UNIVERSIDADE FEDERAL DE VIÇOSA

On the reaction coordinate optimality for reversible aggregation kinetics

Lair Figueiredo Trugilho
Doctor Scientiae

VIÇOSA - MINAS GERAIS
2025

LAIR FIGUEIREDO TRUGILHO

On the reaction coordinate optimality for reversible aggregation kinetics

Thesis submitted to the Physics Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

Adviser: Leandro Gutierrez Rizzi

VIÇOSA - MINAS GERAIS
2025

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
T866o
2025

Trugilho, Lair Figueiredo, 1996-
On the reaction coordinate optimality for reversible aggregation kinetics / Lair Figueiredo Trugilho. – Viçosa, MG, 2025.
1 tese eletrônica (120 f.): il. (algumas color.).

Texto em inglês.
Inclui apêndices.
Orientador: Leandro Gutierrez Rizzi.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Física, 2025.
Referências bibliográficas: f. 97-106.
DOI: https://doi.org/10.47328/ufvbbt.2026.033
Modo de acesso: World Wide Web.

1. Processos de movimento browniano. 2. Físico-química. 3. Agregação (Química). 4. Colóides - Propriedades. 5. Proteínas - Propriedades. 6. Solução (Química) - Propriedades . I. Rizzi, Leandro Gutierrez, 1984-. II. Universidade Federal de Viçosa. Departamento de Física. Programa de Pós-Graduação em Física. III. Título.

CDD 22. ed. 530.475

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

LAIR FIGUEIREDO TRUGILHO

On the reaction coordinate optimality for reversible aggregation kinetics

Thesis submitted to the Physics Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

APPROVED: December 12, 2025.

Assent:

Lair Figueiredo Trugilho
Author

Leandro Gutierrez Rizzi
Adviser

ACKNOWLEDGMENTS

I would like to thank my advisor, Leandro Gutierrez Rizzi, with whom I have been working since the early days of my undergraduate course. I thank him for all discussions, patience, learning experience, and excellence as a tutor. He was fundamental during my complete academic journey.

I am also very grateful to my co-advisors Sergei Krivov and Stefan Auer for the great reception they provided me during the time I spent working with them at the University of Leeds. I learned a lot with both, and their participation was essential for the work we developed.

I thank my entire family for the support provided over all those years, especially my parents, Paulo Fernando Trugilho and Nelzy Luzório Figueiredo Trugilho, for their outstanding dedication and affection.

I also thank all my friends. Their support was fundamental for successfully finishing this part of my academic career. In particular, I am thankful for the wonderful hours we spent in special places, like Teddies, Bar do Cabelo, Bar do Hildeu, and Seu Zé Rock Bar.

I thank all the professors of the Universidade Federal de Viçosa, from which I benefited from an outstanding learning experience.

I also thank all other employees of the Universidade Federal de Viçosa, since without them none of this could have been possible.

I thank the computational structure provided by the Grupo de Investigação de Sistemas Complexos (GISC) of the Universidade Federal de Viçosa and all the people who worked hard to establish it in the past, as well as the ones who maintain it functional in the present.

I thank the fundamental service provided by the public funding agencies in benefit of social and scientific progress. In particular, I thank FUNARBE, FAPEMIG, and CAPES for the student scholarships that I received during my academic career.

This work has been sponsored by the following Brazilian research

agencies: Coordination for the Improvement of Higher Education Personnel (CAPES; Financing code 001), Minas Gerais State Foundation for Research Aid (FAPEMIG) and National Council of Scientific and Technological Development (CNPq).

ABSTRACT

TRUGILHO, Lair Figueiredo, D.Sc., Universidade Federal de Viçosa, December, 2025. **On the reaction coordinate optimality for reversible aggregation kinetics**. Adviser: Leandro Gutierrez Rizzi.

Fluctuation-driven reversible transitions are present in a vast class of finite-sized systems, including atomic cluster transitions, macromolecular aggregation, and protein folding. As a consequence, potential applications related to these transitions span a myriad of scientific areas, from the development of phase change materials in material science to the treatment and prevention of proteinopathies in the biological sciences. Here we are particularly interested in aggregation processes in intrinsically finite macromolecular systems, such as colloidal suspensions and protein solutions. These systems undergo transitions between diluted and aggregated states, and our main objective is to quantitatively describe the kinetics of such transitions. In order to study the kinetic properties of these many-body systems, it is usually necessary to project the highly-dimensional configuration space onto low-dimensional reaction coordinates. One of the main contributions of this thesis is to develop a methodology to accurately determine an optimal coordinate, known as the committor, for many-body systems. With the committor, several kinetic properties are calculated exactly, and we demonstrate the applicability of our approach by considering Monte Carlo simulations of an anisotropic lattice model as well as Brownian simulations of a Lennard-Jones system. In the other main contribution, we employ a suitable separation of timescales in a way that the more usual energy variable can also be used as a reliable reaction coordinate. In particular, we validate our previously introduced analytical expressions for the aggregation and dilution rates, which are based on the microcanonical entropy, via the simulations of the lattice system. As a final contribution, we introduce a method based on mean first-passage times that predicts the microphase separation in aggregating systems with competing interactions.

Keywords: aggregation; optimal reaction coordinates; microcanonical analysis; Monte Carlo simulations; Brownian dynamics simulations

RESUMO

TRUGILHO, Lair Figueiredo, D.Sc., Universidade Federal de Viçosa, dezembro de 2025. **Sobre a otimalidade de coordenadas de reação para cinética de agregação reversível.**. Orientador: Leandro Gutierrez Rizzi.

Transições reversíveis impulsionadas por flutuações são presentes em uma vasta classe de sistemas finitos, incluindo transições em agregados atômicos, agregação macromolecular e enovelamento de proteínas. Como consequência, potenciais aplicações relacionadas com estas transições abrangem uma miríade de áreas científicas, desde o desenvolvimento de materiais de mudança de fase em ciência dos materiais até o tratamento e prevenção de proteinopatias nas ciências biológicas. Aqui estamos particularmente interessados em processos de agregação em sistemas macromoleculares intrinsecamente finitos, tais como suspensões coloidais e soluções de proteínas. Estes sistemas descrevem transições entre estados agregados e diluídos e o nosso objetivo principal é descrever quantitativamente a cinética destas transições. Para estudar a cinética destes sistemas de muitas partículas, é usualmente necessário projetar o espaço de configurações de dimensão elevada em coordenadas de reação com baixa dimensionalidade. Uma das principais contribuições desta tese é o desenvolvimento de uma metodologia para se determinar precisamente uma coordenada de reação ótima, chamada de committor, para estes sistemas de muitas partículas. Com o committor, várias propriedades cinéticas são obtidas exatamente, e nós demonstramos a aplicabilidade da nossa abordagem considerando simulações de Monte Carlo para um modelo de rede anisotrópico e simulações de dinâmica Browniana para um sistema do tipo Lennard-Jones. Na outra contribuição principal, nós empregamos uma separação de escalas de tempo apropriada, de modo que uma variável mais usual, a energia, também pode servir como coordenada de reação confiável. Em particular, validamos expressões analíticas para as taxas de agregação e dissolução propostas em nossos trabalhos anteriores e baseadas na entropia microcanônica, utilizando simulações do modelo de rede. Como uma contribuição final, introduzimos um método, baseado em tempos de primeira passagem, para prever a separação de microfase em sistemas com interações competitivas.

Palavras-chave: agregação; coordenadas de reação ótimas; análise microcanônica; simulações de Monte Carlo; simulações de dinâmica Browniana

List of abbreviations

BD	Brownian dynamics
CV	collective variable
FEP	free-energy profile
LJ	Lennard-Jones
MC	Monte Carlo
MFPT	mean first-passage times
MTPT	mean transition path time
MSD	mean squared displacement
MSM	Markov state model
OP	order parameter
PDE	partial differential equation
PDF	probability density function
RC	reaction coordinate
SALR	short-range attraction and long-range repulsion
SDE	stochastic differential equation
TSD	total squared displacement
TPS	transition path segments

Contents

Part I

Introduction

Chapter 1

Motivation and theoretical background

First-order phase transitions are present in a vast class of materials, so that comprehensive descriptions of these transitions are of fundamental importance for different scientific areas. When the interactions between particles in some material lead to a first-order phase transition, that material can be found in more than one state, and as a result the system may become inhomogeneous, with parts of the material in different phases that coexist, as in the canonical example of water and ice coexisting. When the system is large enough so that the thermodynamic limit applies and is in equilibrium, it will stay inhomogeneous, and the fractions of the system in each phase are equilibrium constants, i.e., do not change in time. Parts of the material will only transform from one phase to another if an external agent happens to add or remove energy to that system, disturbing its equilibrium state. The total amount of energy that has to be exchanged with the system to convert an amount of the material from one phase to the other is called the latent heat [1].

The situation becomes more complicated when dealing with intrinsically finite-sized systems because, at the transition, there is a steady-state probability distribution which is bimodal, and, due to fluctuations, the system can evolve between different states that have similar probabilities. The states related with the maxima in the probability density are stable or metastable states characterized by minima in the system's free-energy profile. That is the case for a number of important problems, such as atomic cluster transitions [2], condensation of small molecules [3], macromolecular aggregation [4, 5, 6], protein crystallization [7], and protein folding [8, 9]. In all of those cases, the system, even in equilibrium, describes transitions between two (or more) states that have distinct physical properties.

In this work we will be concerned with macromolecular aggregation, roughly defined as the self-assembly of many subunits into ordered, finite-sized aggregates [10]. Those subunits can be molecules, colloids, or proteins, for example. Protein aggregation, in particular, has received considerable attention due to its connection with several human diseases, such as Alzheimer's and Parkinson's [11, 12].

As those systems present a large, yet finite, number of degrees of freedom, it is usually necessary to reduce the dimensionality of the problem for describing them. Equilibrium properties, like free-energy barriers and latent heats, can be computed by equilibrium thermostatistics approaches. Kinetic properties, such as mean transition times and reaction rates, on the other hand, are more subtle and are usually studied by projecting the highly dimensional configurational space onto one or a few reaction coordinates (RCs). The remaining large number of degrees of freedom are modeled as noise by a stochastic model, such as the diffusive models described by the Smoluchowski and Fokker-Planck-like equations [13, 14].

Those RCs are usually defined by physical intuition. In the aggregation problem, one normally works with the size of the biggest aggregate or the energy. But for such physically motivated RCs, there is no guarantee that a diffusive model is adequate to describe the system's kinetics. That is because the projection of the original multidimensional dynamics usually introduces non-Markovian effects [15, 16], so that the diffusive approximation may not hold. In principle, such non-Markovian character can be described by the generalized Langevin equation that contains a memory kernel. However, the computation of the kernel is often challenging, and its introduction complicates the simple interpretation of diffusion on a potential (free-energy) landscape. The existence of optimal RCs where a simple one-dimensional diffusive model can be used to quantitatively describe the systems' kinetics is demonstrated in the literature [17, 18, 19], but their computation for multidimensional systems is still an open problem. For systems describing transitions between two boundary states, which is usually the case in first-order phase transitions, the committor function is considered an optimal RC [20]. The committor function, also known as splitting probability, corresponds to the probability that, starting at an arbitrary intermediate state, the system reaches one boundary state without passing in the other one [21]. In Part II we will consider this problem for multimolecular aggregating systems. Particularly, we show that the committor can be determined by combining a systematic way of computing invariant collective coordinates with a nonparametric variational approach [22]. The applicability of the nonparametric method has already been demonstrated for unimolecular protein folding systems, and one of the main contributions of the present thesis (also published in Ref. [23]) is to extend it for multimolecular aggregation. With the computed optimal RC, we illustrate how one can exactly compute several important kinetic properties.

Although committor functions allow one to rigorously describe the kinetics of the systems, their computation is far from trivial and requires rather detailed information about the system, e.g., the temporal evolution of the positions of all particles. In this way, the other main contribution of this thesis is presented in Part III (also published in Ref. [24]), where we inspect in what conditions a more usual coordinate (the energy) can be employed as an RC. We show that by employing a separation of timescale, our previously

derived expressions for the aggregation and dilution rates [25, 26] are also applicable for many-body systems. The usefulness of this approach is that it is based on microcanonical thermostatistics [27], which has been widely used for the equilibrium characterization of finite-sized systems. In this way, it opens the way for readily determining kinetic properties for systems already considered in the literature. Moreover, it makes a more direct connection between kinetic and equilibrium descriptions of aggregating systems, as energy is a coordinate commonly used in thermostatistics approaches.

At last, we present in Part IV another contribution of this thesis, which is related to a way of theoretically computing the phase diagram for systems with competing short-range attraction and longer-range repulsion. Those systems display a rich phase diagram, including gelation and the appearance of equilibrium cluster phases [28]. The latter is characterized by multiple finite-sized clusters that happen to occur at the onset of the so-called microphase separation. In Part IV, we provide a theoretical method for estimating the regions of the phase diagram where such microphase separation takes place, looking only at the pairwise interaction potential.

Part II

Nonparametric determination of the committor

Chapter 2

Specific objectives

The main objective of this part is to describe and validate a method we developed to numerically construct the committor function for multimolecular aggregating systems. As previously discussed, the committor is a key element for describing both the equilibrium and kinetic properties of systems undergoing transitions between two boundary states in the context of stochastic diffusive models. The reason is that this quantity is considered an optimal reaction coordinate since it can be used to compute exactly important kinetic properties, like the flux, mean first-passage times (MFPT), and mean transition path times (MTPT). In this way, we start reviewing some definitions and properties of such diffusive models in Chapter 3, from which the concept of the committor function arises. Then, in Chapter 4, we describe the variational approach as well as how to obtain a set of permutationally invariant collective variables employed for determining the committor. Finally, in Chapter 5 we demonstrate the applicability of the method by considering a lattice gas with anisotropic interactions and a system of Lennard-Jones particles.

Chapter 3

Diffusive models

We start reviewing the properties of one-dimensional (1D) diffusive models, as they provide a clear and instructive way of introducing the matter due to their simplicity. In fact, as we shall discuss in the following, if the 1D committor projection is used, the usefulness of its corresponding 1D diffusive models relies on the fact that some quantities can be computed exactly, even for the multidimensional case of multimolecular systems.

Diffusive models are a family of random processes $\xi(t)$ that satisfies the stochastic differential equations (SDE) of the type [14]

$$\dot{\xi}(t) = f(\xi, t) + g(\xi, t)\zeta(t), \tag{3.1}$$

where the dot stands for time derivatives, $f(\xi, t)$ and $g(\xi, t)$ are arbitrary functions of ξ and t, and $\zeta(t)$ is a particular random process known as Gaussian white noise that has the following statistical properties:

$$\langle \zeta(t) \rangle = 0 \qquad \text{and} \qquad \langle \zeta(t)\zeta(t') \rangle = \delta(t - t'). \tag{3.2}$$

where $\langle ... \rangle$ indicates expected values. Those processes have been used for describing a variety of phenomena in different areas, including theoretical physics [29, 30], biology [31], and applied engineering and finance [32].

3.1 The Langevin equation

In physics, one important diffusive process is the one that describes the velocity $v(t)$ of a tagged particle in solution, subject to the assumed random collisions with the fluid particles that act like a heat bath at temperature T. This process is modeled by the so-called Langevin equation [30, 33, 34]

$$\dot{v}(t) = -\frac{\gamma v(t)}{m} + \frac{\sqrt{2\gamma k_B T}}{m}\zeta(t) \tag{3.3}$$

where the first term represents a systematic drift opposing the particle's movement with a friction coefficient γ, m is the mass of the tagged particle, k_B is Boltzmann's constant, and the second term models the random force due to the collisions with fluid particles. The constant multiplying the white noise sets the amplitude of the thermal fluctuations and is related to γ by the fluctuation-dissipation relation, which ensures that the system relaxes to the proper equilibrium state at long times [14, 30]. The Langevin equation is of the general form of Eq. 3.1 with $f(v,t) = -\gamma v/m$ and $g(v,t) = \sqrt{2\gamma k_B T}/m$.

3.1.1 The overdamped equation

It can be shown that in the long-time limit (or high-friction limit) $\gamma t/m \gg 1$, the inertial term can be neglected in the Langevin equation [14, 35, 36]. With this approximation, the equation can be written as a diffusive process for the position of the tagged particle $x(t)$, which reads

$$\dot{x}(t) = -\frac{V'(x(t))}{\gamma} + \sqrt{\frac{2k_B T}{\gamma}}\zeta(t) \tag{3.4}$$

where we also allowed for an external force to be applied at the particle, which in one dimension can be derived from a potential $F_{\text{ext}}(x) = -\mathrm{d}V(x)/\mathrm{d}x \equiv -V'(x)$. Note that in the case of harmonic potentials $V(x) = kx^2/2$, we have $f(x,t) = -kx/\gamma$ and $g(x,t) = \sqrt{2k_BT/\gamma}$ so that both diffusive models of Eq. 3.3 and 3.4 are indeed of exactly the same form, albeit for different constants, which are known as the Ornstein–Uhlenbeck process.

The diffusion coefficient is defined by the Einstein relation $D = k_BT/\gamma$ so that Eq. 3.4 can be rewritten as

$$\dot{x}(t) = -\frac{D}{k_BT}V'(x(t)) + \sqrt{2D}\zeta(t). \tag{3.5}$$

Dynamics satisfying Eq. 3.5 are known as Brownian, and the simulations that we will consider in Section 5.2 will be based on this equation.

3.2 The Smoluchowski equation

There is a correspondence between the SDE defining a diffusive model and the partial differential equation (PDE) that determines the temporal evolution of the related (conditional) probability density function (PDF). For instance, if the stochastic process is associated with the velocity as in the SDE given by Eq. 3.3, the related equation for the PDF is known as the Fokker-Planck equation. Invoking the SDE-PDE correspondence for the diffusive process defined by Eq. 3.5, one gets the Smoluchowski equation [14]

$$\frac{\partial p(x,t|x_0,t_0)}{\partial t} = D\frac{\partial^2 p(x,t|x_0,t_0)}{\partial x^2} + \frac{D}{k_B T}\frac{\partial}{\partial x}[V'(x)p(x,t|x_0,t_0)] \tag{3.6}$$

where $p(x,t|x_0,t_0)$ represents the probability density at position x and time t given that at the initial time t_0 the particle was at $x = x_0$, that is to say the PDE is subject to the initial condition $p(x,t_0|x_0,t_0) = \delta(x - x_0)$. Eq. 3.6 is a generalization of the diffusion equation for particles diffusing subject to an external potential. Free diffusion is recovered in the absence of external forces when $V'(x) = 0$ so that the second term on the right-hand side of Eq. 3.6 vanishes. In this work we will be restricted to stationary random processes, in which case the conditional PDF is a function of the elapsed time or time difference $(t - t_0)$ alone [14], so we can simply set $t_0 = 0$ and write $p(x,t|x_0,t_0) = p(x,t|x_0)$. We will further drop the x_0 dependence $p(x,t|x_0) \equiv p(x,t)$ for shortening the notation, but keeping in mind that the PDE determines the time evolution of the PDF conditional to the specified initial condition. It is also assumed that, in the limit of long times, the conditional PDF loses its dependence on x_0 and tends to the absolute or unconditional PDF, which for stationary processes is independent of time and equals the equilibrium distribution $p_{\text{eq}}(x)$,

$$\lim_{t\to\infty} p(x,t|x_0) = p_{\text{eq}}(x). \tag{3.7}$$

Such an equilibrium distribution will exist if the potential $V(x)$ is integrable and increases like $|x|^\alpha$ as $x \to \pm\infty$, where α is any positive number [14]. That means that the potential is confining, which will be the situation considered here. An important exception is the case of free diffusion, where the PDF does not converge to an equilibrium one in the long time limit; instead, it vanishes everywhere, and the related stochastic diffusive process (Wiener process) is known to be non-stationary [14]. It is easy to show that the equilibrium canonical distribution $p_{\text{eq}}(x) \propto e^{-\beta V(x)}$ is a stationary solution of Eq.3.6, as should be expected since that equation describes the PDF of a particle in contact with a heat bath at temperature $T = 1/k_B\beta$. Using the equilibrium distribution, one can write the Smoluchowski equation in the following way [13, 31]

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}\left[D(x)p_{\text{eq}}(x)\frac{\partial}{\partial x}\left(\frac{p(x,t)}{p_{\text{eq}}(x)}\right)\right], \tag{3.8}$$

In Eq. 3.8, the diffusion coefficient is allowed to have a position dependence, which in the context of the Brownian model corresponds to a position-dependent friction $\gamma(x)$. However, we note that Eq. 3.8 has also been used to model other systems, where it can be approximately derived under specific conditions, such as stochastic dynamics of Markov-chain Monte Carlo simulations [13].

3.2.1 Stationary solutions, committor functions and fluxes

Under appropriate boundary conditions, other stationary solutions of Eq. 3.8 besides $p_{\mathrm{eq}}(x)$ are also useful. We first note that the Smoluchowski equation can be rewritten as a continuity equation for the probability flow

$$\frac{\partial}{\partial t}p(x,t) = \frac{\partial}{\partial x}\mathcal{J}(x,t), \tag{3.9}$$

being $\mathcal{J}(x,t)$ the probability flux given by (see Eq. 3.8)

$$\mathcal{J}(x,t) = D(x)p_{\mathrm{eq}}(x)\frac{\partial}{\partial x}\left(\frac{p(x,t)}{p_{\mathrm{eq}}(x)}\right). \tag{3.10}$$

Now, in order to define the committor function, we look for stationary or steady-state distributions $p_{\mathrm{ss}}(x)$ that satisfy the boundary conditions [31]

$$p_{\mathrm{ss}}(x_A) = 0, \tag{3.11a}$$

$$p_{\mathrm{ss}}(x_B) = p_{\mathrm{eq}}(x_B) \tag{3.11b}$$

which means that the steady-state distribution maintains the equilibrium distribution at point x_B and satisfies an absorbing boundary state at x_A. For concreteness, we consider $x_A < x_B$, where x_A and x_B will usually discriminate regions in configuration space with different properties, such as a reactant well (A) for $x < x_A$ and a product well (B) for $x > x_B$ on a double-well potential, see e.g. Fig. 3.1. In this way, the formalism described in this subsection is aimed at describing kinetically the transitions between boundary states A and B. For the distribution to be stationary, the flux needs to be constant, $\mathcal{J}(x,t) = J_{AB}$ (independent of x and t), so that $p_{\mathrm{ss}}(x)$ satisfies

$$J_{AB} = D(x)p_{\mathrm{eq}}(x)\frac{\partial}{\partial x}\left(\frac{p_{\mathrm{ss}}(x)}{p_{\mathrm{eq}}(x)}\right) \tag{3.12}$$

where J_{AB} is the equilibrium unidirectional flux between milestones x_A and x_B, which equals the equilibrium number of transitions from A to B, N_{AB}, per unit time [37].

The stead-state distribution $p_{\mathrm{ss}}(x)$ is identified with the distribution of the system moving down in the positional x space and can be calculated from a trajectory $x(t)$ by the method of labeled "walkers" of Ref. [38]. This label specifies which of the boundary states were visited most recently in the course of the trajectory, which can be seen as a "walker" over x. For example, if the system just visited boundary state x_B, it has a label (B) that will only be changed to A when it reaches boundary x_A. This means that x_A acts like a absorbing boundary for walkers with label B, while state x_B acts like an reflecting boundary since label B is not changed when the system reaches x_B. The steady-state PDF for the position of walkers with label B corresponds to $p_{\mathrm{ss}}(x)$ as defined above [13].

The equilibrium and steady-state distributions corresponds, respectively, to the complete PDF $p_{\mathrm{eq}}(x)$ and the partial stationary PDF $p_{\mathrm{ss}}(x)$ computed only with the trajectories where the system has last visited x_B. Thus, the probability $q(x)$ that a "walker" at point x is coming from boundary B rather than A is just the ratio of those PDFs [31]

$$q(x) = \frac{p_{\mathrm{ss}}(x)}{p_{\mathrm{eq}}(x)}, \tag{3.13}$$

and due to time-reversal symmetry,[1] $q(x)$ is also the probability that, starting from any given x, the system will reach boundary B before reaching A. This quantity is known as splitting probability or committor and plays a central role in kinetic theories describing many systems. Using Eq. 3.12 and 3.13, one identifies that $q(x)$ satisfies the committor equation [37]

$$\frac{\partial}{\partial x}\left[D(x)p_{\mathrm{eq}}(x)\frac{\partial q(x)}{\partial x}\right] = 0 \tag{3.14}$$

respecting boundary conditions

$$q(x_A) = 0, \tag{3.15a}$$

$$q(x_B) = 1, \tag{3.15b}$$

in accordance with Eq. 3.11.

Eq. 3.13 and 3.15 can be solved to formally give [17, 37]

$$q(x) = \int_{x_A}^{x} \frac{\mathrm{d}x}{p_{\mathrm{eq}}(x)D(x)} \Bigg/ \int_{x_A}^{x_B} \frac{\mathrm{d}x}{p_{\mathrm{eq}}(x)D(x)} \,. \tag{3.16}$$

Analogously, the steady-state distribution $\hat{p}_{\mathrm{ss}}(x)$ of a system moving up in x (that has last visited x_A) is computed by exchanging the roles of points x_A and x_B as boundary states (Eq. 3.11). Summing both steady-state distributions of the system coming both from x_A and x_B, one recovers the equilibrium distribution [13], $p_{\mathrm{eq}}(x) = p_{\mathrm{ss}}(x) + \hat{p}_{\mathrm{ss}}(x)$. And the reverse committor $\hat{q}(x)$, or probability that starting from x, the system will reach boundary A before reaching B, is $\hat{q} = \hat{p}_{\mathrm{ss}}(x)/p_{\mathrm{eq}}(x)$ (analogous to Eq. 3.13) so that $q(x) + \hat{q}(x) = 1$. This just expresses that the system has to come from one of the two boundaries. As a final remark, we note that quantities such as $q(x)$ and $p_{\mathrm{ss}}(x)$ clearly depend on the defined boundary conditions A and B, but we will not carry this dependence explicitly in the notation for shortening it.

[1] As we will always deal with equilibrium trajectories respecting detailed balance, time-reversal symmetry will always hold [37, 39].

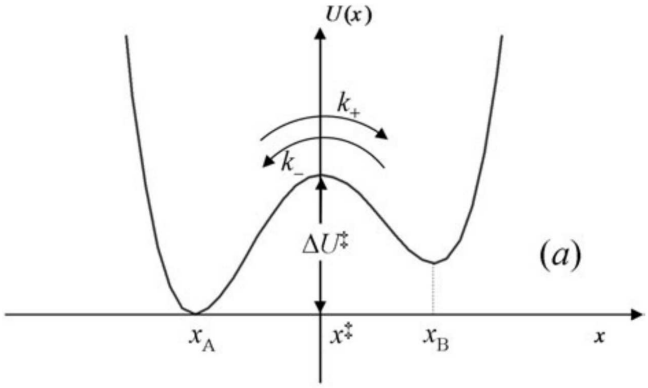

Figure 3.1: An example of a schematic double-well potential $U(x) \equiv V(x)$ with minima at $x = x_A$ and $x = x_B$ extracted from Ref. [31]. The forward escape rate from A to B can be estimated as $k_+ = 1/\tau_{A\to B}$, while the reverse escape rate from B to A can be estimated as $k_- = 1/\tau_{B\to A}$.

