilde{x})^{2n+2m} p(\tilde{x})d\tilde{x}
\]

\[
\Leftrightarrow \left( (2n + m)!D^{(2n+m)} \right)^2 \leq (2n)!D^{(2n)}(2n + 2m)!D^{(2n+2m)}.
\]  

(4.44)

This inequality is fulfilled for \( 2n + m \leq 2 \): while for \( n = 0 \) and \( m = 1 \) the above expression is always fulfilled, for \( m = 0 \), we have equality. For \( n, m \geq 1 \), one can see, that \( D^{(2n)} = 0 \Leftrightarrow D^{(2n+m)} = 0 \) \footnote{since both \( D^{(2n+2m)} = 0 \Rightarrow D^{(2n+m)} = 0 \) and \( D^{(2n)} = 0 \Rightarrow D^{(2n+m)} = 0 \) follow from eq. (4.44)}. This finding is called the Pawula Theorem. It states that the expansion (4.42) either stops after the second term or has to contain an infinite number of terms.

Truncating eq. (4.42) after the second term yields the Fokker-Planck equation.

It can be shown, that the above derivation is straightforward generalizable towards a \( d \)-dimensional system \( \mathbf{x} = \{x_i\} \), yielding the form [67]

\[
\frac{\partial}{\partial t} P(\mathbf{x}, t) = -\sum_i \frac{\partial}{\partial x_i} \left(D^{(1)}_{i}(\mathbf{x}, t)P(\mathbf{x}, t)\right) + \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} \left(D^{(2)}_{ij}(\mathbf{x}, t)P(\mathbf{x}, t)\right).
\]
This expression can also be written in terms of the conditional probability \( p \) which marks the special case \( P(x, t = 0) = \delta(x - x') \). We will stick to this form in the following for convenience, using the abbreviated notation \( p \equiv p(x, t|x', t') \) as Gardiner [66]. We obtain

\[
\frac{\partial}{\partial t} p = - \sum_i \frac{\partial}{\partial x_i} \left( D_i^{(1)}(x, t) p \right) + \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} \left( D_{ij}^{(2)}(x, t) p \right),
\] (4.45)

with the multidimensional Kramers-Moyal coefficients

\[
D_i^{(1)}(y, t) = \lim_{\tau \to 0} \frac{1}{\tau} \langle x_i(t + \tau) - x_i(t) \rangle |_{x(t) \to y(t)},
\] (4.46)

\[
D_{ij}^{(2)}(y, t) = \frac{1}{2} \lim_{\tau \to 0} \frac{1}{\tau} \langle (x_i(t + \tau) - x_i(t))(x_j(t + \tau) - x_j(t)) \rangle |_{x(t) \to y(t)}. \] (4.47)

### 4.2.1. Solutions for small time steps

The Fokker-Planck equation (4.45) can be directly propagated ab initio. While for time-independent drift and diffusion coefficients, the solution of the Fokker-Planck equation is straightforward for comparably small time steps, meaning that drift and diffusion coefficients are presumed to be not time-dependent on the propagation time scale. Approximate solutions can be obtained for time-dependent fields. To illustrate this, we restrain to the one-dimensional case again. The extension towards the multidimensional case is straightforward. Since \( p(x, t = t'|x', t') \equiv \delta(x - x') \), eq. (4.41) can be propagated for small \( \delta t = t - t' \) via [67]

\[
\frac{\partial}{\partial t} p(x, t|x', t') \approx \exp \left[ \frac{\partial}{\partial x} D^{(1)}(x', t') \delta t + \frac{\partial^2}{\partial x^2} D^{(2)}(x', t') \delta t \right] \delta(x - x')
\]

\[
= \frac{1}{2\pi} \int_{-\infty}^{\infty} dk \exp \left( -ikD^{(1)}(x', t')\delta t - k^2 D^{(2)}(x', t')\delta t + ik(x - x') \right)
\]

\[
= \frac{1}{2\sqrt{\pi D^{(2)}(x', t')\delta t}} \exp \left( -\frac{(x - x' - D^{(1)}(x', t')\delta t)^2}{4D^{(2)}(x', t')\delta t} \right). \] (4.48)

Notably, if the coefficients are constant, eq. (4.48) holds for arbitrary time steps.

The above expression is often used to propagate the dynamics of system which obey the Fokker-Planck equation or the associated Langevin equation: For sufficiently short time intervals, the Kramers-Moyal coefficients can be calculated to forecast the time evolution of the system\(^6\). With regards to contents, eq. (4.48) is congruent with the microscopic Langevin equation using the Euler-Maruyama method eq. (4.34).\footnote{Further refinement has been suggested based on the time evolution of the Fokker-Planck equation to account for finite time-corrections [120].}
4.2.2. Relation to Markovian Langevin equations

Eq. (4.45) is an equivalent macroscopic description of a system which obeys the Markovian LE in the Ito picture, since it originates a forward discretization of the system variable $x$ and the Chapman-Kolmogorov property. Formally we can relate both expressions applying eq. (4.35) based on the identity

$$
\frac{d}{dt} \langle f(x(t)) \rangle_{x(t)=x'} = \left( \frac{df(x(t))}{dt} \right)_{x(t)=x'}.
$$

(4.49)

Proceeding conceptually similar to the above derivation of the Fokker-Planck equation, one can recover \[66\]

$$
\frac{\partial}{\partial t} p = \sum_i \frac{\partial}{\partial x_i} [h_i(x)p] + \frac{1}{2} \sum_{i,j} \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} \left[ \{D(x)D^T(x)\}_{ij} p \right].
$$

(4.50)

Notably, the diffusive contribution is invariant under local unitary transformations $U(x)^7$. Therefore, the diffusion field of the corresponding Langevin equation is not uniquely defined. We can identify drift and diffusion fields of mLE1 eqs. (4.24), (4.25) directly from this expression,

$$
h(x, t) = \{D^{(1)}\}(x, t),
$$

(4.51)

$$
\frac{D(x, t)}{2}D(x, t)^T = \{D^{(2)}\}(x, t).
$$

(4.52)

Likewise, the second order LE can be expressed in terms of a Fokker-Planck equation. Considering the mLE2 eq. (4.20) as a system of two first order differential equations yields the corresponding Fokker-Planck equation

$$
\frac{\partial}{\partial t} p = \sum_i \left[ v_i \frac{\partial}{\partial x_i} p + \frac{\partial}{\partial v_i} \{f_i(x, v) + (\gamma(x, v))_i p \} \right]
+ \frac{1}{2} \sum_{i,j} \frac{\partial}{\partial v_i} \frac{\partial}{\partial v_j} \left[ \{K(x, v)K^T(x, v)\}_{ij} p \right],
$$

(4.53)

which explicitly introduces a phase-space dependence of the probability distributions.

The corresponding Fokker-Planck description of the overdamped Langevin equation eq. (4.23) which describes the dynamics of Brownian particles that experience a constant friction $\gamma$ is also referred to as Smoluchowski equation and reads

$$
\frac{\partial}{\partial t} p = \gamma^{-1} \left[ \nabla U'(x) + k_B T \nabla^2 \right] p.
$$

(4.54)

\[7\]since for $\tilde{D}(x) = D(x)U(x)$ follows: $\tilde{D}(x)\tilde{D}^T(x) = D(x(t))U(x(t)U^T(x(t)))D^T(x(t)) = \tilde{D}(x)\tilde{D}^T(x)$

51
4.3. The data-driven Langevin equation (dLE)

The mLE1 as introduced above is applied to propagate the dynamics of a model system provided the field information. A more challenging task is the inverse problem, that is to reconstruct the performance of a mLE1 without explicit knowledge about the fields directly from a time series. A huge number of applications are used to deal with systems that show stochastic behavior, e.g. autoregressive modeling, Markov state modeling and Langevin modeling [58, 73, 75, 76, 79, 121–126]. An approach to model dynamics governed by the Langevin equation was first introduced as an application working in a data based manner by Siegert and Friedrich [71, 72]. The basic idea of this ansatz, which is working on the basis of the mLE1 (4.23) is to locally estimate the drift and diffusion coefficients $h$ and $D$. In the context of biomolecular dynamics, this method was first applied at the example of the hepta-alanine peptide [80]. Prior to the application of the dLE a suitable set of system coordinates has to be established that fulfills the model assumptions required in a system governed by the dLE.

In this section, firstly the basic concept of the first order data-driven Langevin equation (dLE1) will be introduced, an algorithm mimicking the dynamical behavior of the first order mLE1 [80, 127]. This algorithm will be modified and extended towards the second order dLE2 afterwards which mimics the behavior of the more general mLE2. In the remainder of this chapter, we will denote the system coordinate that is propagated by a dLE by $y$ in order to distinguish it from the coordinate denoting the input data, $x$.

4.3.1. The first order data-driven Langevin equation (dLE1)

In order to mimic the dynamics of the mLE1, a data-driven algorithm requires a similar propagation scheme. Following the Ito interpretation, the Euler-Maruyama scheme mentioned above is assumed to propagate the $d$-dimensional system coordinates $y$ of a dLE1 realization and $x$ of the model reference similarly,

$$
x_{m+1} = x_m + h(x_m)\delta t + D(x_m)\delta t \xi_m,
$$

$$
y_{n+1} = y_n + h(y_n)\delta t + D(y_n)\delta t \xi_n,
$$

where $n$ and $m$ denote the discretized time step, respectively. The basic idea for a data-driven algorithm requires a definition of estimators that determine the drift and diffusion fields $h = \{h_i\}$ and $D = \{D_{ij}\}$ in eq. (4.23). In order to estimate fields that propagate $y$ similarly to $x$, we demand

$$
h(y_n) = \langle h(x_m) \rangle,
$$

where $\langle \cdot \rangle$ denotes the average over all available data points.
where the average $\langle . \rangle$ is evaluated locally. If the above average is applied to the mLE1 propagation scheme (4.55), using the noise property (4.21) we get a relation for the drift field,
\[
h(y_n) = (x_{m+1} - x_m)/\delta t. \tag{4.58}
\]

**The Estimator**

In order to determine an estimator for this field, the average is defined as follows: Considering neighboring realizations $x_m$ of our mLE1 reference to the dLE at time $n$, we demand
\[
\langle F(x_m) \rangle = \frac{1}{k} \sum_{m \in I_n^k} F \{ \{ x_{m+i} \} , x_m \}, \tag{4.59}
\]
where $i = 1$ for the dLE1. The parameter $k$ defines the number of nearest neighbors of the input data at time step $n$ at this which are determined by a metric,
\[
\sum_j \Theta \left( |x_j - y_n| - \epsilon \right) = k. \tag{4.60}
\]

$\Theta$ denotes the Heavyside stepfunction in this expression. We use the Euclidean metric for all estimates, thus interpreting $\epsilon$ as the radius of a $d$-dimensional hypersphere.

In order obtain an estimate for the diffusion matrix, one has to consider the second momentum. If we argue similarly to the drift field definition and use the property of the drift field to propagate both mLE and dLE, we obtain
\[
D(y_n)D^T(y_n) = \langle D(x_m)D^T(x_m) \rangle = \langle D(x_m)\xi_m\xi_m^T D^T(x_m) \rangle = \delta t \langle D(x_m)\xi_m\xi_m^T D^T(x_m) \rangle = \delta t \langle (h(y_n) - (x_{m+1} - x_m)/\delta t) (h(y_n) - (x_{m+1} - x_m)/\delta t)^T \rangle = \frac{1}{\delta t} \left( \langle (x_{m+1} - x_m)(x_{m+1} - x_m)^T \rangle / \delta t - h(y_n)h(y_n)^T \right) = \text{Cov}(x_{m+1} - x_m, x_{m+1} - x_m)/\delta t. \tag{4.61}
\]

In the second line, we used eq. (4.22) and the statistical independency of $D$ and $\xi$ to obtain
\[
\langle D_{ij}\xi_k\xi_l D_{ji} \rangle = \langle D_{ik}D_{jl} \rangle \langle \xi_k\xi_l \rangle = \langle D_{ik}D_{jk} \rangle / \delta t. \tag{4.62}
\]
The covariance matrix $\text{Cov}(x, y) = \langle xy^T \rangle - \langle x \rangle \langle y \rangle^T$ in the last line yields an abbreviated form. Eq. (4.61) does not uniquely define the diffusion field. Since the covariance is Hermitian and positive-definite, the Cholesky decomposition is numerically efficient. It is applied thus to compute the diffusion field, which therefore has lower triangular form. The noise coupling is uniquely defined by the decomposition: Since the $i$-th coordinate receives a diffusive displacement

---

*in Einstein summation convention*
4.3. The data-driven Langevin equation (dLE)

\[ \sum_{j=1}^{i} D_{ij} \xi_j \], the first system component is only influenced by the first entry of the noise vector, \( \xi_1 \), while higher components receive contributions from all lower components of the noise vector.

Notably, the definition of the fields eqs. (4.58), (4.61) is consistent with the Kramer-Moyal coefficients (4.46), (4.47), if the propagation time step \( \delta t \) is small enough and neighborhood locality is given. They are referred to in other works [71, 72] and refined for finite time steps [120]. The performance of the dLE1 is further discussed in the next chapter at the example of a simple model.

**Modified Average**

While the original dLE1 basically defines estimates via local averages of its input data that consider differences of next neighbors and their temporal successors, one might also suggest a slight modification: If we assume a dense or at least a sufficiently isotropic neighborhood, the average of the nearest neighbors for reliable statistics (large values of \( k \)) coincides with the actual realization of the dLE, namely

\[ \langle x_m \rangle \equiv y_n. \tag{4.62} \]

This interference slightly modifies the definition of the drift field\(^9\):

\[ h(y_n) \delta t = \langle x_{m+1} \rangle - y_n \tag{4.63} \]

The alternative definition is consistent with (4.58) in the limit of high local density, since \( \lim_{\epsilon \to 0} \langle x_{m+1} - x_m \rangle = \langle x_m \rangle - y_n \). Moreover, the propagation scheme (4.56) becomes

\[ y_{n+1} = \langle x_{m+1} \rangle + \delta t D(y_n) \xi_n, \tag{4.64} \]

which seems intuitively plausible: the displacement of the dLE1 realization simply imitates the average drift displacement of the underlying mLE1 and superimposes a diffusive displacement. The diffusion matrix is then derived analogously to (4.61),

\[ D(y_n) D^T(y_n) = \text{Cov}(x_{m+1} - x_m, x_{m+1} - x_m) / \delta t = \text{Cov}(x_{m+1}, x_{m+1}) / \delta t, \tag{4.65} \]

where the last line directly results from the fixation (4.62) and translational invariance of the covariance matrix. Strikingly, the diffusive displacement explicitly depends only on the average temporal successors of the input data; this underlines the superposition

\(^9\)Note that the average (4.59) still considers nearest neighbors at time step \( n \)!
Chapter 4. Data-driven Langevin modeling

4.3.2. The second order data-driven Langevin equation (dLE2)

As mentioned above, the dLE1 and modified brands of the algorithm are widely used and discussed to all intents and purposes. The main shortcoming of the scheme however is that its applicability is restricted to the strong friction limit. Nevertheless, analogously to the dLE1 scheme, a second order approach can be introduced which defines field estimates to recover the dynamics of the more general mLE2. For this purpose, a discretized scheme to propagate the velocity $w = \dot{y}$ of the dLE2 is used and matched to a similarly discretized scheme propagating the velocity $v$ of the mLE2 input,

$$v_{m+1} = f(x_m)\delta t - \tilde{\Gamma}(x_m)v_m + K(x_m)\xi_m\delta t,$$  \hspace{1cm} (4.66)

$$w_{n+1} = f(y_n)\delta t - \tilde{\Gamma}(y_n)w_n + K(y_n)\xi_n\delta t,$$  \hspace{1cm} (4.67)

where the effective friction field is abbreviated via $\tilde{\Gamma}(x_n) := \delta t\Gamma(x_n) - I$. A definition for the field estimate is established by calculating a time-lagged covariance,

$$\text{Cov}(v_{m+1}, v_m) = \text{Cov} \left( f(x_m)\delta t - \tilde{\Gamma}(x_m)v_n + K(x_m)\xi_m\delta t, v_m \right)$$

$$= -\tilde{\Gamma}(y_n) \text{Cov}(v_m, v_m)$$

$$\Leftrightarrow \tilde{\Gamma}(y_n) = -\text{Cov}(v_{m+1}, v_m)\text{Cov}^{-1}(v_m, v_m)$$  \hspace{1cm} (4.68)

In order to derive the above expression, it is assumed that all mLE2 fields considered for averaging are constant and equal to the respective dLE2 field. (e.g $f(y_n) = f(x_m) \forall m \in \mathbb{T}_k^n$). An expression for the drift field is determined analogously to the overdamped case:

$$f(y_n) = \langle f(x_m) \rangle$$

$$= \langle v_{m+1} \rangle + \tilde{\Gamma}(y_n)\langle v_m \rangle$$  \hspace{1cm} (4.69)

To derive the generalized diffusion field estimate, its introductory to insert the drift field estimator into the model eq. (4.66):

$$v_{m+1} = \langle v_{m+1} \rangle + \tilde{\Gamma}(y_n)(\langle v_m \rangle - v_m) + K(x_m)\xi_m\delta t$$  \hspace{1cm} (4.70)

The basic idea to obtain an expression for $K$ is then similar to the first order case:

$$K(y_n)K^T(y_n)\delta t = \langle K(x_m)K^T(x_m) \rangle$$

$$= \text{Cov}(v_{m+1}, v_{m+1}) - \tilde{\Gamma}(y_n)\text{Cov}(v_m, v_m)\tilde{\Gamma}^T(y_n)$$  \hspace{1cm} (4.71)
4.3. The data-driven Langevin equation (dLE)

If the second covariance on the right hand side of eq. (4.71) is perfectly symmetric\(^{10}\), the expression may be further simplified using eq. (4.68):

\[
K(y_n)K^T(y_n)\delta t = \text{Cov}(v_{m+1}, v_{m+1}) + \tilde{\Gamma}(y_n) \text{Cov}(v_m, v_{m+1})
\]  

(4.72)

So far we only introduced field estimators for a dLE2 that are functions of the velocities \(w, v\). In order to access a propagation scheme for the system coordinate \(y\), we have to agree on a discretization scheme which we apply to \(\dot{x} = v\) and \(\dot{y} = w\) likewise. We do not have to discriminate different schemes in order to solve this equation: If the velocities can be assumed to be continuous functions in time, we do not have to consider the ambiguity of stochastic integration, see Sec. 6.1.1. In order to gain an iterative scheme to propagate system coordinates from eq. (4.67) which only considers realizations of the coordinate itself in the past, we apply a simple backward Euler discretization\(^{11}\),

\[
w_n\delta t = y_n - y_{n-1}.
\]  

(4.73)

Inserting this into eq. (4.67) then yields

\[
y_{n+1} = y_n + \delta t^2 f(y_n) - \tilde{\Gamma}(y_n)(y_n - y_{n-1}) + \delta t^2 K(y_n)\xi_n
\]  

(4.74)

and likewise for eq. (4.66). If we express the different field estimates through system coordinates \((\Delta x_m = x_m - x_{m-1})\), we get

\[
\tilde{\Gamma}(y_n) = -\text{Cov}(\Delta x_{m+1}, \Delta x_m) \text{Cov}^{-1}(\Delta x_m, \Delta x_m),
\]  

(4.75)

\[
\delta t^2 f(y_n) = \langle \Delta x_{m+1} \rangle + \tilde{\Gamma}(y_n)\langle \Delta x_m \rangle,
\]  

(4.76)

\[
K(y_n)K^T(y_n)\delta t^3 = \text{Cov}(\Delta x_{m+1}, \Delta x_{m+1}) - \tilde{\Gamma}(y_n) \text{Cov}(\Delta x_{m+1}, \Delta x_{m})\tilde{\Gamma}^T(y_n).
\]  

(4.77)

Introducing the additional constraint eq. (4.62), the field estimates read

\[
\tilde{\Gamma}(y_n) = \text{Cov}(x_{m+1}, x_{m-1}) \text{Cov}^{-1}(x_{m-1}, x_{m-1}),
\]  

(4.78)

\[
\delta t^2 f(y_n) = \langle x_{m+1} \rangle - y_n + \tilde{\Gamma}(y_n)(y_n - \langle x_{m-1} \rangle),
\]  

(4.79)

\[
K(y_n)K^T(y_n)\delta t^3 = \text{Cov}(x_{m+1}, x_{m+1}) - \tilde{\Gamma}(y_n) \text{Cov}(x_{m-1}, x_{m-1})\tilde{\Gamma}^T(y_n).
\]  

(4.80)

Eq. (4.72) reduces to

\[
K(y_n)K(y_n)^T\delta t^3 = \text{Cov}(x_{m+1}, x_{m+1}) - \tilde{\Gamma}(y_n) \text{Cov}(x_{m-1}, x_{m+1})
\]  

(4.81)

\(^{10}\)which is not necessarily true since it is numerically computed

\(^{11}\)The alternative forward Euler discretization is discussed in Sec. 6.1.1.
Chapter 4. Data-driven Langevin modeling

for this form. This shape nicely illustrates, that the inertia effects included in the effective friction \( \tilde{\Gamma} \) that contribute to the diffusion are related to the past of the system. As already discussed in the context of the dLE1, both versions of the estimators are consistent for sufficient local density. However for \( \Delta x_m \to 0 \), the inversion of the covariance matrix in eq. (4.75) becomes unstable. This also affects the accuracy of the other fields since they receive contributions of the friction field estimate. For this reason, we refer to the latter definitions of the estimators in our application in Ch. 6. A detailed comparison of the different schemes is also provided at that point.