3.2.2 Characteristic times and rates

One problem of particular interest that can be treated with the above formalism is that of the escape rate over a potential barrier by diffusion. Consider the double-well potential $V(x)$ sketched in Fig. 3.1. We consider a particle subject to this potential in the context of the Brownian diffusive model (Eq. 3.4). The potential presents two minima at distances $x = x_A$ and $x = x_B$, so that at these positions the diffusing particle has a tendency to stay bounded due to the forces it experiences. In this way, most of the time the particle is found in the regions close to x_A or close to x_B, which can then define convenient boundary conditions: state A for $x < x_A$ and state B for $x > x_B$. Nevertheless, given that the particle is close to one of the minima, it still can escape the potential barrier due to thermal fluctuations and reach the other minima. Given that the potential is bounded for $x = \pm\infty$, this process will repeat with the particle describing transitions between states A and B. This kind of problem turns out to be an archetype model for different types of thermally activated processes, from chemical reactions [31] to aggregation [26].

The escape rate from A to B can be estimated as the inverse of the mean first-passage time (MFPT) from $x = x_A$ to $x = x_B$. Such MFPT is defined as the mean time that a particle starting from x_A takes to first reach boundary x_B. In the formalism of the one-dimensional Smoluchowski equation, it can be calculated exactly as [37, 40]

$$\tau_{A\to B} = \int_{x_A}^{x_B} \frac{\mathrm{d}x}{p_{\mathrm{eq}}(x)D(x)} \int_{-\infty}^{x} p_{\mathrm{eq}}(x')\mathrm{d}x' \tag{3.17}$$

for $x_A < x_B$. Similarly, the reverse MFPT, or the mean time a particle starting at x_B takes to reach x_A for the first time, is given by

$$\tau_{B\to A} = \int_{x_A}^{x_B} \frac{\mathrm{d}x}{p_{\mathrm{eq}}(x)D(x)} \int_{x}^{\infty} p_{\mathrm{eq}}(x')\mathrm{d}x', \tag{3.18}$$

and its inverse defines the escape rate from state B to A.

For completeness, the mean transition path time (MTPT), $\hat{\tau}$, defined as the mean time the particle takes for transitioning between x_A and x_B without returning to the initial state (the barrier-crossing time), which is equal in both directions (from A to B or vice-versa) [41], can be exactly computed as [41, 42]

$$\hat{\tau} = \frac{\int_{x_A}^{x_B} p_{\mathrm{eq}}(x) \left(\int_{x_A}^{x} \frac{\mathrm{d}x'}{p_{\mathrm{eq}}(x')D(x')} \right) \left(\int_{x}^{x_B} \frac{\mathrm{d}x'}{p_{\mathrm{eq}}(x')D(x')} \right) \mathrm{d}x}{\int_{x_A}^{x_B} \frac{\mathrm{d}x}{p_{\mathrm{eq}}(x)D(x)}} \tag{3.19}$$

A remarkable feature of the 1D diffusive model defined by Eq. 3.5, or equivalently by the Smoluchowski equation (Eq. 3.8), is that all those characteristic times are expressed as closed expressions depending only on $p_{\mathrm{eq}}(x)$ (or $V(x)$) and $D(x)$ (for the more general case of position-dependent diffusion coefficient). We also note that despite the points x_A and x_B being usually chosen near a minimum of the potential, the formalism is general, and they can be chosen arbitrarily as the boundary conditions for the problem at hand. It can be shown that expressions 3.17–3.19 remain exact and of the same form for the committor variable (just substituting $x \to q$, $p_{\mathrm{eq}}(x) \to p_{\mathrm{eq}}(q)$, and $D(x) \to D(q)$ using the transformation $q = q(x)$ (Eq.3.16)). This result is not of much interest in the one-dimensional case, because one can simply work with the expressions in the original x space. But, as it also holds for the multidimensional case where there are no such closed expressions in the original space, the usefulness of the committor as an optimal coordinate becomes apparent, as we shall discuss in detail next.

3.3 The multidimensional case

We now consider the multidimensional case of N interacting (overdamped) Brownian particles in d spatial dimensions. We assume that the force exerted on each particle due to the others can be derived from a potential $V(\mathbf{x}_1, ..., \mathbf{x}_N) = \sum_{i>j} V_{ij}(\mathbf{x}_i, \mathbf{x}_j)$, which is a sum of pairwise terms $V_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ of the interaction between particles i and j at positions $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Let $\vec{X} \equiv (\mathbf{x}_1, ..., \mathbf{x}_N)$ denote such multidimensional configurational space for brevity. There are no external forces acting on the particles, so that, analogously to Eq. 3.5, we have

$$\dot{x}_\mu = -\frac{D_0}{k_B T} \frac{\partial V(\vec{X}(t))}{\partial x_\mu} + \sqrt{2D_0}\zeta_\mu(t) \tag{3.20}$$

for the evolution of any component x_μ of the position of any particle. We used a Greek index that runs for all components $\mu = 1, ..., dN$ in contrast with the Latin index $i, j = 1, ...N$ for particle number. The deterministic part of the force is obtained as the gradient of the potential, and the stochastic force is driven by independent white noises ζ_μ for each

component, so that $\langle \zeta_\mu(t)\zeta_\nu(t')\rangle = \delta_{\mu\nu}\delta(t-t')$. The diffusion coefficient at physical space is still defined by the Einstein relation $D_0 = k_B T/\gamma$, but we add the subscript "0" from now on to differentiate it from the diffusion coefficients defined on more abstract spaces like $D(q)$, which will be important in the following[2].

The set of coupled SDEs 3.20 defines a multidimensional diffusive model [14], and the corresponding PDE, known as the multidimensional Smoluchowski equation, for the evolution of the associated PDF, $p(\vec{X},t)$, is given by [31, 44]

$$\frac{\partial p(\vec{X},t)}{\partial t} = \sum_{\mu=1}^{dN}\left[\frac{D_0}{k_B T}\frac{\partial}{\partial x_\mu}\left(\frac{\partial V(\vec{X})}{\partial x_\mu}p(\vec{X},t)\right) + D_0\frac{\partial^2 p(\vec{X},t)}{\partial x_\mu^2}\right], \tag{3.21}$$

which is the multidimensional analogue of Eq.3.6. In a similar way, the equilibrium distribution is $p_{\mathrm{eq}}(\vec{X}) \propto e^{-\beta V(\vec{X})}$, the canonical distribution.

In this case the committor $q(\vec{X})$, a one-dimensional function of the multidimensional space, satisfies [39]

$$\sum_{\mu=1}^{dN}\frac{\partial}{\partial x_\mu}\left[D_0 p_{\mathrm{eq}}(\vec{X})\frac{\partial q(\vec{X})}{\partial x_\mu}\right] = 0 \tag{3.22}$$

respecting the boundary conditions

$$q(\vec{X}\in A) = 0, \tag{3.23a}$$

$$q(\vec{X}\in B) = 1, \tag{3.23b}$$

where A and B are arbitrary regions in the multidimensional space $\vec{X}$. Eq. 3.22-3.23 are the multidimensional analogues of Eq. 3.14-3.15. $q(\vec{X})$ is the probability that, starting from any point $\vec{X}$ on the multidimensional space, a system evolving through the dynamic equations 3.20 will reach boundary state B before it reaches A.

In contrast to the 1D case, there is no formal solution for Eq. 3.22-3.23, i.e., no analogue to Eq. 3.16. In fact, due to its high complexity, Eq. 3.22-3.23 has only been directly solved for low-dimensional systems [39]. Alternatively, Eq. 3.22 can be expressed as a variational problem [45, 46]

$$\min_q \int \sum_{\mu=1}^{dN}\left(\frac{\partial q(\vec{X})}{\partial x_\mu}\right)^2 p_{\mathrm{eq}}(\vec{X})\mathrm{d}\vec{X} \tag{3.24}$$

which is equivalent to Eq. 3.22 under the same boundary conditions Eq. 3.23. The formu-

[2]Eqs. 3.20-3.22 can be generalized to include a diffusion tensor that is a function of configurational space $\mathbf{D}(\vec{X})$ [43], but we keep the diffusion coefficient D_0 as a scalar constant for simplicity as it describes particle diffusion in an isotropic medium, and also because the Brownian simulations that we will consider in Sec. 5.2 will be based on Eq. 3.20 in this form.

lation of Eq. 3.24 has been used as the starting point for variational methods proposed for computing the committor in multidimensional systems. The majority of which are parametric methods where the function $q(\vec{X})$ is parameterized using arbitrary functional forms containing many parameters, which are varied allowing for the minimization of Eq. 3.24 and a reasonable representation of the full function $q(\vec{X})$ [47]. Specifically, the function $q(\vec{X})$ has been parameterized using artificial neural networks [46, 48, 49] and tensor networks [50]. In contrast, a non-parametric variational approach was proposed in Ref. [22] that deals directly with collective coordinate time series in a way that one needs not to determine the full multidimensional function $q(\vec{X})$ but computes directly the committor time series. In this work we used this last approach, which will be described in detail in Chapter 4.

3.3.1 Generalized free-energy profiles and optimality criterium

In the following, we introduce the generalized cut-based free-energy profiles, from which a stringent validation criterion for a reaction coordinate (RC) that closely approaches the committor can be formulated. We define an RC as an arbitrary one-dimensional projection of the configurational space $r \equiv r(\vec{X})$, of which the committor is a special case that we reserve the letter $q = q(\vec{X})$. First we make a brief revision of conventional histogram-based free-energy profiles, which are related to the equilibrium distribution of a general RC r.

Histogram-based free-energy profile

We consider a long equilibrium trajectory described by a system evolving through the dynamical equations 3.20. The system is characterized by the multidimensional trajectory $\vec{X}(k\Delta t)$ at different times $k\Delta t$, $k \in \mathbb{N}$. We consider a suitable discretization of time so that the trajectory is sampled with a time interval Δt. An arbitrary reaction coordinate time series $r(k\Delta t)$ is then computed by applying the projection over the trajectory $r(k\Delta t) = r(\vec{X}(k\Delta t))$.

The equilibrium distribution $p_{\mathrm{eq}}(r)$ can be expressed in terms of the usual histogram-based free-energy profile $F_H(r, \Delta t)$,

$$p_{\mathrm{eq}}(r) = \frac{e^{-\beta F_H(r,\Delta t)}}{\mathcal{Z}}, \tag{3.25}$$

being $\mathcal{Z}$ the total partition function. Note that the free-energy profile $F_H(r, \Delta t)$ and the equilibrium distribution $p_{\mathrm{eq}}(r)$ of a general coordinate are related exactly in the same way as the potential $V(x)$ and $p_{\mathrm{eq}}(x)$ in the case of a one-dimensional Brownian particle subject to Eq.3.5. Given the reaction coordinate time series $r(k\Delta t)$, $F_H(r, \Delta t)$ can be estimated by binning the coordinate and computing histograms. The associated histogram $Z_H(r, \Delta t)$,

which is also called partial partition function[3], can be computed as the density of points in a bin $[r, r + \Delta r]$, that is,

$$Z_H(r, \Delta t) = \frac{\mathcal{M}_r}{\Delta r}, \tag{3.26}$$

where $\mathcal{M}_r$ is the number of points of the trajectory lying between r and $r + \Delta r$ with Δr the bin size. The estimate of the histogram-based free-energy profile is then given by, see Eq. 3.25,

$$F_H(r, \Delta t) = -k_B T \ln Z_H(r, \Delta t), \tag{3.27}$$

so that the estimate of the equilibrium probability density distribution is

$$p_{\text{eq}}(r) = \frac{Z_H(r, \Delta t)}{\int Z_H(r, \Delta t) \mathrm{d}r} \tag{3.28}$$

with the denominator the estimator of the total partition function.

It is important to note that, since $Z_H(r, \Delta t)$ is a non-normalized histogram that is directly related to the equilibrium distribution, it scales trivially with sampling interval as $Z_H(r, \Delta t) \sim \Delta t^{-1}$. That is because increasing Δt only reduces the number of points considered in the computation. For example, if $\Delta \bar{t} = 2\Delta t$, the points are twice as spaced, so only half the points are considered in computing $Z_H(r, \Delta \bar{t})$, so it would be reduced by a factor of two. This constant only appears because $Z_H(r, \Delta t)$ is not normalized by construction, but of course the normalized equilibrium probability distribution $p_{\text{eq}}(r)$ is not affected at all by such a change of sampling interval. Because of that, we shall drop the Δt dependence of $F_H(r)$, and as the normalization just appears as an additive constant due to the logarithm in Eq. 3.27, we will set that constant so that the minima of $F_H(r)$ lies in the abscissa axis (i.e., the minima is zeroed) when presenting the corresponding histogram-based free-energy profiles.

Cut-based free-energy profile

In addition to the histogram-based free-energy profiles defined by Eq. 3.27, it is also useful to compute cut-based free-energy profiles, with the ordinary one defined as [47]

$$F_{C,0}(r, \Delta t) = -k_B T \ln Z_{C,0}(r, \Delta t) \tag{3.29}$$

with its cut-based partition function $Z_{C,0}(r; \Delta t)$ equating half the total number of transitions through point r in the whole coordinate time series sampled with time interval Δt.

[3]Here we are following the notations used in references [17, 47] where the histogram $Z_H(r, \Delta t)$ is called a partition function. It can be named this way since it corresponds to a sum over the other degrees of freedom, but it is not to be confused with the total partition function $\mathcal{Z}$ obtained by summing it all up.

Mathematically, this can be written as

$$Z_{C,0}(r,\Delta t) = \frac{1}{2}\sum_k \Theta[(r(k\Delta t) - r)(r - r(k\Delta t + \Delta t))], \tag{3.30}$$

where Θ is the Heaviside step function and the sum extends to the complete coordinate time series. Each term in the summation is zero if $r(k\Delta t) > r$ and $r(k\Delta t + \Delta t) > r$ or $r(k\Delta t) < r$ and $r(k\Delta t + \Delta t) < r$. In these cases, there was no transition through point r between subsequent times $k\Delta t$ and $k\Delta t + \Delta t$. On the other hand, if $r(k\Delta t) < r$ and $r(k\Delta t + \Delta t) > r$ or $r(k\Delta t) > r$ and $r(k\Delta t + \Delta t) < r$, the step function returns one, and the passing of the series through r is counted, coming either from lower or higher values of the coordinate.

Note that, differently from the histogram-based free-energy, the cut-based one may change with sampling interval in non-trivial ways, as it is not equivalent to equilibrium properties. In fact, changing Δt will generally modify $Z_{C,0}(r,\Delta t)$ since it will change the connectivity between $r(k\Delta t)$ and $r(k\Delta t + \Delta t)$ for different Δt.

Generalized cut-based free-energy profile

The cut-based profile can be further generalized by summing up the transitions through r with weights equal to the length of the transition raised to the power a [47], i.e., the generalized partition function is written as

$$Z_{C,a}(r,\Delta t) = \frac{1}{2}\sum_k |r(k\Delta t + \Delta t) - r(k\Delta t)|^a \Theta[(r(k\Delta t) - r)(r - r(k\Delta t + \Delta t))], \tag{3.31}$$

and, as usual, the generalized cut-based free-energy profile is defined as

$$F_{C,a}(r,\Delta t) = -k_B T \ln Z_{C,a}(r,\Delta t), \tag{3.32}$$

being Eq. 3.30 the particular case $a = 0$, which justifies the subscript zero in Eq. 3.29 and 3.30. In Fig. 3.2, it is illustrated the construction of $Z_{C,1}(r,\Delta t)$ given an RC time series, and in the Appendix A we also show an algebraic example of the computation of $Z_{C,1}(r,\Delta t)$ for a simple system with only three states.

Optimality criterion and diffusion coefficient

The case of $a = 1$ is of particular interest because a simple but stringent optimality criterion is expressed in terms of $Z_{C,1}(r,\Delta t)$. It states that [17], for the committor RC, q, $Z_{C,1}(q,\Delta t)$ is independent of q and Δt and equals the number of transitions N_{AB} between the regions A and B defining the committor boundary states (Eq. 3.23),

$$Z_{C,1}(q,\Delta t) = N_{AB} \qquad \text{or} \qquad F_{C,1}(q,\Delta t) = -k_B T \ln N_{AB} \tag{3.33}$$

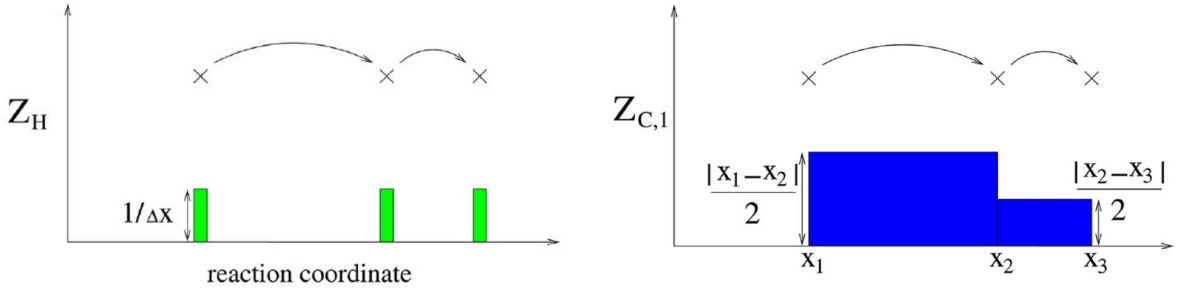

Figure 3.2: Illustration extracted from reference [17] of the computation of the usual histogram-based free-energy Z_H and the generalized cut-based free-energy $Z_{C,1}$ for three consecutive points of a trajectory $x_k \equiv r(k\Delta t)$. Whereas the former is computed summing $1/\Delta x \equiv 1/\Delta r$ for each bin visited, the latter is computed summing half the distance between two consecutive points of the trajectory for all bins between x_i and x_{i+1}.

Thus, Eq. 3.33 serves as validation criteria for a putative coordinate approaching the committor via a variational approach, for example. One simply needs to compute $Z_{C,1}$ for different sampling intervals and see how close it is to the constant value. For putative RCs that deviate from the optimal committor, which we call suboptimal RCs, $Z_{C,1}(r, \Delta t) > N_{AB}$. In fact, for large Δt, some transitions between boundaries may not be detectable due to the low temporal resolution (i.e., the system can visit a boundary and leave it during Δt) so that $Z_{C,1}(q, \Delta t)$ would start to decrease for high enough Δt. However, this issue can be simply rectified by considering the ensemble of transition path segments for computing $Z_{C,1}(q, \Delta t)$ [17]. We explain the implementation of this important technical detail in Appendix A. It is worth mentioning that we will always use such construction for computing the generalized cut-based free-energy profiles, as in this case it is straightforward to apply the validation criteria, since $Z_{C,1}(q, \Delta t)$ is always constant.

In addition, a useful relation can be derived between $Z_{C,1}(r, \Delta t)$ and the diffusion coefficient $D(r, \Delta t)$ along that (now generic) coordinate r [47],

$$D(r, \Delta t) = \frac{Z_{C,1}(r, \Delta t)}{\Delta t Z_H(r, \Delta t)}. \tag{3.34}$$

Eq. 3.34 is derived assuming that the stochastic process along r is diffusive so that the diffusion coefficient should be independent of Δt, as it is for normal (Markovian) diffusion satisfying Eq. 3.5; for details on the derivation of Eq. 3.34, see Ref. [47]. Nevertheless, we can extend it for a general coordinate where deviations of this behavior, i.e., diffusion coefficients varying with Δt, are associated with non-Markovian effects. The more $Z_{C,1}(r, \Delta t)$ varies with Δt, or equivalently $D(r, \Delta t)$ since $Z_H(r, \Delta t) \sim \Delta t^{-1}$, the stronger the non-Markovian effects [47]. On the other hand, $Z_{C,1}(q, \Delta t)$ is independent of Δt for the committor, and so is $D(q, \Delta t)$ according to Eq. 3.34. This indicates that the

projection along the committor minimizes non-Markovian effects, even though it may not be strictly Markovian [43].

3.3.2 Diffusive model along the committor

It is common to associate a one-dimensional projected RC, $r(\vec{X})$, with a 1D diffusive model of the kind described by Eq. 3.5 or 3.8 along r. This can be justified under certain approximations; see, e.g., Ref. [13], where it is done for general stochastic Markov-chain simulations using energy as RC. Besides, such a description has been employed in different contexts, such as protein folding using the number of native contacts as RC [51], membrane transport over a channel with the displacement along the channel as RC [41], and aggregation phenomena, using either the cluster size [52, 53] or energy [26, 54] as RCs. But, in general, the projections along those physically motivated RCs are non-Markovian due to the loss of information associated with dimensionality reduction [15, 16]. Besides, as we will show in this work, neither energy nor the cluster size can be used to quantitatively describe the aggregating systems considered here at arbitrary short time scales. An alternative is to consider more sophisticated stochastic models that describe the non-Markovianity using a generalized Langevin equation with a memory kernel, which can be derived from the Mori-Zwanzig formalism [55]. However, computing the kernel can be challenging in practical situations. And, even if it can be computed, one loses the simple interpretation of diffusion on a potential landscape, described only by the $F_H(r)$ (or $p_{\text{eq}}(r)$) and $D(r)$, which can be readily computed from the coordinate time series via Eq. 3.26-3.27 and 3.34, respectively. The other option, which we will consider here, is to use the committor as an optimal RC. In this case, the simple diffusive description can be used to compute exactly important kinetic quantities using the diffusive model along the committor, as has been demonstrated in a series of recent works [43].

The first kinetic quantity that can be computed exactly with the committor projection is the equilibrium flux J_{AB} between two arbitrary disconnected regions, A and B, in the original multidimensional configurational space. It equals the number of transitions between regions A and B per unit time $J_{AB} = N_{AB}/(\mathcal{Z}\Delta t_0)$ with $\mathcal{Z} = \int Z_H(q, \Delta t_0)\mathrm{d}q$ the normalization constant as in Eq. 3.25 and 3.28, which is just the trajectory length sampled with[4] $\Delta t = \Delta t_0$. From the 1D diffusive model along the committor q, the flux can be computed exactly as [18]

$$J_{AB} = D(q)p_{\text{eq}}(q) = D(q)Z_H(q)/\mathcal{Z}. \tag{3.35}$$

[4] In the following we set the sampling interval dependence of all quantities equal to a given reference Δt_0 and drop the notation about this dependence for functions such as $Z_{C,1}(q, \Delta t_0)$ and $D(q, \Delta t_0)$ because all of those quantities are indeed independent of Δt for the committor.

Note that, using Eq. 3.34, the above equation can be rewritten as

$$J_{AB} = \frac{Z_{C,1}(q)}{\mathcal{Z}\Delta t_0} = \frac{N_{AB}}{\mathcal{Z}\Delta t_0} \tag{3.36}$$

where it is clear that the optimality criteria Eq. 3.33 is equivalent to the fact that the diffusive model along q yields the exact value for J_{AB} via Eq.3.35.

Secondly, the MFPT from A to B can also be computed exactly using the diffusive model along q [56], that is, it can be computed similar to Eq. 3.17 as

$$\tau_{A\to B} = \int_0^1 \frac{\mathrm{d}q}{p_{\mathrm{eq}}(q)D(q)} \int_0^q p_{\mathrm{eq}}(q')\mathrm{d}q' \tag{3.37}$$

but now for a multidimensional system with an arbitrarily complex potential surface $V(\vec{X})$. This MFPT from A to B can also be expressed exactly in terms of the equilibrium mean of the reverse committor $\langle(1-q)\rangle \equiv \int(1-q)p_{\mathrm{eq}}(q)\mathrm{d}q$, as [37, 57]

$$\tau_{A\to B} = \frac{\langle 1-q\rangle}{J_{AB}}. \tag{3.38}$$

Similarly, the reverse MFPT from region B to A can be evaluated as

$$\tau_{B\to A} = \int_0^1 \frac{\mathrm{d}q}{p_{\mathrm{eq}}(q)D(q)} \int_q^1 p_{\mathrm{eq}}(q')\mathrm{d}q' = \frac{\langle q\rangle}{J_{AB}}. \tag{3.39}$$

Finally, the MTPT $\hat{\tau}$, or mean time the system takes for transitioning between A and B without returning to the initial boundary, is exactly computed similar to Eq. 3.19 as [42, 43]

$$\hat{\tau} = \frac{\int_0^1 p_{\mathrm{eq}}(q)\left(\int_0^q \frac{\mathrm{d}q'}{p_{\mathrm{eq}}(q')D(q')}\right)\left(\int_q^1 \frac{\mathrm{d}q'}{p_{\mathrm{eq}}(q')D(q')}\right)\mathrm{d}q}{\int_0^1 \frac{\mathrm{d}q}{p_{\mathrm{eq}}(q)D(q)}} = \frac{\langle q(1-q)\rangle}{J_{AB}}. \tag{3.40}$$

Note that $q(\vec{X})$ is defined under boundary conditions determined by the regions A and B (Eq. 3.23) so that Eq.3.37-3.40 are indeed valid for arbitrary regions; one just needs to redefine the boundary states also for the committor. Of course, in such a case, the committor itself should be recomputed. This can be done more easily if the two new boundaries A' and B' are defined by two points a and b on the (original) committor landscape $q(\vec{X}) = a, b$ $(a < b)$: A' contains all points such that $q < a$ and B' contains all points such that $q > b$. The new committor q' can be simply computed by rescaling the original q as $q' = (q-a)/(b-a)$ for $a < q < b$, $q' = 0$ for $q < a$, and $q' = 1$ for $q > b$ [19, 58]. Besides, the difference between $\tau_{A\to B}$ (Eq. 3.38) and $\tau_{B\to A}$ (Eq. 3.39) arises in the definitions of A and B. For example, exchanging the roles of A and B, one has $q \to (1-q)$ so that the values of both MFPTs are just interchanged, as should be. Here

we will always use the definition $q(\vec{X} \in A) = 0$ and $q(\vec{X} \in B) = 1$. On the other hand, it is clear from Eq. 3.40 that $\hat{\tau}$ is independent of these definitions since it is obtained by the product $q(1-q)$ so that it should be independent of whether the system is going from A to B or vice versa, as it was rigorously demonstrated in Ref. [41].

Eq.3.35-3.40 illustrate the usefulness of the committor RC to describe complex systems with many degrees of freedom. The original multidimensional dynamics on a landscape represented by $V(\vec{X})$ is simplified to a one-dimensional description. More than that, this 1D description is arguably the simplest possible, i.e., diffusive motion along a 1D free-energy landscape. Although that 1D projection is not an exact diffusive model, as the dynamics projected on q do not exactly satisfy a Brownian-like SDE (Eq. 3.5) or equivalently a Smoluchowski-like PDE (Eq. 3.8), the above kinetic properties are computed exactly assuming the diffusive model. It is of particular interest for systems presenting bistability, such as protein folding, chemical reactions, and first-order phase transitions in general. In those cases, the configurational space exhibits two basins that discriminate between states with different properties, e.g., folded and unfolded for proteins, aggregated and diluted phases in first-order aggregation transition, among others. In this way the regions A and B discriminate such states, and the committor serves not only as an order parameter but also describes the kinetics of phase transformation in a simple way.