4.3.3. Overdamped limit of the dLE2

The dLE2 fields can be inserted into eq. (4.74) in order to turn it into a form which is comparable to eq. (4.64),

\[
y_{n+1} = \langle x_{m+1} \rangle - \tilde{\Gamma}(y_n) (\langle x_{m-1} \rangle - y_{n-1}) + \delta t \tilde{K}(y_n) \xi_n,
\]

(4.82)

where \( \tilde{K} = K \delta t \). Apparently the above expression reduces to (4.64) if \( \tilde{\Gamma} = 0 \). In this case, (4.80) resembles (4.65). Notably, this case requires the identity \( \Gamma(y_n) = I \delta t^{-1} \) which is not attainable for space-dependent friction. It even implies a time step dependence of the propagation scheme in the case for constant \( \Gamma \) which is no more capable of correctly propagating the underlying mLE2. Obviously, \( \tilde{\Gamma} = 0 \) is not achievable physically. Nevertheless, we can conduct another intellectual game: Suppose that our dLE input data was generated by a mLE2 and the time step is adopted such that the condition \( \Gamma(y_n) = I \delta t^{-1} \) is fulfilled. The dLE1 would then give an appropriate description of the system. Of course, the preconditions for such a scenario are very restrictive.

The overdamped limit can also be achieved, if \( \langle x_{m-1} \rangle = y_{n-1} \) and \( \text{Cov}(x_{m-1}, x_{m-1}) = 0 \). The latter condition directly results from the first condition: The past of the system realization has to be fixated.

4.3.4. Different scaling of fields for small time steps

As discussed in Sec. 4.1.3, the Ito interpretation of the stochastic integral facilitates an Euler integration of the differential equation, where the Wiener process is propagated as \( dW_n = \xi_n \delta t \). Since both dLE schemes obey this interpretation, they do also recover Ito fields from the underlying data. The data is assumed to obey the discretization eq. (4.55)
4.3. The data-driven Langevin equation (dLE)

or (4.66) depending on the dLE algorithm. We can rewrite these expressions in the form

\[ v_m = h(x_n) + \frac{D(x_n)\xi_n}{\sqrt{\delta t}}, \]

(4.83)

\[ a_m = \frac{v_{m+1} - v_m}{\delta t} + \Gamma(x_m)v_m + \frac{K(x_n)\xi_n}{\sqrt{\delta t}}, \]

(4.84)

where \( \xi \) is a Gaussian-distributed random number. In this form, it is obvious, that in the limit of small propagation time steps, \( \lim_{\delta t \to 0} \), the stochastic contribution diverges in both expressions. This feature is restricting the propagation time step and furthermore the applicability of the dLE. Since it affects the time step dependency of the fields, it has to be considered in addition to the zeroth order approximation\(^{12}\).

4.3.5. Back-calculation of the noise

A useful feature of the dLE is that it has the ability to verify its own applicability in terms of the noise model. Inverting eqs. (4.74) and (4.56), one can back-calculate the contribution of the noise which results for a single propagation time step for both algorithms, exploiting the assumption that the underlying input data is propagated likewise, \( y_n \equiv x_n \). This yields

\[ \xi_{n,dLE1} = D(y_n)^{-1}\left( \frac{y_{n+1} - y_n}{\delta t} - h(y_n) \right), \]

(4.85)

\[ \xi_{n,dLE2} = K(y_n)^{-1}\left( \frac{y_{n+1} - y_n}{\delta t} - f(y_n) + \Gamma(y_n)\frac{y_n - y_{n-1}}{\delta t} \right), \]

(4.86)

respectively. Hence, we set up the propagation scheme to back-calculate the displacement which results from the noise realization from a given time series. We simply apply the dLE to a given time frame \( n \) that possesses a successor (and a predecessor in case of the dLE2) and compute the corresponding noise realization. From the resulting time series, the properties of the noise distribution can be calculated to review whether the input time series fulfills the model assumptions.

4.3.6. Spurious drift related to the dLE

As discussed in Sec. 4.1.4, the Ito interpretation of the stochastic integral introduces an additional deterministic term in case of a space dependent diffusion, if we interpret the Markovian dLE1 as a limiting case of a process which exhibits colored noise. We will delineate later that this effect will not occur for the dLE2 for sufficiently continuous fields,

\(^{12}\)i.e. the field locality assumption which is given by the Kramers-Moyal coefficients eqs. (4.46), (4.47) for the dLE1 and shown in 6.1.2 for the dLE2
see Sec. 6.1.1. However, from the perspective of the dLE1, according to eq. (4.36) the field estimate would read

\[ h_i(y_n) = h_{S,i}(y_n) + \frac{1}{2} \sum_{j,k} D_{k,j}(y_n) \frac{\partial}{\partial x_k} D_{i,j}(y_n), \]  

(4.87)

where the index \( S \) indicates the interpretation of the field \( h \) in the Stratonovich sense.

Although a misinterpretation of Stratonovich does not affect the dLE1 performance, since Ito fields are just self-consistently recovered, a back-calculation of the constituent parts is not possible. In the context of this work, this will only affect a single application of the dLE scheme, see Sec. 8.1.

4.4. Multidimensional system coordinates

The extension to a multidimensional system space is an outstanding feature of the dLE concept in contrast to other approaches [75, 125, 128]. For the one-dimensional case, one could easily derive field estimates a priori by e.g. partitioning the phase space into \( n \) bins, smoothly enough to compute reliable local fields. Therefore, on-the-fly estimation of fields as conducted by the dLE could easily be discarded in this case and the partitioning would anyway be the first step towards the derivation of an analytical field which then in turn facilitates a closed model. Yet, for \( d \) dimensions partitioning would scale like \( n^d \) which is soon not practical anymore. Besides dealing with a memory lack, many regions of the system are not populated at all and will therefore not be accessed by the dLE. This aspect makes it redundant to allocate memory to store fields which could then retrieved by the dLE.

A multidimensional system coordinate facilitates to account for dynamics which are not resolvable in a projection. Unfortunately, it also lacks locality which is required by the estimators since the volume of the system space is expanded. This behavior of the neighborhood can be easily illustrated. Firstly, I want to discuss how the locality of our neighborhood varies for the change from 1 to 2 dimensions. For this purpose, I assume a uniform density distribution in a 2-dimensional system space. A uniform distribution is a plausible assumption for the following reason: If we want to model a system applying the dLE, the bottleneck that determines locality is the highest barrier which we want to resolve on the one hand. Therefore, our locality measure has to be adopted to the sampling density in the related region. On the other hand, for a region where we already possess sufficient statistics, we do not need to consider more input frames. Since the concept of the dLE does not rely on continuous trajectories or Boltzmann distributed input data, we can systematically discard input frames. We explicitly discuss a technique which basically aims to reduce data to some threshold in the next chapter - targeting a uniform
4.4. Multidimensional system coordinates

distribution - at the example of a simple model system. It is a useful application named “data pruning” and is introduced in Sec. 5.4 in order to reduce the computational effort. For the $2d$ system, we have $\tilde{\rho}(V_1,V_2) = N/(L_1L_2) = \tilde{\rho}_0$, where $L_1$ and $L_2$ denote the length of the system coordinates $V_1$, $V_2$, respectively and $N$ the sampling size. The projection onto $V_1$ will also be uniformly distributed, $\rho(V_1) = N/L_1 = \rho_0$. If we want to be comparably local in the 2-dimensional system using the Euclidean metric, we have to demand

$$\pi \epsilon^2 \tilde{\rho}_0 = 2 \epsilon \rho_0,$$

where $\epsilon$ and $\tilde{\epsilon}$ denote the radius of an environment in 1 and 2 dimensions, respectively. In 1 dimension, the data considered is located within the distance $\epsilon$, if we use a neighborhood size $k$ and the identity $2\epsilon/L_1 = k/N$ for the uniform distribution. For 2 dimensions, the circular area with radius $\ \epsilon$ relates like $\pi \epsilon^2/(L_1L_2) = \hat{k}/N$. Altogether, this yields

$$\tilde{\epsilon} = \sqrt{\frac{2\epsilon}{\pi} L_2} = \sqrt{\frac{k}{\pi \rho_0} L_2}.$$

The locality of the neighborhood is related both to the size of the projected dimension and the neighborhood size of the projection, but does not scale linearly. For instance, if we increase the neighborhood size in the projected space by $\Delta \epsilon$, the resulting increase in the two-dimensional space will scale like $\Delta \tilde{\epsilon}^2$. On the other hand, this means that we require a drastic increase of the sampling to attain a comparable locality.

More generally, assuming the same conditions as above and a similar extent $L$ of the system space for each dimension, one can compare a $\tilde{d}$ dimensional system to a $d$ dimensional projection, $d < \tilde{d}$. This yields

$$\tilde{\epsilon} \sim \epsilon^{d/\tilde{d}} L^{1-d/\tilde{d}},$$

$$\leftrightarrow \tilde{\epsilon} \sim k^{1/d} L,$$

where we used $\epsilon^d \sim kL^d$. The last expression shows, that an increase of the statistics in the projected system of dimension $d$ basically does not affect the locality of the estimate in the $\tilde{d}$ dimensional system for comparably large neighborhood sizes. Still, the features of the 2-dimensional example remain in the generalized case.

A sample of the neighborhood is shown in Figure 4.1 at the example of the AIB9 peptide. The radius which contains a fixed neighborhood of size $k = 20$ is marked by the black circles. Neighbors were casted in the full system space which contains five dimensions for this particular system. The figure shows a projection of the FEL onto the PCs $V_1, V_2$ and the full neighborhood along with their followers for five probe regions. Therefore,

\[\text{The application of AIB9 for Langevin modeling is discussed in Sec. 6.2}\]
Figure 4.1.: Neighborhood samples of $\text{AIB}_0$ as used by the dLE1, projected onto $F(V_1, V_2)$: Shown are different extracts of the FEL projection, indicated by the white rectangles. For each of them, the point $y_n$ for which a field estimate shall be computed (black dot) is shown along with $k = 20$ local neighbors (red circles) which are considered along with their followers (red squares). The black circle indicates the projection of the five-dimensional sphere of radius $\epsilon$ which contains the neighborhood. The black square denotes the center of the neighbors. Their alignment relative to $y_n$ quantifies the isotropy or the compromise eq. (4.62).
testing points are retained from these regions which are indicated by the black dots. For example, point 1 is drawn from the center of one major transition path, while point 5 presents a realization in one of the two most stable states. Depending on the sampling of the respective environment, of course the locality of the estimates varies. In fact, locality is not provided at all for sparsely sampled regions. Exemplary for this feature in Figure 4.1 is point 4, which is located in one of the most sparsely sampled regions. Since the probability of the system to pass this region is anyway small, the bias which such an event would introduce to the overall performance of the dLE is low.

The distribution of the neighbors, indicated by the red circles, and their followers, indicated by the red squares, visualizes also the consequences of the respective neighborhood on the field estimates, at least for the dLE1: while for points 1 to 3, the average orientation of the neighborhood-displacements, denoted by the gray arrows, is rather distinct with some exceptions, this measure is rather zero for points 4 and 5. The dynamics in the latter regions are preliminarily governed by diffusion, while drift is more distinct in the first three examples. Besides, it is obvious that we still lack statistics even in the 2d projection to get reliable, converged estimates.

Figure 4.1 also shows the quality of the assumption eq. (4.62) that was invented to essentially stabilize the dLE2 numerically. The black squares denote the arithmetic mean of the neighborhood. The deviation from the sample points (the black dots) quantifies the anisotropy of the estimate. Except for point 4 which represents the most undersampled region in the system space, the approximation eq. (4.62) is rather appropriate. That implies, that the modified average results in an implementation of the dLE which is consistent with the default setup. The issue of neighborhood locality is further discussed, particularly related to the dLE2 algorithm, in Sec. 6.1.5.

4.5. Summary

In this chapter, I introduced the microscopic Langevin equation, its form in the Markovian limit and the alternative macroscopic formalism to describe the propagation of probability densities, the Fokker-Planck equation. I embedded the first order data-driven Langevin algorithm in the context of modeling systems in a data-based manner which has been shown to recover the properties of a multidimensional set of collective coordinates under certain assumptions. The algorithm is further extended towards a second order data-driven Langevin equation. In the remainder of this chapter, features of both dLE algorithms are discussed. A remarkable property of the algorithm is that it is able to check the validity of some model assumptions, namely the noise realization on the basis of the input data. This feature facilitates the dLE to at least work in a self-consistent
manner if the field estimates are converged and is further discussed and demonstrated in Ch. 6.
Since the locality assumption is fundamental to the concept of the dLE, the dimensionality of the system coordinate is limited. The last section of this chapter focused on the qualitative behavior of the locality measure as a function of the dimensionality of the system space.
Chapter 5

Application of the dLE1 to a simple model

In this chapter, the performance of the dLE1 will be further investigated. Besides the requirement, that a set of system coordinates has to match the model assumptions of the dLE1, a major bottleneck of the dLE1 is the accuracy of the local estimates. This property determines how reliable the assumption of a local drift field in (4.57) is and also determines the accuracy of the diffusion field estimate. Since we want to examine the impact of this and other properties on the overall performance of the algorithm, we rely on a reference. Without loss of generality, we therefore restrained our analysis to a one-dimensional model featuring two states [127]. Although this model lacks the challenges of a multidimensional system space, it is suited to investigate the parameters which settle the performance of the algorithm in a differentiated manner. In fact, the simple model already illustrates many generalizable properties that apply to arbitrary systems, if one considers it as an outtake of a more versatile system, representing just a set of two neighboring states which interact along a transition barrier. After introducing the double well model, a model-based ensemble of trajectories is generated which in turn provides the input for the data based algorithm. Various scenarios that influence the accuracy of the dLE are discussed for this setup. Furthermore, a modified setup of the first order algorithm is introduced, allowing for the description of a system which is governed by the second order Langevin equation. Finally, an efficient scheme for input data reduction is demonstrated.

5.1. One-dimensional double well model

In order to generate a benchmark system which shall be reproduced by the dLE algorithm, we first establish a time-series by propagating a system coordinate. The single mass-weighted system coordinate \( x = \sqrt{m}x' \), where \( x' \) is a space coordinate) is governed by the mLE1, eq. (4.23), and satisfies the fluctuation-dissipation theorem, eq. (4.27). It
5.1. One-dimensional double well model

reads

\[ \dot{x} = -\frac{1}{\gamma} \frac{d}{dx} U(x) + \sqrt{\frac{2k_B T}{\gamma}} \xi(t), \tag{5.1} \]

where the friction constant \( \gamma \) is an independent parameter in our model and \( U(x) \) describes the potential energy. We choose a simple polynomial ansatz for \( U(x) \). Defining two minima at \( x_0, x_1 \) and the barrier maximum at \( x = 0 \), we get

\[ U'(x) = \alpha x(x - x_0)(x - x_1) \]

and therefore a general potential ansatz:

\[ \Delta U(x) = \alpha \left( \frac{x^4}{4} - \frac{x_0 + x_1}{3} x^3 + \frac{x_0 x_1}{2} x^2 \right). \tag{5.2} \]

The parameter \( \alpha \) and the positions of the minima were chosen such that the potential barriers are within the typical range that we observed for biomolecular systems, \( \Delta U_{0\rightarrow1} \approx 3.5k_B T \) and \( \Delta U_{1\rightarrow0} \approx 2.5k_B T \). To accomplish this, we chose \( x_0 = -2.7, x_1 = 2.4 \) and \( \alpha = 0.28 \).

For simplicity, we strictly confine two different states, 0 for \( x < 0 \), and 1 for \( x > 0 \). A constant friction implies that the potential energy is directly proportional to the free energy. Thus, the equilibrium distributions are analytically accessible,

\[ P_0 = \frac{1}{Z} \int_{-\infty}^{0} \exp(-U(x)/k_B T) dx \tag{5.3} \]

and \( P_1 = 1 - P_0 \) with \( Z = \int_{-\infty}^{\infty} \exp(-U(x)/k_B T) dx \). For our parameters, we obtain \( P_0 \approx 2/3 \) for the population of the ground state and \( P_1 \approx 1/3 \) for the population of the excited state. The model potential is illustrated in Figure 5.1.

The friction constant is directly related to the diffusion field, \( D \sim \gamma^{-1/2} \) and sets the rate of barrier crossings. It was adjusted in order to achieve a sufficient number of transitions \( 0 \leftrightarrow 1 \). By switching the energy units\(^1\) to \( \text{ps}^{-1} \), \( \gamma \) becomes dimensionless.

We set \( \gamma = 1900 \), which corresponds to the overdamped limit as shown later. For the system temperature \( T = 300 \text{ K} \), this system of units corresponds to a thermal energy of \( k_B T \approx 38 \text{ ps}^{-1} \). Therefore, we get a diffusion constant \( D \approx 0.2 \text{ ps}^{-1/2} \).

Since the time scale of fluctuations within a single state is clearly separated from the time scale on which intrastate transitions take place, the autocorrelation function becomes mono-exponential, \( C(t) := C(x, t) \sim \exp(-t/\tau) \) (see eq. (2.20)). On the other hand, Kramers’ reaction rate theory directly relates the transition times to the potential barrier height. Eq. (2.29) yields the characteristic transition times \( \tau_{0\rightarrow1} = 1/k_{0\rightarrow1} \approx 3.9 \text{ ns} \) and \( \tau_{1\rightarrow0} \approx 1.5 \text{ ns} \). The characteristic decay time \( \tau \) of the autocorrelation function can be approximated by the average mean first passage time of both processes, \( \tau \approx 1.1 \text{ ns} \). This

\(^1\)Writing the energy in inverse wavenumbers, we obtain \( E[J] \approx 2 \cdot 10^{-23} E[\text{cm}^{-1}] \). Further relating inverse wavenumbers to inverse picoseconds yields \( E[\text{cm}^{-1}] \approx 5.31E[\text{ps}^{-1}] \).

66
result is shown in Figure 5.2 (b), compared to the autocorrelation function measured from mLE1 simulations. The insert shows deviations of the mono-exponential curve for short lag times that arise due to fast intrastate fluctuations.

5.1.1. Model trajectories

In order to generate reference mLE1 data which will be used for subsequent dLE runs, we applied an Euler-Maruyama integration of eq. (5.1) to generate ten runs of length $N_m = 50\,000$, using a time step $\delta t = 1\,$ps. Thus, we obtained an overall statistics of 500 ns. All simulations were started at $x = -1$ with different initial seeds. Ensemble averages are shown in Figure 5.2 (a), (b) compared with the analytical model. Individual probability distributions and autocorrelations are detailed in Figure 5.2 (c) , (d). Although each single run does not exhibit convergence concerning neither statistical nor dynamical properties, the ensemble average quantitatively recovers the model system. In order to compare reference mLE1 and dLE1 runs, we will always show ensemble averages and standard deviations. Furthermore, scalar quantities will be used to describe both statistical and dynamical performance. To summarize the statistical accuracy, we calculate $P_0$. This quantity is shown in Figure 5.2 (e) (blue curve) as a function of the length of individual mLE1 runs $N_m$. The error bars clearly indicate huge fluctuations in the ensemble at low statistics. Besides, the ground state is favored. This is due to the initial conditions that we specified. The average number of transitions is depicted in Figure 5.2 e by the orange, dashed line. It basically shows, that for $\gtrapprox 25$ transitions, the average population
$P_0$ is converged towards the expectation value set by our model and the standard errors remain constant. For larger statistics, there are only no major changes. If we exploit full statistics, the average population of state $0$ is $P_0 = 0.74 \pm 0.09$.

The dynamical accuracy is measured via the decay time $\tau$ assuming a mono-exponential autocorrelation decay,

$$C(\tau) = \frac{1}{e}. \quad (5.4)$$

While we observe perfect agreement for full statistics, $\tau = 1.1 \pm 0.22\text{ns}$, the trend is systematic for too few statistics, as shown in Figure 5.2 (f). For little statistics, interstate transitions do not take place at all and the autocorrelation decay reveals intrastate fluctuations. For medium statistics, the amount of transitions is not sufficient to converge the autocorrelation estimate. This causes the fluctuations of the ensemble average. Even for values $N_m > 100\,000$, where $P_0$ is already converged, $\tau$ is less robust.

## 5.2. dLE1 performance

The mLE1 runs were used as input data for subsequent dLE1 runs. To gain comparable statistics, we generated an ensemble of dLE1 trajectories, 10 runs of length $N_d = 50\,000$ points each. Every single dLE1 run uses the full mLE1 statistics as input and is initiated at $x = -1$. In order to achieve converged estimators, we figured an optimal neighborhood size of $k = 50$ for this model, as shown in Figure 5.4 (c),(d). This setup yields $P_0 = 0.69 \pm 0.05$ and $\tau = 1.02 \pm 0.12\text{ns}$. Incidentally, the spread of both statistical and dynamical quantities is less for the dLE1 ensemble than for the appropriate mLE1 ensemble while both are comparably accurate.

To evaluate the performance of the dLE1 algorithm, we now varied the parameters that influence the algorithm.

### 5.2.1. Restricting the length of dLE1 runs

The performance of the dLE1 is of course limited by the input data and hence convergence cannot arise faster than for the mLE1. Yet, reduced dLE1 sampling on the full input data is another adjustment screw. Figure 5.4 (a), (b) on page 73 show, that the dLE1 runs not necessarily have to be of similar length as the input data. The ensemble averages of $P_0$ and $\tau$ are already qualitatively converged for $N_d \gtrsim 100\,000$, which is the same minimal length as for the model runs. This simulation time corresponds to about 25 transitions as a minimum time window to qualitatively recover the system properties. Nevertheless, the standard deviations of the ensemble are rather large for dLE1 runs. For an insufficient simulation length $N_d$, the initial conditions of the dLE1 become important, which favor

---

4If we allow for non-continuous input data, this no longer proves true, see Sec. 5.4
Figure 5.2.: Langevin simulations of a 1D double well model: (a) The potential energy (orange) along with the corresponding probability distribution (black) is compared to the total probability distribution of the 8 mLE1 runs. (b) Analytical autocorrelation (black) compared with the averaged autocorrelation computed from the mLE1 runs. (c) Detailed probability distributions of individual mLE1 runs and (d) corresponding autocorrelation. (e) $P_0$ and (f) characteristic autocorrelation decay time $\tau$ as a function of overall length $N_m$ of mLE1 runs. While the blue curves indicate measured quantities on mLE1 data, the red curves report on the same quantities of subsequent dLE1 runs of length $N_d = 50000$ each. Black dotted lines indicate the correct values. The orange dotted line (e) reports the number of transitions that occur in mLE1 runs (right label).
the ground state 0. The autocorrelation decay is increased in the limit of too short dLE1 simulations since only intrastate fluctuations remain.

5.2.2. Non-locality of estimates

In order to crucially enhance the performance of the dLE1, the accuracy of the field estimates has to be modified. This is rather simple to achieve by modifying either the amount of local input data or the extent of the local neighborhood. An effect which arises in both cases is what I will refer to as non-locality of the estimates in the following. In this section, I will distinguish mLE1 and dLE1 according to the nomenclature Ch. 4 denoting the system coordinate with $x$ and $y$, respectively. It is explained in Figure 5.3 schematically for an extract of a potential curve that highlights already several features of the algorithm: The blue dot indicates a realization of the one-dimensional system coordinate $y_n$ which is propagated for a single time step by the dLE using a neighborhood size of $k = 4$. The red dots betoken the neighborhood $x_m$ of this point, taken from an mLE1 trajectory. They are sketched along with their temporal successors $x_{m+1}$, indicated by the red crosses. The average of the displacements which the neighborhood experiences determines the temporal follower of the dLE1 realization, as indicated by the blue cross. The follower that results for the modified average, namely the average of the $x_{m+1}$, is also indicated by a black cross (see Sec. 4.3.1 for the definition). Since the considered neighborhood consists of points which represent different regions along the potential surface in this exaggerated example, they will induce a bias to the estimate.

The left plot shows an example where the neighboring points to the left will shift the drift field average towards higher absolute values since they are governed by the potential in that specific region, while the neighborhood to the right will shrink the absolute of the average. Both effects will not compensate in general. If we assume a Boltzmann distributed sampling of the phase space, this non-locality of the neighborhood will further introduce an anisotropy, since the density of the neighborhood will be comparably increased in the energetically lower regions. This will virtually relocate the system center towards these regions, as indicated by the average of the neighborhood on the x-axis. As a consequence, the gradient of the potential energy surface will be shifted from the perspective of the dLE realization. This is indicated by the dotted line. In general, this feature of the dLE explains the broadening of the local minima along the potential energy landscape. It is apparent in the context of multidimensional dLE modeling, see Figure 6.6 on page 91.

Similarly to the drift estimate, the diffusion estimate is affected by non-locality, as shown in the right plot of Figure 5.3: In this illustration a realization is shown in which the resulting drift displacement is zero and we focus on the diffusive displacement. If we prepare $y_n$ on top of the potential barrier, a non-local neighborhood will increase the
Figure 5.3.: Scheme to illustrate non-locality: The blue dot represents a dLE realization of a system coordinate on a potential energy surface, which is propagated according to the local information provided by the neighborhood of red dots that represent mLE points. The mean of all mLE neighbors is marked along the x-axis. Followers of the dLE points are represented by red crosses, the resulting displacement for the dLE is also indicated by a blue cross for the default algorithm and by black crosses for the modified version.

The variance of displacements. This is, since the \( x_m \) considered are representing regions that feature large drift values. Consequently, the diffusion field will be overestimated. We encounter these effects only for very low local statistics in our simple model system. They will occur more noticeably in the next chapters, where the dLE will be applied to a system of higher complexity. For the double well model, there are two simple ways to reduce local statistics. These are explained separately in the following.

**Restricting the input data**

Firstly, we tested the behavior of the dLE1 in relation to the input data. The local density crucially influences the accuracy of the field estimates for a constant neighborhood size until we are in the local limit. By decreasing the length of the mLE1 runs even-handedly, we reduced the local statistics and also shifted the weight to the initial conditions of the input data. The resulting performance is monitored in Figure 5.2 (e), (f) by the red curve: Conspicuously, the dLE1 is imitating the bias which was introduced by finite sampling through the mLE1 input data. This is especially visible in the population \( P_0 \), where both mean and variance of the ensemble are recovered and converge likewise for increasing mLE1 statistics. The quality of the autocorrelation time \( \tau \) recovers the low sampling features of the mLE1: for very low samplings, interstate transitions occur sparsely and the estimates in the barrier range are biased towards the ground state. Thus the autocorrelation is dominated by intrastate fluctuations. The convergence of the dLE1 mostly relies on sufficient sampling of the barrier and is therefore lagging behind the mLE.
ensemble as a function of $N_m$.

**Varying the number of neighbors $k$**

The size of the neighborhood directly influences the field estimates. Due to the law of large numbers, the convergence of estimates is expected to scale as $1/\sqrt{k}$. However, opposing effects are determining the accuracy of the estimate. For large values of $k$, statistics determining the field estimates are sufficient, but the neighborhood is not necessarily local. The locality approximation eq. (4.57) for instance may not be fulfilled which introduces a bias. On the other hand, locality requires small neighborhoods especially in regions of the system coordinate which are sparsely sampled. The effect of sparse sampling is already included in the last section, where local densities are reduced by limitation of the input data $N_m$. In order to avoid the two contradicting effects mentioned above, we varied the number of neighbors, keeping the ratio to the input statistics $N_m/k = 10^4$ constant. The resulting trend is shown for variation of $k$ in Figure 5.4 (c), (d): The estimates $P_0$ and $\tau$ are converged quite rapidly at $k \approx 50$.

**5.2.3. Varying the propagation time step**

So far we only discussed the performance of the dLE1, applying the same time step on which the mLE1 input was integrated. Since the time step was selected to achieve a stable implementation of the Euler-Maruyama scheme and the dLE1 algorithm is using the propagation time step of the model, for sufficient local densities, both schemes are performing consistently. Yet in a realistic system where we do not have any knowledge about the underlying model, the choice of the integration time step of the dLE1 is not straightforward. In order to match the model assumptions eqs. (4.21) and (4.22), the propagation time step is restricted to a minimum in order to guarantee an instantaneous decay of the bath correlations. Although a larger time step would seem preferable since it reduces the overall simulation time of the dLE1, an upper bound for the integration time step is set by the discretization and implied convergence of the integration scheme. The difficulty of the latter issue arises already in our simple model system for increased time steps of the dLE1 and is shown in Figure 5.4 (e), (f): For a marginal increase of $\delta t$, both $P_0$ and $\tau$ remain nearly constant. While a permanent shift that overpopulates state 0 appears for larger time steps, the transition time firstly increases and finally decays back towards the initial value. This effect is deceptive: Although the dLE1 recovers both observables that we introduced, the model is not correctly propagating the system dynamics. We rather designed a special case where different errors compensate each other. Due to the large time step, the slope of the overall drift is underestimated. The diffusive displacement is overestimated in the very top barrier region, while it is vastly underestimated close to
Figure 5.4.: Performance of dLE1 on a 1D double well model: The quality of the dLE1 is indicated by the accuracy of $P_0$ and $\tau$, as measured from dLE1 data (red lines). The analytical benchmark is given by the black dotted lines. Varied are the length of dLE1 runs $N_d$ (a,b), number of neighbors $k$ (c,d) and the propagation time step $\delta t$ (e,f).
5.3. Handling inertial effects

The minima. Thus, the dLE1 recovers a completely different picture of the potential energy landscape on one the hand and an increased entropic barrier due to the diffusion field on the other hand. Drift and diffusion field as recovered by the dLE1 are monitored for different time steps in Figure 5.5.

Figure 5.5.: Time-step dependence of drift and diffusion fields: Shown are the drift (left) and diffusion fields (right) as recovered from a single dLE run for different time steps in ps. The values provided by the model are indicated by the black lines.

5.3. Handling inertial effects

As detailed in Sec. 4.3.1, the dLE1 is operating on a set of system coordinates which follow a Langevin description in the overdamped limit and thus obey eq. (4.23). However, more generalized systems show inertial effects. For the one-dimensional model, the appropriate description of a generalized Markovian mLE2 reads

\[
\dot{v} = -U'(x) - \gamma v + \sqrt{2k_B T \gamma} \xi(t) \quad (5.5)
\]

\[
\dot{x} = v. \quad (5.6)
\]

Thus the velocity trajectory is described by a combination of deterministic and stochastic processes. For comparably small values of \(\gamma\), the autocorrelation decay of the velocity is no longer instantaneous. Synonymously, the acceleration is not negligible.

We integrated eqs. (5.5) and (5.6) with an Euler-Maruyama scheme for different values of \(\gamma\). Notably, the simple scheme requires an adjustment of the integration time step: the discretization of eq. (5.5) requires \(\gamma \delta t < 1\). This feature effectively increases the computational effort of the mLE2 propagation for large friction values, in particular for the overdamped limit. Figure 5.6 shows the resulting local accelerations \(\langle a \rangle = \langle \dot{v} \rangle\) that we measured along the system coordinate: While \(\gamma = 100\) is still representative for the
overdamped regime as \( \langle a \rangle \) is negligible, lower values of \( \gamma \) introduce significant inertial effects. These do not affect the populations, however the dynamics of the system is altered. As shown in Figure 5.6 (b) for a friction value of \( \gamma = 40 \), a dLE1 applied on the nonoverdamped system described by the mLE2 does not recover the correct decay of the autocorrelation. A dependency \( \tau(\gamma) \) is already included in eq. (2.29). A simple modification of the dLE1 scheme however may be utilized to mimic the mLE2 behavior: Firstly, we apply an embedding of the mLE2 trajectory writing \((x_m, v_m)^T = (x_m, (x_m - x_{m-1})/\delta t)^T\) in order to synthesize the representation of eqs. (5.5), (5.6). This results in the mLE1 scheme

\[
\frac{1}{\delta t} \begin{pmatrix} x_{m+1} - x_m \\ v_{m+1} - v_m \end{pmatrix} = \begin{pmatrix} v_m \\ h(x_m, v_m) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & D(x_m, v_m) \end{pmatrix} \xi_m.
\]  

The dLE1 propagates the identical scheme and is thus able to recover the dynamics.

The performance of this embedding scheme is indicated by the green line in Figure 5.6 (b). As expected, the autocorrelation quantitatively matches the underlying mLE2 result. Embedding of the input data prior to a dLE1 run is a powerful tool to improve the performance of the dLE1 algorithm. Besides this particular application, delay-embedding is a conventional scheme to improve the accuracy of all dLE algorithms. It will be further discussed in 7.2.1 where it is applied to a genuine system.

Conveniently, the dLE1 algorithm provides a tool to test whether the underlying Markovian model is overdamped or not. A simple approach to check the validity of the dLE1 algorithm is to check the compatibility of the noise model. Since the algorithm relies on a delta-correlated noise model, a simple test of the noise model, as explained in Sec. 4.3.5, reveals the presence of inertial effects, see Figure 5.6 (c): While the dLE1 applied on embedded mLE2 data shows instantaneous decay in the noise autocorrelation, the non-prepared data set decays much slower. This can be derived analytically: If we suppose stationarity, the velocity has to vanish in the infinite past, \( \lim_{t \to -\infty} v(t) = 0 \). Integrating eq. (5.5) then yields

\[
\int_{-\infty}^{t} ds \dot{v}(s) = v(t) = \int_{-\infty}^{t} ds \exp (-\gamma (t - s)) \left(-U'(x(s)) + \sqrt{2k_B T \gamma} \xi(s)\right)
\]

\[
\Leftrightarrow \dot{x} = \int_{0}^{\infty} ds \exp (-\gamma s) \left(-U'(x(t - s)) + \sqrt{2k_B T \gamma} \xi(t - s)\right).
\]

\[
= - \int_{0}^{\infty} \exp (-\gamma s) U'(x(t - s))ds + F_s(t),
\]

where \( F_s(t) = \int_{0}^{\infty} ds \sqrt{2k_B T \gamma} \xi(t - s) \) is the stochastic force term. The first term on the right hand side of eq. (5.8) can only be described by means of a memory kernel. This can be easily seen if one compares the expression with eq. (4.13). An equivalent first order Langevin equation thus requires to be non-Markovian.
If we increase the friction, the mLE2 description of our double well model becomes more and more consistent with the overdamped mLE1 description. Therefore, the ordinary dLE1 description becomes gradually consistent with the embedded version. This is shown in Figure 5.6 (d). We finally observe the limit of strong friction for $\gamma \geq 100$. Due to computational effort that arises from the increased propagation time-step, it is advantageous to describe the model using an mLE1 beyond that limit.
5.4. Reducing the input via data pruning

So far, we only discussed low-sampling effects and their impact on the dLE1 performance. The most common bias that alters the field estimate is introduced by insufficient sampling. Besides a systematic bias, non-locality is a common consequence of this factor. While enhanced sampling can be conducted to circumvent sparse sampling, excessive sampling provides more input data to be considered in the neighborhood-search which effectively slows down the algorithm. Technically, some regions in our system space are even oversampled. Generally, these regions correspond to states that possess a high metastability and are thus sampled rather densely. Since the dLE algorithm does not rely on Boltzmann-distributed input data and continuous input trajectories, we do not rely on dispensable local sampling. In our particular model system, we oversampled the two minima, while the barrier regions are just visited sufficiently to converge the field estimates. We therefore introduce a systematic way to reduce the input data locally which we refer to as data pruning.

In order to obtain a reduced sample, we firstly divide our coordinate space into equidistant bins. For each bin we set a maximum number of frames $N_b$ of the input time-series that are recorded in this bin and discard the remaining time frames. This issue turns out to be quite challenging, since there are different ways to select subsets from a raw input data set of length $N_m$. Since the algorithm relies both on local neighbors and their temporal successors, many short trajectories actually contain huge amounts of waste, as the trajectory ends are not serviceable. The number of effectively usable data points $N_p$ gathered by data pruning may differ up to a factor of two from the overall size of reduced data. On the other hand, few continuous trajectory pieces are questionable in other respects. This is envisioned the easiest by considering the input data which contains inertial effects in Sec. 5.3: The information of trajectory points considered does no longer only depend on the associated system coordinate but also on its state of motion, since the system conserves its momentum at least partially. Hence, if we for instance only record correlated data points during the pruning procedure, we will introduce a bias to our field estimates. For our designed example, this issue may be of course circumvented by embedding. However, in general we have no knowledge about the underlying model. Therefore, the most reliable strategy for data pruning is to randomly pick trajectory pieces. To prevent effects that arise from time correlation, a trajectory can also be divided into several pieces initially. For each segment, a separate data pruning is applied then before the resulting sample is finally combined.

Figure 5.7 (a) shows the resulting frequency with respect to the Boltzmann-distributed initial sample ($N_m = 500000$) if we apply data pruning using 50 bins along the $x$-coordinate ($-4.5 < x < 4.5$). Even if we request an upper limit for the population
5.4. Reducing the input via data pruning

Figure 5.7.: Pruning of input data for the 1D double well model: If we reduce local densities for the dLE1 estimates reducing the mLE1 data (a), the performance of the dLE1 does not necessarily have to degrade. The static measure \( P_0 \) (b) and \( \tau \) (c) are shown as a function of effective input points \( N_p \) after pruning. The fraction \( N_m/N_p \) as a function of the barrier \( \Delta U_{0\rightarrow 1} \) (d) shows a speed-up of the dLE1 due to pruning (red), compared to a simple model dependency based on Kramers’ theory (see text).
Chapter 5. Application of the dLE1 to a simple model

...of single bins, the reduced sample does not show a uniform upper limit, since we have to carry along successors of the cast. The peaks at $-4 < x < -3$ and $3 < x < 4$ contain the successors of the sparsely sampled edges of the coordinate space. For values of $N_p$ that induce local densities which are comparable to the local population along the barrier or even underneath, the resulting distribution becomes nearly uniform. The dependency on the effective number of input data $N_p$ is shown for both static and dynamic quantities $P_0$ and $\tau$ in Figure 5.7 (b), (c): Both values converge quite rapidly, for $N_p \gtrsim 5 \cdot 10^4$, almost no variation is recognizable anymore. This finding confirms that the crucial point which affects overall convergence of the dLE1 is the barrier sampling. Only for lower values of $N_p$ a bias is introduced which primarily concerns the accuracy of the dynamical picture.

As mentioned above, data pruning is desirable to gain a speed-up for the dLE1 algorithm. The fraction $N_m/N_p$ of input data points and residual points after pruning quantifies this speed-up. For our particular system, this yields a value of $\approx 7$. Of course, it is hard to evaluate this rating for arbitrary systems. For our particular model, a two state system with a single barrier, we can investigate the speed-up as a function of barrier height. In order to conduct this, we varied the parameter $\alpha$ in (5.2) to increase the barrier $\Delta U_{0\rightarrow1}$ by factors of $1.25$, $1.5$, $1.75$ and $2$. For each of these modified potentials including the original one, we performed mLE1 simulations and recorded trajectories, until 50 barrier crossings were performed. Subsequently, we applied pruning to data sets until the dLE1 applied on reduced data recovered these 50 transitions within the same sampling time, respectively. From this we determined the speed-up $N_m/N_p$. The dependency is indicated in Figure 5.7 (d) as a function of the barrier height $\Delta U_{0\rightarrow1}$ by the red curve. Notably, the speed-up due to data pruning is much more proliferated for increasing barrier heights. This observation is intuitive, since the system spends disproportionally more time in the states 0 and 1 while the barrier is sampled rarely.