3.3.3 Coordinate transformations and unitary diffusion coefficient

A diffusive model along a generic RC is characterized by two quantities: the free-energy profile $F_H(r)$ and the diffusion coefficient $D(r, \Delta t)$. But as soon as both $F_H(r)$ and $D(r, \Delta t)$ are known, it is possible to perform a transformation of coordinates to obtain a rescaled coordinate $\tilde{r}$ where the diffusion coefficient is unitary $D(\tilde{r}, \Delta t) = 1$. This is convenient since for the rescaled coordinate the free-energy profile alone contains all kinetic information. Such a transformation can always be performed for one-dimensional diffusive models [59, 60]. This is a consequence of the fact that the cut-based free-energy profile $F_{C,0}$ (or $Z_{C,0}$) is invariant under an arbitrary continuous invertible transformation of the coordinate space [59]. Similar to $Z_{C,1}$, $Z_{C,0}$ can also be related to the diffusion coefficient [59]

$$D(r, \Delta t) = \frac{\pi}{\Delta t}\left[\frac{Z_{C,0}(r, \Delta t)}{Z_H(r, \Delta t)}\right]^2. \tag{3.41}$$

Particularly, we consider the transformation $\tilde{q} = \tilde{q}(q)$ with q being the committor and $\tilde{q}$ being the rescaled coordinate where $D(\tilde{q}, \Delta t) = 1$. Due to its invariance, $Z_{C,0}(q, \Delta t) = Z_{C,0}(\tilde{q}, \Delta t)$. On the other hand, Z_H is not invariant under such transformation since it satisfies [59] $Z_H(q, \Delta t)\mathrm{d}q = Z_H(\tilde{q}, \Delta t)\mathrm{d}\tilde{q}$. Thus, one can write:

$$\frac{\mathrm{d}\tilde{q}}{\mathrm{d}q} = \frac{Z_H(q, \Delta t)}{Z_H(\tilde{q}, \Delta t)} \tag{3.42}$$

and, from Eq. 3.41 with $D(\tilde{q}, \Delta t) = 1$ at any particular Δt, one can replace $Z_H(\tilde{q}, \Delta t)$ by $Z_{C,0}(\tilde{q}, \Delta t)$, which differ only by a constant factor for the $\tilde{q}$ coordinate, so that

$$\frac{\mathrm{d}\tilde{q}}{\mathrm{d}q} = \sqrt{\frac{\Delta t}{\pi}} \frac{Z_H(q, \Delta t)}{Z_{C,0}(\tilde{q}, \Delta t)}. \tag{3.43}$$

Finally, using the invariance $Z_{C,0}(\tilde{q}, \Delta t) = Z_{C,0}(q, \Delta t)$, the rescaled coordinate can be obtained as:

$$\tilde{q}(q) = \sqrt{\frac{\Delta t}{\pi}} \int_0^q \frac{Z_H(q', \Delta t)}{Z_{C,0}(q', \Delta t)} \mathrm{d}q'. \tag{3.44}$$

This last result can be rewritten using the relations between $Z_{C,0}$ and $Z_{C,1}$ with the diffusion coefficient, Eqs. 3.41 and 3.34, respectively, as

$$\tilde{q}(q) = \sqrt{\Delta t} \int_0^q \sqrt{\frac{Z_H(q', \Delta t)}{Z_{C,1}(q', \Delta t)}} \mathrm{d}q', \tag{3.45}$$

from which one easily sees that, because $Z_H(q, \Delta t) \sim \Delta t^{-1}$ and $Z_{C,1}(q, \Delta t)$ is in fact independent of Δt, the transformation is also independent of Δt for the optimal coordinate q. On the other hand, for suboptimal coordinates, such transformation will depend on Δt as it is also the diffusion coefficient. As a rule, we will always present the free-energy profiles as a function of the rescaled coordinates throughout this work so that the kinetic description is facilitated.

3.3.4 Markov state model (MSM)

In general, the PDE describing an stochastic process can be approximated by a discrete jump process described by a master equation [36]. In particular, a Markovian process, such as the one described by the Smoluchowski equation, can be approximated by a discrete jump process known as a Markov State Model (MSM) [43]. In that description one assumes a suitable discretization of the configuration space where a state is denoted by the index i, and the master equation describes the time evolution of the conditional probability $P_{i|i_0}(t|t_0)$ of the system being found in state i at time t given that it was in state i_0 at the initial time t_0, which can be written as [35, 43]

$$P_i(t + \Delta t) = \sum_j P_{ij}(\Delta t) P_j(t), \tag{3.46}$$

where we consider a stationary process where the conditional probabilities depend only on the time difference so that we can fix $t_0 = 0$ and drop the dependence of the initial condition ($P_{i|i_0}(t|t_0) \equiv P_i(t)$) in the same way we did for the PDFs in the case of continuous processes. We employ the case of discrete time so that, given a suitable time interval Δt, Eq.3.46 expresses the probabilities $P_i(t + \Delta t)$ at the next time $(t + \Delta t)$ as functions of

the probabilities $P_i(t)$ at current time t. $P_{ij}(\Delta t)$ is the probability of transitioning from state j to i after Δt. It can be computed as $P_{ij}(\Delta t) = n_{ij}(\Delta t)/n_j$, counting the number of transitions $n_{ij}(\Delta t)$ from j to i after Δt and the equilibrium number of times, n_j, state j is visited, if a long trajectory sampling sufficiently all nodes i is available. We assume that an equilibrium probability P_i^{eq} independent of time is reached in the long time limit (similar to Eq.3.7) so that we have $P_i^{\text{eq}} = \sum_j P_{ij}(\Delta t) P_j^{\text{eq}}$ and $n_i \propto P_i^{\text{eq}}$.

Eq. 3.46 is derived from fundamental properties of Markov processes [35] and can be viewed as the discrete version of the Smoluchowski equation, e.g., Eq. 3.21. The fact that Eq. 3.46 depends on the quantities $P_{ij}(\Delta t)$, which can be computed directly from trajectories time series if such suitable discretization can be performed, rather than the complicated multidimensional function $V(\vec{X})$ indicates an advantage of this formalism compared to that derived from Eq. 3.21. Indeed, the MSM formalism has become popular in recent decades and was successfully applied to a variety of systems [61]. However, we note that it is also susceptible to the so-called "curse of dimensionality", as for the highly dimensional case it becomes almost impossible to sample all states i in order to compute P_{ij} with sufficient accuracy. Given that, we anticipate that there will be no need to construct the MSM for describing the multimolecular systems considered in this work. Nevertheless, this theoretical construction allows one to express and operate with quantities that are computed directly from trajectories, following the formalism derived in Refs. [22, 43, 62], which is of particular usefulness for computing committor functions.

Given the MSM defined by the master equation Eq.3.46, the committor q_i at any state i is computed by solving the system of linear equations [63, 64]

$$q_i = \sum_j P_{ji}(\Delta t) q_j \tag{3.47}$$

subject to the boundary conditions

$$q_A = 0, \tag{3.48a}$$

$$q_B = 1, \tag{3.48b}$$

which are the MSM analogues of Eq. 3.22-3.23. Although $P_{ij}(\Delta t)$ clearly depends on Δt, the committor itself is independent of Δt, i.e., Eq. 3.47 is valid for any timescale [17]. The above committor equations can also be formulated in a variational way; the committor function is obtained by minimizing the total squared displacement (TSD) [17, 62]

$$\min_{q_i} \sum_{ij} n_{ij}(\Delta t)(q_i - q_j)^2. \tag{3.49}$$

It can be easily shown by direct differentiation (plus the detailed balance condition $n_{ij} = n_{ji}$) that Eq. 3.49 is equivalent to Eq. 3.47-3.48 for the same boundary conditions.

Eq. 3.49 can be viewed as the MSM analogue of Eq. 3.24. It can be reformulated directly in terms of the RC equilibrium time series $r(k\Delta t) = r(\vec{X}(k\Delta t))$ as defined in Sec. 3.3.1, or equivalently $r(k\Delta t) = r_i(k\Delta t)$ in the notation of the MSM. In any case, the RC is a function of configurational space ($\vec{X}$ or i), and its time series are the values of this projection at different times $k\Delta t$ along a trajectory. The variational principle can then be rewritten as [62]

$$\min_r \sum_k [r(k\Delta t + \Delta t) - r(k\Delta t)]^2, \tag{3.50}$$

where the sum runs all over the trajectory and with the same boundary conditions

$$r(\vec{X} \in A) = 0, \tag{3.51a}$$

$$r(\vec{X} \in B) = 1, \tag{3.51b}$$

Eq. 3.50-3.51 express the committor variational principle in a simple way in terms of RC time series: among all possible reaction coordinates time series $r(k\Delta t)$ satisfying the specified boundary conditions Eq.3.51, the committor time series $q(k\Delta t)$ is the one that minimizes the TSD. The theoretical minimum of the TSD, which we denote as Δr^2 for brevity, is attained when $r = q$ and corresponds to $\Delta q^2 = 2N_{AB}$ [17]. This expected result can be used as an indication of the validity of the committor estimates obtained during optimization schemes. A more stringent validation criterion is, however, the $Z_{C,1}$ test described in Sec.3.3.1, since it states that $Z_{C,1}(r, \Delta t)$ is constant (equals N_{AB}) for $r = q$ and thus allows one to inspect "spatial" (r) and temporal dependencies [58]. Both criteria are interrelated, as it can be shown [17] that $\Delta r^2(\Delta t) = 2\int Z_{C,1}(r, \Delta t)\mathrm{d}r$, that is, the TSD criteria is averaged over r. Specifying for the committor, one sees that both criteria are related:

$$\Delta q^2(\Delta t) = 2\int_0^1 Z_{C,1}(q, \Delta t)\mathrm{d}q = 2N_{AB}. \tag{3.52}$$

In practice, the variational principle can be used in the following way to numerically construct the committor RC. One can define a functional parametric form in terms of the RC $r = r(\vec{X}, \vec{\alpha})$, where in addition to the dependence on some features of the configurational space $\vec{X}$, one also allows the RC to depend on several parameters denoted by $\vec{\alpha}$. Then, one can numerically optimize that functional form, minimizing 3.50 with respect to $\vec{\alpha}$. That strategy was employed for the protein folding problem in Ref. [47], where the RC was defined in terms of the number of native contacts between atoms in the protein (the $\vec{X}$ dependence). To determine whether a native contact is formed or not, a set of distance thresholds needs to be defined, which was used as the set of parameters $\vec{\alpha}$ in Ref. [47]. The resulting coordinate provides a much better kinetic description compared with commonly physically motivated RCs, but it is still suboptimal as tested through the $Z_{C,1}$ validation criteria [17]. This illustrates the challenges of this kind of parametric optimization scheme; if the functional form $r = r(\vec{X}, \vec{\alpha})$ is too restrictive, it

will not be possible to reach the global minimum in Eq. 3.50. Besides that, selecting such a functional form likely demands extensive expertise within the system. Artificial neuron networks have risen as candidates to approximate complicated multidimensional functions and have been applied to the committor optimization problem [46, 48, 49] due to their flexible functional form. On the other hand, Eq. 3.50 indicates that one can deal directly with the RC time series, avoiding the difficult task of selecting the full functional form $q(\vec{X})$, and instead computing directly the committor time series $q(k\Delta t)$, from which the diffusive model can be constructed and used to exactly compute kinetic properties. This approach was introduced in Ref. [22], and we will extend it for multimolecular aggregating systems in Chapter 4.

As a final remark, we note that although the MSM can be viewed as a discrete approximation of the multidimensional PDE for diffusing interacting particles, where the continuous treatment looks more appropriate, it has the advantage of being applicable for any Markovian process. In particular, the optimization scheme of Eq. 3.50 remains valid for any equilibrium trajectory (obeying detailed balance) that can be described as a Markovian stochastic process at a suitable timescale Δt, with trajectories satisfying Brownian dynamics described by the multidimensional Smoluchowski equation as a particular case. It is also important to note that, at a more microscopic level, a molecular system might be well described by the deterministic Newtonian equations. In this case, and also for systems described by the full (i.e., not overdamped) Langevin equations, the dynamics is Markovian on phase space rather than configurational space [62]. That is, the multidimensional process consisting of the momenta and positions together composes a Markovian process, not the process defined by the positions alone, which is Markovian only in the overdamped limit. In this way, for describing those systems with the formalism we employ here considering only the configurational space, one should consider a sufficiently high temporal scale Δt where the system loses the memory about the momenta, i.e., the temporal scale where the Smoluchowski equation holds. Fortunately, for many molecular systems, such a suitable Δt is rather low. For example, the analysis has been successfully applied for atomistic Newtonian simulations of protein folding at typical timescales where the trajectories are stored, which are usually already sufficiently higher than the simulation time step, where the non-Markovianity in configurational space should appear; see e.g., Ref. [47, 58]. Importantly, such timescales are much smaller (up to orders of magnitude) than the ones necessary for a description using physically motivated RCs to become approximately Markovian, as demonstrated in Ref. [47, 58] for the folding problem and as we shall illustrate for the aggregating phenomena in this work. In this way, the diffusive models along the committors we were able to compute allow for a description with much higher temporal resolution when compared with most physically motivated RCs.

Chapter 4

The variational nonparametric approach

Related publication

Nonparametric determination of the committor in multimolecular systems [23]
L. F. Trugilho, S. Auer, L. G. Rizzi, S. V. Krivov
J. Chem. Theory Comput. **21**, 10080 (2025)

Now that we have introduced the theoretical formalism needed for constructing the diffusive model along the optimal committor coordinate, we turn to the problem of obtaining such an RC in practice for multimolecular systems, with the starting point being the variational principle of Eq.3.50-3.51. We assume as input a long equilibrium multidimensional trajectory of the system's configurations denoted by $\vec{X}(k\Delta t)$, i.e., the positions of each molecule at each time $k\Delta t$. Close to the transition temperature such trajectory transits between the boundary states A and B, e.g., aggregated and diluted states.

4.1 The nonparametric iterative method

As described in Ref. [22], a suitable way of implementing Eq.3.50 is in the following. First, one defines a seed coordinate time series $r_0(k\Delta t) = r_0(\vec{X}(k\Delta t))$ respecting the boundary conditions 3.51. To do so, one usually defines an order parameter (OP), which should separate states A and B. In Sec. 4.3 we will detail the procedures we used for that, but for now we assume that such determination of boundaries is feasible. In this way we can define $r_0(\vec{X} \in A) = 0$ and $r_0(\vec{X} \in B) = 1$ in accordance with Eq. 3.51. The remaining trajectory points can be simply initiated with a constant value between zero and one, e.g., $r_0 = 1/2$. Thus, the time series of the seed coordinate $r_0(k\Delta t)$ is completely prepared.

The nonparametric method is then implemented by iteratively updating the putative RC time series $r_m(k\Delta t) = r_m(\vec{X}(k\Delta t))$ at the m-th iteration as

$$r_{m+1}(\vec{X}(k\Delta t)) = r_m(\vec{X}(k\Delta t)) + \delta r(\vec{X}(k\Delta t)), \tag{4.1}$$

where $\delta r(\vec{X}(k\Delta t))$ denote a variation of the putative RC time series, which is a function of the configurational trajectory at the m-th iteration.

The variation is taken of the form of a low-dimensional l-th degree polynomial,

$$\delta r(r_m, x_m) = \tilde{I}_b \sum_{i=0}^{l} \sum_{j=0}^{l-i} a_{ij} (r_m)^i (x_m)^j, \tag{4.2}$$

where we used a shortened notation for the involved time series $r_m \equiv r_m(k\Delta t) = r_m(\vec{X}(k\Delta t))$. Also, $x_m \equiv x_m(k\Delta t) = x_m(\vec{X}(k\Delta t))$ represents the time series of a randomly chosen coordinate of $\vec{X}$ at the m-th iteration. The set of coordinates $\{x_m\}$ plays a pivotal role in the optimization scheme as they carry the dynamical information from $\vec{X}$, which is then "mixed" with the current putative RC at each iteration step by Eq. 4.1-4.2, and will be detailed in Sec. 4.2. $\tilde{I}_b = \tilde{I}_b(k\Delta t)$ is an indicator function, which is zero at the boundaries A and B and one otherwise, which only ensures that the variation is null on the boundaries. Finally, the polynomial coefficients a_{ij} are chosen at each step following the variational principle (Eq. 3.50) like

$$\min_{\{a_{ij}\}} \left\{ \sum_k [r_{m+1}(k\Delta t + \Delta t) - r_{m+1}(k\Delta t)]^2 \right\}, \tag{4.3}$$

with the sum running over all trajectory steps. In this way the putative RC is progressively optimized at each iteration step, and it is expected that the RC becomes close to the committor. Closeness to the committor can be inferred by the value of the TSD computed like in Eq. 3.50 at each iteration, and, more rigorously with the $Z_{C,1}$ validation criteria Eq. 3.33, which is applied usually after the iterative process converges. In practice, we found that it may be useful to modulate the variation by a common envelope function (which is multiplied at Eq. 4.2) in order to focus optimization at different values of the RC. In particular, we used sigmoid functions $1/(1 + e^{\pm(r_m - \hat{r}_m)/b})$ to focus optimization either for values $r_m \gtrsim \hat{r}_m$ or $r_m \lesssim \hat{r}_m$ at each iteration where $b = 0.01$ (used in this work) defines the scale. This is easily implemented by drawing $\hat{r}_m$ and the sign of the sigmoid function periodically after some iterations are performed. The reason for this choice is to alternate the focus of optimization between the two basins (A and B), where optimization is known to shrink [58]. We also note that for this kind of variation, the coefficients a_{ij} can be computed exactly in an efficient way; the explicit equations can be found at Ref. [62].

It is important to stress the fundamental difference of the method described above

from the more common approaches, which aim to parameterize the full multidimensional function $q(\vec{X})$. Since the present method deals directly with the trajectories updated at each iteration, there is no need to approximate the full function $q(\vec{X})$ in a parameterized way; in this sense the present method is nonparametric. One directly computes the committor time series $q(k\Delta t)$, from which the diffusive model is readily constructed by means of Eq. 3.27 for the free-energy profile and Eq. 3.34 for the diffusion coefficient. Parametric approaches have the advantage that they allow for directly evaluating the committor values for new configurations or trajectories once function $q(\vec{X})$ is determined at the training phase, e.g., using an artificial neuron network [46, 49, 48]. This ability can, however, be undermined when dealing with systems' dynamics dominated by rare events and/or with extremely high-dimensional configurational space. In those cases, the critical hypothesis that the sampled configurations used in training are representative of the entire configurational space [65] is likely to break [66]. On the other hand, the nonparametric approach does not rely on that hypothesis because it focuses on computing the committor for the actually sampled configurations that belong to the trajectory. Moreover, the nonparametric method is expected to be more flexible, requiring less system-specific expertise, as there are no assumptions about the functional form of $q(\vec{X})$, at the expense of requiring rerunning optimization once new trajectories are available [66].

What remains is to define the set of coordinates $\{x_m\}$. Such a set is required to fulfill two main requisites. First, they should carry the necessary information about the system's dynamics and about the possible conformations on the multidimensional landscape. In particular, it should be possible to distinguish between the boundary states A and B and to track the progress of the system while transitioning between them. To fulfill this requirement, one usually works with coordinates constructed from the positions of all the system's constituents, e.g., atoms, molecules, which determine by definition the space $\vec{X}$. But one cannot use directly the Cartesian coordinates of each particle due to the second requisite, which requires that those coordinates respect all the system's symmetries [58]. The reason for this second requisite is that any resulting RC should be invariant under operations respecting the physical symmetries. For example, if a system not subjected to external forces is translated, its state has not changed at all, and so should be any RC. And as the RC is constructed from the set $\{x_m\}$ through Eq. 4.1-4.2, the easiest way of fulfilling this physical requirement is to ensure that all coordinates in $\{x_m\}$ respect such symmetries.

The nonparametric approach has been successfully applied to particle-based unimolecular systems, specifically protein folding simulations [58, 67]. Those molecular systems obey translational and rotational symmetries. In this way, one can simply define $\{x_m\}$ as the set of collective variables (CVs) determined by the pairwise distance between particles in the protein chain, since all distances between parts of the chain are invariant under translations and rotations. At each step of the iteration, one simply randomly

chooses the particle pair to use the related distance time series as x_m in Eq. 4.2 [58, 67]. Such a set of CVs is clearly invariant under the involved symmetry while preserving detailed information of the configurational space, as the distances are computed directly from all particles's positions. A particular goal of the present work is to extend this methodology for multimolecular or many-body systems, in particular for the aggregating systems that we will consider as benchmarks. In those cases there is one additional symmetry requirement, as the system is also invariant under the permutation of identical particles. That is because exchanging two identical particles do not affect at all the state of the system so that both configurations must have the same RC. The strategy employed is detailed in Sec. 4.2.

4.2 Permutationally invariant CVs

Here we present the method developed in the current work to construct an informative set of CVs invariant under the symmetries respected by a general many-body system, namely translational, rotational, and permutational symmetries. In particular, such a set can be used as $\{x_m\}$ for optimizing the committor RC. In a broader sense, CV discovery has also been an important tool for describing complex system dynamics, particularly those dominated by rare events, where CV-based biasing simulation techniques are popular [68]. To ensure translational and rotational symmetries, we follow commonly used approaches and work with pairwise distances between the constituents [58, 67]. Permutational symmetry has been more involved to ensure, and there are two main strategies [68]. One can sum all the involved features (distances here) over all identical particles (or pairs in this case) of some configuration so that the resulting CV is invariant under particle exchange. A second option is to perform a permutationally invariant ordering; for example, one can sort all the pairwise distances occurring in some configuration of the system. Although the distances are computed considering specific particle pairs, all sorted distances, i.e., the smallest, second smallest, and so on, are CVs invariant under permutation of particles. In any of those cases, the time evolution of such CVs gives its time series.

The former option is likely to be more coarse-grained, in the sense that it sums all distances into only one CV. Moreover, ours tests indicate that a set with only one CV is unlikely to provide the necessary flexibility for optimizing a complex function such as the committor. That can be improved by summing functions of the pairwise distances with free parameters that, when varied, yield a collection of CVs, one for each value of the parameter. We tried such an approach for defining the set of CVs $\{x_m\}$ to be used in the committor optimization (Eq. 4.2) for some of our aggregating model systems. As shown in Appendix B, the $F_{C,1}$ validation test indicates that the resulting putative committor RC was suboptimal, we believe due to the coarse-grained character of those CVs. Slightly better committor estimates were obtained using sorted distances. In contrast

with the former case, one now has a huge number of CVs, namely $N(N-1)/2$ of such sorted distances. This is an advantage of the ordering case, since a set of CVs composed of all those distances certainly carries much more detailed information about the system's configurational space, as well as its dynamics. But some difficulties also arise due to the huge number of CVs. One cannot deal with all of those time series simultaneously due to hardware (memory) limitations. And it is also difficult to identify the more informative ones to work with out of that bunch of CVs distributed over a 1D-like structure with $N(N-1)/2$ entries. On the other hand, we noticed that considering a small but representative set of those CVs led to slightly better committor estimates, while still suboptimal as we show in Appendix B. Inspired by that, we implemented a more informative way of sorting distances.

We considered the symmetric matrix $\mathbf{d}$ whose $N \times N$ elements are the distances between any two molecules [1]. i and j, d_{ij}. First, the two-dimensional matrix is sorted along one of its axes (see example below). Mathematically, the elements of the new matrix $\hat{\mathbf{d}}$, after sorting along the first index, have the property $0 = \hat{d}_{1j} < \hat{d}_{2j}, ..., < \hat{d}_{Nj}$ for all j ranging from 1 to N. Each molecule corresponds to a specific index j, and the related column contains the distance of the j-th labeled molecule to all others in a sorted way. As these columns are molecule dependent, none of the distances inside them are invariant under permutations. Even so, one can sort it again along the remaining index in order to get a completely sorted matrix $\bar{\mathbf{d}}$ where, in addition to the previous inequality, it also has $\bar{d}_{i1} < \bar{d}_{i2}, ..., < \bar{d}_{iN}$ for every i defining a row. Now, neither row i nor column j is molecule dependent, and all the N^2 distances of the resulting matrix are invariant under permutations. So, by repeating this procedure for each distance matrix at each time step, the time series of each matrix element $d_{ij}(k\Delta t)$ is an invariant CV.

Although we have virtually the same huge number of CVs in this new way of sorting compared to the more common previous one, the new one turned out to produce more informative CVs and be more suitable for the optimization procedure. That is because one can work with one axis of $\bar{\mathbf{d}}$ during some iteration steps but then subsequently continue the optimization with other axes to improve the committor time series estimate while avoiding memory issues. Note that for performing such ordering, the full matrix needs to be considered, but one can store only one axis at a time. When considering the next axis, one just needs to repeat the procedure. As far as our tests went, and as we will demonstrate later, by using about ten axes, we were able to construct the putative optimized RCs, which closely approach the committor for the considered model systems, as measured by the $F_{C,1}$ validation test. Importantly, for the lattice system that we consider in Sec. 5.1, it was not possible to achieve that degree of accuracy with the more usual way of computing permutationally invariant CVs described before (i.e., summing and usual ordering), as we

[1] If the system is simulated using periodic (spatial) boundary conditions, the distances d_{ij} should be calculated taking this into account, i.e., using the minimum image convention [69]

detail in the Appendix B. Moreover, the proposed way of sorting turns out to be more instructive since it allows comparing axes to find which are the best ones. For example, we noted that axes with a much higher index (higher than about $N/3$) are much less informative, as using them gives much worse committor estimates, so that one readily rules out a considerable fraction of the CVs. This can be physically explained since much higher axes contain only the highest distances between molecules, which contain mostly information about diluted molecules, but not about the aggregated ones, so these CVs are not expected to distinguish well the diluted and aggregated states of the system and thus to track the transitions between them.

To illustrate the procedure, let us consider a hypothetical system with only $N = 5$ molecules, with the symmetric matrix of distances between molecules $\mathbf{d}$ given by

$$\mathbf{d} = \begin{bmatrix} 0 & 101 & 89 & 52 & 38 \\ 101 & 0 & 12 & 50 & 62 \\ 89 & 12 & 0 & 39 & 51 \\ 52 & 50 & 39 & 0 & 14 \\ 38 & 62 & 51 & 14 & 0 \end{bmatrix}. \tag{4.4}$$

Now, sorting inside each column one gets $\hat{\mathbf{d}}$:

$$\hat{\mathbf{d}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 38 & 12 & 12 & 14 & 14 \\ 52 & 50 & 39 & 39 & 38 \\ 89 & 62 & 51 & 50 & 51 \\ 101 & 101 & 89 & 52 & 62 \end{bmatrix}. \tag{4.5}$$

Finally, sorting inside each row one gets the final completely sorted matrix $\bar{\mathbf{d}}$:

$$\bar{\mathbf{d}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 12 & 12 & 14 & 14 & 38 \\ 38 & 39 & 39 & 50 & 52 \\ 50 & 51 & 51 & 62 & 89 \\ 52 & 62 & 89 & 101 & 101 \end{bmatrix}. \tag{4.6}$$

Taking fixed positions, say $\bar{d}_{2,5}$ here, and repeating the sorting procedure for every frame in the trajectory, one obtains $\bar{d}_{2,5}(k\Delta t)$, a permutationally invariant CV time series.