A simple theoretical consideration gives a rough estimate about the speed-up: the time the system needs to perform a constant and sufficient number of barrier crossings is approximately linear to the amount of data $N_m$ that we need to sample a priori and is given by eq. (2.29). On the other hand, data pruning reduces the amount of input data globally to the local density along the barrier. If we assume comparable sampling of the barrier region for all barrier heights, the reduced input data $N_p$ is thus independent of $\Delta U_{0\rightarrow1}$. The resulting relation $N_m/N_p \sim \exp(\Delta U/k_B T)$ is indicated in Figure 5.7 by the green line, where the offset is aligned with the measured speed-up for the default potential. The deviation of the prediction from the actual trend stems from the fact, that the system more frequently populates the vicinity of the transition state at $x = 0$. 

79
5.5. Summary

In this chapter, the performance of the dLE1 algorithm was discussed by means of a simple one-dimensional model which features many properties of a realistic system. We suggested two simple measures that consider statistical and dynamical properties to evaluate the accuracy of the dLE1, while varying parameters which affect the algorithm. For instance, to provoke insufficient statistics in the field estimates, we reduced the amount of input data. A variation of the neighborhood size additionally affects the convergence of the estimators. Typical effects of non-locality are illustrated. We examined the time step dependence of the fields by varying the propagation time step of the dLE1. We further showed, that the dLE1 is able to account for inertial effects. Therefore, we embedded input data that we generated from a second order mLE and simultaneously propagated two first order schemes.

Lastly, we introduced a concept which profits from the property that the dLE does not rely on continuous Boltzmann-distributed input data. What we name “data pruning” systematically removes local statistics from areas in which the local density exceeds the required amount to converge the field estimates. In the application of our double well model, we could reduce our extensive sampling to a uniform distribution, discarding about 85% of the overall input points. The resulting speed-up clearly depends on the specific system. The particular feature that the dLE does not rely on continuous input data will be further exploited in Ch. 7.
So far, I discussed the performance of the dLE1 based on a simple one-dimensional model. Although this model highlights many effects which are generalizable, it lacks features that occur only in the multidimensional case. For instance, the decomposition of the covariance matrix to determine the diffusion field comes into play. Besides, multidimensional FELs exhibit a more complex neighborhood, especially for realistic systems which feature a more versatile and rugged shape. The problem of neighborhood locality therefore plays a more important role. Due to finite sampling, there are regions in system space which may not be sufficiently sampled to guarantee reliable estimates.

In contrast to the designed model in the last part, an unknown system poses the additional challenge that it lacks a reference which enables to verify whether the right field estimates are obtained. Notably, the values of the field estimates strictly depend on the realization of the underlying stochastic differential equation. For this reason, we propagated the model similarly in the last chapter. Since the dLE will always recover fields according to an Euler discretization, it presumes the underlying model to be working in an identical manner.

In order to test the performance of the dLE, one can check whether it both fulfills the requirements of the model and recovers the measurable properties of the underlying system consistently. We show, that a realistic set of system coordinates may not be described by the dLE1 but requires the application of the more general dLE2 algorithm.

This chapter is organized as follows: Firstly, the dLE2 will be reviewed in detail. We focus on the difference that arises from the modified average eq. (4.62) and compare the algorithm with and without this assumption. Besides, the fields of the dLE2 will be briefly discussed at the example of the double well potential which is introduced in the last chapter. Before we proceed with the application to the AIB9 peptide system, it is shown how the quality of the dLE can be diagnosed on the basis of the noise model.
6.1. Characteristics of the dLE2

The dLE2 algorithm was already discussed in Sec. 4.3.2. It was derived there analogously to the first order algorithm by the discretization of the velocity for a single time step following the Euler-Maruyama scheme. A backward Euler discretization was used to introduce the system coordinates, see eq. (4.73). The justification why we did not propagate the system coordinates using a forward Euler discretization was not yet provided.

6.1.1. Forward discretization of system coordinates

In this section we explicitly consider a forward discretization of the system coordinates,

\[ w_n \delta t = y_{n+1} - y_n. \]  

(6.1)

For this case, the resulting discretization of an alternative dLE2 scheme reads

\[ y_{n+2} = y_{n+1} + \delta t^2 f(y_n) - \tilde{\Gamma}(y_n)(y_{n+1} - y_n) + \delta t^2 K(y_n)\xi_n, \]  

(6.2)

and the field estimates for our preferred algorithm yield

\[ \tilde{\Gamma}(y_n) = \text{Cov}(x_{m+2}, x_m) \text{Cov}^{-1}(x_m, x_m), \]  

(6.3)

\[ \delta t^2 f(y_n) = \langle x_{m+2} \rangle - y_n + \tilde{\Gamma}(y_n)(y_n - \langle x_m \rangle), \]  

(6.4)

\[ K(y_n)K(y_n)^T \delta t^3 = \text{Cov}(x_{m+1}, x_{m+1}) - \tilde{\Gamma}(y_n) \text{Cov}(x_{m-1}, x_m)\tilde{\Gamma}(y_n). \]  

(6.5)

It can be seen that the neighborhood information provided for the dLE2 is shifted by one time step into the future. This intervention strongly affects the field locality, since we consider followers of followers of our neighborhood. In fact, the implementation is not functional since we consider statistics that show increased fluctuations, as visualized in Figure 6.1.

Another interpretation of the discretization eq. (6.2) is obtained, if we introduce the shift \( n \rightarrow n - 1 \). It yields the field estimates

\[ \tilde{\Gamma}(y_{n-1}) = \tilde{\Gamma}(y_n), \quad \tilde{f}(y_{n-1}) = f(y_n), \quad \tilde{K}(y_{n-1}) = K(y_n), \]

denoted by the “hat” in relation to the dLE2 field estimates introduced on page 56. The only difference is the time frame which we assign to our fields. They are all scheduled one time step prior to the current dLE2 frame. For sufficiently small time steps or constant fields, both interpretations coincide. This observation is equivalent with the statement that the dLE2 velocities \( \mathbf{w} \) are continuous. Therefore all interpretations of the integral
Figure 6.1.: Different implementations of system coordinates for the dLE2: Field estimates for two different dLE2 implementations at a single point (black dot) are shown along with $k = 10$ local neighbors (red circles) which are considered along with their followers (red squares). The black circle indicates the projection of the five-dimensional sphere of radius $\epsilon$ which contains the neighborhood. In contrast to the first order algorithm (see page 61), an additional point of the neighborhood time series is considered (blue triangle) while a forward Euler discretization of the dLE2 coordinate $y$ (a) introduces an increased non-locality to the estimators in eqs. (6.3), (6.4) and (6.5) since followers of followers are considered, the preferential backward discretization introduced in Sec. 4.3.2 considers both followers and predecessors (b).

$\int_{t}^{t+\delta t} dt' \mu(t') \quad \text{coincide. Consequently the Ito shift is not present in the dLE2} \ [118, 119]$. The physical interpretation which arises from this misalignment for finite time steps is not relevant for the dLE2 algorithm since we do not explicitly use the time step information of the fields.

### 6.1.2. Zeroth order estimates

So far, we did not show whether and to what extent the non-locality of fields affects the accuracy of the estimators defined in Sec. 4.3.2. We now assume that fields are locally varying such that we can describe these variations by a Taylor expansion around the dLE2 fields for each individual neighborhood point $x_m$. The drift field for example then reads

$$f(x_m) = f(y_n) + (\Delta_m \nabla) f(y_n) + O(\Delta_m^2), \quad (6.6)$$

where we expand with respect to the distance

$$\Delta_m = x_m - y_n. \quad (6.7)$$
If we apply this expansion to all fields, insert it into the mLE2 propagation scheme eq. (4.66) and neglect contributions of second order, we get

\[
v_{m+1} = f(y_n)\delta t - \tilde{\Gamma}(y_n)v_m + K(y_n)\xi_m \delta t \\
+ \left(\Delta_m \nabla f(y_n)\right) \delta t - \left(\Delta_n \nabla \tilde{\Gamma}(y_n)\right)v_m \delta t + \left(\Delta_m \nabla K(y_n)\right)\xi_m \delta t.
\] (6.8)

This expression can now be used to proceed similarly to Sec. 4.3.2. Besides the usual model assumptions, we also explicitly use that both system coordinates and velocities are statistically independent of the noise variable, namely \(\langle x_m \xi_n \rangle = 0\) and \(\langle v_m \xi_n \rangle = 0\).

Averaging over eq. (6.8) yields an expression for the dLE2 drift field,

\[
f(y_n) = \langle v_{m+1} \rangle + \tilde{\Gamma}(y_n)\langle v_m \rangle \\
- \left(\left(\Delta_m \nabla \right) f(y_n)\right) \delta t + \langle F(\tilde{\Gamma}) \rangle.
\] (6.9)

While the first line in this expression resembles the dLE2 estimate derived in Sec. 4.3.2, we obtain two additional terms each of which alters the estimator differently: The first term in the second line is - apart from the locally varying gradient - dominated by the mean distance of the neighborhood points from the dLE2 realization, \(\langle \Delta_m \rangle = \langle x_m \rangle - y_n\).

The second term is proportional to the expression \(\langle \Delta_m v_m^T \rangle = \langle (x_m - y_n)(x_m - x_{m-1})^T \rangle\) which most likely vanishes if neighborhood locality is provided, \(x_m \rightarrow y_n\). Analogously, expressions for friction and diffusion fields can be derived:

\[
\tilde{\Gamma}(y_n) = - \text{Cov}(v_{m+1}, v_m) \text{Cov}^{-1}(v_m, v_m) \\
+ \left[\text{Cov}(F(f), v_m) - \text{Cov}(F(\tilde{\Gamma}), v_m)\right] \text{Cov}^{-1}(v_m, v_m) \] (6.10)

\[
K(y_n)K^T(y_n) = \text{Cov}(v_{m+1}, v_{m+1}) - \tilde{\Gamma}(y_n)\text{Cov}(v_m, v_m)\tilde{\Gamma}^T(y_n) \\
+ \sum_{i \in f} \left(\text{Cov}(F(i), v_m)\tilde{\Gamma}^T_n - \tilde{\Gamma}_n \text{Cov}(v_m, F(i))\right) \\
+ \text{Cov}(F(K), \xi_m)\tilde{\Gamma}^T_n - \tilde{\Gamma}_n \text{Cov}(\xi_m, F(K)).
\] (6.11)

Exploiting the noise properties, we get \(\text{Cov}(F(K), \xi_m) = (\langle \Delta_m \rangle \nabla) K(y_n)\) for the last line. Thus, at least this contribution is also influenced by the mean distance of the neighborhood. Yet, in order to have vanishing contributions due to derivatives of the fields, we have to demand that derivatives are negligible in general. In summary, we at least rely on sufficiently constant fields within a neighborhood to guarantee a quantitative match of the dLE2 performance.
6.1.3. Fields

We used the double well model from Sec. 5.1 to qualitatively recover fields by means of the dLE2. Therefore, were generated input data by using an mLE2. For this purpose, the mLE2 is implemented similarly to the dLE2 (eq. (4.74)), using a propagation time step of $\delta t = 0.001$ ps. In order to sample our model in the nonoverdamped regime, we set the friction constant $\gamma = 30$ and applied the fluctuation-dissipation theorem eq. (4.26) to obtain $K = 60/k_B T_0$.

In contrast to the first order algorithm, we require an increased neighborhood size to converge the estimates and thus also an increased local density. We used data pruning (Sec. 5.4) to generate about 5.6 million convertible, almost uniformly distributed data points as dLE2 input. The optimal dLE2 configuration for this data set is using a neighborhood size of $k = 300$. All fields are shown in Figure 6.2, each compared to the reference. We computed local averages for the field estimates, using 100 bins (blue dots) in comparison with the specification (black line). The neighborhood size is $k = 300$ except for (b), where we used $k = 50$. The green bars indicate the standard error.

Figure 6.2.: Field estimates of the dLE2 applied to a 1D double well: Shown are the estimates derived by the dLE2 from an underlying mLE2 time series, discussed in Sec. 6.1.3. Drift $f$ (a,b), friction $\Gamma$ (c) and diffusion $K$ (d) fields as measured by the dLE2 are indicated as averaged quantities by the blue dots (using 100 bins) in comparison with the specification (black line). The neighborhood size is $k = 300$ except for (b), where we used $k = 50$. The green bars indicate the standard error.

---

$^1$ where units are chosen according to Sec. 5.1 and $T_0 = 300$ K
points) and also show the standard errors for each bin (green bars). The latter indicate that an on-the-fly calculation of field estimates is rather unreliable at least for the determination of drift and friction fields. Comparing Figure 6.2 (a) and (b) demonstrates, that an increased neighborhood size decreases the standard errors if we are still local, as we would expect from the law of large numbers. For \( k = 300 \), the fluctuations of the field estimate do not affect the performance of the dLE2 for the following reasons. On the one hand, these fluctuations are exceeded by stochastic fluctuations. On the other hand, they disappear in the averages since local events appear frequently for a converged dLE2 run. Besides, the deviations of the friction field estimate do not occur explicitly in the algorithm, where the rescaled and shifted friction field \( \tilde{\Gamma}(x_n) = \delta t \Gamma(x_n) - I \) is applied. They are amplified by the time step.

### 6.1.4. Noise model

The moments defining the noise model are shown in Figure 6.3 (a), where we applied the dLE2-feature to test the forecast, see Sec. 4.3.5. Mean and variance of the noise variable are calculated similarly to the fields in a binned manner. They are consistent with the mLE2 as expected for \( k = 300 \) (blue lines).

If we increase the neighborhood size and therefore provoke nonlocal estimates, we can observe that the noise model assumptions are no longer fulfilled. This is exemplified for the neighborhood sizes \( k = 50, 900 \) and 2000. While we observe \( \langle \xi \rangle \approx 0 \) even if we provide few statistics for the estimators, the noise variance significantly deviates from the required value if the neighborhood is no longer local. Strikingly, the accuracy of the noise model is strictly correlated with the field estimates. This is shown explicitly for the friction field estimate are shown in (b).

![Figure 6.3: Noise model of the dLE2 applied to the 1D double well model: (a) shows mean \( \langle \xi \rangle \) and variance \( \langle \xi^2 \rangle - \langle \xi \rangle^2 \) of the noise, locally averaged using 100 bins for the neighborhood sizes \( k = 50 \) (red), 300 (blue), 900 (green) and 2000 (orange). The respective averages for the friction field estimate are shown in (b).](image-url)
in Figure 6.3 (b). While we recover the fields of the model for local estimates, a narrower noise distribution coincides with an increased friction field estimate.

6.1.5. Different implementations

In Sec. 4.3.2, we already discussed the dLE2 algorithm including the modification $\langle x_m \rangle = y_n$. It has the advantage to not suffer from small differences and can thus be numerically more stable. Yet, it is more restrictive than the other implementation. It is illustrative to consider the Kramers-Moyal coefficients eq. (4.46), (4.47) again to clarify the differences of both dLE2 implementations. For example, the second coefficient which defines the diffusion field reads

$$D^2(y) = \lim_{\tau \to 0} \frac{1}{\tau} \langle (x(t + \tau) - x(t)) (x(t + \tau) - x(t)) \rangle |_{x(t) \to y(t)},$$

(6.12)

for a one-dimensional system. Notably, the timescale on which the limit of $\tau$ is carried out still has to exceed the timescale of the bath fluctuations, $\tau \to \delta t$. Every realization $x$ within our neighborhood has to be representative for the dynamics in $y$. The average is evaluated at the sharp value $y(t)$ after the limit is applied. Therefore, the modified average introduces an additional constraint.

The difference of both algorithms is contained in eq. (6.7). We can write

$$\delta t v_m = x_m - x_{m-1} = y_n - x_{m-1} + \Delta_m$$

(6.13)

$$\delta t v_{m+1} = x_{m+1} - x_{m} = -y_n + x_{m+1} - \Delta_m$$

(6.14)

to express the velocity as a convergent expansion in $\Delta_m$. For negligible $\Delta_m$, both conventions coincide.\(^2\)

The differences of both algorithms can be monitored for the one-dimensional double well model: Since we derived field estimates for our constant fields from pruned input data, the deviations are globally constant as well. This was already shown in Figure 6.3. We can compute the global average to summarize the field estimates in a single quantity. The result is shown for the variance of the noise as a function of the neighborhood size $k$ in Figure 6.4 (a). While for small values of $k$, we observe deviations due to convergence of the estimators, for large values the non-locality causes the modified algorithm to collapse. The other algorithm is not sensitive to this behavior.

\(^2\)Applying the expressions eqs. (6.13),(6.14) to the velocity-dependent dLE2 field estimates eqs. (4.68),(4.69), (4.71) directly yields the system coordinate-dependent expressions eqs. (4.78),(4.79), (4.80).
6.2. Modeling AIB$_9$ in chloroform

After discussing the dLE2 algorithm in more detail, I will dedicate the remainder of this chapter to an application. The peptide system introduced in Ch. 3, solvated in chloroform, is perfectly suited to demonstrate the dLE2 performance [129]. We decided to apply the dLE2 in chloroform solute to the MD input data generated at $T = 320$ K. In comparison to the runs at lower temperature we have an improved sampling of the essential conformational states. In contrast to higher temperatures we have an enhanced resolution of the FEL profile.

The deviation of both algorithms can be quantified by a velocity locality measure,

$$
\mu = \left( \frac{\langle (\delta t v_m + \Delta_m)^2 \rangle}{\langle \Delta_m^2 \rangle} \right)^{-1} = \frac{\langle (x_m - y_n)^2 \rangle}{\langle (x_{m+1} - y_n)^2 \rangle}.
$$

(6.15)

If the difference $\Delta_m$ is negligible, this expression vanishes. Thus, for $\mu \ll 1$ both algorithms perform likewise. For the double well system, the velocity locality measure can be varied by altering the neighborhood size. Aligning it with the averaged variance of the noise estimate, one can demonstrate the influence of this quantity on the testable model assumption. This is illustrated in Figure 6.4 (b). It shows the averaged noise variance as a function of the velocity locality. The deviations for a small neighborhood size is again caused by the insufficient convergence of the estimates. For values of $10^{-4} \lesssim \mu \lesssim 10^{-3}$, we observe a quantitative agreement of both algorithms. However, if the neighborhood locality is violated the dLE2 performance of the algorithm that uses the modified average is corrupted.

6.2. Modeling AIB$_9$ in chloroform

Figure 6.4.: Comparison of different dLE2 algorithms: Shown are globally averaged variances of the noise estimate as a function the neighborhood size $k$ (a) and the velocity locality $\mu$ (eq. (6.15)) (b) for the dLE2 algorithm with (blue triangles) and without (red squares) the assumption $\langle x_m \rangle = y_n$.
6.2.1. Preparing the input data

To recapitulate, the overall input data provided includes eight trajectories of 2 µs length each, yielding altogether $16 \cdot 10^6$ input points since we used a time step of $\delta t = 1$ ps. A proper set of system coordinates is established by a dPCA. The autocorrelation reveals a clear time scale separation of the PCs 1 to 5, as shown on page 24. The fastest transitions that we want to resolve qualitatively along the FEL of the five system coordinates are the $l \leftrightarrow r$ transitions which occur on a timescale of about 50 ns. Still, in order to faithfully recover the barriers, we have to propagate our system on a timescale that also resolves the $0 \leftrightarrow 1$ dynamics which occur merely two orders of magnitude more frequently. It was figured, that the time step $\delta t = 1$ ps suited best. This will be justified in Sec. 6.2.3.

As already noted in Sec. 3.1.2, AIB$_9$ features predominantly two pathways which are arranged at the edges of the FEL projection onto the first two principal components $V_1$, $V_2$. On the other hand, the majority of states is rather sampled, see Figure 6.5: while right-handed and left-handed helical conformation occupy about 80% of the overall statistics, the eight states which account for the most probable transition paths are already sampled an order of magnitude less frequently. Few uncommon states are sampled even another two orders of magnitude less frequently. Looking at the transitions, we see that some of these states are accessed only once in a single trajectory and are comparably stable thereafter. Therefore, we lack the statistical evidence whether they are exceeded at all on the time scale of our reference data. What is more serious is the fact that we lack statistics to get reliable estimates for the associated transition states.