We expect that the above procedure may be useful for computing a set of invariant CVs for an extensive class of systems where the pairwise distances carry sufficient information about the configurational space and dynamics. This is likely to be true for molecular aggregation, as we will demonstrate in this work, but maybe also for other many-body self-assembly processes, such as micellization [10]. In the case of colloidal and

macromolecular aggregation, we will show here that such a set of CVs is suitable for the RC optimization $\{x_m\}$ implemented through Eq. 4.1-4.2, but it might also be useful for other tasks related to CV discovering [68]. Finally, it is worth mentioning that despite it being non-trivial to identify a suitable set of CVs to be used in nonparametric optimization for an arbitrary system, it is still much less challenging than conventional parametric approaches [47, 50]. Besides the identification of CVs, the implementation of parametric approaches also requires the multidimensional mapping from those features to the low-dimensional RC, e.g., $q(\vec{X})$ for the committor, which is arguably a much harder task. While artificial neuron networks have emerged as putative candidates for performing such mapping [46, 48, 49], most of the architectures still need to be supplied with invariant CVs, as most architectures are blind with respect to physical symmetries. Methods for ensuring such symmetries at the architecture level have been developed more recently [68], so that one would supply the neuron network directly with the Cartesian coordinates; see e.g., Refs. [70, 71] where such networks were applied to the committor problem. However, none of those methods have yet provided such high precision (as measured by stringent validation criteria) for approximating the committor in multimolecular systems, as we provide in this work.

4.3 Boundary conditions and soft committor

In this section, we give more details about the task of defining the boundary conditions, as our approach takes them as inputs. In particular, we define a way of computing an order parameter (OP) that takes explicitly the original multidimensional dynamics into account, which we call the soft committor function.

The usual way of defining such boundary conditions is to define an OP by physical intuition. The size of the largest cluster n and energy E are the most common physically motivated OPs for aggregation phenomena. One can then define states A and B based on the free-energy profile along the OP ($\mathcal{O}$), which should be bimodal with each minimum corresponding to the different macroscopic states. For example, one selects a value of the OP, $\mathcal{O}_A$, near the first minimum such that trajectory frames with $\mathcal{O} < \mathcal{O}_A$ are assumed to belong to the aggregated state A. Similarly, one selects $\mathcal{O}_B$ near the second minimum such that trajectory frames respecting $\mathcal{O} > \mathcal{O}_B$ belong to the diluted state B (we will always assume $\mathcal{O}_B > \mathcal{O}_A$). This is illustrated for both the potential energy E and the (negative of the) size of the largest aggregate n in Fig. 4.1 (for the lattice system, which we will consider in Sec. 5.1). We use the convention of negative n because in this way state B defined by $(-n > -n_B)$, i.e., smaller cluster sizes in a diluted state, is similar to the state B defined by energy $E > E_B$, since there is an attractive interaction between particles so that the higher the energy lesser interacting are the molecules. Similarly, configutations with $(E < E_A)$ or $(-n < -n_A)$ should denote the presence of larger aggregates in the

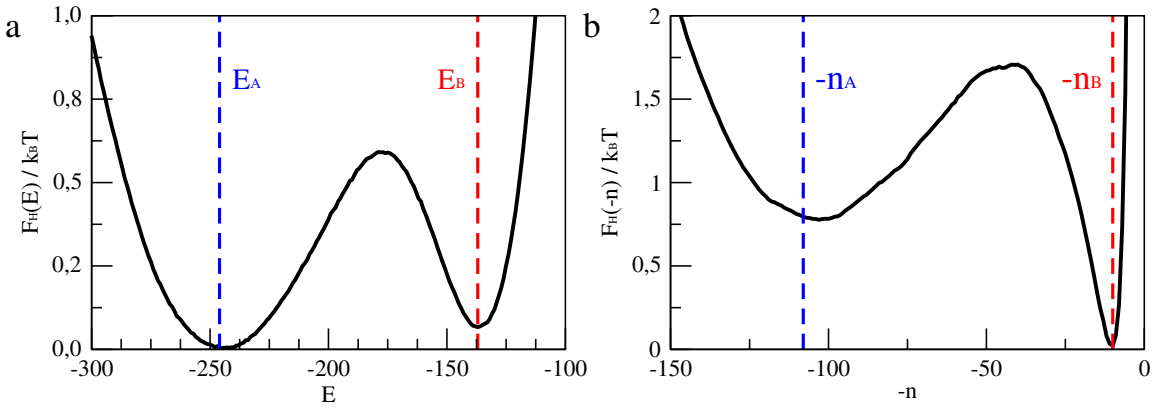

Figure 4.1: Free-energy profile along two common OPs for aggregating systems. (a) For the energy E, all configurations with $E < E_A$ define the aggregated state A, and all states $E > E_B$ define the diluted state B. (b) For the size of the biggest aggregate n, all configurations with $-n < -n_A$ define the aggregated state A, and all states $-n > -n_B$ define the diluted state B.

system.

Although the above-described procedure looks straightforward, it can be far from trivial for complex multidimensional systems. As an example, for the lattice model considered in Sec. 5.1, such a strategy generates somewhat different states A and B when E or n are taken as OPs, although the free-energy profiles along both clearly display a bimodal behavior, as can be seen in Fig. 4.1. The number of measured transitions N_{AB} along a trajectory, in particular, might differ by a significant amount depending on the chosen OP. Thus, we analyze the situation in more detail here. Consider a system described by the MSM drafted in Fig. 4.2 with a few nodes. Each point in that figure corresponds to a node i that represents a suitable discretization of the configurational space. The system can move through those nodes with probabilities defined by P_{ij} represented by the connections between the points. Then, the committor function can be found by solving Eq. 3.47 under the boundary conditions $q_A = 0$ and $q_B = 1$. In the top panel of Fig. 4.2, we illustrate this situation, which we call "rigid" boundary conditions, since the selected A and B nodes are assigned the exact boundary values. The situation described before with the order parameters is of this type, since all points with $\mathcal{O} < \mathcal{O}_A$ respect this boundary condition, i.e., all the microscopic states defined by $\mathcal{O} < \mathcal{O}_A$ are merged in a single node A such that $q_A = 0$, and, similarly, all the microscopic states with $\mathcal{O} > \mathcal{O}_B$ are merged in a single state with $q_B = 1$. This situation determines the usual committor function defined earlier, e.g., Eq. 3.47-3.48.

However, one can alternatively define "soft" boundary conditions illustrated in the bottom panel of Fig. 4.2. Instead of assigning one (or a collection of nodes) to the boundary states, one can define regions (denoted by A' and B' in the figure) of configurational space containing some nodes, and that will not serve as "rigid" boundaries. Instead, all nodes inside regions A' and B' are connected respectively to the artificial nodes

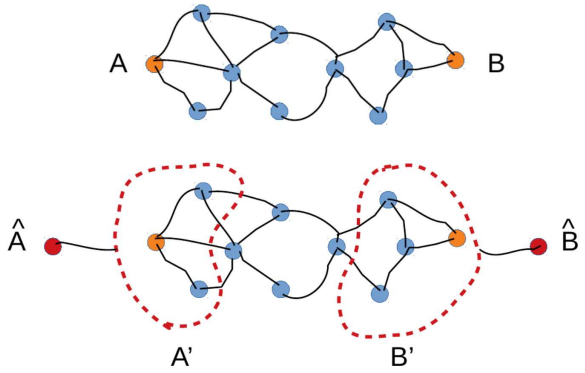

Figure 4.2: Schematic representation of rigid (top panel) and soft (bottom panel) boundary conditions defined for an MSM. The points represent the nodes of such MSM, while the connections between them denote possible transitions with non-zero probabilities Pij. Figure extracted from Ref. [23].

$\hat{A}$ and $\hat{B}$, which will indeed serve as the boundaries ($q_{\hat{A}} = 0$ and $q_{\hat{B}} = 1$), as represented in Fig. 4.2. Note that the nodes inside regions A' and B' are not merged together as they do not define the boundaries themselves; in particular, transitions between them are still considered. The boundaries $\hat{A}$ and $\hat{B}$ are called artificial since they are not present in the original network. They are defined by the free parameters $\lambda_A = P_{\hat{A}A'}$ and $\lambda_B = P_{\hat{B}B'}$, which establish the connection between the nodes inside A' (or B') to the boundaries $\hat{A}$ (or $\hat{B}$). Thus, Eq. 3.47-3.48 are defined in the same manner, but for the extended network. The solution is what we call the soft committor function, which is dependent on λ_A and λ_B. The aim of this construction is to determine an improved OP, which stratifies nodes based on their closeness to the boundaries and can be used to determine boundary conditions in a more robust manner. Regions A' and B' can be chosen with a usual OP that just weakly separates boundary states, since they do not represent the final choice of boundaries. In fact, our tests for the lattice system indicated that the constructed soft committor based on either E of n for defining the regions A' and B' fixes the issue related to observing different number of transitions.

In practical, soft committor can be estimated with a similar variational approach to the one derived for the usual committor in Sec. 4.1. One can use the same iterative process defined by Eq. 4.1 with equivalent variations like in Eq. 4.2. The only difference is that the a_{ij} coefficients are determined by minimizing the TSD for the extended network, which reads

$$\min_{\{a_{ij}\}}\Bigg\{\sum_{k}\Big\{[r_{m+1}(k\Delta t+\Delta t)-r_{m+1}(k\Delta t)]^2 \\ +2I_{A'}(k\Delta t)\lambda_A[r_{m+1}(k\Delta t)]^2 \\ +2I_{B'}(k\Delta t)\lambda_B[r_{m+1}(k\Delta t)-1]^2\Big\}\Bigg\} \tag{4.7}$$

where k again runs over all the time series. The difference lies in the two additional terms related to the artificial transitions between nodes inside regions A' or B' and their corresponding artificial nodes now defining the boundaries. $I_{A'}$ and $I_{B'}$ are indicator functions, which can be defined using any usual OP that just roughly separates the boundaries as

$$I_{A'}(k\Delta t)=\begin{cases}1, & \mathcal{O}(k\Delta t)<\mathcal{O}_{A'}\\ 0, & \mathcal{O}(k\Delta t)>\mathcal{O}_{A'}\end{cases} \tag{4.8}$$

and, analogously for $I_{B'}$, that is, $I_{B'}=1$ for $\mathcal{O}>\mathcal{O}_B$, or zero, otherwise. In this way, when summing up over all the trajectory in Eq. 4.7, these two additional terms select the number of times the system enters regions A' and B', which, multiplied by λ_A and λ_B, respectively, results in the number of artificial transitions to each boundary. The squared displacement related to those transitions is just $[r_{m+1}(k\Delta t)-0]^2$ for transitions to boundary $\hat{A}$ ($q_{\hat{A}}=0$) and $[r_{m+1}(k\Delta t)-1]^2$ for transitions to boundary $\hat{B}$ ($q_{\hat{B}}=1$). The factor of two accounts for the fact that we consider transitions in both directions $n_{A'\hat{A}}=n_{\hat{A}A'}$ to preserve detailed balance, and similarly for boundary $\hat{B}$, that is, $n_{B'\hat{B}}=n_{\hat{B}B'}$.

It is worth noting that the soft committor time series can then be constructed using the same set of permutationally invariant CVs as $\{x_m\}$. It can be used, in turn, as an OP to fix the boundaries in a "rigid"-like way for optimizing the usual committor function via Eq. 4.1-4.3. This OP is expected to better separate boundary states since it takes into account explicitly the multidimensional configurational space through $\{x_m\}$. Moreover, it is optimized in an unconstrained way, as all transitions between nodes are accounted for in the first term of 4.7, so the resulting OP stratifies all nodes, including those close to the multidimensional basins. Indeed, as we find that, while directly using E or n as OPs results in rather different states A and B, the constructed soft committor OPs based on either E or n to define the regions A' and B' separate almost equivalently the boundary states (i.e., with almost the same number of transitions between boundaries), indicating the robustness of the soft committor construction.

Finally, we note that similar soft boundary conditions have been used for parametric optimization schemes, where it is usually convenient to convert the original constrained variational formulation (usual committor definition) to an unconstrained optimization

problem (soft-committor-like); see e.g., Ref. [50]. Those approaches, however, employ this strategy as a mathematical trick to directly approximate the usual committor function. In contrast, the methodology described here aims to construct a better OP, from which the boundary conditions of the usual committor function can be determined more robustly subsequently.

Chapter 5

Applications to model systems

Related publication

Nonparametric determination of the committor in multimolecular systems [23]
L. F. Trugilho, S. Auer, L. G. Rizzi, S. V. Krivov
J. Chem. Theory Comput. **21**, 10080 (2025)

In this chapter, we apply the formalism of the nonparametric approach of Chapter 4 for computing the committor functions in two systems, which are prototype models for aggregation. First, in Sec. 5.1, we consider Monte Carlo (MC) simulations of a 2D lattice gas with anisotropic interactions. Then, in Sec. 5.2, we consider Brownian simulations of a 3D Lennard-Jones system. This results are also published in Ref. [23]. In both cases the committor functions are obtained almost exactly, i.e., within statistical error, which is a rare achievement in the field of committor estimation, particularly for many-body multimolecular systems. In those systems, the committor is a high-dimensional function, and most methods for estimating it are challenged by the curse of dimensionality, i.e., it becomes almost impossible to sample the full configurational space, let alone solve the high dimensional partial differential equation defining the committor (Eq. 3.22). Besides, in systems undergoing transitions between two boundaries, the dynamics is generally dominated by rare events, which further complicates the task. By applying our method for two different multimolecular systems in which one observes fluctuation-driven rare aggregation events related to a first-order phase transitions, we hope to showcase the robustness and generality of the approach. Particularly, the systems are embedded in different spatial dimensions, and the dynamics are simulated with different protocols. Moreover, we illustrate how the estimated committor provides a simple and precise kinetic description of the considered systems, as already predicted by theory.

The different protocols for simulating each system's dynamics are worth some attention. Brownian dynamics (BD) is usually more appealing for kinetic studies since it simulates a real model of the dynamics between particles. That is, one explicitly numerically integrates Eq. 3.20 to obtain the temporal evolution of the positions of each

particle, which interact with each other and with an implicit solvent that acts as a thermal bath at a fixed temperature. Moreover, this is precisely the kind of system where all the theoretical frameworks derived in Chapter 3 are expected to hold in an exact manner. That is because Eq. 3.20 are indeed the equations of motion from which all the framework can be derived, e.g., the Smoluchowski equation, the committor definition. In this way, a system simulated by BD represents an ideal benchmark system for testing the variational approach of Chapter 4 for computing the committor. In contrast, the Monte Carlo (MC) protocol is designed to sample the system's configurational space in a way that it approaches some target equilibrium distribution in the long time limit. To do so, an arbitrary stochastic dynamics can be employed, which is not necessarily related to the system's real dynamics. Under specific conditions, however, it can be viewed as the microscopic model for generating the system dynamics. In any case, the applicability of our approach for a system driven by such stochastic dynamics highlights the generality of the method. But there is more. As we discussed in Sec. 3.3.4, the theoretical formalism of Chapter 3 can be translated to a discrete version valid for general Markov State Models respecting detailed balance. In addition, the majority of MC simulations are indeed constructed by Markov chains respecting detailed balance. Thus, most MC simulations (including the ones we consider here) can be seen as particular cases in the broader context of the MSM formalism.

5.1 Monte Carlo simulations of an anisotropic lattice model

Here we consider Metropolis Monte Carlo simulations of a lattice gas system with N particles/molecules disposed on a 2D square lattice of size L. Although MC simulations do not necessarily describe the real dynamics of the system as mentioned earlier, there is evidence that it can be used to infer the real dynamics if only local physical movements are used [72, 73]. In this case the Metropolis scheme can be viewed as a way of generating the highly-dimensional dynamics of the system with aggregating particles in the canonical ensemble [74]. In this way, we implemented the stochastic dynamics by local Kawasaki movements [75], where the particles can move only to their nearest neighbor unoccupied sites.

5.1.1 Model

The lattice model is defined as in our previous work [76]. Besides their positions in the lattice, each of the N particles has an extra degree of freedom that determines their orientations, which can be of two types, as represented by horizontal or vertical arrows in Fig. 5.1. Only nearest neighbor molecules attract each other, and the magnitude of the

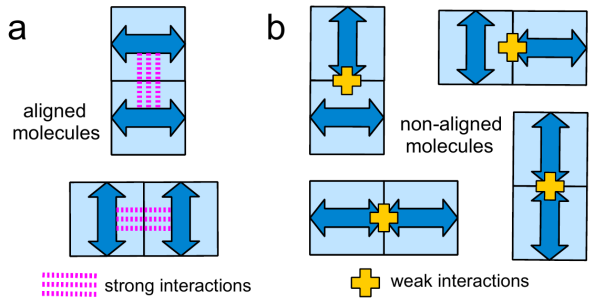

Figure 5.1: Illustration of the possible interacting configurations between two nearest neighbor molecules in the 2D lattice model. The molecules are represented by squares with arrows that indicate their orientation. (a) Examples of relative orientations that lead to stronger attraction, aligned molecules. (b) Examples of relative orientations that lead to weaker attraction, non-aligned molecules. Figure extracted from Ref. [76].

interaction can be stronger ψ_s or weaker ψ_w depending on the relative orientation of the pair, as detailed in Fig. 5.1. In this way, the anisotropy of the interaction can be defined by the factor $\xi \equiv \psi_s/\psi_w$. Particularly, we fixed ψ_w to unity, which determines the energy scale, and ψ_s is a multiple ($\xi \geq 1$) of the former. There is also excluded volume, so two or more particles cannot occupy the same site simultaneously. We define then a time step as an MC sweep, which corresponds to N attempts to rotate (change orientation) and move (to randomly selected neighbors) the particles in the lattice. Each particular attempt (rotate or move) is accepted or rejected according to the usual Metropolis algorithm [77]. In this way, the equilibrium distribution of those systems corresponds to the canonical ensemble.

This model was proposed to investigate the influence of the described anisotropic interactions in the aggregation phenomena, and it is similar to the ones studied in Ref. [78, 79, 80, 81]. The motivation is the aggregation of proteins into highly oriented fibrillar states, known as amyloid fibrils [12]. In those systems, interactions of different magnitudes are present, such as stronger directional hydrogen bonds and weaker hydrophobic interactions [82].

Simulation details

All results considered here were obtained at a fixed concentration of $\rho \equiv N/L^2 = 10^{-2}$. At this concentration, the model undergoes first-order transitions at temperatures close to T^*. The values of T^* for each ξ were determined in Ref. [76] and are tabulated in the Appendix C. For the results presented next, we consider systems with $L = 200$ containing $N = 400$ molecules at $T = T^*(\xi)$. The data production consisted of six simulations of length 10^8 MC steps (MCs) for each considered ξ. Configurations were saved every $\Delta t_0 = 400$MCs, which thus corresponds to the highest temporal resolution available. In the Appendix C, we also show additional results for larger systems (up to

$N = 1024$) that corroborate the main conclusions presented here. All systems considered were simulated using periodic boundary conditions.

5.1.2 Isotropic case

We first consider the isotropic case ($\xi = 1$) as a detailed exemplification for applying our approach. As the methodology employed for the anisotropic systems is almost identical, we combine the results obtained for systems with different ξ more directly at Section 5.1.3.

Soft committor vs. physically motivated OP

In Sec. 4.3 we have already anticipated that, regarding the lattice model, different choices of physically motivated OPs, namely energy E and the size of the largest aggregate n, resulted in somewhat different states A (aggregated) and B (diluted). Because of that we computed the soft committor q_s to serve as an OP, which is expected to be more robust, as discussed in Sec. 4.3. It tuned out that the states separated by the soft committor OP are more close to the states separated by n. Because of that and also because n is more used as OP and RC for describing aggregating systems than E, we decided to perform a comparison between n and our computed commitor q. Here we illustrate the computation of the soft committor and show how it compares to n as OPs.

Now we turn to the soft committor construction. First, we need a seed OP to define the regions A' and B' like in Fig. 4.2. We could use n for that, but to illustrate the generality of the construction, we pick one out of the $N \times N$ elements of the sorted matrix $\bar{\mathbf{d}}$ defining the set $\{x_m\}$ of permutationally invariant CVs, namely $\bar{d}_{7,190}$. The exact values of the row and column indices are not important; the CV just needs to roughly separate aggregated and diluted states, which can be inspected by checking for the appearance of bimodal curves in histogram-based free-energy profile along the CV. As can be seen in Fig. 5.2, two regions can be readily identified from the histogram-based profile $F_H(\bar{d}_{7,190})/k_BT$ defined for $\bar{d}_{7,190}$. Regions A' and B' were defined then according to the marked regions in Fig. 5.2. In fact, one also has plenty of flexibility for defining such regions. For example, even choosing one unique divider near the top of the barrier, such that all small values of $\bar{d}_{7,190}$ are considered in A' and the large values of $\bar{d}_{7,190}$ are in B' (i.e., every point on the trajectory will belong to either A' or B') results in a soft committor OP that separates similar states (data not shown).

Having defined the regions A' and B', we performed the soft committor variational scheme (Eq. 4.7) using only the seventh axis $\bar{d}_{7,j}$ of the sorted matrix as the input for the randomly chosen CVs $\{x_m\}$ at the m-th iteration step. This was sufficient to get a soft committor OP, q_s, that separates well aggregated A and diluted B states, which is shown in Fig. 5.3(a). In that figure we show the results for the rescaled $\tilde{q}_s$ coordinate where the

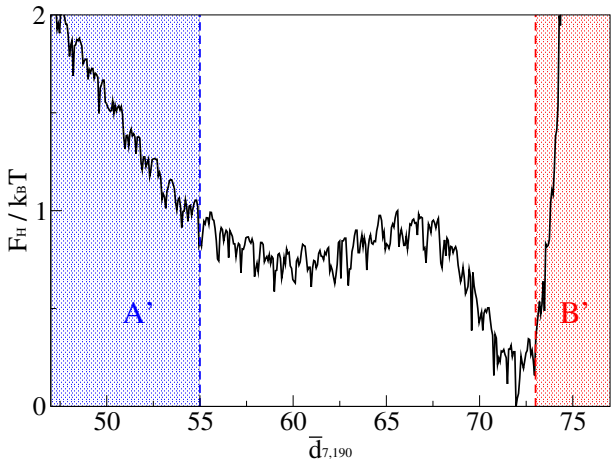

Figure 5.2: Free-energy profile along a selected CV $\bar{d}_{7,190}$ extracted from the sorted matrix. Regions A' and B' as indicated were used for the subsequent computation of the soft committor OP.

diffusion coefficient is unitary $D(\tilde{q}_s, \Delta t_0) = 1$, as discussed in Sec. 3.3.3. We also show in Fig. 5.3(b) the free-energy profile along the (negative of) the n. $\tilde{q}_s$ and n are the OPs that we used to define the boundary states A a and B as indicated in Fig. 5.3. $\tilde{q}_s$ was used for defining the boundaries for the proposed committor function that we computed, while we used directly n for defining boundaries for the n-based coordinate that we define below in order to compare both coordinates. Note that the soft committor $\tilde{q}_s$ is computed with the unconstrained variational scheme with soft boundary conditions (Eq. 4.7), but then the committor functions q are always defined employing $\tilde{q}_s$ as an OP to define usual rigid-like boundary conditions, since q is defined by the constrained variational scheme (Eq. 4.1-4.3).

Committor vs. physically motivated RC

To perform the comparison between our proposed coordinate and the physically motivated n coordinate, it is instructive to perform the one-dimensional transformation $n \to q(n)$, where the estimate for the committor $q(n)$ is evaluated ignoring any other degrees of freedom. This can be done in at least three ways. First, one can numerically integrate Eq. 3.16 for the desired coordinate n. Note that the equilibrium PDF $p_{\text{eq}}(n)$ (or equivalently the free-energy profile $F_H(n)$) and diffusion coefficients $D(n, \Delta t_0)$ can be straightforwardly computed from the coordinate time series $n(k\Delta t)$ obtained from the simulations via Eq. 3.27 and Eq. 3.34, respectively. Second, one could discretize the coordinate (In fact, n is already discrete) and construct an MSM, i.e., compute the transition probabilities $P_{nn'}(\Delta t_0)$ from $n(k\Delta t)$, which is also straightforward in one dimension [83]. Then one can easily solve Eq. 3.47 for the committor. As a third option, one can perform the nonparametric optimization scheme employing the time series $n(k\Delta t)$ as the unique set $\{x_m\}$ of CVs in Eq. 4.1-4.3. In this way, the optimization will converge to $q(n)$ as

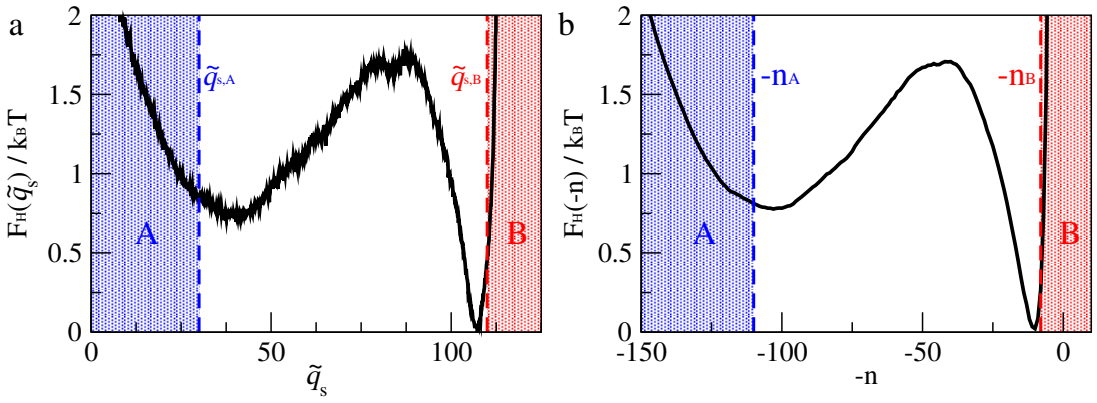

Figure 5.3: Free-energy profiles along the two OPs. (a) The soft committor $\tilde{q}_s$ was used for defining boundary conditions for our proposed coordinate $q = q(\vec{X})$. $\tilde{q}_s$ were computed via the optimization scheme of Eq. 4.7 with $m = 4 \times 10^3$ iteration steps, employing as the set of CVs the elements of the seventh axis $\bar{d}_{7,j}$ of the sorted matrix, and with the regions A' and B' as defined as in Fig. 5.2. Then, regions A (with $\tilde{q}_s < \tilde{q}_{s,A} = 30$) and B (with $\tilde{q}_s > \tilde{q}_{s,B} = 110$) are used to determine the boundary states for the full committor function that we show next. (b) The n coordinate was used for defining the boundaries for the n-based coordinate. Then, regions A (with $-n < -n_A = -110$) and B (with $-n > -n_B = -8$) define the boundaries for the n-based RC.

there is no additional information available to improve the RC optimization procedure. Our tests indicate that all options resulted in similar results for the function $q(n)$ and in the following we consider the results obtained from the third one. In any case, it is worth noting that $q(n)$ represents an estimate of the committor obtained assuming that the original coordinate n is already described by a Markovian process. That is, it is associated with a 1D Brownian-like diffusive process in the first option or with an MSM in the second one. If the assumption that the projection over n is described by a Markovian process is true, n would be already an optimal coordinate and $q(n)$ would closely approaches the full multidimensional committor $q = q(\vec{X})$. This can be easily checked with the $F_{C,1}$ validation criteria described in Sec. 3.3.1. That is, one computes $F_{C,1}(q(n), \Delta t)$ for various Δt, which should be independent of q and Δt for the true committor. This procedure can be used to check the optimality of any putative RC [22].

Committor and validation criteria

Now we turn to the computation of the committor function $q \equiv q(\vec{X})$ and compare it with the n-based coordinate $q(n)$. Both RCs are obtained via the optimization scheme Eq. 4.1-4.3. $q(n)$ was obtained using only $n(k\Delta t)$ as the set of CVs $\{x_m\}$, while $q \equiv q(\vec{X})$ was obtained by performing the optimization using eight axes of the sorted matrix uniformly distributed between $\bar{d}_{7,j}$ and $\bar{d}_{42,j}$. In the first case, the optimization rapidly converges after about $m = 20$ iteration steps. For the bigger set of CVs based on the sorted matrix, the scheme converged after about $m = 10^4$ iteration steps for each axis, which

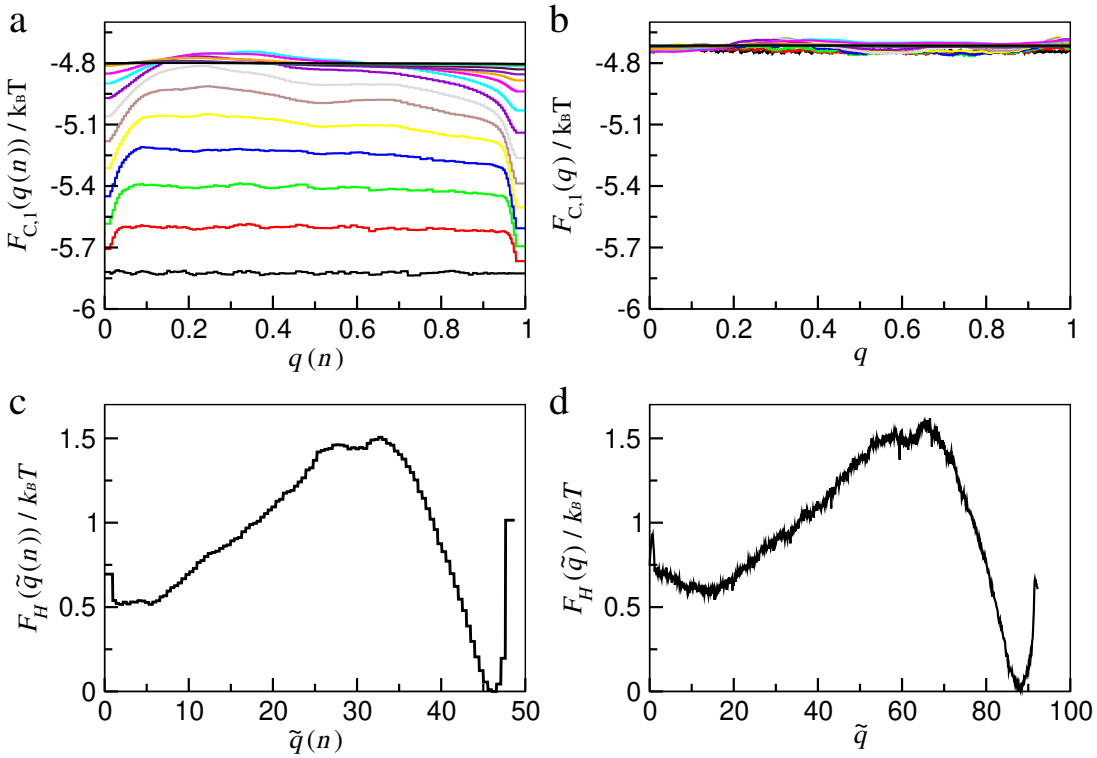

Figure 5.4: Generalized $F_{C,1}(r, \Delta t)/k_BT$ and histogram-based $F_H(r)/k_BT$ free-energy profiles for the n-based RC $q(n)$ and the optimized committor $q = q(\vec{X})$. Panels (a) and (b) show the $F_{C,1}$-based committor validation tests for both RCs, where different colors correspond to different $\Delta t = 2^i \Delta t_0$, with $i = 0, 1, 2, ..., 15$ (from the bottom to the top). In panels (c) and (d), the usual profiles are presented as a function of the rescaled RCs where $D(r, \Delta t_0) = 1$. Figure adapted from Ref. [23]

took only about 230 seconds for each axis using an RTX 460 GPU. The more demanding part of our protocol is the computation of the sorted matrix itself, which takes about 30 minutes for a system with $N = 400$ particles. In the way we implemented it, that needs to be recomputed for each axis so that one need not store multiple axes time series and avoid memory limitations, but that can be avoided depending on the memory available. Boundary states A and B were defined directly using n as OP for $q(n)$ and using the soft committor OP for the proposed q, as indicated in Fig. 5.3.

In Fig. 5.4(a) we show the $F_{C,1}$ validation criteria applied to the n-based coordinate. As can be seen in that figure, the profiles change considerably with sampling interval Δt. $F_{C,1}(q(n)\Delta t)/k_BT$ increases from about -5.8 to -4.8 as Δt increases from Δt_0 (black curve) to $2^{15}\Delta t_0$. Since the transformation n to $q(n)$ does not alter the optimality of the RC [17] the n reaction coordinate did not pass the validation test discussed in Sec 3.3.1. On the other hand, it can be seen in Fig. 5.4(b) that our proposed coordinate successfully passes the validation test, as the profiles $F_{C,1}(q, \Delta t)/k_BT$ just fluctuate around the expected value, that is, $F_{C,1}(q, \Delta t)/k_BT \approx -\ln N_{AB} \approx -4.72$ for all considered Δt, with the statistical uncertainty roughly estimated as $1/\sqrt{2N_{AB}} \approx 0.07$. Based on Eq. 3.34, one

has, for the q RC that the diffusion coefficient is indeed independent of Δt, just like it is for simple diffusion. In contrast, the fact that $F_{C,1}(q(n), \Delta t)$ changes with Δt indicates that, for the n-based coordinate, $D(q(n), \Delta t)$ should change considerably with sampling interval, signaling strong non-Markovian effects. In particular, kinetic quantities computed through a diffusive model over n will depend on Δt. For now our aim is to analyze the kinetics of the system at the highest temporal resolution available, Δt_0, to compare the RCs as rigorously as possible. Even so, as it can be seen in Fig. 5.4(a), at longer sampling intervals the generalized profiles converge to the expected values, indicating a way to perform a suitable separation of timescales for recovering kinetic information from suboptimal coordinates, like n or E. This will be further examined in Part III.

In addition, in Fig. 5.4(c) and (d), we show the conventional histogram-based free-energy profiles for the rescaled $\tilde{q}(n)$ and $\tilde{q}$ coordinates, respectively. We note that although the free-energy barriers are almost of the same height, $\tilde{q}(n)$ spans a narrower range between the valleys observed in the profiles when compared to $\tilde{q}$. As these RCs are rescaled to have a unitary diffusion coefficient $D(\tilde{q}(n), \Delta t_0) = 1$ and $D(\tilde{q}, \Delta t_0) = 1$, the smaller separation $\Delta\tilde{q}(n) = \tilde{q}(n)_B - \tilde{q}(n)_A \approx 45$ on the $q(n)$ profile compared to $\Delta\tilde{q} = \tilde{q}_B - \tilde{q}_A \approx 90$ on the $\tilde{q}$ profile indicates that the kinetics predicted by the diffusive model over $\tilde{q}(n)$ is faster, as we will confirm in the following. This can alternatively be inferred from the fact that $F_{C,1}(q(n), \Delta t_0)/k_BT = -\ln Z_{C,1}(q(n), \Delta t_0)$ is lower when compared to $F_{C,1}(q, \Delta t_0)/k_BT = -\ln Z_{C,1}(q, \Delta t_0)$ at Δt_0, as can be seen in Fig. 5.4(a) and (b). Thus, we have $Z_{C,1}(q(n), \Delta t_0) > Z_{C,1}(q, \Delta t_0)$ (see Eq. 3.32) so that the diffusion coefficient is also higher for $q(n)$ at Δt_0 due to Eq. 3.34.

5.1.3 Anisotropy dependence

Similar procedures were employed to obtain the committor free-energy landscape for the anisotropic systems with $\xi = 3$, 5, and 7. Particularly, a similar number of axes of the sorted matrix were needed for the optimization step. Moreover, the best axes have a similar index range for all systems considered (i.e, $i = 5$ to 40).

Diffusive models

The fundamental appeal of the committor RC is that it allows for an exact kinetic description of the system by means of the arguably simplest stochastic process, namely 1D diffusive model. In our approach we determined directly the committor RC time series, from which the diffusive model can be straightforwardly constructed. One just needs to compute the free-energy profile and diffusion coefficient as functions of the RC, which is done via Eq. 3.27 and 3.34, respectively. In other words, one just needs to compute the usual histogram $Z_H(q)$ (Eq. 3.26) and the generalized partition function $Z_{C,1}(q, \Delta t_0)$ (Eq. 3.31). From the diffusive model, one can compute the flux J_{AB}, or equivalently the

Table 5.1: Estimates for the number of transitions N_{AB} and the MFPT $\tau_{B\to A}$ obtained from both the n-based RC $q(n)$ and the full commitor $q \equiv q(\vec{X})$ for systems with different anisotropies ξ. All times are given in units of $10^3 \times \Delta t_0$.

	$\xi = 1$	$\xi = 3$	$\xi = 5$	$\xi = 7$
N_{AB} from Eq. 3.52 with $q(n)$	336	171	161	114
N_{AB} from time series of $q(n)$	122	55	40	28
$\tau_{B\to A}$ from Eq. 3.39 with $q(n)$	1.8	3.3	3.6	6.2
$\tau_{B\to A}$ from time series of $q(n)$	4.7	9.7	13	22
N_{AB} from Eq. 3.52 with q	113	52	38	26
N_{AB} from time series of q	112	53	37	27
$\tau_{B\to A}$ from Eq. 3.39 with q	5.4	11	15	26
$\tau_{B\to A}$ from time series of q	5.3	10	14	25

number of transitions N_{AB} between aggregated and diluted states via Eq. 3.35-3.36. In fact, due to the relation between $Z_{C,1}$ and the TSD (Eq. 3.52) we computed N_{AB} directly from the TSD Δq^2 since $\Delta q^2 = 2N_{AB}$. The values of N_{AB} computed from the TSD of a generic RC Δr^2 can then be compared with the true values extracted directly from the trajectories, i.e., simply counting the number of transitions. Particularly, we did these for both the n-based coordinates $q(n)$ and our proposed committor RC q, for systems with different anisotropies ξ, as indicated in Table 5.1. We also include the estimates of the MFPT $\tau_{B\to A}$ computed from the diffusive models via Eq. 3.39 to see how they compare with the estimates obtained directly from the trajectories. The results indicate that the $q(n)$ coordinate overestimates the number of transitions by about three to four times the actual values, while the MFPTs are underestimated by a similar factor. On the other hand, for the q coordinate, both ways of estimating N_{AB} and $\tau_{B\to A}$ yield essentially the same results within statistical uncertainty, which corroborates that our coordinate closely approaches the full committor $q(\vec{X})$ for all considered systems.

As one observes from the results presented in Table 5.1 that there is a slight discrepancy between N_{AB} computed from the time series of both coordinates $q(n)$ and q. It is worth clarifying that this occurred because of the different choice of OPs for defining the boundary states A and B. Recalling that for the n-based coordinate, we used directly the n time series as OP, while for q, we used the soft committor OP. As the differences in N_{AB} computed from the times series are small, we conclude that both OPs separate similar boundary states for all considered systems. Notably, using the potential energy E as OP, we got a substantial difference between the measured N_{AB} directly from the trajectories, with the N_{AB} computed from the energy time series about two times the number of transitions obtained from the n time series for the isotropic case. This was the main reason we employed the soft committor in this work, in an attempt to define the boundaries in a more robust way. It turned out that n seems to be a better OP than E for these systems. On the other hand, performing the soft committor construction with

Table 5.2: Estimates for the MFPT $\tau_{A\to B}$ and the MTPT $\hat{\tau}$ obtained from both the n-based RC $q(n)$ and the commitor q for systems with different anisotropies ξ. All times are given in units of $10^3 \times \Delta t_0$.

	$\xi = 1$	$\xi = 3$	$\xi = 5$	$\xi = 7$
$(T^*)^{-1}$	2.73087	1.61744	1.13764	0.89222
$\tau_{A\to B}$ from Eq. 3.37 with $q(n)$	2.9	5.9	6.1	7.6
$\tau_{A\to B}$ from time series of $q(n)$	7.6	19	26	32
$\hat{\tau}$ from Eq. 3.40 with $q(n)$	0.38	0.6	0.7	1.5
$\hat{\tau}$ from time series of $q(n)$	0.94	1.5	2.6	6.1
$\tau_{A\to B}$ from Eq. 3.37 with q	7.9	18	25	31
$\tau_{A\to B}$ from time series of q	8.0	19	27	30
$\hat{\tau}$ from Eq. 3.40 with q	1.2	1.9	2.9	6.8
$\hat{\tau}$ from time series of q	1.2	1.9	2.9	8.1

either E, n, or an element of the sorted matrix (like we did here) for defining the regions A' and B' did not lead to significant discrepancies in the resulting soft committor OP, which indicates the robustness of that construction.

In addition, in Table 5.2 we show two other characteristic times, namely the MFPT from aggregated to diluted states $\tau_{A\to B}$ and the MTPT $\hat{\tau}$. Again, we show these quantities computed from the diffusive model to see how they compare with the results obtained directly from the time series for both $q(n)$ and q RCs. Similarly to what was observed for the other kinetic quantities, the diffusive models over $q(n)$ result in kinetics that are about three to four times faster than the true value computed directly from the $q(n)$ time series, while for the proposed optimum coordinate q, the estimated values are virtually the same as the values measured from the q time series. In Table 5.2 we also list the values of the inverse temperature $(T^*)^{-1}$ used in the Metropolis MC simulations for each system, in units where k_B and ψ_w are unitary. In the Appendix C (Table C.1) we show similar results for larger isotropic systems, which also confirms the superiority of the diffusive model along q. All results suggest that our constructed coordinate closely approaches the multidimensional function $q(\vec{X})$. In contrast, it is clear that the diffusive model along n does not accurately describe the kinetics of aggregation in any of the cases.

Finally, we show in Fig. 5.5 the histogram-based free-energy profiles for systems with different anisotropies ξ. All profiles display a two-state bimodal behavior characteristic of systems undergoing first-order phase transitions, which are already known to occur in those systems [76]. The profiles are shown as functions of the rescaled coordinates $\tilde{q}$, where the diffusion coefficient is unitary. In this way, besides describing equilibrium features, the profiles contain all kinetic information available from the diffusive models. In particular, we note that the free-energy barriers between aggregated and diluted states exhibit a non-monotonic behavior with ξ, attaining a maximum for $\xi = 3$. Interestingly, this characteristic is similar to the one observed for the barriers determined from energy

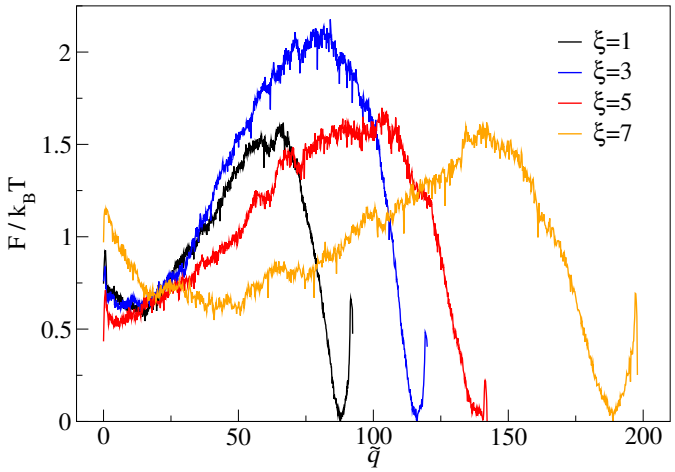

Figure 5.5: Histogram-based free-energy profiles $F_H(\tilde{q})/k_BT$ along the rescaled committor $\tilde{q}$ for different anisotropies ξ. Figure adapted from Ref. [23].

related free-energy profiles extracted directly from the density of states [76]. Nevertheless, the kinetics of the phase transformations is monotonically lowered with increasing anisotropy, as can be checked from the results in Tables 5.1 and 5.2, e.g., N_{AB} decreases [1] while $\tau_{B\to A}$, $\tau_{A\to B}$ and $\hat{\tau}$ increase with ξ. This fact is somewhat counterintuitive but can be explained by examining the profiles in more detail. Differently from the barriers, the separation between boundary states increases monotonically with ξ. And as $D(\tilde{q}) = 1$, this corresponds to a decrease in diffusivity, compensating the reduction of the barrier for $\xi > 3$, in a way that anisotropy reduces overall kinetics speed.

[1] N_{AB} can be taken directly as a measure of the phase transformation kinetics since all results presented in Table 5.1 were computed with time series of of the same length.

5.2 Brownian simulations of a Lennard-Jones system

In this section we consider Brownian simulations of a system with $N = 100$ particles interacting through a Lennard-Jones (LJ) potential in a cubic box of volume $V = L^3$ in a way that the concentration is $\rho = 10^{-2}$ in length units where the diameter of each particle is unitary. The simulation is performed by numerically integrating the overdamped SDE given by Eq. 3.20 at a fixed temperature T. Thus, similar to the simulations considered in Sec.5.1, this system is also in the canonical (NVT) ensemble, but now in three spatial dimensions.

5.2.1 Model

The particles inside the cubic box interact via a pairwise potential V_{ij} that depends only on the distance d_{ij} between the pair of particles i and j computed using the minimum image convention [69], which is of the Lennard-Jones type

$$V_{ij}(d_{ij}) = 4\epsilon \left[\left(\frac{\sigma}{d_{ij}} \right)^{2\alpha} - \left(\frac{\sigma}{d_{ij}} \right)^{\alpha} \right] \tag{5.1}$$

where σ is the diameter of each particle, i.e., the minimum distance, which is set to unity, ϵ defines the depth of the potential and energy scale (we set $\epsilon = 5$ arbitrary units of energy). The first term in Eq.5.1 corresponds to an excluded volume-like continuous repulsive term, while the second term accounts for the attraction of the Van der Waals type. For the usual Lennard-Jones potential, the exponent α is equal to six. Here we allow for the generalization known as the Mie potential [84], that allows one to modify the range of the attractive part of the potential by varying α [85]. This generalization is useful for colloidal particles where the range of attraction can be varied by adding polymers of different sizes to the colloidal suspension [86]. Here we set $\alpha = 10$ because we were following Ref. [28] that was aiming to describe the experimental data of Ref. [87]. In this way the potential $V(\vec{X})$ is written in the form we detailed in Sec. 3.3:

$$V(\vec{X}) = \sum_{i>j} V_{ij}(d_{ij}). \tag{5.2}$$

Besides interacting with each other via the potential of Eq. 5.2, the N particles are immersed in an implicit solution at temperature T with friction coefficient γ, as described in Sec. 3.3. The model is thus a prototype model for a suspension of colloidal particles with diffusion coefficient equal to $D_0 = k_B T/\gamma$.

Simulations details

The set of $3N$ couple equations, each described by Eq. 3.20, can be numerically integrated employing Euler's method [30, 88]

$$x_\mu(t+\delta t) = x_\mu(t) - \frac{D_0}{k_B T}\frac{\partial V(\vec{X}(t))}{\partial x_\mu}\delta t + \sqrt{2D_0\delta t}\mathcal{N}(0,1), \tag{5.3}$$

where x_μ corresponds to the Cartesian coordinates of all particles like in Sec. 3.3 ($\mu = 1,..,3N$) and $\mathcal{N}(0,1)$ are random numbers drawn from a Gaussian distribution with zero mean and unitary variance, which is independent for different μ and time steps. The force exerted at each particle $\partial V(\vec{X}(t))/\partial x_\mu$ at each component can be easily computed by differentiating Eq. 5.1-5.2. In practice, we defined a radius of cutoff $r_c = 4\sigma$ so that only particles closer than it r_c contribute to the sum in Eq. 5.2. We fixed the diffusion coefficient at $D_0 = 0.005$ and employed a time discretization of $\delta t = 0.01$. In this way Eq. 5.3 can be iteratively resolved to give the time series of the positions of all particles, namely the multidimensional trajectory[2] $\vec{X}(\ell\delta t)$. By testing different temperatures and looking for a bimodal stationary distribution through the energy histograms $Z_H(E, \Delta t_0)$, we found that the system undergoes a first-order transitions at a temperature $T^* = 1.92$ (assuming $k_B = 1$). In this way we perform five long equilibrium simulations at those conditions for the data production (each of length 6×10^9 integration steps). Configurations were saved at every $\Delta t_0 = 8 \times 10^3 \delta t$ integration steps, which corresponds to the highest temporal resolution available for the $Z_{C,1}$-based analyses and the determination of the optimal committor RC.

5.2.2 Committor vs physically motivated RC

Boundary conditions

We defined the boundary states A and B to perform the variational optimization procedures of Chapter 4 in a similar way that we did for the lattice systems in Sec. 5.1. In this way, boundary conditions are defined using the soft committor OP for the proposed $q \equiv q(\vec{X})$, while for the $q(n)$ coordinate, we simply use n directly as OP. Both boundary states for q and $q(n)$ are defined in usual "rigid" boundary conditions. Only the soft committor OP was computed with the unconstrained variational scheme of Eq. 4.7 with soft boundaries. As a seed coordinate to determine regions A' and B' for computing the soft committor, we again used one element of the sorted matrix, namely $\bar{d}_{14,70}$. Then, the variational scheme of Eq. 4.7 was performed using as a set of CVs $\{x_m\}$ one axis of the sorted matrix $\bar{d}_{14,j}$.

[2] We decided to use δt for the integration step, as the notation Δt is reserved for multiples of Δt_0, that is, $\Delta t = k\Delta t_0$ with $k \geq 1$, so it keeps the notation of the previous sections.

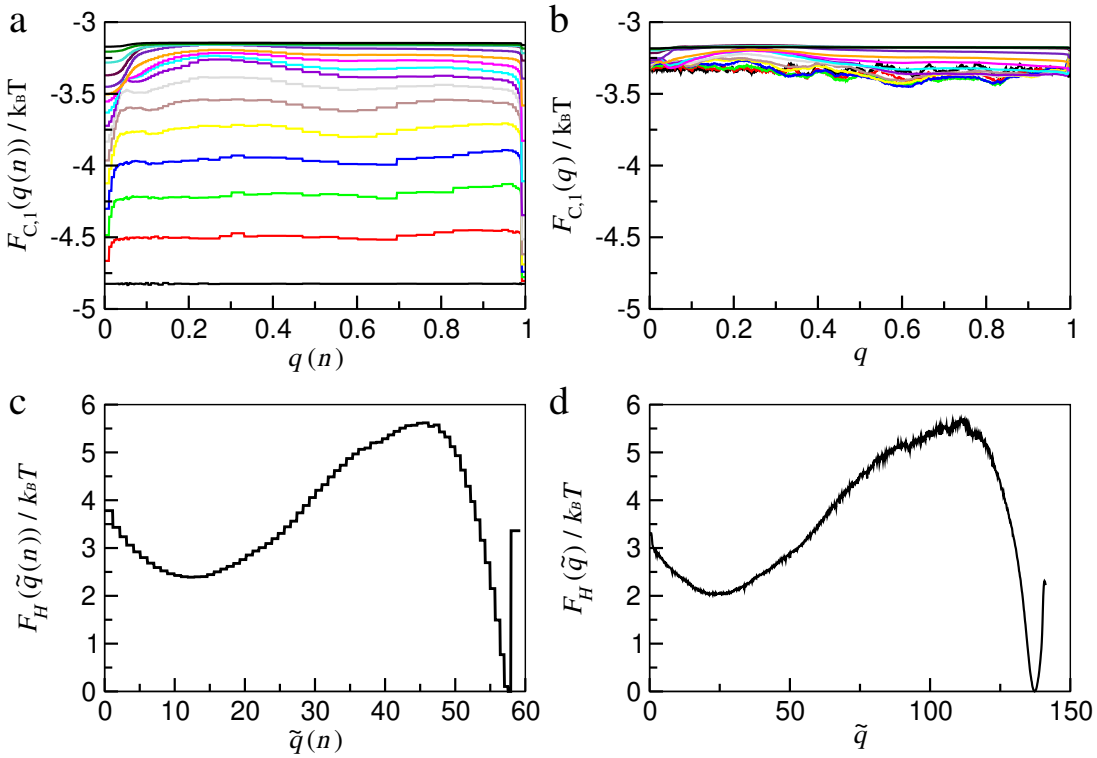

Figure 5.6: Generalized $F_{C,1}(r, \Delta t)/k_BT$ and histogram-based $F_H(r)/k_BT$ free-energy profiles for the n-based RC $q(n)$ and the optimized committor $q = q(\vec{X})$. Panels (a) and (b) show the $F_{C,1}$-based committor validation tests for both RCs, where different colors correspond to different $\Delta t = 2^i \Delta t_0$, with $i = 0, 1, 2, ..., 15$ (from the bottom to the top). In panels (c) and (d), the usual profiles are presented as a function of the rescaled RCs where $D(r, \Delta t_0) = 1$.

Committor and validation criteria

In Fig. 5.6 we show the histogram-based free-energy profiles and $Z_{C,1}$-based validation test for the $q(n)$ and q coordinates obtained for the LJ system. The $q(n)$ was again obtained by performing the committor variational approach Eq. 4.1-4.3, taking the n time series alone as the set of CV $\{x_m\}$. And for the proposed optimized coordinate, we used eight axes of the sorted matrix between $\bar{d}_{4,j}$ and $\bar{d}_{35,j}$. And again, our preliminary tests indicated that axes with much higher index give much poorer optimized coordinates, so one can readily rule out the CVs with index $\bar{d}_{ij}$, $i \gtrsim N/3$. In Fig. 5.6(a) we show the validation test for the n-based coordinate. The behavior is similar to what we got for the lattice systems. $F_{C,1}(q(n), \Delta t)/k_BT$ changes considerably with sampling interval from about -4.8 to -3.2 as Δt increases from Δt_0 (black curve) to $2^{15}\Delta t_0$ and did not satisfy the $F_{C,1}$-based committor validation criteria. Conversely, for the q RC, $F_{C,1}(q, \Delta t)$ only fluctuates near the expected value $F_{C,1}(q, \Delta t)/k_BT \approx -\ln N_{AB} \approx -3.2$ for all considered Δt, up to statistical uncertainty, roughly estimated as $1/\sqrt{2N_{AB}} \approx 0.15$.