Restricting and pruning

Regions which show insufficient sampling are able to easily introduce a bias to the field estimates. Moreover, the corresponding transitions are not significant at all. Therefore, we removed them from the input data along with the transition trajectory which connects the associated core region with the next more stable state. An efficient way to conduct this is to use the data pruning procedure introduced in Sec. 5.4 as follows: The particular negligible states are resolved in the central region of the FEL, projected onto the first two principal components $V_1$, $V_2$. We are binning relatively to the projection along $V_1$, $V_2$ and randomly pick points from the initial input data to fill the bins up to some threshold. Particularly, $50 \times 50$ bins are used to establish the input data for the dLE while the maximum threshold per bin is set to 20000 points. This upper limit affects only the most stable states since the sampling along the broad barrier region is already extenuated. The amount of input data which will be considered by the dLE is reduced to about 1 million points in this way.
6.2. Modeling AIB₉ in chloroform

Figure 6.5.: Sampling of AIB₉ in chloroform at 320K: The different product states which occur are aligned as in the projection of the FEL along $V_1, V_2$. The node sizes correspond to the population of different states, while the thickness and opacity of the arrows scale with the number of transitions between different nodes. They are labeled for the major transition paths.

6.2.2. dLE2 performance

We preferred the dLE2 algorithm that uses the modified average to compute the field estimates since it showed the better performance. In particular, the autocorrelation decay along the system coordinates was closer to the MD reference. Besides, the velocity locality allows to apply this version of the algorithm, since the displacement in a single time step exceeds the average distance to the neighborhood by far. In order to recover the performance of the underlying MD, a comparable ensemble of dLE trajectories is generated: Both the original dLE1 and the dLE2 where applied to the reduced input data to generate 10 different dLE trajectories of 2µs length each. Half of these were initiated in state $L$ while the other half was initiated in $R$. The field estimates were figured to converge and still be reasonably local for a neighborhood size of $k = 300$ points.

The two-dimensional projection of the FEL along the first PCs is shown in Figure 6.6. It clarifies that the dLE2 qualitatively reproduces the structural information of the major metastable states while the dLE1 fails to recover the structure of the projection. The broadening of states is a typical feature of the dLE, explained in Sec. 5.2.2. Compared to the MD reference, the accuracy of the dLE performance is shown in Figure 6.7 in terms of autocorrelation and FEL projection along the first PCs $V_1$ to $V_3$, respectively. It reveals

---

3 using the field estimates eqs. (4.58) and (4.58)
that compared to the reference which is indicated by the black lines the dLE2 recovers both dynamics and statistics far better than the dLE1. This finding is most distinct for the first PC: While the dLE2 autocorrelation decay lies within the standard error of the MD reference, the dLE1 overestimates the dynamics along $V_1$ by one order of magnitude. Likewise, the barrier height of the FEL projection is underestimated along $V_1$. These two effects are correlated. They are caused by an increased diffusion estimate in the barrier region. Besides a broadening of the distribution of the metastable states, the transition barriers are systematically underestimated by the dLE1. The dLE1 also overestimates the autocorrelation decay by an order of magnitude for higher PCs, while the deviation of FEL projections is not that conspicuous for higher components, as shown in Figure 6.8. Nevertheless, the small error bars indicate a systematic bias. Compared to the dLE2, the dLE1 shows less resolution of the ruggedness of FEL projections.

The dLE2 performance nicely reproduces the dynamics of the MD reference for the first two PCs and only slightly deviates for $V_3$. The latter finding is also caused by rare events of the MD ensemble which do not access since the input data does not provide the transition anymore: $V_3 \approx -1.2$ corresponds to a state which is rather stable but rarely accessed during the transition $L \leftrightarrow R$. Since the dLE2 misses this transition, it also deviates on the corresponding time scale as reflected in the autocorrelation.

We have to keep in mind, that the overall sampling of the system space as represented by the MD data, is still not converged: for example, the symmetry $F(V_1) = F(-V_1)$ which arises from the achirality of the AIB$_9$ is not apparent in Figure 6.7 (b). The projection
Figure 6.7.: Comparison of different dLE algorithms: Autocorrelations (left column) and projections of the FEL (right column) onto PCs 1 – 3 (rows) are shown for dLE1 (green) and dLE2 (red) compared to the MD reference (black). Bars account for the standard error.
onto $V_2$ that resolves the sampling of the two major transition pathways also lacks symmetry. Comparably large error bars of the reference indicate this statistical bias: the autocorrelation estimate along the first PC for instance deviates from the reference for large lag times due to the sparsely visited states along the transition barrier.

**Accuracy for single transitions**

Since both autocorrelations and FEL projections along single PCs are highly averaged quantities, they might easily mantle crucial deviations of the dynamics for single transitions. This property can easily entail a wrong interpretation of the system dynamics. Yet, we can identify the different states of our dLE trajectory by applying the relative clustering method suggested in Sec. 2.5.2 and deriving properties of particular states in order to compare dLE and MD performance more quantitatively.

The metastability of a single state for a given lag time denotes the probability of the system remaining in this particular state. Figure 6.9 shows this self-transition probability for a lag time $\tau = 1$ ns. As mentioned above, the reference remains stable for the major states, which reflects in the black bars: While $L$ and $R$ feature a metastability of $T_{ii} \approx 1$, states which are entered along a sequence of transitions, for example along the two major transition paths, are less stable. The asymmetry of the sampling is also recognizable in
6.2. Modeling AlB₉ in chloroform

this transition.

A comparison of the different dLE algorithms with the MD benchmark shows, that the
dLE2 generally samples much more stable states and is therefore closer to the reference. Particularly, the stability of the ground states is already underestimated by the dLE1 while the stability of the remaining comparably stable states is systematically underestimated at least by an order of magnitude.

Although the dLE2 also in many cases underestimates the metastability, the deviations are significantly lower. The state llrrrl is for example not stable at all for the dLE1 on the given lag time but it appears stable for the dLE2. Occasionally, even the dLE2 is quantitatively far off the reference, e.g. the state rrrlr appears stable for both dLE algorithms.

We have to recall that the sampling provided by the MD is not converged and causes a bias to the estimators.

Complementary to the metastability, the transition rates \( k_{ij} \) that express the number of transitions from states \( i \) to \( j \) for a given lag time are a measure to quantify the connectivity of different states. They are plotted for transitions along the major pathways in Figure 6.10, using a lag time of \( \tau = 500 \) ps. It is shown, that the transition probability of the MD reference is commonly lower than predicted by both dLE algorithms, especially for relaxations to the ground state in terms of absolute deviations. Since the latter are related to the transition states which are best sampled, this finding reasonably favors the dLE2 algorithm. While the dLE1 is not able to even qualitatively reproduce the performance of the reference, the dLE2 functions convincingly better also for less frequent transitions. The most important transitions which predominantly account for the transition path time along the barrier are in good agreement, such as rrrll \( \rightarrow \) rrrrl.
6.2.3. Justification of the dLE2

Although I presented many indicators which show that a dLE2 propagation of the system coordinates of AIB\textsubscript{9} is preferable, I did not argue why the dLE1 fails so far. To recall, it is based on a restricted form of the LE which does not consider inertial effects. The latter are reflected amongst others by local accelerations,

\[ \ddot{x}_i = \frac{\langle x_{i,m+1} + x_{i,m-1} - 2x_{i,m} \rangle}{2\delta t(f_i)} \]  

(6.16)

In this expression, \( \langle . \rangle \) is a local average. \( f_i \) denotes the \( i \)-th component of the gradient of the free energy, \( f_i = \partial F(x)/\partial x_i \), so that \( \ddot{x}_i \) is measured in units of an average force. Therefore, \( \ddot{x}_i \) gives a reasonable estimate for the influence of an external force onto the system coordinates. The first component \( \ddot{x}_1 \) is shown in Figure 6.11 (a), projected onto the first PC. It illustrates, that the acceleration is at least locally not negligible. Notably, this measure does not rely on any model since it is directly computed from the MD trajectory.

The effective friction field \( \tilde{\Gamma}(x_n) := \delta t\Gamma(x_n) - I \) which is estimated to propagate the dLE2 also represents a quantity that indicates whether we rely on a second order model: As detailed in Sec. 4.3.3, for \( \tilde{\Gamma} = 0 \) the dLE2 algorithm reduces to the overdamped dLE1. Considering the projection onto \( V_1 \), Figure 6.11 (b) shows, that \( \tilde{\Gamma}_{11} \) significantly deviates from zero, most distinctly in the barrier region where local energies are comparably high.
Since the dLE2 recovers the system dynamics qualitatively and self-consistently, this finding shows that the dLE1 assumptions are not fulfilled. Although both aspects support the predication that the dLE2 is more applicable than the dLE1, the model assumptions still have to be fulfilled in order to guarantee that the dLE2 is functioning. Most notably, the noise realization has to match the model assumptions. The noise distribution has to be of zero mean and unit variance. The first property is fulfilled for both dLE1 and dLE2, while the variance $\langle \xi^2 \rangle - \langle \xi \rangle^2 \approx \langle \xi^2 \rangle$ deviates for both models, as shown in Figure 6.11 (c). While we recover the required noise model for the dLE2, the dLE1 strictly underestimates the noise variance. In order to obey a Markovian Langevin equation, the noise has to be uncorrelated. As carried out in Ch. 4.3, both dLE1 and dLE2 explicitly rely on the property $\langle \xi_i(\delta t)\xi_j(0) \rangle = 0$. This property holds for
both algorithms for all $i \neq j$. For $i = j$, the result is shown in Figure 6.11 d: While the dLE1 even shows an anti-correlated behavior which is decreasing in absolute value for higher system coordinates, the dLE2 fairly aligns to the required property. This indicator is the most striking evidence that the dLE1 is not applicable. The explanation for the misbehavior is rather simple: the dLE1 is not able to conserve the momentum of the underlying data but misinterprets the data instead, blaming this feature to the noise. In particular, frequent oscillations within the major populated states are a deficiency of the dLE: it permanently switches the direction since it is not able to consider inertial effects, which is mirrored in the anti-correlated noise model.

6.3. Conclusion

In this chapter, we extensively discussed the features and capabilities of the dLE2 algorithm which was introduced in Ch. 4. At first, the actual implementation is justified. It is shown, that the algorithm is consistent with a second order model LE in the limit of sufficiently constant local fields. We applied it to the simple one-dimensional double well system which was delineated in the last chapter and showed, that it relies on enhanced statistics compared to the dLE1 to provide reliable results. In order to diagnose whether the dLE2 aligns with the underlying model, we showed and discussed how conclusions about the model can be drawn from the properties of the noise variable. We also compared our two different implementations of the algorithm and illustrated the differences in the context of the double well model.

The remainder of this chapter was focusing on the application of the dLE2 algorithm to a biomolecular system, the AIB$_9$ peptide which we discussed in Ch. 3. We showed, that in the application to the system, particularly exposed to chloroform solvate, the overdamped dLE1 algorithm does not satisfy the requirements of the system. This finding could be attributed to inertia effects which we could directly measure from the MD data. It should be noted that the functioning of the dLE2 does not rely on the source of these inertia effects, for instance, if they are not inherent to the full system but arise from dimensionality reduction and projection towards the reduced system space. The dLE2 algorithm was shown to fulfill the model assumptions and consider the addressed effects to superiorly reproduce both statistics and dynamics of the system. It represents a significant generalization to describe the dynamics of a more complex system.
So far, the dLE algorithms to model the behavior of a reduced set of system coordinates have been introduced. Two different dLE schemes were presented and compared both for a well-established model scheme and a small peptide system. It was shown that the performance crucially depends on local statistics and the conformity of the system coordinates to the particular model assumptions. Nevertheless, so far we only reproduced the underlying data.

A promising purpose of the dLE is based on the property, that it does not rely on continuous input data. The concept of data pruning was already discussed in Sec. 5.4, where the performance of the dLE1 is shown to not be affected at all by the distribution of the input data. A reduction of the input data does not affect the local field estimates as long as the local statistics are sufficient everywhere in system space. Although this scheme introduces a speed-up to the data-driven algorithm, it relies on extensive MD simulations covering the entire accessible state space to basically reproduce the picture of the underlying data at most. Contrary to the functioning of data pruning, another approach will be introduced in this chapter which does not rely on dynamical information about the whole state space generated by a continuous MD run [110]. We suggest an enhanced protocol that is based on a combination of a well-known sampling method [130] and conventional MD simulations. It will be introduced and discussed by the example of the AIB9 peptide system. The performance of this protocol will be evaluated in comparison with a conventional setup, meaning dLE based on long continuous MD runs.

### 7.1. Enhanced sampling methods

MD simulations are limited to millisecond time scales even for small protein systems due to limited computational power. However, conformational dynamics of interest occur on larger time scales that are not accessible with conventional MD simulations. Hence, many recent works focus on enhanced sampling techniques [131, 132]. While some of these techniques only explore the conformational space at the cost of dynamical information
or introduce at least a bias to the dynamics [133–135], other more elaborate schemes are able to regain dynamical properties of the underlying system [58]. Most enhanced sampling protocols work on a reduced conformational space and are therefore crucially dependent on the selected subset of collective coordinates. Many of these methods are stochastically driven and introduce a bias to the potential energy surface on which they are applied [136–138]. One of the widely used schemes - Replica-exchange [18, 139, 140] - exploits correlation of dynamics and thermal energy: An ensemble of realizations of the same system at different temperatures is simulated in parallel. A Metropolis criteria is then applied to interchange the different runs without perturbing the statistical ensemble. Since the realizations at higher temperatures are able to overcome potential barriers more rapidly, this setup allows to accelerate the overall system dynamics. Variations of this particular concept are dealing with other ensembles but operate in a similar manner [141–143]. Although methods like the above mentioned seem promising since they significantly decrease the overall simulation time, they still require huge resources as several simulations have to be performed.

### 7.2. dLE based enhanced sampling

In order to impose an enhanced sampling scheme that recovers the dynamical connectivity on the basis of the dLE, we decided to first of all apply a simple tool which is capable of generating stable conformations named CONCOORD\(^1\) [130]. It features an algorithm which explores the conformational space of peptide or protein systems by varying internal coordinates like angles. Therefore, geometric restraints are considered. Hence, it already operates in a reduced conformational space and is able to traverse barriers that are geometrically accessible.

Once we applied CONCOORD to generate different stable structures, these structures are in turn used to run short MD simulations. The basic idea of this approach is that the resulting ensemble of short MD simulations sufficiently covers the accessible conformational space to guarantee accurate field estimates for a subsequent dLE run. In fact, the term “sufficiently” is not well defined for an unknown system. Apart from converged estimators and conformity of the noise model (eqs. (4.21),(4.22)), it is hard to figure out, whether a statistical bias is introduced since the dLE algorithm simply adopts this bias. The dLE is at best able to interpolate the local information and establish a dynamical model. In Figure 7.1, the CONCOORD-supported protocol is indicated in a rough schematic picture by green arrows. In order to evaluate it, the protocol was applied to our small peptide system AIB\(_9\) solvated in water. Alongside, we also applied the standard

\(^1\)from CONstraints to COORDinates
dLE protocol, using eight long continuous MD simulations directly as input for subsequent dLE runs. This benchmark is indicated by the blue arrow in Figure 7.1. A compromise between both schemes is also considered, indicated by the red arrows: From a long continuous MD run, several frames at different positions in the conformational space are used as starting structures for independent short MD runs which in turn are then used as dLE input. These different workflows are discussed and compared extensively in the following.

![Figure 7.1: dLE based Enhanced Sampling scheme](image)

Figure 7.1: dLE based Enhanced Sampling scheme: Our standard dLE protocol uses 8 trajectories generated by independent long continuous MD simulations directly as input for a subsequent dLE run as indicated by the blue arrow, the enhanced sampling is indicated by green and an intermediate protocol by red arrows. The different workflows are discussed in more detail in Sec. 7.2.

### 7.2.1. Delay embedding

In order to prepare any input data for the dLE, a time delayed embedding was applied to temporally connected time frames. This practice is well known in time series analysis [144, 145] and works applied prior to a dPCA similarly to a low pass filter for our purposes. Given a set of discretized system coordinates \( x_n = \{x_{i,n}\} \), where \( n \) denotes the time step, a \( d \) dimensional delayed embedding simply yields

\[
\tilde{x}_n = \left( x_n^T, x_{n-1}^T, \ldots, x_{n-d}^T \right)^T.
\]  

(7.1)

Thus, delay embedding effectively increases the dimensionality of the system by a factor of \( d + 1 \). From a physical perspective, it introduces information about the system’s past to each time step, which is in turn considered by a subsequent coordinate transformation.
In the concrete application to AIB₉, we apply it directly to the central dihedral angles along the peptide backbone and further process the data utilizing the dPCA. The resulting measures for both autocorrelations and FELs are shown in Figure 7.2 (a). Short trajectory pieces for both the raw dPCA data and the delay embedded version with embedding dimension \( d = 3 \) are demonstrated. The feature introduced by low pass filtering is clearly visible in this representation. It is demonstrated quantitatively in Figure 7.2 (b) in terms of the autocorrelation decay within the time lag \( \delta t = 10 \) ps: While for the raw data the autocorrelation decay of the fourth PC indicates the presence of uncorrelated noise, embedding degrades this feature. Up to an embedding dimension of \( d = 3 \), the system coordinates which we use for a further dLE processing are enhanced inasmuch as the highfrequent oscillations along the trajectory are reduced.

Since the dPCA exhibits the feature that the maximum range \( 2R \) of a collective coordinate is delimited as a function of the number \( N \) of dihedral angels, \( R \leq \sqrt{N} \), delay embedding so to speak stretches the spread along a principal component. In the particular case of AIB₉ the first PC remains stable for any delay embedding dimension. Hence, its relative range \( R_1(d)/R \) remains constant for any embedding dimension \( d \). The first PC therefore
relates to the dPCA without embedding following

$$R_1(d) = \frac{R_1(0)}{d+1}. \quad (7.2)$$

The FELs in Figure 7.2 (c),(d) are rescaled so that they align with the raw dPCA. Along the first component delay embedding results in a stabilization of the different intermediate states along the potential barrier. The absence of noise locally decreases the barrier. This information is hidden in the autocorrelation decay which mostly reflects events within the two most stable states. Along the fourth component, a different set of intermediate states evolves at the expense of other states. Furthermore, the most stable state is more distinct for high embedding dimensions.

7.2.2. MD reference

Similarly to the system in chloroform, extensive MD simulations were performed as a basis for further processing. 8 MD simulations of the system at 300K were conducted that were each of a length of 4 $\mu$s. Both as a simple test and in order to optimize the performance of the dLE1 setup, the algorithm is first applied to the full MD data. The structure of the system containing five PCs resembles that of the system in chloroform, see 3.1.1. Therefore the same criteria for the choice of the system size are suitable in this context. In contrast to the peptide system in chloroform, we do not observe the metastable intermediate state for AIB$_9$ in water, see page 24. Due to the delay embedding, the spread of the FEL along the system coordinates is stretched as mentioned above: The most stable conformations $L$ and $R$ appear at $V_1 \approx -5$ and $V_1 \approx 5$. Compared to the chloroform solvent, the longest time scale of the dynamics is amplified as reflected in the autocorrelation decay, $C_1(\tau) = 1/e$. We observe a characteristic decay time of $\tau \approx 290$ ns in average for the MD data, while the comparable scheme for chloroform at 300K highlights a decay time of $\tau \approx 135$ns, which qualifies the doubled simulation time. It is observed, that the system spends $\approx 80\%$ of the simulation time in either of the two most stable conformations, $L$ or $R$, which is in accordance with the peptide solvated in chloroform.

7.2.3. Comparison of different sampling schemes

For AIB$_9$ in solvated in water, the system is observed to occupy only left- and right-handed conformations for each of the inner residues considered and not the intermediate state 1 which we observed for the system solvated in chloroform. This is why the time scale of the dynamics which has to be resolved is comparably larger. This allows for propagating the time series using the time step of $\delta t = 10$ ps, as already indicated in Figure 7.2 (b). The neighborhood size that we used to compute adequate local estimates (eq. (4.59))
was set to $k = 50$. We use this setup for all dLE1 applications in this chapter. In order to reach comparable statistics, the dLE1 ensemble also contains 8 runs of length 4µs for each of the protocols itemized in the following.

**Native dLE1**

Applied directly to the full MD input data, the dLE1 setup qualitatively mimics the dynamics of the reference. This is indicated in Figure 7.3 by the blue lines for both FELs and autocorrelations along $V_1, V_2$. The autocorrelation decay time $\tau \approx 235$ ns lies within the standard deviation of the MD ensemble. Fluctuations are slightly increased for the dLE1 ensemble. The match in the autocorrelation decay is even better for the second PC. In terms of FEL projections, the dLE1 ensemble even quantitatively matches the benchmark: All intermediate states along the first and second PC are recovered true to original and even the barriers align with the trend of the MD input. Notably, the asymmetry which is intrinsic to the MD along the first PC also appears in the dLE1. The performance is similarly accurate for higher PCs, as shown in Figure 7.4.