In Fig.5.6(c) and (d), we show the free-energy profiles along the rescaled coordinate where the diffusion coefficient is unitary, $D(\tilde{q}(n), \Delta t_0) = 1$ and $D(\tilde{q}, \Delta t_0) = 1$. The

Table 5.3: Estimates for the number of transitions N_{AB}, MTPT $\hat{\tau}$, and MFPTs, $\tau_{A\to B}$ and $\tau_{B\to A}$, obtained from both the n-based RC $q(n)$ and the full commitor $q \equiv q(\vec{X})$ for the Lennard-Jones system. All times are given in units of $10^4 \times \Delta t_0$.

N_{AB} from Eq. 3.52 with $q(n)$	125	$\tau_{A\to B}$ from Eq. 3.37 with $q(n)$	1.6
N_{AB} from time series of $q(n)$	23	$\tau_{A\to B}$ from time series of $q(n)$	10
$\hat{\tau}$ from Eq. 3.40 with $q(n)$	0.11	$\tau_{B\to A}$ form Eq. 3.39 with $q(n)$	1.4
$\hat{\tau}$ from time series of $q(n)$	0.41	$\tau_{B\to A}$ from time series of $q(n)$	7.0
N_{AB} from Eq. 3.52 with q	28	$\tau_{A\to B}$ from Eq. 3.37 with q	7.1
N_{AB} from time series of q	24	$\tau_{A\to B}$ from time series of q	9.9
$\hat{\tau}$ from Eq. 3.40 with q	0.34	$\tau_{B\to A}$ form Eq. 3.39 with q	6.3
$\hat{\tau}$ from time series of q	0.37	$\tau_{B\to A}$ from time series of q	6.8

results are also very similar to the ones of the lattice systems. The smaller separation between aggregated and diluted states on the $\tilde{q}(n)$ coordinate indicates that the diffusive model over that coordinate gives faster kinetics than the obtained committor coordinate $\tilde{q}$, although the free-energy barriers are again almost equal for both RCs.

Diffusive models

Finally, we show in Table 5.3 the comparison between the kinetic properties theoretically predicted from the diffusive models, as well as their counterparts measured directly from the simulation trajectories, for both $q(n)$ and q. These results confirm that the diffusive model along n results in kinetics of about five times faster than the correct values. On the other hand, the diffusive model along the proposed optimal RC q describes accurately the kinetic properties. Although the differences between the estimated and measured properties seem slightly greater in comparison to what we got for the lattice systems, this is probably due to the worse statistics of the present case, with only $N_{AB} = 24$ transitions observed. The associated free-energy barriers are considerably higher than the lattice systems, which makes it more difficult to obtain a large number of transitions N_{AB} in the simulations. In any case, these results illustrate the successfulness of our variational optimization method when describing the phase transition kinetics in highly-dimensional systems dominated by rare events

Part III

Kinetic approach from microcanonical thermostatistics

Chapter 6

Specific objectives

In Part II, we developed an approach for computing the committor for aggregating systems. From the committor time series, a diffusive model can be straightforwardly constructed, from which important kinetic properties are computed exactly. For this reason the committor can be seen as an exact or optimal RC. Now we turn to the problem of trying to recover kinetic information from suboptimal coordinates, where the dynamics cannot be described by a simple diffusive model at arbitrary timescales. The importance of this study comes from the fact that one does not always have access to the detailed information that we assumed known in Part II. That is, for computing the committor, we have used the sorted matrix $\bar{\mathbf{d}}$, which can only be constructed if one knows the positions of every particle as a function of time (the full configurational space $\vec{X}(t)$). Although it is nowadays common to obtain that kind of trajectory from simulation data, that detailed information may be missing in various cases, such as experimental data. Specifically, we will be particularly interested in constructing a kinetic approach where energy will serve as the RC, but that might be extended for other RCs as well. From a theoretical point of view, this could be advantageous since usual thermostatistical theories usually deal with energy in a usual way. Of particular interest is the microcanonical thermostatistics that we review in Chapter 7, where energy is the natural independent variable. Then, in Chapter 8, we will detail the proposed kinetic approach, which is based on the microcanonical analysis, and apply it to the same lattice model that was considered in Part II.

Chapter 7

Microcanonical thermostatistics

Our main motivation to work with energy E as a reaction coordinate for describing the system's kinetics comes from the fact that this is a natural coordinate for thermostatiscal equilibrium theories. Standard equilibrium theories do not deal with kinetic properties. We would like then to connect usual thermostatiscal quantities, such as latent heats and free-energy barriers, with kinetic properties. This is already partially achieved by constructing a diffusive model along an arbitrary coordinate free-energy profile. In this case the kinetic description is done assuming diffusive motion on a potential that is just the equilibrium free-energy profile (FEP). It is particularly useful considering the committor as a RC because the related diffusive model is quantitatively accurate by construction, as we described in Part II. But if we could employ a diffusive model along the energy, the connection would be more direct, particularly with the microcanonical thermostatistics.

The microcanonical ensemble is described by the Boltzmann definition of the entropy $S(N, V, E)$,

$$S(N, V, E) = k_B \ln \Omega(N, V, E). \tag{7.1}$$

The usual thermodynamic extensive control parameters are the number of molecules in the system N, the volume V, and the internal energy E, which determines the macroscopic state of the system in the microcanonical ensemble. The number of microscopic states $\Omega(N, V, E)$, or microcanonical partition function, is then computed as a function of these parameters. Eq. 7.1 relates it to the entropy, building the fundamental thermodynamic relation $S(N, V, E)$, from which all equilibrium properties can be derived [89]. As we will be solely interested in the dependence of the entropy on energy, always keeping N and V fixed, we will drop their dependence from now on.

Besides providing the usual thermodynamic description from $S(E)$, the microcanonical approach has been able to describe phase transitions in finite-sized systems by means of the inflection point analysis [90] of $S(E)$ and its derivatives $b(E) = k_B^{-1}(\partial S(E)/\partial E)$ (microcanonical inverse temperature), $\gamma(E) = (\mathrm{d}b(E)/\mathrm{d}E)$, and so on. This includes the classification scheme for the order of a phase transition based on the analysis of the

derivatives of different degrees [91], in the spirit of the Ehrenfest classification scheme [92]. Moreover, the microcanonical ensemble has been supported to be the more suitable one for describing finite-sized systems [27]. One of the reasons is that fluctuations play a major role in finite systems. For example, these systems can transit between two boundary states with different properties (e.g., aggregated and diluted) exclusively due to intrinsic fluctuations. Although the microcanonical ensemble is the one defined at fixed energy, the microcanonical analysis can be used to study systems in other ensembles as well, particularly the canonical, which we will consider here, in the following sense. From the microcanonical entropy, one can compute the canonical equilibrium distribution $p_{\text{eq}}(E)$ (or the free-energy profile $F_H(E)$) at any temperature straightforwardly, as we will illustrate next. Conversely, this cannot be done in a direct way from the canonical partition function (or the Helmholtz free-energy). This is because the canonical ensemble only describes the mean values of homogeneous systems averaged out over the fluctuations as a function of the intensive variable, which is the temperature T [27]. For example, from the Helmholtz free-energy at T, one can compute the mean energy at that temperature. But one has no direct access to the energy distribution $p_{\text{eq}}(E)$. This is also valid for any other ensemble obtained by the Laplace transform of the microcanonical ensemble, exchanging one (or more) extensive variables (NVE) by the related intensive variable (chemical potential μ, pressure P, and temperature T). Even if one is committed to trying to perform the inverse Laplace transform to recover the microcanonical information from other ensembles, this is challenging due to the high sensitivity of that transform, mainly at first-order phase transitions where heterogeneous configurations (e.g., where phase separation occurs) get exponentially suppressed [27].

The most difficult task for constructing the microcanonical ensemble is the computation of $\Omega(E)$, which rarely can be done analytically for interacting systems [93]. Fortunately, various methods for estimating numerically $\Omega(E)$ with high accuracy were developed over the past three decades. Those methods are called flat-histogram algorithms, including weighted histogram [94], multicanonical [95], entropic sampling [96], broad histogram [97], Wang-Landau [98], and the ones based on statistical temperature [99, 100]. With those methods, the microcanonical equilibrium characterization of different finite-sized systems, including particle condensation [101], polymer aggregation [54] and adsorption [102], protein folding [103], and aggregation [104, 105], have been performed. In this way, kinetic properties could be readily extracted from those systems if a kinetic theory based on the energy coordinate is available.

As we will be interested in simulations on the canonical ensemble, we formalize how the equilibrium canonical distribution $p_{\text{eq}}(E;\beta)$ can be evaluated at an arbitrary temperature[1] $T = 1/k_B\beta$ from $\Omega(E)$. This is straightforwardly done by the definition of

[1]As we deal with both canonical and microcanonical variables, we set the following notation. The microcanonical quantities, like temperature $T(E)$ and its inverse $b(E) = 1/k_BT(E)$, are always written as

the canonical PDF as [93]

$$p_{\text{eq}}(E;\beta) = \frac{\Omega(E)e^{-\beta E}}{\mathcal{Z}(\beta)} = \frac{e^{-\beta(E-TS(E))}}{\mathcal{Z}(\beta)} \equiv \frac{e^{-\beta F_H(E;\beta)}}{\mathcal{Z}(\beta)} \tag{7.2}$$

where we have used Eq. 7.1, $\mathcal{Z}(\beta) = \sum_E \Omega(E)e^{-\beta E}$ is the canonical partition function, and we defined the canonical free-energy profile as

$$F_H(E;T) = E - TS(E), \tag{7.3}$$

so that we have a relation similar to Eq. 3.25. $F_H(E;\beta)$ is an energy-dependent Helmholtz-like free-energy at the specified temperature but should not be confused with the complete canonical Helmholtz free-energy which is a function of temperature only (besides N and V) and which is computed by summing over all energies ($-k_BT\ln\mathcal{Z}(\beta)$). The subscript H indicates that this FEP can also be obtained from the (canonical) histogram $Z_H(E;\beta)$ at a given temperature. But note that if $S(E)$ is known, $F_H(E;\beta)$ can be computed directly for any temperature. As it will not be necessary to compute the normalization constant $\mathcal{Z}(\beta)$ the free-energy profile is indeed defined up to an additive constant, which we can set so that its minimum is zeroed, as we have done throughout Part 4.

In Fig. 7.1(a), we show the microcanonical inverse temperature $b(E)$ for the same anisotropic lattice considered in Sec. 5.1 with $\xi = 3$. This result was obtained with the multicanonical algorithm in one of our previous work [76]. It is clear that $b(E)$ displays the S-shaped behavior characteristic of first-order phase transitions [27]. As $b(E)$ is given by

$$b(E) = \frac{1}{k_B}\left(\frac{\partial S}{\partial E}\right)_{N,V}, \tag{7.4}$$

the regions where $b(E)$ is a crescent function of E correspond to regions where the entropy is a convex function of E ($\partial^2 S/\partial E^2 > 0$), the so-called convex intruder. In the thermodynamic limit, the convex region is suppressed, converging to a linear function determined by the Maxwell construction, which implicates in phase separation. For finite-sized systems, the presence of the convex region in $S(E)$ implicates the appearance of a bimodal canonical PDF, which is the signal of first-order phase transitions in finite systems [27], in analogy with thermodynamic phase separation. An equal area Maxwell construction can be performed in $b(E)$ to obtain the (canonical) transition temperature T^*. This criterion states that the area between the curve $b(E)$ (or $T(E)$) and the straight line $b(E) = \beta^*$ (or $T(E) = T^*$) is equal for both the regions above and below the straight line, as represented in Fig. 7.1(a). It is equivalent to requiring that the canonical equilibrium distribution to display peaks of equal height near each of the maxima corresponding to both phases [106].

a function of E (the independent variable of the microcanonical ensemble). The canonical counterparts T and $\beta = 1/k_BT$ are the independent variables of the canonical ensemble and thus have no dependence. In contrast, the canonical PDF is a function of E for each β, so we write $p_{\text{eq}}(E;\beta)$.

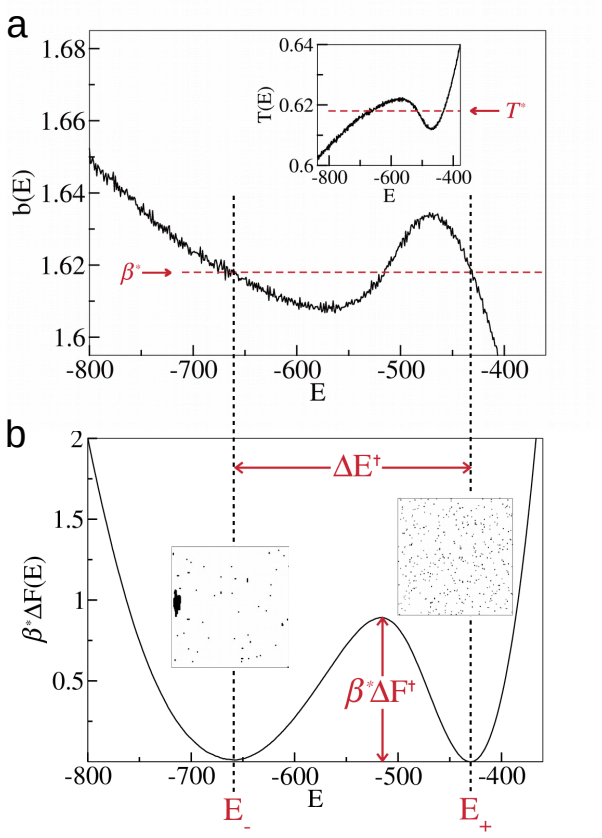

Figure 7.1: (a) Inverse of the microcanonical temperature $b(E)$ and the microcanonical temperature $T(E)$ on the inset obtained via the multicanonical algorithm. The dashed red lines indicate the values at the transition β^* and T^* evaluated by the Maxwell construction. (b) Free-energy profile obtained through Eq.7.3 at the transition temperature T^*. $\Delta E^{\ddagger}$ represents the latent heat and $\beta^*\Delta F^{\dagger}$ the free-energy barrier. The insets represent typical configurations of states near each basin close to E_a and E_d. The results were obtained for the lattice model described in Sec. 5.1 with $\xi = 3$ and $\rho = 10^{-2}$. Figure adapted from Ref. [76].

This can be observed in Fig. 7.1(b), where we show the FEP obtained through Eq. 7.3 at the transition temperature, and it can be seen that its two minima have the same depth. We reserve the notation $\beta^*\Delta F(E;\beta^*)$ for the free-energy profile at the (inverse) transition temperature $\beta^* = k_B^{-1}(S(E_d) - S(E_a))/(E_d - E_a)$, which is defined as

$$\beta^*\Delta F(E;\beta^*) = \beta^*(F_H(E,\beta^*) - F_H(E_{a(d)};\beta^*)) \tag{7.5}$$

where $E_{a(d)}$ could mean either E_a or E_d, which correspond to the values of energy where the above energy-based FEP has its minima. Since at the transition temperature the two minima have equal depth, one can subtract any $F_H(E_a;\beta^*) = F_H(E_b;\beta^*)$ in Eq. 7.5 so that both minima of $\beta^*\Delta F(E;\beta^*)$ lie on the abscissa axis. This property will be useful in Sec. 8.1.

We also show in Fig. 7.1(b) the two main quantities that characterizes the FEP,

namely the latent heat $\Delta E^{\ddagger} = E_d - E_a$ and the free-energy barrier $\beta^* \Delta F^{\dagger}$, which can be easily computed from the profile. From now on, we specialize our notation to make reference to aggregated a and diluted d states rather than general A and B states that we have used previously. Since E describes the potential energy between attractively interacting particles, E_a refers to the lower value of E at which the free-energy has a minimum, while E_d refers to the basin with higher energy. In the inset of Fig. 7.1(b), there are typical configurations of each phase, a state with an aggregate at an energy close to E_a and a completely diluted state with energy near E_d. Note that the formed aggregate is not spherical but elongated due to the anisotropic interactions present in the model. Finally, it is worth saying that for temperatures close to but different from T^*, the two minima will be imbalanced. While for higher temperatures the diluted phase will be more stable $F_H(E_d; \beta) < F_H(E_a; \beta)$, the contrary will happen for lower temperatures.

As last, we clarify that although in the usual textbook formulation of the microcanonical ensemble the entropy is a function of the total energy, e.g., potential plus kinetic energy, it may be useful to work only with the potential energy. In spite of being eventually equivalent in terms of transition temperature [93], many numerical approaches available to compute $\Omega(E)$ are based on schemes which usually employ the potential energy only (one computes the number of states for a fixed potential energy value). In this case, one is, in fact, considering the conformational microcanonical ensemble [54, 93]. In particular, the Metropolis MC algorithm is usually performed only considering potential energy. And, specially for systems of colloidal particles diffusing in an overdamped regime which we are interest in, the potential energy does not depend on the momenta, and thus the kinetic contribution to the microcanonical partition function can be performed analytically [107, 108]. In any case, it is more convenient to work only with potential energy when considering MC simulations, so we will restrict ourselves to the conformational microcanonical ensemble here. It seems that for many aggregating systems, the physical properties obtained using those two definitions, such as transition temperatures and free-energy barriers, are indeed pretty much the same [54, 93].

Chapter 8

Kinetic approach

Related publication

A density of states-based approach to determine temperature-dependent aggregation rates

[24]

L. F. Trugilho, S. Auer, L. G. Rizzi

J. Chem. Phys. **161**, 051101 (2024)

We now turn to the problem of extracting kinetic information from a free-energy profile as a function of the (potential) energy, such as the one displayed at Fig. 7.1(b). In contrast, most kinetic approaches based on the classical nucleation theory (CNT) employ the largest cluster size n as a RC. The most simple ones assume an Arrhenius-like expression, where the nucleation rate is considered proportional to $e^{-\beta\Delta G^\dagger}$, where $\Delta G^\dagger$ is an effective activation energy or the work of formation of a cluster of critical size [25, 52]. Those approaches are inspired by the pioneering works of Gibbs and others [52, 109]. From a theoretical point of view, several proposals have been derived for computing ΔG as a function of n, from which $\Delta G^\dagger$ could be extracted, e.g., finding the maximum of the function $\Delta G(n)$ where the critical nucleus size is assumed to be found [52]. Some complications arise because those theories are dependent on hypotheses regarding the shape of the forming aggregates [52]. Also, for complex systems, even the definition of the cluster size and interface may be challenging [54]. Usually, the aggregates are assumed to have a spherical symmetry, which is unlikely to hold when anisotropic interactions are important [78, 79].

In this context the authors of Ref. [54] proposed estimating free-energy barriers directly from the profiles obtained from the microcanonical analysis, like in Fig. 7.1(b). As the authors of this reference claimed, the nucleation rate could then be estimated with an Arrhenius-like formula

$$j(\beta) = Ke^{-\beta\Delta F^\dagger}, \tag{8.1}$$

where K is the so-called kinetic factor and should incorporate the kinetic details of the nucleation, such as diffusion and attachment frequencies [54]. Although it is mentioned

in Ref. [54] that $\Delta F^{\dagger}$ clearly should depend on temperature, that dependence was not explored in that work.

In our previous works [25, 26], the dependence of the aggregation and dilution rates on the temperature (or $\beta = 1/k_B T$) was fully incorporated into expressions that, unlike Eq. 8.1, were derived directly from the FEP obtained through microcanonical analysis. This was done, just like in Kramers's seminal work [110], assuming that the motion along the RC (energy in our case) satisfies the Smoluchowski diffusion equation (Eq. 3.6). In fact, in our approach it is sufficient that the Eq. 3.17 and 3.18 for the MFPTs hold for the energy coordinate. Now we understand that this assumption meant that energy should be an optimal coordinate, in the sense that for the optimal committor RC, Eq. 3.17, and 3.18 are satisfied exactly. In other words, the energy should be equivalent to the committor RC. In that case the function $q(E)$ obtained through Eq. 3.16 or 3.47 would be equal to the multidimensional committor $q(\vec{X})$. As it happens, for many-body system, this is unlikely to happen since the projection of the multidimensional space onto a 1D RC usually introduces memory effects [16, 109]. We have already showed this for n in Chapter 5, but a similar behavior can be expected for E for the models studied here. In this way, the main goal of the present work [24] is to establish a suitable separation of timescales so that the expressions for the temperature-dependent rates can be applied to describe aggregation transitions in many-body systems. Particularly, we will illustrate it for the lattice system of Sec. 5.1.

8.1 Analytical expressions for the rates

In this section we review the approach introduced in Ref. [25] and [26] to derive analytical expressions for the temperature-dependent aggregation and dilution rates.

In order to make a pedagogic derivation of the analytical expressions, we first show in Fig. 8.1(a) and (b) the energy-based free-energy profiles $\beta^* \Delta F(E)$ obtained for the lattice systems described in Sec. 5.1 at their transition temperatures T^* for two different ξ (both with $L = 200$ and $N = 400$, so that $\rho = 10^{-2}$. At panel (a), we display the FEP obtained from multicanonical simulations (i.e., $N_r = 500$ recursion steps of $N_s = 10^6$ MC sweeps each) for an isotropic system ($\xi = 1$), while in panel (b) we show the same but for an anisotropic system ($\xi = 3$). Note the higher free-energy barrier and latent heat for the latter. Besides $\beta^* \Delta F^{\dagger}$ and $\Delta E^{\dagger}$, the profiles are characterized by the energy differences $\Delta E_a^{\ddagger} = E^* - E_a$ and $\Delta E_d^{\ddagger} = E^* - E_d$ between the energy at the top of the barrier E^* and the energies at the basins E_a and E_d, as well as the curvature-related parameters $\gamma^* \equiv \gamma(E^*) = (\partial b(E)/\partial E)|_{E=E^*} > 0$ and $\gamma_{a(d)} \equiv -\gamma(E_{a(d)}) = -(\partial b(E)/\partial E)|_{E=E_{a(d)}} > 0$. The fact that $\gamma(E^*) > 0$ is a consequence of the convex intruder near E^*, which is related to the thermodynamic instability of such states. Conversely, $\gamma(E_{a(d)}) < 0$ indicates the thermodynamic stability of the basins. In any case we define all curvature parameters as

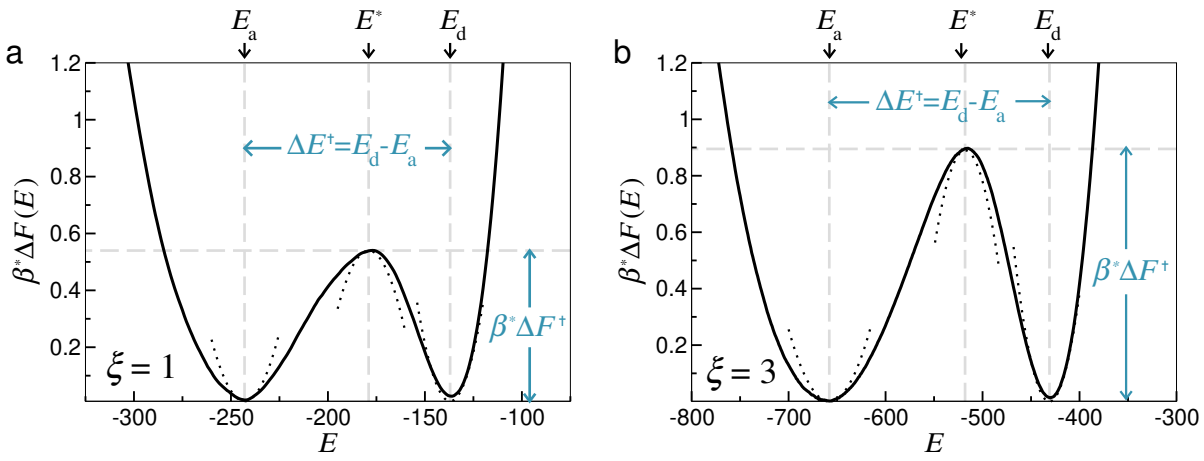

Figure 8.1: Free-energy profile $\beta^*\Delta F(E)$ for (a) an isotropic ($\xi = 1$) and (b) an anisotropic ($\xi = 3$) system. Free-energy barriers and latent heats are indicated by $\beta^*\Delta F^\dagger$ and $\Delta E^\dagger$, respectively. E_a and E_d are the energies at both minima, while E^* is the energy at the local maximum. The dotted lines correspond to quadratic functions near each of the extrema (Eq. 8.2-8.3). To illustrate their suitability, the curvature parameters γ_a, γ_d, and γ^* were determined from the fit of the kinetic data of Fig. 8.4, rather than directly fitting the quadratic functions near the extrema of the FEP (compare Table 8.1). Figure extracted from Ref. [24].

positive in the above relations. Note also that $\gamma(E)$ is defined as the second derivative of the microcanonical entropy $S(E)/k_B$ so it also corresponds to the second derivative of $(-\beta^*\Delta F(E;\beta^*))$ through Eqs. 7.3 and 7.5. In that way, γ^*, γ_a, and γ_d are directly related to the curvatures of the FEPs near E^*, E_a, and E_d, respectively. Both the energy differences and the curvatures can be used to describe the asymmetry of the free-energy profiles. For example, from Fig. 8.1 it is clear that the curvature near E_d is greater than the one around E_a (i.e., $\gamma_d > \gamma_a$) for both values of ξ. In a close relationship one has $|\Delta E_d^\ddagger| < |\Delta E_a^\ddagger|$ for both ξ.

The above-defined energy differences and curvature parameters are implicated in the proposed expressions for the temperature-dependent rates. Particularly, we assume that the FEP near its local maximum E^* can be approximated by the quadratic expression

$$\beta^*\Delta F(E;\beta^*) \approx \beta^*\Delta F^\dagger - (\gamma^*/2)(E - E^*)^2, \tag{8.2}$$

where the barrier $\beta^*\Delta F^\dagger$ is just the value of the profile when $E = E^*$ and γ^* the module of the second derivative as defined above (the first derivative vanishes near the extrema). Similarly, around the minima $E_{a(d)}$, the profile can be approximated by

$$\beta^*\Delta F(E;\beta^*) = (\gamma_{a(d)}/2)(E - E_{a(d)})^2, \tag{8.3}$$

in this case $\beta^*\Delta F(E = E_{a(d)};\beta^*)$ is zero by construction, and $\gamma_{a(d)}$ are the moduli of the second derivatives near each minimum as above (near the minima the profiles have

positive curvature as opposed to the maximum). Expressions 8.2 and 8.3 indicate a simple way to estimate γ^*, γ_a, and γ_d. One can perform a quadratic fit of those expressions near the respective extrema of $\beta^*\Delta F(E;\beta^*)$, as illustrated in Fig. 8.1.