Figure 7.3.: Comparison of enhanced sampling schemes: FELs (first row) and autocorrelations (second row) are shown for different dLE1 protocols (see Figure 7.1): The black line indicates the MD result, while the blue line shows the performance of the dLE1 directly applied to MD data. Furthermore, the performance for short MD trajectories generated both from continuous MD runs (red line) and by CONCOORD (green line) is shown.
Short trajectories from continuous MD

As a minor variation of the simple dLE1 protocol that aims to establish an enhanced sampling scheme, a single continuous run was drawn from the MD ensemble at first. This run was used in a second step to randomly pick starting positions, imposing only the constraint that they have to be more or less homogeneously distributed along the FEL $F(V_1, V_2)$, see Figure 7.6. That was accomplished by a simple binning along the two-dimensional FEL projection. Since a single MD run sparsely samples the configurational space especially in the broadened barrier region $-4 \lesssim V_1 \lesssim 4$, it by far lacks statistical information, compared to the whole ensemble. The majority of transitions occur along the main pathways, see page 26. Therefore, the minor populated states which are located centrally in the $V_1-V_2$ projection may not even be visited once in a single run. Altogether, about 1650 starting configurations were sampled. The corresponding set of structures was then used for subsequent short MD runs. For this purpose, a short equilibration run was conducted initially to avoid a bias due to non-equilibrated input data. This was followed by a short continuous MD run of 7.5 ns length. Hence, the overall statistics of these short MD runs contains 12.375 µs of exploitable input data for subsequent dLE1 runs. The idea of this simple MD sampling scheme is to avoid spending computational time in the most stable configurations, since the available statistics in these regions already exceeds the input required to guarantee a successful dLE performance. In contrast to the data pruning scheme introduced in 5.4, the purpose of this interference is not to accelerate
the dLE performance but to save computational time when generating input data. In this particular setup, we initially only save one eighth of the simulation time by picking only one continuous MD run. The subsequent short runs can be scheduled in parallel to save real time. In case of our AIB₉ peptide system, we initialize the system in the barrier region, where it will finally relax towards either state L or R, normally within a time scale of a few nanoseconds. Eventually, these short samples are more likely to access the less metastable states in the barrier region since they are already initialized in their vicinity, both locally and energetically. A sample of these short runs is visualized in Figure 7.5 (a). The resulting density distribution in the same projection is biased towards the two most

Figure 7.5.: Short trajectory samples generated for enhanced sampling protocols: Shown are representative examples of the short trajectories generated from (a) the continuous MD run and (b) the starting configurations generated by CONCOORD. Absolute populations of the dLE₁ input data are shown projected along the first (c) and second (d) PC for both cases (green lines for the CONCOORD setup) in relation to the full MD ensemble (blue lines).

stable states like the full MD reference, but contains much more local sampling in the
barrier region. In this region, the local densities provided by the short MD runs exceed the statistics of the reference, see Figure 7.5 (c), (d).

Before generating a dLE1 ensemble from this data, it is processed by delay embedding and projected using the eigenvectors derived from the full MD ensemble in order to compare the results. The performance of the dLE1 is indicated in Figure 7.3 by the red lines; it resembles the performance of the native dLE1 in every respect, e.g. the mean autocorrelation decay of the first component, $\tau \approx 220$ ns is consistent with the native dLE1 result. The symmetry of the FEL-projection along $V_1$ is enhanced, which we would expect for our peptide system since it was designed symmetrically. The largest fluctuations that we observe in the vicinity of the ground state along the $V_2$ axis are induced by a small bias due to the limited length of short trajectories. This is the case for the following reason: Our short trajectories are not able to converge the sampling of single excitations (e.g. $L = llll \rightarrow lrrr$, see Figure 7.5 (c), (d)) but in the majority of cases enter the most stable states.

Short trajectories from CONCOORD

The deficiency of the workflow introduced in the section above is that the sampling scheme introduced above still relies on a continuous MD simulation which has to surpass even the major potential barriers and sample the whole accessible phase space. In our specific peptide system, the absence of clearly disconnected states fortifies the success of the simple sampling scheme. In general, there is no method to monitor whether all import states are visited to reliably construct a dLE1 model. The CONCOORD algorithm at least offers a
method to access conformationally stable states [130]. Compared to the more elaborate enhanced sampling methods mentioned in the beginning of this chapter, its advantage is that it operates much faster for our purpose. We simply aim to generate stable structures without any a priori knowledge of the phase space.

Following the recommendation in the documentation, we decided to take the “yamber” configuration to set the Van-der-Waals parameters. The remaining setup was set to default. The proceeding of the algorithm works as follows: given a starting structure, all interactions are identified at first. Retaining geometric restraints, the remaining degrees of freedom are then identified. The stable conformations are generated afterwards by randomly varying the set of parameters and subsequently applying corrections in order to fulfill geometric constraints in an iterative manner. In our specific application to the AIB9 peptide, an additional restraint was imposed. The root mean square difference to all previously determined structures was set to be at least 0.8 Å. This additional constraint avoids the creation of similar configurations. We started with the configuration corresponding to state \(L\) ending up with \(\approx 5000\) stable configurations this way. The stability of these structures under the conditions of an MD simulation was further evaluated by applying an energy minimization to the system after adding the solvent and finally running MD simulations. An abort of the MD setup indicated instability for most of these structures. We ended up with only \(\approx 650\) stable configurations which are shown projected on the \(V_1 - V_2\) plane in Figure 7.6: High concentrations of these starting structures are located in the most stable regions along the major transition pathways. Notably, many stable conformations were generated in the central region of the FEL. They only seem to cluster within the perspective of the projection. In fact, the system space volume is drastically increased in this area.

For all stable configurations, trajectories of 7.5 ns length each were generated, respectively. Altogether, they yield an overall input data size of \(\approx 5\ \mu\text{s}\) which is usable for the dLE1 application. The data was prepared analogously to the MD in order to execute the dLE1 using the same system space and thus generate results which are comparable to the other workflows. The results of the dLE1 ensemble are indicated by the green lines in Figure 7.3: Although the projections of the free energy landscapes qualitatively recover at least the positions of the different metastable states, the barrier heights are for example systematically underestimated. Consistently, the dynamics along the first two PCs are systematically overestimated. This finding is easy to explain: in fact, besides the metastable states that occur in both our continuous MD ensemble and the short MD runs, CONCOORD generates structures which are not accessible, at least within the time scale of our simulation. These structures exhibit a clear separation from the dynamically accessible phase space being separated by a high potential barrier. States which we generated via CONCOORD are still stable under the energy minimization which is conducted
Chapter 7. Enhanced sampling of AlB₉ using the dLE

before the short MD trajectories are launched. The short trajectory samples which we started from these particular structures still eventually enter the common phase space and introduce a bias to the local estimates of the dLE. A representative sample of the short trajectories is shown in Figure 7.5 (b): the short MD trajectories exhibit an enhanced sampling of different areas in the phase space and also follow different transition paths. For instance, we observed a disproportional percentage of trajectories which relax towards the ground state following not the main transition paths. Some trajectories follow paths which do not exist in the reference. Fast relaxations from the initial structures effectively accelerate the overall dynamics which are recorded in the autocorrelations in Figure 7.3. This observation suggests, that an enhanced sampling scheme based on a method that does not include any information about accessibility of the phase space is ambiguous to deal with. The fortitude of the CONCOORD algorithm - to work quickly - lacks the guarantee of dynamic connectivity.

7.2.4. Alignment of system coordinates

Another conceptional challenge about the CONCOORD algorithm which did not come to the fore yet is that it does not solve the problem of establishing a set of system coordinates. So far, applying the eigenvectors \( v^{(i)} \) of the full MD ensemble in order to project all conceptional approaches onto the same eigenspace, served to compare the results of different schemes. Yet, without a-priori knowledge about connectivity and dynamics, it is rather difficult to attain the appropriate set of system coordinates. At least for the application of AlB₉ it can be shown that both enhanced sampling schemes are able to recover the system coordinates irrespective of the full MD ensemble. For this purpose, we measured the overlap of the individual system coordinates in terms of the quadratic scalar product of the respectively normalized eigenvectors,

\[
\left( v^{(i)}_{\text{fullMD}} \cdot v^{(j)}_{\text{shortMD}} \right)^2,
\]

where the \( v^{(j)} \) are calculated independently from the short MD runs. The overlap is shown in Figure 7.7 (b) for short trajectory pieces generated from CONCOORD starting structures. It shows, that especially the first three system components are aligned nearly parallel to the full MD reference. For the next two components, this congruence slightly decays, fourth and fifth PC together fairly match the reference being slightly tilted against each other.

The comparison of an independent dPCA applied on the short trajectories that were launched from a single continuous run and the dPCA of the full MD ensemble is shown in Figure 7.7 (a). It further supports the picture that - at least for AlB₉ - short trajectory samples are sufficient to recover a reliable set of system coordinates. In fact, the agreement
of the third PC even seems improved for the setup which uses CONCOORD. This finding is slightly misleading, since contributions to the third PC are shifted to fourth and fifth PC. However, the transformation is quite sensitive to these PCs, since the fluctuations are comparable for higher PCs while the first two PCs are rather distinct. At least for

Figure 7.7.: Comparison of independent dPCAs: Shown is the overlap of the normalized eigenvectors, following eq. (7.3) that are established for a dPCA on the full MD ensemble and short MD trajectories, independently. The axes labels denote the corresponding eigenvalues/PCs. The range is restricted to the system coordinates. (a) refers to the enhanced sampling scheme which uses CONCOORD, (b) refers to the simpler enhanced sampling scheme, see Sec. 7.2.3.

the example of AIB$_9$, the basic requirement is a sufficient length of each of the short trajectories. Since all runs within the broad barrier, that separates the states $L$ and $R$, are sampled long enough to eventually relax towards one of the two minima, the major transition paths along the barrier (see page 28) are fairly sampled. They determine the alignment of the system coordinates, while sparsely sampled states do not contribute substantially. For a more sophisticated system structure, the preconditions for a successful formulation of the system are more complicated, at least if a crude picture of the important dynamics is not provided.

### 7.2.5. Convergence of sampling schemes

Up to now we considered the full input data that we generated by running short MD simulations according to two different setups. We showed, that a reduced amount of input data is already sufficient to reproduce both dynamics and statistics at least for the simpler scheme. Since the short trajectories perform an increased sampling of the less metastable states and barriers, the local field estimates are reliable although the input
data is only one third of the full MD data. Yet, one can further reduce the amount of data considered by the dLE1 to see where the algorithm breaks down. Hence, for both enhanced sampling schemes, we evaluate the performance of the dLE1 varying either the number of short MD trajectories or the overall length of each trajectory while keeping the other quantity constant. For each reduced data input \( N_m \) a new set of 8 dLE1 runs is performed. In order to rate the quality of the dLE1, we proceeded analogously to our crafted double well model in Ch. 5. While the characteristic time \( \tau \) of the autocorrelation decay to \( C_1(\tau) = 1/e \) along the first PC determines the quality of the dynamics, the average population of a ground state,

\[
P_0 = \frac{P(V_1 > 4) + P(V_1 < -4)}{2}
\]  

(7.4)

rates the statistical accuracy. The resulting values for the MD reference, namely the 8 continuous runs of 4 \( \mu \)s length determine the reference values, \( P_0 \approx 0.41 \) and \( \tau \approx 0.29 \) \( \mu \)s.

The number of short MD trajectories was reduced by arbitrarily picking random trajectories and removing them from the overall data input considered by the dLE1. Figure 7.8.

Figure 7.8.: dLE1 performance for modified short trajectory samples generated by continuous MD runs: The exploitable input for the dLE1 algorithm was reduced by either varying the number of input trajectories (left column) or the length of each individual trajectory (right column). The quality is evaluated concerning statistics (first row) and dynamics (second row). \( P_0 \) and \( \tau \) are defined in Sec. 7.2.5.

111
(a), (c) show the effects on $P_0$ and $\tau$ for our simple enhanced sampling scheme: while the dynamical properties remain nearly unaffected until $N_m \approx 4 \cdot 10^5$ input points, the dLE1 experiences a small systematic bias concerning the ground state populations already if only few short trajectories are neglected. This threshold corresponds to $\approx 600$ trajectories. Beyond this threshold, we lack the sufficient amount of sampling, which is mostly recognizable in the fluctuations of the autocorrelation decay. For the independent enhanced sampling scheme which uses the starting structures generated by CONCOORD, one has to consider that we measured a bias anyway. The trend in Figure 7.9 (a),(c) fortify the perception, that this bias is not caused by insufficient sampling. The ground state population $P_0$ remains rather unchanged even upon using only $\approx 100$ sample trajectories as data input and likewise does the autocorrelation decay time. This confirms the previously explained suspicion, that the subset of states which was generated is not representative for the accessible system space. The performance of the dLE1 rather changes for a reduced amount of input trajectories. Only if we reduce the input to $\lesssim 100$ trajectories, the results of the algorithm are further deformed: For this limit, we introduce a bias that favors the ground state which is then increasingly populated and the autocorrelation decay now simply reflects intrastate dynamics. This effect is similar to the double well model (see Sec. 5.2).

Varying the length of the short MD trajectories, one can spot the minimum simulation time which is required to adequately sample the relevant system space. As already mentioned above, a minimum sampling length is needed to sample the most prominent transitions. These settle the robustness of the system coordinates that would be evaluated by an independent dPCA. For the simple enhanced sampling model which is triggered by the continuous MD run, Figure 7.8 (b), (d) show, that already little reduction of the length reduces the overall quality of the dLE1. A reduced sampling of the transitions introduces a bias which stabilizes the ground states continuously with decreasing trajectory length. This behavior occurs due to the lack of sampling events which favor the transition of leaving the ground state for very low sampling. In the same manner, the autocorrelation decays less rapidly since less transitions occur. For the CONCOORD induced short MD trajectories on the other hand, no trend is observable as shown in Figure 7.9 (b), (d). Since no recognizable change appears, it is also not feasible to evaluate the performance of the CONCOORD setup for an unknown system. At least the fact that no convergence is apparent for the quantities $P_0$ and $\tau$ reveals that a systematic bias was introduced already beforehand by the CONCOORD setup.
Figure 7.9.: dLE1 performance for modified short trajectory samples generated by CONCOORD: The exploitable input for the dLE1 algorithm was reduced by either varying the number of input trajectories (left column) or the length of each individual trajectory (right column). The quality is evaluated concerning statistics (first row) and dynamics (second row). $P_0$ and $\tau$ are defined in Sec. 7.2.5.
7.3. Summary

In this chapter, a promising application of the dLE was introduced. It enables to describe dynamics that happen on a millisecond timescale by providing short and spatially overlapping trajectories propagated on a nanosecond timescale. The dLE is central to this approach since it requires only local information. This strategy has the advantage that simulations can be executed in parallel. Our approach is thereby ranking among other enhanced sampling methods which aim to explore the FEL of biomolecular systems and to recover functional dynamics avoiding extensive MD simulations [58, 146].

Technically, the short MD runs require information about conformational space which is occupied by the biomolecular system.

To generate these structures, we suggested two different schemes and compared their overall performance in combination with the dLE1. While the first one relied directly on continuous MD data, we applied the CONCOORD sampling method to generate structures without providing any a-priori knowledge of the system.
In Ch. 6 I introduced and discussed the performance of the different dLE algorithms and the application to the AIB$_9$ system. It was discussed extensively how the dLE2 is applied in order to recover the dynamical information of a five-dimensional system coordinate on the basis of this small peptide system. A first application of the dLE algorithm was introduced in the last chapter, where we employed the dLE1 to regain global dynamical information from local statistics. The setting allows to combine more elaborate sampling schemes with the dLE and is still in need of improvement.

There are other possible applications of the algorithm that I want to mention in this chapter. Two different approaches based on the dLE algorithm are explained in this chapter.

The first one of these is a scheme for temperature rescaling. It aims to recover the dynamics of an underlying system at low temperatures working on the basis of input data at high temperatures. This approach is computationally advantageous if we assume, that the system space is explored more rapidly at high temperatures. Advanced sampling methods like “Parallel tempering” which I already mentioned in the last chapter are based on a similar idea [18, 139].

The second application is to construct a simple model by exploiting the information that the dLE provides about fields. The intent of this approach is to obtain analytical expressions for all fields which govern the respective LE and extract information about the physics of the system.

### 8.1. Temperature rescaling

In Ch. 4, the fluctuation-dissipation theorem eq. (4.19) was introduced which relates the fluctuations of the stochastic force to the temperature $T$. It therefore assumes a Boltzmann-distributed bath. If we further demand, that the system is described by a potential which is independent on $T$, the first order LE describing the system dynamics would only introduce an amplification of the stochastic force. However, a space-dependent
friction introduces a temperature dependency to the drift field: Since the friction field is following the Stratonovich calculus if we assume the Markovian LE as a limiting case according to Sec. 4.1.4, we have to consider the spurious drift in our Euler interpretation of the dLE1, see sec 4.1.3. This issue was also recovered by Ermak et al. [147], for example. In the second order scheme, however, we do not observe the spurious drift for sufficiently smooth fields\cite{66, 118} and are able to separate the temperature dependence. The general form of the second order dLE2 propagation step then reads

\[
y_{n+2} = y_{n+1} + \delta t^2 f(y_n) - \bar{\Gamma}(y_n)(y_{n+1} - y_n) + \delta t^2 \bar{K}(y_n)\alpha(T)\xi_n,
\]

which resembles eq. (6.2). The temperature dependency is introduced only in \(\alpha(T) = \sqrt{T/T_0}\), since we absorbed the thermal fluctuations of the reference in

\[
\bar{K}(y_n) = F(\Gamma(y_n), y_n)k_B T_0.
\]

An important assumption in this ansatz is, that \(F\) is only a function of the system space coordinate. This requires the local validity of the fluctuation-dissipation theorem, eq. (4.26). However, if we operate in internal coordinates like dihedral angles and apply a temperature rescaling, it is not clear whether the equipartition theorem holds which directly affects \(F\). In Cartesian space, we generally have the equipartition theorem \(\langle vv^T \rangle = M^{-1}k_B T\), which is obviously invariant for unitary transformations. For an equivalent description of a system in angular coordinates however, the distribution does not necessarily have to be a constant function of the angles and thus \(F\) might also feature a more complex picture.

8.1.1. A first application to dialanine

If we assume the validity of the assumptions introduced above, we can now easily propagate data at a single temperature on the basis of the dLE2 eq. (8.1) to forecast the dynamical behavior at other temperatures. A first application for the concept of temperature rescaling is given in the following on the basis of the small peptide system dialanine. This system was recently simulated for different temperatures in the context of a Bachelor thesis. MD simulations of this peptide solvated in water for the temperatures \(T = 260, 280, 300, 310, 320, 330\) and \(350\) K were generated. Each of these simulations were of an overall length of \(500\) ns. The trajectory was recorded using a time step of \(\delta t = 0.02\) ps. The two dihedral angles represent an excellent system coordinate, since they show a

\[\text{this is discussed in Sec. 6.1.1 particularly at the difference of forward and backward-propagation of the system coordinates}\]
clear time scale separation from the remaining degrees of freedom. A projection of the FEL along these coordinates is shown in Figure 8.2 (a) for the temperature 260 K. It exhibits four local minima corresponding to the left- and right-handed helical conformation $\alpha_L, \alpha_R$, the $\beta$ structure, and an additional conformation, $P_{II}$, which is adjacent to the state $\beta$. The most frequent transitions occur between $P_{II}$ and $\beta$ on a timescale of few ps. The rarest event on the other hand is the transition $P_{II} \leftrightarrow \alpha_L$, which happens on the timescale of tens of nanoseconds. For instance, only 3 transitions were observed for the lowest system temperature.

In order to qualitatively test temperature rescaling applying the dLE2, we concentrated on the most important transition $\alpha_R \leftrightarrow P_{II}$. Since the dLE2 algorithm so far is not implemented to handle circular coordinates, we applied a coordinate shift, aligning the least populated regions with the edges of the FEL. This is shown in Figure 8.1 (b) for the temperature $T = 350$ K. Comparing the FEL at both temperatures also reveals that the free energy barriers significantly change.

We used the circles delineated in Figure 8.1 (b) to define core regions for a dynamic clustering as described in Sec. 2.5. From the transition rates that we computed for the event $\alpha_R \leftrightarrow P_{II}$ we calculated transition times. These are represented for the MD data in Figure 8.2 by the black line. The representation visualizes that the temperature dependency is qualitatively matching the Arrhenius relation, $\ln \tau \sim 1/T$.

For all temperatures, we applied a similar setup of the dLE2, using the propagation time step of $\delta t = 0.02$ ps and a neighborhood size of $k = 200$ for the field estimates. This setup was found to optimize the dLE2 performance at $T = 300$ K. The resulting transition.

Figure 8.1.: FEL of Dialanine along $\phi, \psi$: Shown are the FELs for $T = 260$ (a) and 350K (b). The right plot is shifted in coordinate space thus that the highest free energy barriers are located at the edges. The circles denote the core regions corresponding to different conformations.
8.1. Temperature rescaling

Figure 8.2.: Transition times for temperature rescaling based on the dLE2: Shown are the transition times $\tau = 1/k$ for the transitions $\alpha_R \rightarrow P_{II}$ (a) and $P_{II} \rightarrow \alpha_R$ (b). The MD results are indicated by the black lines, results of the respective dLE2 runs by blue lines. The red lines correspond to a temperature rescaled dLE2 using the reference temperature $T = 300$ K.

Times are indicated in Figure 8.2 by the blue line. While the trend is predicted truthfully, we observe drastic deviations for the highest and lowest temperatures. Concerning the high temperatures, the deviations are explainable by the interference which we induced, allowing no transitions across the higher barrier separating both states. For the lower temperature on the other hand, the propagation time step and a different convergence of the estimators might affect the performance.

We used the MD run at 300 K to adapt the dLE2 and run simulations of similar length, applying temperature rescaling to forecast the dynamical behavior at all temperatures. The result is indicated by the red line in Figure 8.2. The trend of the rescaled dLE2 resembles the dLE2 which we applied directly to the MD data. At least for the transition $\alpha \rightarrow P_{II}$, both approaches are aligned. Only for the other transition, the default dLE2 performs significantly better. This finding suggests, that at least some of the model assumptions are violated: The above assumed fluctuation theorem might not hold. The fields might exhibit a temperature dependency, introducing for example an increased ruggedness at lower temperatures. In Figure 8.3, the populations of the clusters associated with the states $\alpha$ and $P_{II}$ are shown as a function of temperature. The MD curve, indicated by the black line, shows a clear trend: While the population of state $\alpha$ increases with temperature, the population of $P_{II}$ declines. The dLE2 applied directly to the MD trajectory at the same temperature is shown to recover this trend. For very low temperatures, it performs less accurate which implies that the setup is less suitable. The rescaled dLE2 clearly fails to recover the trend for state $\alpha$ and is also less accurately predicting the population of state $P_{II}$.

Nevertheless, we observe an impressive agreement of both dLE approaches. Further re-
8.2. Modeling dLE fields

In order to propagate the dynamics of a system coordinate, both dLE algorithms calculate fields on-the-fly. Thus, along with the free energy landscape, they provide information which is useful to learn from the dLE algorithm. If we have a detailed understanding of fields, we can start to learn about their effect on the dynamics along the free energy landscapes, e.g. how conformational peptide dynamics or even protein folding take place. A purpose of this approach would also be to understand how fields behave depending on for example the ruggedness of the free energy landscape or the dimensionality of the system space. Particularly, it would be interesting to figure out whether and to what extent fields vary for each of these different processes.

As already mentioned, saving the information provided by different fields can be rather cumbersome. Especially for a multidimensional system coordinate, the fact that the system size increases exponentially with the number of dimensions rapidly limits the concept of storing field information in a binned system space.

Yet, in low-dimensional projections, field estimates of the binned system space provide significant information about the behavior of the system. This is shown in Figure 8.4 for the projections of the hepta-alanine peptide [31] onto PCs $V_1$, $V_3$, using a time step of $\delta t = 1$ ps. The drift field indicates the transition paths in the projected system.
Figure 8.4.: 2D fields estimated by the dLE1 algorithm: Shown are the fields derived from a dLE1 run applied to the dPCA of hepta-alanine, using five PCs. All fields are projected along the first two PCs $V_1, V_2$. a) The drift field is indicated by a vector field, projected onto the FEL. The remaining plots show the diffusion field components $D_{11}$ (b), $D_{21}$ (c) and $D_{22}$ (d).
space. The diffusion field component $D_{11}$ increases in barrier regions of the system which indicates an increased dynamical flexibility. The other components of the diffusion field together characterize the flexibility of the system concerning the second PC, since $\langle V_2^2 \rangle - \langle V_2 \rangle^2 = D_{21}^2 + D_{22}^2$. As shown particularly in Figure 8.4 (b), (d), the dLE1 observes an increased diffusive behavior in the barrier regions of the FEL, especially along the first component. The off-diagonal element $D_{21}$ which couples first and second component of the noise vector almost vanishes, compared to the diagonal element $D_{22}$.

8.2.1. The coupled harmonic oscillator

Apart from the fact, that the extent of field information we have to store makes dealing with multidimensional systems rather intricate, a simple model is preferable for several reasons: On the one hand, an analytical form provides a systematic ansatz to describe a system which is based on a Langevin equation. On the other hand, a plausible model facilitates a physical understanding of the system coordinates.

In this section, I will introduce a simple approach which approximates the global fields by an analytical model. The particular model is motivated by the empirical valence bond (EVB) method [148–150]. Its basic idea is to characterize fields by coupled harmonic potentials. The example of the double well potential, page 66, provides a simple illustration: We firstly approximate both states represented in this model by harmonic functions, $V_i(x_i) = U_i(x_i) + U_i''(x_i)(x_i - x_{i-1})^2$, $i \in \{0, 1\}$, where $U_i'' = \alpha x_i(x_i - x_{i-1})$ denotes the curvature in the minima of this specific system. This yields a function min $(U_1(x), U_2(x))$ which by construction fairly matches the reference close to the minima and intersects in the vicinity of the transition state, see Figure 8.5 (a). At this point, the EVB model comes into play: It introduces the system state as a superposition of the two states 0 and 1, using the Hamiltonian

$$H(x) = \begin{pmatrix} V_0(x) & V_{01} \\ V_{10} & V_1(x) \end{pmatrix}. \quad (8.1)$$

$H(x)$ describes the joint system, where the off-diagonal elements $V_{01}$ and $V_{10}$ induce a coupling between both states. This coupling will be assumed both constant and symmetric in the following ($V_{ij} = V_{ji}$). The eigenvalues of this Hamiltonian correspond to ground

footnote{2 assuming it is described by a Euler-Maruyama scheme using the time step of the underlying trajectory}
8.2. Modeling dLE fields

and excited state $E(x)$, $E'(x)$, respectively. They read

\[
E(x) = \frac{1}{2}(V_0(x) + V_1(x)) - \frac{1}{2}\sqrt{(V_0(x) - V_1(x))^2 + 4V_{01}^2},
\]

(8.2)

\[
E'(x) = \frac{1}{2}(V_0(x) + V_1(x)) + \frac{1}{2}\sqrt{(V_0(x) - V_1(x))^2 + 4V_{01}^2}.
\]

(8.3)

Both eigenvalues are exemplified shown in Figure 8.5 (a). The coupling strength deter-

mines the gap between both eigenvalues. What is more important for our purposes is that it also relates to the energy barriers of our ground state, $\Delta E_{0 \rightarrow 1}, \Delta E_{1 \rightarrow 0}$. This leads to our application of the EVB model as a fitting scheme: By adjusting the coupling strength and the parameters of the harmonic potential, we want to fit a model of coupled harmonic oscillators to a field.

As shown in Figure 8.5 (b), an increased coupling reduces the distance and provokes a broadening of the minima in $E(x)$. It further varies the offset. In order to successfully apply a fitting procedure, we have to vary the parameters that determine each of the harmonic wells considered as well. For the particular double well model, this would result in $2 \cdot 3 + 1 = 7$ parameters.

8.2.2. A generalizable fitting algorithm

The basic idea of a generalized fitting algorithm based on the EVB model is to start with a predefined number of states as an initial guess, which are only weakly coupled to other states. We proceed by optimizing the parameters in order to fit the model as close as possible to a reference as accurately as possible. I will introduce this procedure at
the example of the FEL\textsuperscript{3} of hepta-alanine in the following. It will be shown to be also applicable to different manifolds as the Langevin fields.

A rough structure of the workflow which will be described in more detail later reads:

1. Initialize harmonic potentials

2. Identify neighborhoods (link states which are coupled to each other)

3. Run fitting procedure iteratively to adjust model to the underlying manifold

The general Hamiltonian for a $d$-dimensional system coordinate $\mathbf{x} = \{x_1,...,x_d\}$ reads

$$H(x) = \begin{pmatrix}
    \ddots & \ddots & \ddots \\
    V_i(x) & \ddots & V_{i,j} \\
    \ddots & \ddots & V_i(x) \\
    V_{i,j} & \ddots & \ddots \\
    \end{pmatrix}, \tag{8.4}$$

where potential functions along the diagonal read

$$V_i(x) = \sum_{j=1}^d \omega_{i,j} (x_j - x_{i,j})^2 + \epsilon_i \tag{8.5}$$

with the curvature vector $\omega_i$, the center $x_i$ and the offset $\epsilon_i$.

**Initialize harmonic potentials**

The first step of the algorithm is to declare an initial set of $N$ harmonic potentials $V_i(x)$. For this purpose, we consider information about the metastable states of our small peptide system. As detailed in Sec. 2.5.1, we can determine the population of different states directly on the basis of geometrical core definitions.\textsuperscript{4} It turned out that $N = 32$ stable states can be completely distinguished, using five PCs of a dPCA. Assuming a Gaussian distribution of the probabilities along each coordinate $j$ with variance $\sigma_{i,j}^2$, the curvature $\omega_{i0,j} = (2\sigma_{i,j}^2)^{-1}$ can be directly calculated. The mean $x_{i0}$ is set to the local minimum of the multidimensional FEL within the state boundaries. The additional index 0 denotes that these are initial parameters. The offset $\epsilon_{i0}$ is calculated with respect to the most stable state, for which we set $\epsilon_{i,\text{min}} = 0$ according to the convention that $G(\max_{\mathbf{x}} P(\mathbf{x})) = 0$ in Sec. 2.2.

\textsuperscript{3}where we account for the free energy in units of $k_B T$

\textsuperscript{4}a more elaborate method which was applied to this peptide system found a similar state definitions [151]
Identifying neighborhoods

Once an initial set of states is determined, we can fill the diagonal of the Hamiltonian eq. (8.4). Yet, before we can start to optimize our lowest eigenvalue $E(x)$, the off-diagonal coupling constants have to be introduced to mark states which are connected. Our idea is to connect only states which share a common barrier. Therefore, for each pair of states $i,j$ we simply check whether the projection of the manifold along the coordinate $x_{j,0} - x_{i,0}$ shows exactly one single local maximum between $x_{i,0}$ and $x_{j,0}$. The procedure to calculate the manifold along relative coordinates is explained below in Sec. 8.2.2. In addition, a distance cutoff is set. Each pair of states which complies with these criteria is accepted as a neighborhood and a small coupling constant is set in such a way that it does not significantly modify the positions of the minima of both states. All neighborhoods are recorded in a list.

Iterative fitting

Once a list of neighborhoods is assigned, a fitting procedure is applied which adjusts the free parameters of the system in an iterative manner, solving a minimization problem. Altogether, we introduced $N(2d + 1) - 1$ free parameters just to describe the harmonic wells. In addition, we get at most $N^2/2 - N$ additional parameters if all states are coupled pairwise. Since this implies a huge parameter space, a conventional optimization is rather difficult. For this reason, we apply an optimization method which only considers neighborhoods. It operates in two steps: Firstly, neighborhoods are randomly picked from the list. For each particular neighborhood, the $4d + 2$ associated free parameters are slightly varied in order to minimize the measure

$$\Delta\Delta E \equiv (\Delta E_{i\to j} - \Delta F_{i\to j})^2 + (\Delta E_{j\to i} - \Delta F_{j\to i})^2,$$

where $\Delta F$ denotes the target barrier height which was determined along with the neighborhoods. This is illustrated in Figure 8.6. The procedure is then repeated until all neighborhoods have been adjusted once. Then, the position of each minimum of the model is compared with the target position $x_{i0}, \epsilon_{i0}$ and corrected proportionally to the deviation in order to prevent shifts. Both steps of this procedure are repeated until $\sum_\text{neighborhoods} \Delta\Delta E$ is below a limit or does no longer vary significantly for further iterations.

Calculating manifolds along relative coordinates

In order to both cast neighborhoods and measure the barrier heights $\Delta F_{i\to j}, \Delta F_{j\to i}$ for each neighborhood $i,j$, we invented a tool to measure the manifold along the projection
Chapter 8. An outlook: Further applications

Figure 8.6.: Minimization of a single neighborhood: This figure illustrates the projection of the system along the coordinate $y = x_i + \alpha(x_j - x_i), \, 0 < \alpha < 1$ for a pair of states $i, j$. While the red curve denotes the manifold of the target, the black curve shows the projection of the harmonic model. The minimization of eq. (8.6) adapts the model to the target.

onto the relative coordinate $x_{j,\text{min}} - x_{i,\text{min}}$. The index $\text{min}$ denotes, that these coordinates are the local minima of $E(x)$ which are closest to the minimum of the respective states. For both casting the neighborhoods and the barrier heights of the reference manifold, they coincide with the minima of the harmonic potentials, $x_{i,\text{min}} = x_i^0$. In order to determine the barrier heights of the model in later iterations, $x_{i,\text{min}}$ generally deviates and is determined by a simple steepest descent method.

Once the relative coordinate is established, we consider only the range between both minima,

$$r = \{x_{i,\text{min}} + \alpha(x_{j,\text{min}} - x_{i,\text{min}}) \mid 0 < \alpha < 1\}. \quad (8.7)$$

For all neighborhoods, this interval is intersected in $n$ pieces so that $\alpha = l/(n - 1), \, l \in \{0, ..., n - 1\}$ and $r = r(l)$. The value $f$ of the manifold along this binned range is now calculated from the surrounding points of the reference $x$ as

$$f(l) = \frac{\sum \Theta(||r(l) - x|| < \epsilon) g(x)}{\sum \Theta(||r(l) - x|| < \epsilon)}. \quad (8.8)$$

Here, $\Theta$ denotes the Heavyside stepfunction and $\mid \cdot \mid$ the Euclidean metric. For the FEL, we have $g(x) = x$, while for a field $g$ the value of the respective field which is adopted at the reference point $x$ is considered. This is conceptually similar to the dLE. For the examples in Sec. 8.2.3, we used $n = 21$ and $\epsilon = 1/n$. We accept pairs of points as neighborhood if we observe a single local maximum. The latter is defined as the highest
value within a range of $i \pm 3$ intervals in order to account for the ruggedness of the FEL. Finally, the differences of maximum and the values at $l = 0$ and $l = n - 1$ define $\Delta F_{i \rightarrow j}$ and $\Delta F_{j \rightarrow i}$ or rather the difference of the logarithmic values in case of the the FEL (in units of $k_B T$).

The distribution of barrier heights which we measured for all neighborhoods is shown along with an example of a projection of the FEL in Figure 8.7.

![Figure 8.7.: Computing the FEL along projections: The left picture shows an example of a projection of the FEL of hepta-alanine along a relative coordinate, see eq. (8.7). The distribution of free energy barriers for the overall 80 neighborhoods is shown on the right plot.](image)

### 8.2.3. Examples

After introducing the workflow of the algorithm, it is instructive to illustrate its performance. I will restrain to the FEL of the hepta-alanine peptide system. The capability of the above introduced algorithm will be presented both for a one-dimensional and a five-dimensional projection of the FEL obtained from a dPCA. It was already shown that the five-dimensional picture is able to resolve the full connectivity [151]. Besides, the dLE1 was able to qualitatively recover the dynamics [80].

#### 1D modeling

As a first simple example, I consider the projection of the FEL of hepta-alanine along the first PC. We find 7 local minima, see Figure 8.8. Besides, we identify 6 neighborhoods so that we can write down a tridiagonal Hamiltonian, if we number the wells with increasing $V_1$.

The algorithm adjusts the model to the reference by shifting the offset of each harmonic well since an increased coupling naturally lowers the positions of the local minima in $E(V_1)$. This affects the leftmost state at $V_1 \approx -2.3$ the least which is weakly coupled to its right neighbor.
Chapter 8. An outlook: Further applications

Figure 8.8.: Modeling the FEL of hepta-alanine along PC 1: The reference is indicated by the rugged green line, while the converged fit is displayed by the brown line. The harmonic potentials represent the diagonal entries of the Hamiltonian 8.4 for the converged picture.

5D modeling The simple one-dimensional example does not necessarily require many features of the algorithm, but we also applied it to the five-dimensional FEL. Initializing the 32 states is rather straightforward on the basis of the 5d-trajectory. The identification of neighborhoods is visualized in Figure 8.9 (a): It shows the projection of the FEL along $V_1$, $V_2$. Metastable states are numbered and white lines indicate the coordinates which connect the centers of different states in the projection, according to the above introduced neighborhood criteria. All of these white lines represent neighborhoods. Cuts along these lines in the five-dimensional picture are used to adjust the model.

Figure 8.9 (a) is also instructive to see that some states would not be considered neighboring according to the above criteria in the two-dimensional projection. For instance, state 5 and state 6 are not separated by a single local maximum along the direct connection. The same holds for states 13 and 16. In the five-dimensional space however, we again recover the full connectivity, identifying 80 neighborhoods.

The list of neighborhoods is in turn used to apply the iterative fitting. We stopped iterating if the average deviation of the model barrier heights from the reference is less than 3%. The result of our fitting procedure is again shown in the projection along the first two PCs in Figure 8.9 (b). As a result of our model we observe a broadening of the different states compared to the reference in the projection. Yet, compared to the FEL which results from a dLE1, as shown in Figure 8.9 (c), the difference is smaller. Since we observe a broadening of metastable states by the dLE anyway due to effects discussed in Sec. 5.2.2, the simple harmonic model seems to work reasonably well.
Figure 8.9.: Modeling the five-dimensional FEL of hepta-alanine: Shown are projections of the FEL along $V_1$, $V_2$ from (a) MD data, (b) the harmonic model and (c) an optimized dLE1 run on the five-dimensional MD data. In the first plot, states are numbered and neighborhoods which were identified only in the two-dimensional projection are denoted by white lines.
8.2.4. Future perspectives

A more interesting application of the fitting algorithm introduced in this chapter is to construct an analytical model for all fields that determine the dLE. Subsequently, the established fields can be used to propagate a model Langevin equation. Figure 8.4 shows, that the positions of the minima of the diffusion field coincide with the minima of the FEL. Determining the drift field components on the other hand is more cumbersome. A possible approach is to assume that the potential energy surface which originates the drift fields can be approximated by the FEL.

Once a model is determined and a mLE working on this model fairly recovers the dynamics of the underlying system, the parameters of the final model can be varied in order to see how the system would react. For example, transitions can be regulated or single conformations can be made inaccessible.

The algorithm can be easily extended to describe the fields of the more general dLE2, even in combination with the concept of temperature rescaling if the necessary assumptions apply.
Conclusions

Understanding the dynamics of biomolecular systems is a major challenge. More than 50 years after Feynman’s statement that “everything that living things do can be understood in terms of the jiggings and wigglings of atoms” [152], the complex dynamics of even comparably small proteins is still not fully understood. While MD simulations as a classical semiheuristical ansatz paved the way for a detailed analysis of the microscopic dynamics in proteins, the interpretation and analysis of the data remains an additional considerable task.

Based on a proper choice of collective coordinates, the free energy landscape provides a helpful picture to identify metastable conformational states and understand molecular dynamics. We suggested a new picture of the free energy landscape of the peptide system of AIB$_9$ [107]. Although the system is rather small in terms of biomolecular scales, we showed that it exhibits a multitude of states which cannot be differentiated on the basis of an ordinary set of collective coordinates. The various states could rather be identified using different perspectives which are representative for the associated processes. Moreover, we were able to relate these processes to one another, assigning them to a hierarchical model: The most frequent transition that corresponds to the forming and breaking of hydrogen bonds is more than four orders of magnitude separated from the slowest process, an overall conformational change of the helical structure. To our knowledge, this is the first demonstration of the concrete mechanism of a hierarchical energy landscape. Furthermore, recognizing a similar mechanism for each temperature, we observed that the hierarchical model facilitates a dynamical transition of the peptide.