Now, using Eq. 7.3 and 7.5, it is possible to express the microcanonical entropy $S(E)$ in terms of the FEP at the transition temperature $\beta^*\Delta F(E;\beta^*)$

$$\frac{S(E)}{k_B} = \frac{S(E_a)}{k_B} + \beta^*(E - E_a) - \beta^*\Delta F(E;\beta^*), \tag{8.4}$$

where either E_a or E_d could be used in Eq. 7.5 since $F_H(E_a;\beta^*) = F_H(E_d;\beta^*)$, so we fixed E_a to shorten the notation. Note that Eq. 8.4 is a formal relation since our free-energy profiles are computed from $S(E)$, not the contrary. Indeed, estimating the microcanonical entropy from canonical profiles usually requires combining several profiles at different temperatures to obtain the desired range of energy by means of reweighting techniques [94, 111], since each individual profile goes to zero quickly beyond the mean energy at each temperature. Nevertheless, Eq. 8.4 combined with 7.2 allows one to express the equilibrium canonical distribution $p_{\mathrm{eq}}(E;\beta)$ (at a different temperature β) in terms of $\beta^*\Delta F(E;\beta^*)$

$$p_{\mathrm{eq}}(E;\beta) = \frac{1}{\mathcal{Z}(\beta)} \exp\left[\frac{S(E_a)}{k_B} - \beta^* E_a\right] \exp\left[E(\beta^* - \beta) - \beta^*\Delta F(E;\beta^*)\right]. \tag{8.5}$$

We note that, if $\beta = \beta^*$, Eq. 8.5 reduces to Eq. 7.2. This expression for $p_{eq}(E;\beta)$ is valid for any β, but in the following we will consider temperatures sufficiently close to β^*, where the system makes transitions between aggregated and diluted states and the rates can be evaluated analytically. With that, it is then possible to write down approximated expressions for the equilibrium distribution for energies near E^* ($p^*(E;\beta)$), E_a ($p_a(E;\beta)$), and E_d ($p_d(E;\beta)$) using the quadratic approximations. For $E \sim E^*$, we substitute Eq. 8.2 into Eq. 8.5, and after some algebra, one gets [25]

$$p^*(E;\beta) \approx \frac{\Gamma^*(\beta)}{\mathcal{Z}(\beta)} \exp\left\{\frac{\gamma^*}{2}\left[(E - E^*) - \frac{(\beta - \beta^*)}{\gamma^*}\right]^2\right\} \tag{8.6}$$

with

$$\Gamma^*(\beta) = \exp\left[\frac{S(E_a)}{k_B} - \beta^* E_a - \beta^*\Delta F^\dagger - \frac{((\beta - \beta^*) + \gamma^* E^*)^2}{2\gamma^*} + \frac{\gamma^*(E^*)^2}{2}\right]. \tag{8.7}$$

Similarly, substituting Eq. 8.3 into Eq. 8.5, one finds for $E \sim E_a$ (or $E \sim E_d$) that [25]

$$p_{a(d)}(E;\beta) \approx \frac{\Gamma_{a(d)}(\beta)}{\mathcal{Z}(\beta)} \exp\left\{-\frac{\gamma_{a(d)}}{2}\left[(E - E_{a(d)}) + \frac{(\beta - \beta^*)}{\gamma_{a(d)}}\right]^2\right\} \tag{8.8}$$

with

$$\Gamma_{a(d)}(\beta) = \exp\left[\frac{S(E_a)}{k_B} - \beta^* E_a + \frac{((\beta-\beta^*) - \gamma_{a(d)}E_{a(d)})^2}{2\gamma_{a(d)}} - \frac{\gamma_{a(d)}(E_{a(d)})^2}{2}\right]. \tag{8.9}$$

Note that we intentionally expressed $p^*(E;\beta)$ and $p_{a(d)}(E;\beta)$ as the multiplication of two exponentials, besides the normalization constant. The factors Γ^* and $\Gamma_{a(d)}$ depend only on temperature and constants that are extracted from the energy-dependent free-energy profile at the transition temperature, while the energy dependence is restricted to the remaining exponential, and it is written as a quadratic form that will be essential for calculating the Gaussian integrals that appear next.

Now, the rates can be estimated as the inverse of the MFPT's [109]. The aggregation rate is defined as $k_a = 1/\tau_{d\to a}$ and the dilution rate as $k_d = \tau_{a\to d}$. Thus, assuming that the expressions for the MFPT (Eq. 3.17 and 3.18) are valid for the energy coordinate, one can approximate $\tau_{a\to d}$ as

$$\tau_{a\to d} = \int_{E_a}^{E_d} \frac{\mathrm{d}E}{p_{\mathrm{eq}}(E;\beta)D(E)} \int_{-\infty}^{E} p_{\mathrm{eq}}(E';\beta)\mathrm{d}E' \approx \frac{1}{D^*}\int_{E_a}^{E_d} \frac{\mathrm{d}E}{p^*(E;\beta)} \int_{-\infty}^{E^*} p_a(E;\beta)\mathrm{d}E \tag{8.10}$$

where some commonly used approximations were used. First, we decoupled the integrals, assuming that each is dominated by the distributions near E^* (for the first one) and near E_a for the second one. That is because we assume that $p_{\mathrm{eq}}(E;\beta)$ has a minimum near E^* (so its inverse is maximized) while $p_{\mathrm{eq}}(E;\beta)$ has a maximum near E_a. This should be valid for β close enough to β^* so that the FEPs have a similar overall format with two basins separated by a local maximum. Besides, we assume an effective coordinate-independent diffusion coefficient D^*. Analogously, $\tau_{d\to a}$ is approximated by

$$\tau_{d\to a}(\beta) = \int_{E_a}^{E_d} \frac{\mathrm{d}E}{p_{\mathrm{eq}}(E,\beta)D(E)} \int_{E}^{\infty} p_{\mathrm{eq}}(E';\beta)\mathrm{d}E' \approx \frac{1}{D^*}\int_{E_a}^{E_d} \frac{\mathrm{d}E}{p^*(E;\beta)} \int_{E^*}^{\infty} p_d(E;\beta)\mathrm{d}E \tag{8.11}$$

under similar assumptions.

In this way one can evaluate the Gaussian integrals as detailed in the Appendix D (see also Ref. [25] and [26]). After doing that, the temperature-dependent rates can be finally expressed as

$$k_a(\beta) = \frac{1}{\tau_{d\to a}} \approx A_d \exp\left[-\Delta E_d^{\ddagger}(\beta-\beta^*) - \frac{\bar{\gamma}_d}{2}(\beta-\beta^*)^2\right] \tag{8.12}$$

and

$$k_d(\beta) = \frac{1}{\tau_{a\to d}} \approx A_a \exp\left[-\Delta E_a^{\ddagger}(\beta-\beta^*) - \frac{\bar{\gamma}_a}{2}(\beta-\beta^*)^2\right], \tag{8.13}$$

where $\bar{\gamma}_{a(d)} = (\gamma^*)^{-1} + (\gamma_{a(d)})^{-1}$ and the energy differences were already defined as $\Delta E_{a(d)}^{\ddagger} =$

$E^* - E_{a(d)}$. We also define the equilibrium constant k_{eq} as

$$k_{\text{eq}}(\beta) = \frac{k_a}{k_d} \approx A \exp\left[\Delta E^\dagger(\beta - \beta^*) + \frac{\Delta\bar{\gamma}}{2}(\beta - \beta^*)^2\right] \tag{8.14}$$

where $\Delta\bar{\gamma} = \bar{\gamma}_a - \bar{\gamma}_d = (\gamma_a)^{-1} - (\gamma_d)^{-1}$ and $\Delta E^\dagger = \Delta E_a^\ddagger - \Delta E_d^\ddagger = E_d - E_a$ is just the microcanonical latent heat as defined earlier.

The pre-factors $A_{a(d)}$ are also completely defined but are slightly different depending on the exact approximation used to evaluate the Gaussian integrals. For low free-energy barriers, it turned out that the more suitable approximation leads to [25]

$$A_{a(d)} = \frac{D^*}{|\Delta E_{a(d)}^\ddagger|}\sqrt{\frac{\gamma^*}{2\pi}}e^{-\beta^*\Delta F^\dagger} \tag{8.15}$$

in a way that $A = A_d/A_a = |\Delta E_a^\ddagger|/|\Delta E_d^\ddagger|$. While for high barriers, a Kramers-like approximation turned out to be more adequate [26]

$$A_{a(d)} = \frac{D^*}{2\pi}\sqrt{\gamma^*\gamma_{a(d)}}e^{-\beta^*\Delta F^\dagger} \tag{8.16}$$

so that $A = A_d/A_a = \sqrt{\gamma_d/\gamma_a}$. See the Appendix D for the details of both approximations for the Gaussian integrals. We emphasize that, although Eqs. 8.12-8.14 look like expansions around $\beta \approx \beta^*$, that is not the case; the expansions were made on energy E around the peaks and valley of the FEP and the second order on the difference $(\beta - \beta^*)$ is really were it stops. Interestingly, for $\beta = \beta^*$, one recovers the usual employed Arrhenius-like formulas, such as Eq. 8.1 with a barrier of $\beta^*\Delta F^\dagger$ for both k_a and k_d. In the high barrier case, the pre-factors A_a and A_d are indeed exactly the same as the one derived by Kramers for the escape over a (high) potential barrier in the high friction limit [109, 110]. It should be like this since the high-friction limit corresponds to the Smoluchowski diffusive model from which the expressions for the MFPT used here are derived. Moreover, at $\beta = \beta^*$, the barrier is exactly $\beta^*\Delta F^\dagger$ for both processes (aggregation and dilution), so both the pre-factors and the barriers are in accordance with this well-known result. The novelty of the expressions is to capture the temperature dependence that arises due to the imbalance of the free-energy profile at temperatures close to but different from β^* in an analytical way. At those temperatures the barrier is not just $\beta^*\Delta F^\dagger$, so the rates cannot be proportional to $\beta^*\Delta F^\dagger$ with the same pre-factors anymore. Eq. 8.12 and 8.13 provide the needed corrections.

Note that all quantities that appear in the expressions for the rates 8.12-8.16 can be calculated directly from the free-energy profile $\beta^*\Delta F(E;\beta^*)$, except for the effective diffusion coefficient in energy space D^*. This is expected since the diffusive model depends on both the potential, of which the FEP plays the role, and the diffusion coefficient, which

should be estimated independently. It is worth mentioning that, although the FEPs are different at different temperatures, the approach is able to express the rates only in terms of parameters related to the profile at β^* because the different profiles (or equivalently probability distributions) are related by Eq. 8.5 as a consequence of the fundamental thermodynamic relations Eq. 7.1 and 7.2. In Ref. [25] and [26], the expressions for the rates were tested by means of stochastic simulations directly on energy space that were diffusive by construction. In this way, these previous works evaluated all approximations described in this section, except the diffusive hypothesis, and that is what we will be interested in the next sections. In particular, in Sec. 8.2 we will describe a methodology for estimating D^*, and then in Sec. 8.3 we will test the complete approach for the case where the stochastic process in energy space is generated as a projection of the highly-dimensional lattice system.

A few more words about the validations performed in Ref. [25] and [26] should be helpful for understanding at each point we were with those previous works and what is the advancement of the present work [24]. In Ref. [25], the validity of the temperature-dependent expressions for the rates was validated for mathematically constructed FEPs by means of the stochastic diffusive simulations on energy space. In Ref. [26], we considered (the rare case of) a theoretical model of a physical system with aggregating particles from which the microcanonical entropy can be computed exactly, but the validation of the expressions was done again with similar one-dimensional stochastic simulations. More importantly, in Ref. [25], experiments of ice nucleation and protein folding could be well described by expressions based on Eq. 8.12- 8.14. Unfortunately, one has no access to the thermodynamic coefficients, such as the energy differences and curvature parameters that appear in those equations for most experimental data. Because of that, a comprehensive validation of the temperature-dependent expressions for the rates for systems evolving dynamically at the particle level was still missing, and here we fill this gap by performing simulations of the lattice systems, for which those coefficients are available. In this way, here we perform an quantitative comparison between the estimates of the coefficients obtained from the theoretical expressions and the values extracted directly from the FEPs. Moreover, as already anticipated that the dynamics projected over low-dimensional RCs such as energy usually have strong non-Markovian effects, we describe a methodology to determine the timescale where the non-Markovian effects are mitigated so that the diffusive hypothesis holds.

8.2 Timescale separation

As discussed in Sec. 3.3.1, the $F_{C,1}$ validation test can be employed to inspect whether a reaction coordinate has strong non-Markovian effects. If the dynamics projected on the RC is Markovian, the $F_{C,1}$ profiles are independent of sampling interval Δt so that

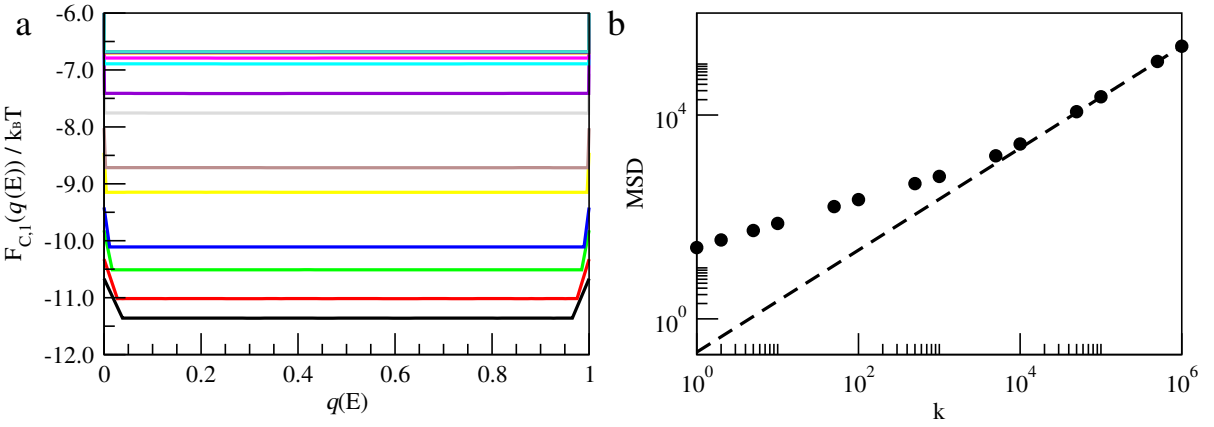

Figure 8.2: (a) $F_{C,1}$-based validation test for the E-based coordinate $q(E)$. Different colors correspond to different sampling intervals $\Delta t_k = k\Delta t_0$, with increasing k from bottom to top curves (b) Mean squared displacement evaluated with TPS as a function of k. Note that the MSD displays a linear behavior (dashed line) for high enough Δt_k, that is, $k \approx 5 \times 10^3$, where the generalized profiles $F_{C,1}(q(E), \Delta t_k)$ are indeed almost independent of Δt_k. Results for the isotropic case with $\xi = 1$.

the diffusion coefficient given by Eq. 3.34 is also independent of observational timescale. The more the profile varies with Δt, the stronger the non-Markovian effects [47]. As in Sec. 5.1.2, we first perform the transformation $E \to q(E)$ to apply the validation test because in this way $F_{C,1}$ is independent of the RC for each Δt, like $F_{C,1}(q(n), \Delta t)$ is for n. Here we constructed the MSM for E and then used Eq. 3.47 to compute $q(E)$. In Fig. 8.2(a), we show $F_{C,1}(q(E), \Delta t)$ for several time intervals $\Delta t_k = k\Delta t_0$, and it is clear that the dynamics projected on E has strong non-Markovian effects, similar to n. In this figure we present results for the isotropic case with $\xi = 1$, which qualitative behavior is similar to the anisotropic case with $\xi = 3$. The profiles $F_{C,1}(q(E), \Delta t_k)$ only become approximately independent of sampling interval for[1] Δt_k for $\Delta t_k \approx 5 \times 10^3 \Delta t_0$. Hence, it is clear that the dynamics might only be described as approximately Markovian from this timescale and above.

In Fig. 8.2(b) we show the mean squared displacement (MSD) (also computed with transition path segments) as a function of k. For low k, the MSD presents a subdiffusive behavior, which means that it increases sublinearly with k. Only from about $k \approx 5 \times 10^3$ the MSD does follow a linear diffusive behavior given by $2J_{AB}(\Delta E^\dagger)^2 k\Delta t_0$, with J_{AB} being the equilibrium flux between A and B (see Appendix A for details). This also indicates that only from about that timescale the non-Markovian effects can be neglected.

Besides, in Eqs. 8.15 and 8.16 we assumed an energy-independent D^* so that we

[1]Here as well as in the following section we set $\Delta t_0 = 20$ MCs, which is smaller than the sampling interval used in Sec.5.1, because we used a different set of simulations as detailed in Sec. 8.3

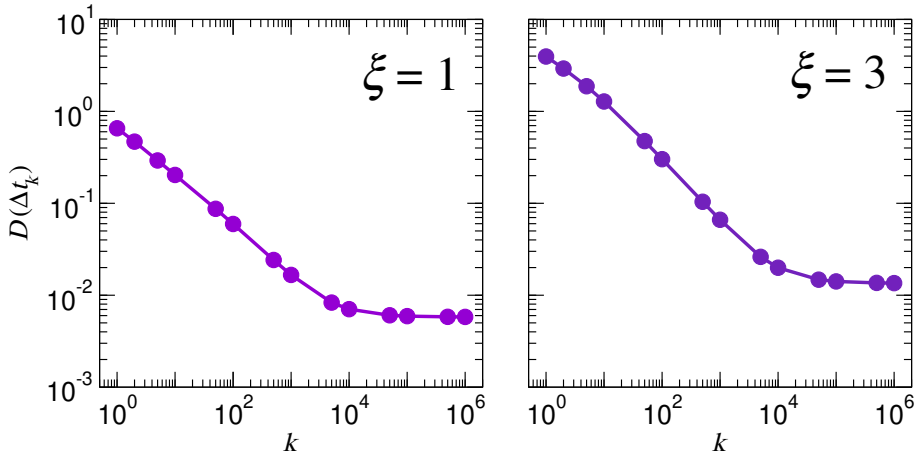

Figure 8.3: Sampling interval dependent diffusion coefficient $D(\Delta t_k)$ as a function of k for both (a) an isotropic system with $\xi = 1$ and (b) an anisotropic system with $\xi = 3$. Figure extracted from Ref. [24].

consider the following average

$$D(\Delta t_k) = \int_{E_a}^{E_d} p_{\mathrm{eq}}(E;\beta^*) D(E, \Delta t_k; \beta^*) \mathrm{d}E. \tag{8.17}$$

where $p_{\mathrm{eq}}(E;\beta^*) \propto e^{-\beta^* \Delta F(E;\beta^*)}$ is the canonical probability distribution at the transition temperature $T^* = 1/k_B\beta^*$, and $D(E, \Delta t_k; \beta^*)$ is computed from Eq. 3.34 also at T^*.

We note that such a definition for the diffusion coefficient generally indicates that, due to non-Markovian effects, $D(\Delta t_k)$ should be dependent on the sampling interval Δt_k, as can be verified in Fig. 8.3, but it becomes approximately independent of Δt_k from about $k \approx 5 \times 10^3$. In that figure we also show the result for the anisotropic case with $\xi = 3$, from where it can be seen that for both cases the diffusive behavior begins at similar timescales. We used the fact that $D(\Delta t_k)$ is constant at later timescales to define the effective diffusion coefficient D^* to be used in expressions we derived for the rates, i.e., Eq. 8.12-8.16, that is

$$D^* = \lim_{\Delta t_k \to \infty} D(\Delta t_k). \tag{8.18}$$

As discussed in detail in Appendix A, the effective diffusion coefficient D^* defined as above equals the temporal derivative of the MSD at later times, that is, in the limit of large Δt_k.

Importantly, it was assumed that the diffusion coefficient is independent of temperature, so the approximate expressions for the rates are expected to hold for temperatures close to the transition temperature β^*, as D^* is computed at β^*. With this definition of D^*, all quantities that are required to compute the rates are determined, and we can now test the validity of Eqs. 8.12-8.16, as we do next.

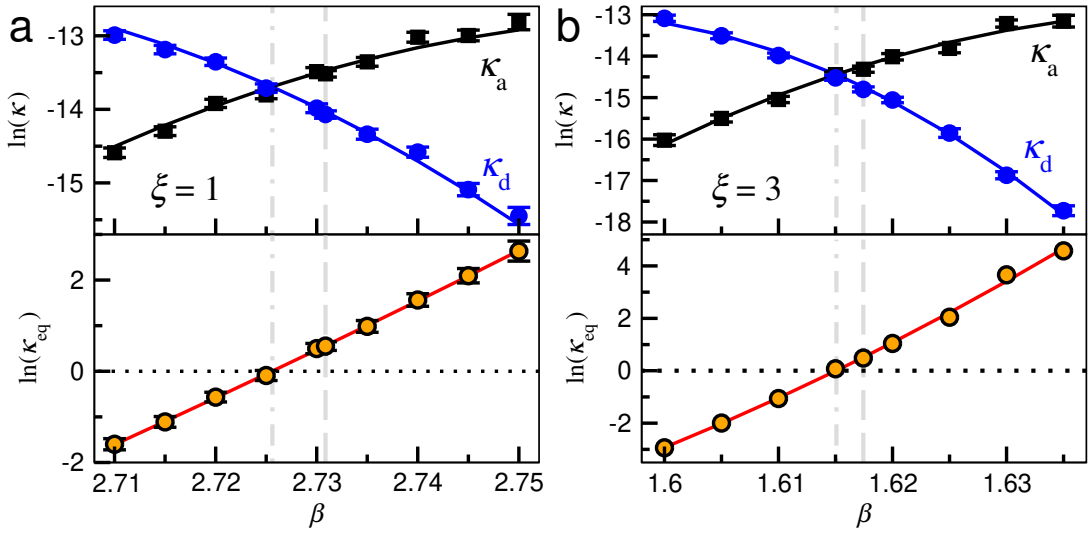

Figure 8.4: Aggregation rate k_a and dilution rate k_d, as well as the equilibrium constant $k_{\rm eq} = k_a/k_d$ as a function of the inverse of temperature β, for both (a) isotropic and (b) anisotropic systems. The symbols represent the values computed directly from the energy time series (simulation data), while the continuous lines represent the fits of Eq. 8.12-8.14 to the numerical data. The fits were done with the pre-factor given by Eq. 8.15 in (a) and Eq. 8.16 in (b). The vertical dashed and dotted-dashed lines (in gray) indicate the location of β^* and β_m, respectively. Figure adapted from Ref. [24].

8.3 Validation of the expressions for the rates

Now we turn our attention to the numerical validation of the analytical expressions we obtained for the temperature-dependent aggregation and dilution rates given by 8.12-8.16. For that, we performed canonical Metropolis MC simulations at different temperatures close to T^*. All simulations were similar to the ones considered in Sec. 5.1. In particular, we considered a system of size $L = 200$ containing $N = 400$ particles. For each temperature, we computed the MFPTs between energies E_a and E_d directly from the energy time series. All time units are given in reference to one Monte Carlo step[2]. The aggregation rate k_a and the dilution rate k_d were then evaluated as the inverse of the MFPTs $\tau_{d\to a}$ and $\tau_{a\to d}$, respectively. The simulations initiate with the system at the diluted phase with randomly placed particles and the analysis of computing the mean first-passage times starts as soon as the first transition between energies E_d and E_a is completed. For each temperature, we considered a minimum number of simulations so that at least 10^2 roundtrips between E_a and E_d had occurred. Then, the MFPTs are estimated by taking the mean, and we show the respective rates represented by the symbols in Fig. 8.4 for both $\xi = 1$ (a) and $\xi = 3$ (b). We also show the equilibrium constant as defined by Eq. 8.14. One interesting feature is that both k_a and k_d clearly display non-Arrhenius behavior; even so, $k_{\rm eq}$ seems to be linear in the mono-log plot.

[2]Although the energy time series were saved with a time-step of $\Delta t_0 = 20$ MCs, all time units in the Chapter were converted to unitary MC step.

Table 8.1: Parameters obtained both directly from the FEPs displayed in Fig. 8.1 and indirectly from the fits of Eq. 8.12, 8.13, and 8.14 to the numerical data presented in Fig. 8.4.

	$\xi = 1$		$\xi = 3$	
	from FEP	from fit	from FEP	from fit
$\beta^*\Delta F^\dagger$	0.53	–	0.88	–
β^*	2.73087	–	1.61744	–
$\Delta E^\ddagger$	106	107	227	217
$\Delta E_a^\ddagger$	64	68	140	132
$\Delta E_d^\ddagger$	-42	-39	-87	-85
γ^*	0.7×10^{-3}	1.6×10^{-3}	4.0×10^{-4}	6.8×10^{-4}
γ_a	0.8×10^{-3}	1.5×10^{-3}	1.5×10^{-4}	2.9×10^{-4}
γ_d	2.0×10^{-3}	2.4×10^{-3}	7.2×10^{-4}	8.0×10^{-4}

Next we validate the temperature-dependent expressions for the rates by fitting Eqs. 8.12- 8.16 to the numerical data presented in Fig. 8.4 as follows. We take the values of β^* and $\beta^*\Delta F^\dagger$ as the ones calculated from the microcanonical analysis, e.g., as extracted from Fig. 8.1. We also set the value of D^* as the value determined from Fig. 8.3 as its later time limit (see Eq. 8.18). Then, we performed the fits of the analytical expressions to the numerical data by letting the energy differences $\Delta E_{a(d)}^\ddagger$ and curvatures $\gamma_{a(d)}$ and γ^* acting as free-parameters. Note that those parameters also appear at the pre-factors 8.15-8.16 so that the fits are performed in a self-consistent manner. As can be seen in Fig. 8.4, the continuous lines obtained from the fits describe well the numerical data for the rates from both ξ. For the isotropic system we used the expressions for the pre-factor more suitable for low free-energy barriers (Eq. 8.15), while for the anisotropic system we used Eq. 8.16, which is more suitable for high barriers.

In Table 8.1, we show the comparison between the obtained parameters extracted both directly from the free-energy profiles displayed in Fig. 8.1 and indirectly from the fits of Eqs. 8.12-8.16 to the numerical data presented in Fig. 8.4. The results are quite similar given the approximate character of Eqs. 8.12-8.16. It is worth noting that the curvature-related parameters show higher deviations since the FEPs are not perfectly quadratic near E_a, E_d, and E^*. Nevertheless, the values extracted from the fits to the kinetic data can describe well the curvatures near the extrema of the FEPs in Fig. 8.1. In that figure, the dotted lines represent the quadratic approximations near the extrema but with the curvatures obtained from the fits of the analytical expressions for the rates to the kinetic data displayed in Fig. 8.4. As can be seen from Fig. 8.1, the values of $\gamma_{a(d)}$ and γ^* extracted from the kinetic data describe the curvatures of the FEP near its extrema reasonably well, despite the small deviations from the best values extracted directly from the FEPs; see Table 8.1.

It is worth noting that the definition of the transition temperature T^* corresponds

to the temperature where the free-energy profile presents two minima (near E_a and E_d) with equal height. However, according to Eqs. 8.12-8.16, that temperature should only agree with the temperature T_m where $k_a = k_d$ if the FEPs are completely symmetric, i.e., $\Delta E_a^{\ddagger} = \Delta E_d^{\dagger}$ and $\gamma_a = \gamma_d$. Nevertheless, by requiring that $k_{\rm eq} = 1$ at $\beta = \beta_m$, it can be shown that the two temperatures should be related by $\beta_m = \beta^* + \delta$, where the correction term is given by

$$\delta = \frac{1}{\Delta E^{\dagger}} \ln\left(\frac{|\Delta E_d^{\ddagger}|}{|\Delta E_a^{\ddagger}|}\right) \tag{8.19}$$

for low free-energy barriers [25] and by

$$\delta = \frac{1}{2\Delta E^{\dagger}} \ln\left(\frac{\gamma_a}{\gamma_d}\right) \tag{8.20}$$

for high free-energy barriers [26].