The main focus of this work was to study the data-driven Langevin equation (dLE), a conceptional approach to recover the characteristics of a low-dimensional time series by extracting local dynamical information. The latter is calculated locally and on-the-fly to estimate fields, which are then in turn used to propagate dynamics.

The research was conducted with respect to several directions: We developed an important extension of the dLE scheme which allows to treat inertia effects. To this end, we proposed a procedure to estimate fields from an underlying time series, which allows for
interpretation and recovery of the features of nonoverdamped systems, the second order Langevin equation (dLE2) [129]. In order to optimize the performance of this algorithm, we compared various implementations of the dLE2. We further showed that the first order dLE1 is a limiting case of the dLE2. Both the dLE1 and the dLE2 algorithm have been shown to decisively depend on several factors. For example, we rely on sufficient sampling of the input data to accurately recover local dynamical information. The choice of the propagation time step is an important factor as well as the amount of local statistics in order to achieve reliable field estimates. Extensive analysis was conducted to investigate the validity, applicability and convergence of both the dLE1 [127] and the dLE2 scheme concerning all of these aspects. To establish reliable benchmarks, we designed a simple model that features various characteristics of a realistic biomolecular system. On the basis of this model, we also applied and tested modifications that affect the input data set on which the dLE is based upon: “Data pruning” produces a multitude of short disconnected trajectories, thus reducing the overall amount of input data considered by the dLE, which results in a substantial speed-up of the algorithm. Delay embedding of the input data renders the possibility to shift intrinsic noise of the system coordinates towards the bath. Besides, it allows to consider inertia effects within the dLE1 by imitating the properties of a second order algorithm.

We prepared a low-dimensional time series of the initially discussed AIB9 peptide system to retrace its dynamics with the help of the dLE algorithm [129]. For this system, we showed that inertia effects can already be measured along the collective coordinates considered. We showed that the requirements of the model are only fulfilled by the dLE2 algorithm. For this purpose, we exploited the fortitude of the dLE to monitor its own practicability: By back-calculating the noise realization, we were able to test the validity of the noise model assumptions. Consequentially, we showed that the dLE2 outperforms the dLE1 in every respect: For instance, in the first order algorithm the metastability of conformational states is highly underestimated and structural variations are predicted to occur much more frequently. With respect to these aspects, the dLE2 fairly resembles the MD reference at least for the most relevant and sufficiently sampled states and the associated transitions.

Aside from establishing a technical guideline on how to construct a dLE model and verify its functionality, we also demonstrated several promising applications. For example, we used the dLE to forecast the long-time behavior of the AIB9 system from short trajectory samples. The idea of this approach is the following: Firstly, the system space is sampled to generate starting structures. In the next step, these starting structures are used to create short, partially overlapping MD trajectories. Subsequently, the dLE is applied to interpolate between the short local runs based on the local dynamical information. Using a simple setup of the CONCOORD algorithm to generate starting structures, we were
able to predict microsecond dynamics from nanosecond trajectories, thus covering three orders of magnitude in terms of the timescale. Since the input data can be generated in parallel, this application introduces an advanced sampling scheme which we consider capable of competing with other comparable approaches like Markov state modeling. In an outlook, we presented two further applications of the dLE in detail. On the one hand, temperature rescaling within the dLE algorithm is supposed to predict dynamics at low temperatures if information about the system at high temperature is provided. On the other hand, a model of coupled harmonic oscillators can be applied to the field estimates that are recovered by the dLE in order to determine an analytical model that completely describes the particular system. A variety of further studies can be conducted on the basis of this model. Particular conformational transitions can be inhibited for instance by altering the regulatory parameters of the model in order to study the response of the system.

Although the dLE concept is still in its infancy, we showed that it provides a powerful tool to describe and predict functional dynamics of biomolecular systems. We believe that the dLE represents a competitive alternative to well-established methods which helps to further understand biomolecular processes, such as protein folding and molecular recognition.
A

Rates, transitions and first passage times

A.1. Transition and rate matrix, normalization

The transition matrix $T(\tau)$ contains the transition probabilities between different states. The elements $T_{ij}$ denote the probabilities to switch from state $j$ to $i$ in a single propagation step of length $\tau$. For a probability distribution of $d$ states $P(t) = (P_1(t), \ldots, P_d(t))^T$ at time $t$, we have the propagation scheme

$$P(t + n\tau) = T(n\tau)P(t)$$  \hspace{1cm} (A.1)

$$\Delta t \equiv n\tau \iff \dot{P}(t) = T(\Delta t) - I \Delta t P(t).$$

Here, we already assumed the special case of Markovianity, where the transition matrix factorizes on the observed lag time $\tau$: $T(n\tau) = T^n(\tau)$. The relation above is a master equation,

$$\dot{P}(t) = M(\Delta t)P(t),$$  \hspace{1cm} (A.2)

where $M$ denotes the according rate matrix. We identify the relation

$$(T(\Delta t) - I)/\Delta t = M(\Delta t).$$  \hspace{1cm} (A.3)

In the following, we will skip the time step dependency of quantities for convenience. If we consider the normalization $\sum_i T_{ij} = 1$ and the rate definition $k_{ij} := T_{ij}/\Delta t$, we can relate both matrices and identify the rate matrix $M$,

$$T_{ii} - 1 = -\sum_{i \neq j} T_{ij} = -\Delta t \sum_{i \neq j} k_{ij} \quad \Rightarrow M_{ii} = -\sum_{i \neq j} k_{ij}, \quad i = j$$  \hspace{1cm} (A.4)

$$T_{ij} = \Delta tk_{ij} \quad \Rightarrow M_{ij} = k_{ij}, \quad i \neq j.$$  \hspace{1cm} (A.5)

Notably, the column normalization of the transition matrix is central to this relation.

On the other hand, a transposed transition matrix $\tilde{T}$ contains transition probabilities
from state \(i\) to \(j\). If we assume a row normalization, we can relate \(\tilde{T}_{ij} = T_{ji}\). \(T\) is directly connected with the transposed rate matrix via (A.3), \(\tilde{T} - I = M^T \Delta t\). Following eqs. (A.4), (A.5) and writing \(\tilde{k}_{ij} = \tilde{T}_{ij}/\Delta t\), we obtain

\[
M_{ii} = M_{ii}^T = -\sum_{j \neq i} \tilde{k}_{ij} = -\sum_{i \neq j} k_{ij} \quad \text{(A.6)}
\]

\[
M_{ij}^T = \tilde{k}_{ij} = k_{ji} \Rightarrow M_{ij} = k_{ij}/\Delta t.
\]

Notably, both interpretations recover the same rate matrix.

### A.2. Continuity equation, stationarity and detailed balance

To fulfill the continuity equation, we have to require \(|P| = \sum_i P_i = \text{const.}\) for the master equation eq. (A.2). This yields the constraint

\[
0 = \sum_i \dot{P}_i = \sum_i \sum_j M_{ij} P_j
\]

\[
= \sum_i \sum_j \left(-\delta_{ij} \sum_{l \neq j} k_{il} + (1 - \delta_{ij}) k_{ji}\right) P_j
\]

\[
= \sum_i \sum_j \left(-\delta_{ij} \sum_{l \neq j} k_{il} + k_{ji}\right) P_j
\]

\[
= \sum_i \left(\sum_j k_{ji} P_j - \sum_j k_{ij} P_i\right). \quad \text{(A.7)}
\]

In the stationary solution, we even demand \(\dot{P}_i = 0 \ \forall i\). Hence, the term in brackets in the last row has to vanish for every component. This balance condition is further confined in case of detailed balance, where we demand \(k_{ji} P_j = k_{ij} P_i\) for all pairs of states \(i\) and \(j\).

### A.3. Mean first passage times (MFPT) and transition times

#### A.3.1. The MFPT

The mean first passage time (MFPT) \(\tau_{mftp}\) is defined as the first moment of the first passage time, which in the one-dimensional case reads [67]

\[
\tau_{mftp} = \frac{1}{N} \sum_{x_1 \leq x' \leq x_2} \int_0^\infty w(x',t)tdt, \quad \text{(A.8)}
\]
where \( w(x, t) \) denotes the distribution of first passage times \( t \). This equation describes the mean time a particle initially positioned at \( x = x' \) takes to reach the edge of a domain \( x_1 < x < x_2 \). If we consider this domain as a state, we have to sum over all possible realizations inside that domain in addition:

\[
\tau_{mfpt} = \sum_{x_1 \leq x' \leq x_2} \int_{x_1}^{x_2} p(x, x')dx.
\]

Here, \( p(x, x') = -\int_0^\infty t\dot{P}(x, t|x', 0)dt \), where \( P \) denotes the probability density to reach \( x \) at time \( t \), being initially positioned at \( x(t = 0) = x' \). Thus \( p \) describes the propagation times for each position in the domain. Therefore, calculating the MFPT analytically requires detailed knowledge about the system properties. Numerically, calculating the MFPT is equivalent to averaging over all times and all realizations of particles in a given region.

### A.3.2. Two domains and relation to the transition time

In the special case of only two domains separated by a common barrier (e.g. a one-dimensional double well system with two absorbing boundaries and a single local maximum), the MFPT is evaluated (for a discretized time series) as:

\[
\frac{\tau_{mfpt}}{\delta t} = \frac{1}{n} \sum_{i=0}^{n-1} (n - i)\delta t = \frac{1}{n} n^2 - \frac{(n - 1)n}{2} = \frac{n + 1}{2} = \frac{\tau_{wait}}{2},
\]

where the summation accounts for a time \( n\delta t \) which the system spends in one domain before it reaches the barrier.

The transition time \( \tau_{trans,ij} \), which describes the transition from state \( i \) to \( j \), is defined as the inverse transition rate,

\[
\tau_{trans,ij} = \frac{1}{k_{ij}}.
\]

For this particular event, it yields

\[
\tau_{trans,ij} = \frac{\delta t (1 + n)}{1} = \tau_{wait} = 2\tau_{mfpt}.
\]

### A.3.3. more than two domains

The problem with more than two domains is visualized Figure A.1: The calculation of the mean first passage time also accounts for the time spent in other states apart from state 2 and 3. We have to consider these events. The MFPT for this escape problem is now the mean time spend in a state several times before state 3 is reached. The figure illustrates, that we have to consider each interval \( t_i \). We account for all these intervals...
A.3. Mean first passage times (MFPT) and transition times

Figure A.1.: MFPT for a system with 3 clusters: calculating the mean first passage time for the transition 2 → 3 has to consider that the system populates domain 1 occasionally.

considering the central time frame and finally “add up the pieces” (the small horizontal red lines shall indicate contributions). For this particular example, we get

\[
\frac{t_1}{2} + T_1 + t_2 + T_2 + t_3 + \frac{t_2}{2} + T_2 + t_3 + \frac{t_3}{2},
\]

where each contribution is listed in a separate row for convenience. This idea can be easily generalized for \( n \) clusters and \( m \) intervals spent in other clusters than the considered clusters in an iterative algorithm: Let \( t_i \) denote time intervals spent in the initial cluster and \( T_i \) the time intervals spent in other clusters than the target cluster and the initial cluster. The mean first passage time (all times in multiples of the time step) can be calculated via

\[
\tau_{mfpt} = 0
\]

\[
i = 1
\]

while not in target cluster do

if leave initial cluster not to other cluster except target cluster then

\[
\tau_{mfpt} = t_i/2 + (i - 1) \cdot t_i + i \cdot T_i
\]

\[
i + +
\]

end if

if leave initial cluster to target cluster then

\[
\tau_{mfpt} = t_i/2 + (i - 1)t_i
\]

end if

if leave other cluster except initial cluster to target cluster then

\[
\tau_{mfpt} = i \cdot (t_i + T_i)
\]

end if

end while
In the last loop, we also considered the event, that the final state can also be indirectly accessed. Notably, this short algorithm calculates the MFPT for a single escape event. In order to consider several events, we introduce a weight that accounts for the number of time frames spent in the initial state and normalize by the sum of weights in the end.

### A.3.4. Application to AIB$_9$

The algorithm introduced above can be used to calculate the distribution of mean first passage times of the single escape problem introduced in the last chapter. They are monitored for the different processes of the hierarchical model (Sec. 3.2) in Figure A.2.

Figure A.2.: Distributions of first passage times of AIB$_9$ for the temperatures 280, 300 and 350K as mentioned in Sec. 3.1. All distributions are binned on a logarithmic time scale. The well-sampled events $b \rightarrow f$ (a) and $0 \rightarrow 1$ (b) are discretized using 10 bins for each decade, while sparse events $l \leftrightarrow r$ (c) and $L \leftrightarrow R$ are binned using 30 bins in total.
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Time scales of different sampling schemes</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>The FEL for protein folding</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>Ramachandran plot and corresponding FEL</td>
<td>11</td>
</tr>
<tr>
<td>2.2</td>
<td>Illustration of PCA</td>
<td>13</td>
</tr>
<tr>
<td>2.3</td>
<td>Relevance of PCs: cumulative fluctuations and 1D Free Energy projections</td>
<td>14</td>
</tr>
<tr>
<td>2.4</td>
<td>Scheme to illustrate loss of connectivity information by projection</td>
<td>16</td>
</tr>
<tr>
<td>2.5</td>
<td>Scheme to illustrate dynamical clustering</td>
<td>17</td>
</tr>
<tr>
<td>3.1</td>
<td>Uncorrelated fast fluctuations of terminal residues</td>
<td>23</td>
</tr>
<tr>
<td>3.2</td>
<td>dPCA of AIB$_9$ solvated in water</td>
<td>23</td>
</tr>
<tr>
<td>3.3</td>
<td>Autocorrelations of AIB$_9$ solvated in chloroform</td>
<td>24</td>
</tr>
<tr>
<td>3.4</td>
<td>State definition in the AIB$_9$ peptide system</td>
<td>24</td>
</tr>
<tr>
<td>3.5</td>
<td>Transition networks for AIB$_9$ at 300 K</td>
<td>26</td>
</tr>
<tr>
<td>3.6</td>
<td>Transition networks for AIB$_9$ at 260 K and 350 K solvated in CHCl$_3$</td>
<td>27</td>
</tr>
<tr>
<td>3.7</td>
<td>Markovianity test for AIB$_9$ at 300 K</td>
<td>28</td>
</tr>
<tr>
<td>3.8</td>
<td>Illustration of the Hierarchical model describing AIB$_9$</td>
<td>29</td>
</tr>
<tr>
<td>3.9</td>
<td>Collective coordinates illustrating tiers of hierarchical dynamics in AIB$_9$</td>
<td>31</td>
</tr>
<tr>
<td>3.10</td>
<td>Hydrogen bond opening and closing in AIB$_9$</td>
<td>33</td>
</tr>
<tr>
<td>3.11</td>
<td>Model scheme for 01 transition and population of states</td>
<td>34</td>
</tr>
<tr>
<td>3.12</td>
<td>Rates describing competing processes concerning 0 → 1 transition</td>
<td>35</td>
</tr>
<tr>
<td>3.13</td>
<td>Time resolved 0 → 1 transition</td>
<td>35</td>
</tr>
<tr>
<td>3.14</td>
<td>Transition rates of hierarchical model</td>
<td>37</td>
</tr>
<tr>
<td>3.15</td>
<td>Rates describing competing processes concerning the l ↔ r transition</td>
<td>38</td>
</tr>
<tr>
<td>3.16</td>
<td>Non-Markovianity of the l ↔ r transition</td>
<td>38</td>
</tr>
<tr>
<td>4.1</td>
<td>Neighborhood samples of AIB$_9$ as used by the dLE1, projected onto $F(V_1, V_2)$</td>
<td>61</td>
</tr>
<tr>
<td>5.1</td>
<td>1D double well model potential</td>
<td>67</td>
</tr>
<tr>
<td>5.2</td>
<td>Langevin simulations of a 1D double well model</td>
<td>69</td>
</tr>
<tr>
<td>5.3</td>
<td>Scheme to illustrate non-locality</td>
<td>71</td>
</tr>
<tr>
<td>5.4</td>
<td>Performance of dLE1 on a 1D double well model</td>
<td>73</td>
</tr>
<tr>
<td>5.5</td>
<td>Time-step dependence of drift and diffusion fields</td>
<td>74</td>
</tr>
<tr>
<td>5.6</td>
<td>Application of the dLE1 on a 1D model that contains inertial effects</td>
<td>76</td>
</tr>
<tr>
<td>5.7</td>
<td>Pruning of input data for the 1D double well model</td>
<td>78</td>
</tr>
<tr>
<td>6.1</td>
<td>Different implementations of system coordinates for the dLE2</td>
<td>83</td>
</tr>
<tr>
<td>Figure</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>6.2</td>
<td>Field estimates of the dLE2 applied to a 1D double well</td>
<td>85</td>
</tr>
<tr>
<td>6.3</td>
<td>Noise model of the dLE2 applied to the 1D double well model</td>
<td>86</td>
</tr>
<tr>
<td>6.4</td>
<td>Comparison of different dLE2 algorithms</td>
<td>88</td>
</tr>
<tr>
<td>6.5</td>
<td>Sampling of AIB9 in chloroform at 320K</td>
<td>90</td>
</tr>
<tr>
<td>6.6</td>
<td>2D projections of the FEL along PCs $V_1, V_2$ for dLE1 and dLE2</td>
<td>91</td>
</tr>
<tr>
<td>6.7</td>
<td>Comparison of different dLE algorithms</td>
<td>92</td>
</tr>
<tr>
<td>6.8</td>
<td>Comparison of different dLE algorithms</td>
<td>93</td>
</tr>
<tr>
<td>6.9</td>
<td>Metastabilities of most stable states</td>
<td>94</td>
</tr>
<tr>
<td>6.10</td>
<td>Rates of individual transitions</td>
<td>95</td>
</tr>
<tr>
<td>6.11</td>
<td>Validity of the dLE model for AIB9 in chloroform</td>
<td>96</td>
</tr>
<tr>
<td>7.1</td>
<td>dLE based Enhanced Sampling scheme</td>
<td>101</td>
</tr>
<tr>
<td>7.2</td>
<td>Delay embedding and dPCA</td>
<td>102</td>
</tr>
<tr>
<td>7.3</td>
<td>Comparison of enhanced sampling schemes</td>
<td>104</td>
</tr>
<tr>
<td>7.4</td>
<td>Comparison of enhanced sampling schemes for higher PCs</td>
<td>105</td>
</tr>
<tr>
<td>7.5</td>
<td>Short trajectory samples generated for enhanced sampling protocols</td>
<td>106</td>
</tr>
<tr>
<td>7.6</td>
<td>Starting points of enhanced sampling protocols</td>
<td>107</td>
</tr>
<tr>
<td>7.7</td>
<td>Comparison of independent dPCAs</td>
<td>110</td>
</tr>
<tr>
<td>7.8</td>
<td>dLE1 performance for modified short trajectory samples generated by continuous MD runs</td>
<td>111</td>
</tr>
<tr>
<td>7.9</td>
<td>dLE1 performance for modified short trajectory samples generated by CONCOORD</td>
<td>113</td>
</tr>
<tr>
<td>8.1</td>
<td>FEL of Dialanine along $\phi, \psi$</td>
<td>117</td>
</tr>
<tr>
<td>8.2</td>
<td>Transition times for temperature rescaling based on the dLE2</td>
<td>118</td>
</tr>
<tr>
<td>8.3</td>
<td>Populations for temperature rescaling based on the dLE2</td>
<td>119</td>
</tr>
<tr>
<td>8.4</td>
<td>2D fields estimated by the dLE1 algorithm</td>
<td>120</td>
</tr>
<tr>
<td>8.5</td>
<td>Modeling a double well with two coupled harmonic potentials</td>
<td>122</td>
</tr>
<tr>
<td>8.6</td>
<td>Minimization of a single neighborhood</td>
<td>125</td>
</tr>
<tr>
<td>8.7</td>
<td>Computing the FEL along projections</td>
<td>126</td>
</tr>
<tr>
<td>8.8</td>
<td>Modeling the FEL of hepta-alanine along PC 1</td>
<td>127</td>
</tr>
<tr>
<td>8.9</td>
<td>Modeling the five-dimensional FEL of hepta-alanine</td>
<td>128</td>
</tr>
<tr>
<td>A.1</td>
<td>MFPT for a system with 3 clusters</td>
<td>138</td>
</tr>
<tr>
<td>A.2</td>
<td>Distributions of first passage times of AIB9</td>
<td>139</td>
</tr>
</tbody>
</table>
Bibliography