For both $\xi = 1$ and $\xi = 3$, we have that $\gamma_d > \gamma_a$ and $|\Delta E_d^{\ddagger}| < |\Delta E_a^{\ddagger}|$, that is, the profiles are narrower near the diluted basin close to E_d. Because of that, Eq. 8.19-8.20 indicates that $\delta < 0$. In Fig. 8.4 we also show the values of β_m, indicated by the vertical dashed-dotted lines, computed through Eq. 8.19-8.20. As can be seen in the figure, β_m corresponds with the values of β where $k_{\rm eq} = 1$ for both $\xi = 1$ and $\xi = 3$, corroborating both expressions for δ. In physical terms, the temperature T_m can be related to the situation where both phases have equal steady-state probability, and that corresponds to the equal area criteria under the maxima of the FEP [25, 112]. In this way, the fact that $T_m > T^*$ ($\beta_m < \beta^*$) implies that, at $T = T_m$, the diluted phase basin of the FEP will have a slightly deeper depth. This compensates for the fact that the profile is narrower at energies near the diluted basin (see Fig. 8.1), and that leads to similar steady-state probabilities and thus, equal aggregation and dilution rates at $T = T_m$.

In conclusion, by performing a suitable separation of timescales, we demonstrate that our previously derived expressions for the temperature-dependent rates can quantitatively describe reversible aggregation kinetics in many-body systems evolving at the particle level. Importantly, the values of the curvatures and energy differences extracted directly from the FEPs were obtained by the multicanonical simulations, from which the microcanonical entropies (and thus the FEPs) were obtained, while the canonical simulations used to validate the expressions for the rates were performed via Metropolis scheme. We note that the approach developed in the Chapter may be used to infer kinetic properties of other (more realistic) aggregating finite-sized systems, many of which have microcanonical characterizations already available in the literature [54, 101, 102, 103, 104, 105].

Part IV

Microphase separation in systems with competing interactions

Chapter 9

Specific objectives

In Parts. II and III, our focus was to describe the kinetics of systems undergoing first-order aggregation transitions. We considered systems that interact via two contributions: excluded volume and an attractive potential that drives aggregation. In this part we consider systems interacting with a more complicated potential, where in addition to the attractive term, there is also a long-range repulsion, and our focus is on identifying the temperature T_m at which aggregation occurs. The competition between attraction and repulsion leads to a slightly different aggregation process called microphase separation. Those systems display a rich phase diagram, and we are particularly interested in localizing the occurrence of this crossover transition that can happen depending on the ranges and strengths of the attractive and repulsive parts of the potential. In the regions of the phase diagram where microphase separation occurs, an equilibrium cluster phase is observed, where multiple aggregates of a characteristic size, determined by the competition between attractive and repulsive forces, coexist in equilibrium with the background gas of particles. In particular, our objective in this Part IV is to introduce and validate a stochastic method we developed to determine *a priori* the regions of the phase diagram where microphase separation takes place, which is based on MFPTs and considers only the pairwise interactions between particles.

Chapter 10

Phase diagram determination

10.1 Motivation

In this chapter the focus is on a rather different problem in comparison with the problems explored in the rest of this thesis. Until now, we have considered systems interacting via two contributions. One repulsive excluded volume type interaction, which is extremely short-ranged, and one attractive interaction, which is longer-ranged compared to the former. This is true both for the lattice systems, where particles cannot occupy the same site simultaneously while nearest neighbor particles attract each other, and for the Lennard-Jones (LJ) system described by the potential of Eq. 5.1, where each of the terms corresponds to the excluded volume interaction and the more long-ranged attraction. It has been well known since the van der Waals [113] that systems of this kind undergo first-order phase transitions at low enough temperatures, and we have studied the kinetic properties of those transitions in preceding chapters, for both the lattice and LJ systems.

When, in addition to excluded volume and attractive interactions, there is another repulsive interaction of sufficient strength and long-range character, different phenomena can happen. These systems have been referred to as short-range attraction and long-range repulsion (SALR) interacting systems in the literature. From a physical point of view, an interaction potential where the repulsive contribution has a range longer than the attractive contribution can appear due to electrostatic interactions between the particles. For example, colloidal particles in solution usually present an effective net surface charge due to the release of counterions [114]. In those situations the potential is dependent on the fraction of dissociated counterions, which in turn depends on the concentration of the colloidal particles as well as the local distribution of all ions [115]. This drastically complicates the analysis of those systems so that it is common to employ simpler approaches based on the mean field Poisson-Boltzmann equation. Under certain circumstances it is possible to write an effective pairwise potential for the interaction between two colloidal particles that is mediated by the counterions, which effectively

screen the repulsion between the like-charge colloids. This screened repulsive interaction can be modeled then by a Yukawa potential [114, 115]. Thus, one can consider the complete potential as a sum of the LJ-like (Eq. 5.1) and Yukawa terms

$$V(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{2\alpha} - \left(\frac{\sigma}{r}\right)^{\alpha} \right] + A\frac{e^{-r/\xi_D}}{r/\xi_D} \tag{10.1}$$

where A is a positive constant that determines the magnitude of the long-range repulsion and ξ_D is the Debye screening length, which controls the range of this repulsive part of the potential. The remaining parameters are exactly the same that were defined in Sec. 5.2 for the pure LJ system. Particularly, we choose $\sigma = 1$ and $\epsilon = 5$, which set the length and energy scales, as well as $\alpha = 10$, just like in Sec. 5.2. To shorten the notation we use here r for the distance between two interacting particles instead of d_{ij} that we have used previously.

As long as the system is in this regime of screened electrostatic interaction, the repulsive term is not, in fact, long-ranged according to the commonly used definitions of long-range interactions (see, e.g., Ref. [116]), as it decreases exponentially. However, the important point here is that it can be more long-ranged than the attractive part for intermediate separations depending on the value of ξ_D. Albeit being a rather simplified model for the interaction between charged macromolecules, it turns out that the potential described by Eq. 10.1 is a benchmark potential for systems with competing interactions and has been widely used in the literature [114]. One of the reasons is that by tuning the parameters α and ξ_D, one can change the range of the attractive and repulsive parts of the potential, respectively [86, 117]. Moreover, by selecting the ratio between ϵ and A, one determines the relative strength of the two parts [117]. Physically, the range and strength of the attraction can be varied by adding polymers of different sizes to the charged colloidal suspension [28, 86]. In the case of the repulsive term, that can be done by adding different salt concentrations to the solution [114].

One striking feature of systems with competing interactions contrasted with simple fluids of the Van der Waals type is that, depending on the strength and range of the extra repulsive interaction compared to the attractive term, the first-order phase transition line is preempted by the so-called microphase separation. Under this condition the usual bulk or macrophase separation is substituted by the formation of a modulated phase with a characteristic length scale determined by the competition between the different interactions [28, 118]. There exists a myriad of such phases, from cluster phases at low concentrations to the formation of stripes [119] and gels at higher concentrations [28, 114]. The reason for such behavior is that the attractive part drives the system towards aggregation, like for simple fluids. However, the repulsive long-range part opposes the increase of aggregates beyond a characteristic size, so that macrophase separation is inhibited [120]. At low concentrations, various clusters of characteristic size and shape are then formed

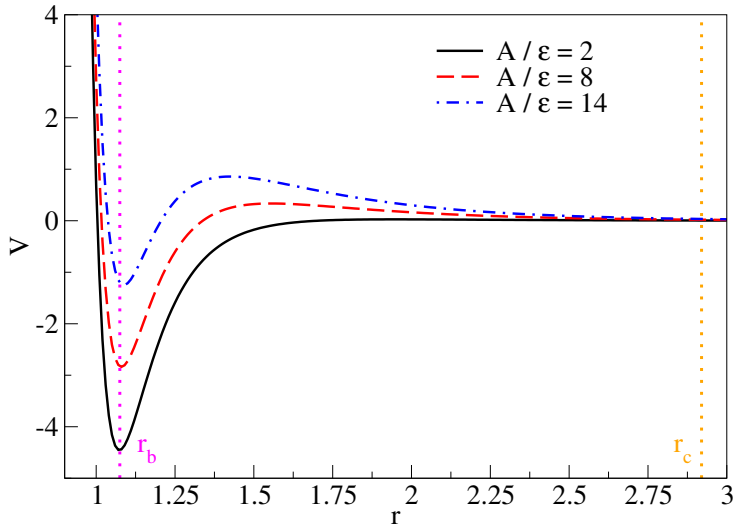

Figure 10.1: Plot of the pairwise interaction $V(r)$ given by Eq. 10.1 for different values of A/ϵ with the other parameters given by $\alpha = 10$, $\xi_D = 0.5\sigma$ and $\sigma = 1$. The vertical dotted magenta and orange lines indicate the values of r close to the minima of the potentials (r_b) and where the potentials become approximately constant (r_c), respectively.

and stay in equilibrium with the background gas phase, which is known as the equilibrium cluster phase [114]. At higher concentrations, stripe-like patterns are formed since the cluster cannot grow spherically due to the repulsion [28]. Particularly, percolating gel phases of nearly one-dimensional shape can be found [28]. Here in this thesis, however, we are interested in the aggregation process that drives the formation of the cluster phase at low concentrations.

In Fig. 10.1 we show the pairwise potential of Eq. 10.1 as a function of the particles' separation r, for different values of the fraction A/ϵ. We acknowledge that it should be interesting to consider different values of the screening length ξ_D in the range between 0.5σ and 2σ, which are generally considered in the literature [117], but in order to validate our approach, we used only the value $\xi_D = 0.5\sigma$, because we start our study following Ref. [28]. In any case, the plots on Fig. 10.1 indicate that $V(r)$ tends to zero from above in the limit of large separations, not from below as in the case of simple fluids, e.g., subject to the usual LJ potential. This implies that, at intermediate separations r, the repulsive part of $V(r)$ is stronger than its attractive contribution. This is the characteristic behavior of systems with competing interactions [119]. Also, if the magnitude of A is increased with respect to ϵ, this effect is magnified. Conversely, $V(r)$ tends to the LJ potential as A goes to zero.

For large ratios A/ϵ, Fig. 10.1 indicates that the potential displays a form that resembles the classical problem of escape from a potential barrier that we introduced in Sec. 3.2.2. This inspired us to use some of the ideas described in Sec. 3.2.2 to try to infer some information about the phase behavior of a many-body system interacting through a pairwise potential like the ones in SALR systems with large enough A/ϵ, only

from the properties of the pairwise potential. This is the aim of the present chapter, and we will detail the proposed method in the following. But, before doing that, it is worth emphasizing why that should be important. In the case of simple fluids with only excluded volume and the attractive short-range isotropic potential, there exists the law of corresponding states derived by Van der Waals and later extended by Noro and Frenkel for colloidal systems [121]. That principle states that for different functional forms of the attractive potentials, there is a correspondence between the thermodynamic states compared at the same virial coefficient. This made it possible to accommodate different systems with short-range attraction on a universal phase diagram containing the first-order macrophase coexistence line. However, in principle, the hypothesis underlying such a law of corresponding states does not cover systems with longer-range repulsion. Because of that, there has been interest in formulating alternative approaches that could determine the location of regions of the phase diagram where microphase separation occurs [114]. For example, it has been proposed that the line of macrophase coexistence of the underlying reference potential, which contains only the attractive part, could indicate the regions of microphase separation for the complete competing potential [122]. Here we propose an alternative approach that relies only on the kinetics of association and dissociation of a single particle subjected to the SALR potential $V(r)$ to infer the location of the line that delimits the onset of microphase separation in the phase diagram of many-body charged colloidal systems.

10.2 Theoretical approach

Our theoretical approach assumes a fixed particle at the origin of the coordinate system $r = 0$. This particle generates the potential $V(r)$ given by Eq. 10.1 at distance r, where a second particle is subject to it. Only the second particle is allowed to move[1]. We consider the escape rate (i.e., the dissociation rate) of the moving particle, which, as discussed in Sec. 3.2.2, can be estimated by the inverse of the MFPT. To make it more realistic, we consider a system in three spatial dimensions. Although the expressions for the MFPTs 3.17 and 3.18 are valid for 1D diffusive models, a three-dimensional problem with spherical symmetry can be expressed as a near one-dimensional process over the radial coordinate r in spherical coordinates [40]. One simple way to see that is to note that the probability $(p_{\mathrm{eq}}(r)\mathrm{d}r)$ of finding the particle between r and $r+\mathrm{d}r$ can be found by integrating, over the angular coordinates θ and ϕ, the probability of finding the particle

[1]This approximation is important not only because the analytical expressions for the MFPTs are only available for 1D or nearly 1D problems (e.g., due to the radial symmetry), but also because our aim is to determine the phase diagram from the properties of the pairwise potential $V(r)$ of a single source.

at an infinitesimal volume $\mathrm{d}\hat{V} = r^2 \mathrm{d}r\mathrm{d}\theta\mathrm{d}\phi$ placed at a distance r. That is,

$$p_{\mathrm{eq}}(r)\mathrm{d}r = \int_0^{2\pi} \mathrm{d}\phi \int_0^{\pi} \mathrm{d}\theta r^2 p_{\mathrm{eq}}(r,\theta,\phi)\mathrm{d}r \tag{10.2}$$

where $p_{\mathrm{eq}}(r,\theta,\phi)$ is the equilibrium probability distribution in a three-dimensional space, while $p_{\mathrm{eq}}(r)$ is the probability distribution along the radial coordinate r. But due to the symmetry of the problem, i.e., $p_{\mathrm{eq}}(r,\theta,\phi) = e^{-\beta V(r)}/\int e^{-\beta V(r)}\mathrm{d}\hat{V}$, the angular integrations can be performed in a way that $p_{\mathrm{eq}}(r)$ is given by

$$p_{\mathrm{eq}}(r) = \frac{r^2 e^{-\beta V(r)}}{\int_0^{\infty} r^2 e^{-\beta V(r)}\mathrm{d}r}. \tag{10.3}$$

In this way, Eq. 3.17-3.18 can be directly used for r, with[2] $p_{\mathrm{eq}}(r;T) \propto r^2 e^{-\beta V(r)}$. Note that $p_{\mathrm{eq}}(r;T)$ carries the temperature dependence through $\beta = 1/k_B T$. Moreover, the diffusion coefficient in this case is given by $D(r) = D_0$ independent of the coordinate because we aim to describe a system similar to the one considered in Sec. 5.2, but now with the potential defined in Eq. 10.1. We set $D_0 = 0.005$, like in Sec. 5.2. And, to illustrate and validate our method, we set our objective to estimate theoretically the temperature T_m where the many-body system undergoes a microphase as a function of A/ϵ.

The idea is that we can compute both the dissociation $k_d(T)$ and the association rate $k_a(T)$ as a function of temperature T. Then T_m can be found by the criterion that $k_a(T_m) = k_d(T_m)$. One should note, although this may seem like our previous approaches, here we are applying that criterion directly to the stochastic diffusion of a particle in a potential, not to the more complex many-body-related free-energy profiles obtained in the previous chapters. The interpretation is that due to the form of the pairwise potential, which presents a barrier, one can define a problem with two states, associated and dissociated. Thus, when $k_a(T_m) = k_d(T_m)$, the particles have similar probabilities of being found associated or dissociated, indicating the location of T_m. At high temperatures $k_a < k_d$, and the particles are much more likely to dissociate, leading the system to a diluted state. Conversely, at low temperatures, $k_a > k_d$, and the particles tend to associate, leading to the clustered states. In any case, one has to acknowledge that the approach is of approximate character since we assume only interactions between two particles, neglecting many-body interactions.

For the potentials displayed in Fig. 10.1 with high enough A/ϵ, $k_d(T)$ can be

[2]Note that because the potential $V(r)$ is not bounded for $r \to +\infty$, the normalization in Eq. 10.3 is not mathematically well defined, but as we detail next the physical requirements of a system confined in a finite volume can fix that issue. See also Ref. [40] for a rigorous proof on how the factor r^2 appears in Eq. 3.17-3.18 for a 3D system with spherical symmetry.

estimated directly by the inverse of the MFPT $\tau_{b\to c}$

$$k_d(T) = \frac{1}{\tau_{b\to c}(T)}, \tag{10.4}$$

where the subscript b makes reference to r_b, a radius where the particle is assumed to be associated and is taken at the minimum of the potential[3] as represented in Fig. 10.1. Moreover, subscript c refers to r_c, which is defined as a cutoff radius from which the potential is almost constant and the corresponding force is assumed to be zero. In this way, $k_d(T)$ can be computed with Eq. 3.17 for the defined boundaries, r_b and r_c, with $p_{\text{eq}}(r)$ given by Eq. 10.3. Note that the lower limit of the inner integral in Eq. 3.17 is zero since the SALR potential defined by Eq. 10.1 diverges at the origin, so that $p_{\text{eq}}(r) = 0$ for $r < 0$ and the integral is well defined.

One problem arises if one tries to define the association rate $k_a(T)$ by the reverse MFPT, $\tau_{c\to b}$, using Eq. 3.18. The upper limit of the inner integral in Eq. 3.18 is infinite so that the integral diverges because the potential $V(r)$ is constant for $r \to +\infty$. In the situations that we will consider, the potential that a particle is subject to is, in fact, bounded by the volume of the system, which is finite. In principle, one could simply define a cutoff for that integral at $r \approx L$, with L being the linear size of the simulation box, for example. However, it turned out that one can choose a more informative way of computing $k_a(T)$, using not only the size of the system but also the information about the system's number concentration $\rho = N/V$, even when considering the simpler problem of two particles interacting in a finite volume. We write the association rate as the reciprocal of the sum of two characteristic times

$$k_a(T) = \frac{1}{\tau_{c,N} + \tau'_{c\to b}(T)}, \tag{10.5}$$

where the MFPT $\tau'_{c\to b}(T)$ is computed with a cutoff on the inner integral of Eq. 3.18 at the value of r_c (not the linear size), i.e.,

$$\tau'_{c\to b}(T) = \int_{r_b}^{r_c} \frac{\mathrm{d}r}{p_{\text{eq}}(r;T)D_0} \int_{r}^{r_c} p_{\text{eq}}(r';T)\mathrm{d}r', \tag{10.6}$$

while $\tau_{c,N}$ is a collision time, which should depend on the number of particles N of the many-body system. Here we assume that the collision time between one and the remaining $(N-1)$ particles in the system, $\tau_{c,N}$, can be estimated by dividing the two-particle collision time by $(N-1)$

$$\tau_{c,N} = \frac{\tau_{c,2}}{N-1}. \tag{10.7}$$

Now, $\tau_{c,2}$ can be computed assuming that a particle is initially at a position r_0 between r_c and R_L, with the latter a characteristic distance related to the linear dimension

[3] The exact values of r_b depend on A/ϵ, but the differences are small as can be seen in Fig. 10.1.

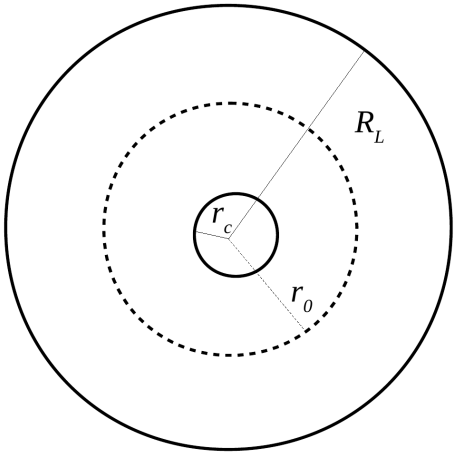

Figure 10.2: Schematic representation of the distances r_c (where the potential becomes almost flat), r_0 (mean distance between particles in a system with concentration ρ) and R_L (characteristic size of a system with volume V)

of the system, as represented in Fig.10.2. We assume that the potential is constant in the range $r_c < r < R_L$ so that one can compute the collision time again via Eq. 3.18 with the appropriate boundaries, but now completely analytically since $e^{-\beta V(r)}$ cancels out, i.e.,

$$\tau_{c,2} = \tau_{0\to c} = \frac{1}{D_0}\int_{r_c}^{r_0}\frac{\mathrm{d}r}{r^2}\int_{r}^{R_L} r'^2\mathrm{d}r' \tag{10.8}$$

now with the cutoff at the characteristic linear dimension of the system. The double integral of Eq. 10.8 can be easily performed and it yields

$$\tau_{c,2} = \frac{1}{3D_0}\left[R_L^3\left(\frac{1}{r_c}-\frac{1}{r_0}\right)+\frac{1}{2}\left(r_c^2-r_0^2\right)\right]. \tag{10.9}$$

What remains is to set the values of r_c, r_0, and R_L. For R_L, one can simply estimate it as a characteristic length so that the volume $V = L^3$ of a cubic simulation box equals the volume $V = 4\pi R_L^3/3$ of a sphere with radius R_L, and one gets

$$R_L = \left(\frac{3}{4\pi}\right)^{1/3} L \approx 0.62L. \tag{10.10}$$

We note that assuming spherical symmetry to compute the integral in Eq. 10.8 is an approximation when considering systems of different geometries. The advantage is that one has a closed and simple form for $\tau_{c,2}$. Besides, this should not be a major problem given the approximate character of the approach.

Now, to incorporate the effect of the number concentration $\rho = N/V$ into the value of r_0 (and $\tau_{c,2}$), we consider the association process between only two particles occurring in an effective small volume $V_0 = 4\pi r_0^3/3$. Here, V_0 is a characteristic volume that is estimated by considering that the two particles' system has the same number

concentration as the many-body system, that is, $\rho = 2/V_0$, hence one has that

$$r_0 = \left(\frac{3}{2\pi\rho}\right)^{1/3} \tag{10.11}$$

so that r_0 is defined as the mean distance between two particles, which carries the dependence on the number concentration ρ. Now we can give a clearer interpretation of Eq. 10.5. The association rate $k_a(T)$ is defined by the reciprocal of the sum between $\tau_{c,N}$ and $\tau'_{c\to b}$. The first term is the collision or hit time for a particle starting at a distance r_0 to arrive at the border of the region with radius r_c, where the potential ceases to be constant and becomes $V(r)$, i.e., for $r < r_c$. This process is governed by free diffusion, and the influence of other particles is taken into account indirectly by the denominator in Eq. 10.7. When the particle gets to r_c, we then estimate the time it takes to go from r_c to the bounded "state" at $r = r_b$, namely $\tau'_{c\to b}$, now under the influence of the potential. Note that placing the cutoff on the inner integral of Eq. 10.6 at r_c, $\tau'_{c\to b}$ is the MFPT computed without allowing the particle to go beyond the region $r > r_c$. This is justified because the time it spends in the region $r > r_c$ is already taken into account by the collision time.

We have already defined the expressions for r_0 and R_L, so it only remains to set a value for the cutoff distance r_c. The determination of this parameter is a little more subtle since it should be the distance where the potential becomes constant, but $V(r)$ changes continuously with r. Because of that, we consider a range for r_c, which should be system dependent. For example, in the systems studied below we consider $\rho = N/V \approx 6.25 \times 10^{-3}\sigma^{-3}$ (i.e., $N = 100$ and $V = L^3 = 1.6 \times 10^4\sigma^3$) so that Eq. 10.11 yields $r_0 \approx 4.2\sigma$, and, as the scheme in Fig. 10.2 indicates, r_c should be smaller than r_0. In this way, we consider values of r_c in the interval between $r = 3\sigma$ and $r = 4\sigma$, since, as can be seen in Fig. 10.1, a minimum distance r where the potential $V(r)$ seems to become constant is around $r = 3\sigma$. By defining such a interval, one can estimate a range of temperatures where T_m is expected to lie within, as we demonstrate in the following for the systems considered here.

10.3 Results and validation

Here we show the computed association and dissociation rates as described in Sec. 10.2, and also the estimates of T_m defined by the temperature where both rates are equal. To validate the theoretical estimates for T_m we performed many-body Brownian simulations for systems with $N = 100$ particles in a volume $V = L^3 = 1.6 \times 10^4\sigma^3$ with $\sigma = 1$. The simulations were performed numerically integrating Eq. 5.3 in the same way we had done for the pure LJ system in Sec. 5.2, but now for the pairwise potential given by Eq. 10.1.

In Fig. 10.3 we show the theoretically computed association $k_a(T)$ and dissociation

$k_d(T)$ rates as a function of temperature T for the ratios $A/\epsilon = 8$ in panels (a) and (b), and $A/\epsilon = 14$ in panels (d) and (e). While in Figs. 10.3(a) and (d) we used $r_c = 3$, in Figs.(b) and (e) we considered $r_c = 4$ for comparison. As can be seen in this figure, no drastic changes are observed in the rates for the different values of r_c, although the values of T_m obtained from the condition that $k_a(T_m) = k_d(T_m)$ are not really the same, as indicated by the dashed lines, red for $r_c = 3$ and orange for $r_c = 4$. Besides, in panels (c) and (f) of Fig. 10.3, we show the temperature-dependent specific heat $C_V(T)$ computed from the many-body simulations to corroborate the theoretically estimated transition temperatures for each value of A/ϵ. The specific heats were computed in the usual way from the variance of the energy E, considering the time series of E at each temperature, that is,

$$C_V(T) = \frac{\langle E^2 \rangle - \langle E \rangle^2}{NT^2}. \tag{10.12}$$

As can be seen in Figs. 10.3(c) and (f), the specific heats have a maximum near the values of T_m estimated by the semi-analytical method we introduced in the previous section. The maxima in the graphs of $C_V(T)$ indicate larger energy fluctuations near that range of temperatures, suggesting that the crossover transition from a diluted state to the equilibrium cluster phase occurs at temperatures close to the values of T_m predicted from our new method. As detailed in the following, this scenario of microphase separation was also verified qualitatively by looking the typical configurations obtained at different temperatures around T_m. Although the dependence of T_m on the parameter r_c can be seen as a weakness of the approach, it turned out that considering two limiting values for r_c is useful, because in this way it is possible to define a range of temperatures where the true value of T_m should lie. This is justified as intermediate values of r_c give predictions inside that temperature range (data not shown). Moreover, the range of the theoretically computed T_m is relatively small, at least for the system considered here. And, given the approximate character of the approach, any of the values presented in Fig. 10.3 for the evaluated T_m describes fairly well the crossover region in each of the systems considered. In fact, for finite systems in general, transitions occur not in a single temperature but instead in a range of temperatures (see e.g., Sec. 8.1 and 8.3). In this way, the proposed approach can be valuable for the situations where the information about the crossover region is scarce. Thus, through a simple and fast manner, one can estimate a limited range of temperatures where the microphase separation is expected to happen. Then, a more detailed analysis can be performed in that restricted region of temperatures.

Finally, we show in Fig. 10.4 the theoretical phase diagram, i.e., the estimated temperatures T_m as a function of A/ϵ, obtained using both $r_c = 3$ and $r_c = 4$, as dashed lines. Moreover, we also show the crossover temperatures determined from the maxima of $C_V(T)$ obtained from the many-body Brownian simulations (filled orange symbols). It can be seen that for the values of A/ϵ considered, the symbols lie between the two dashed

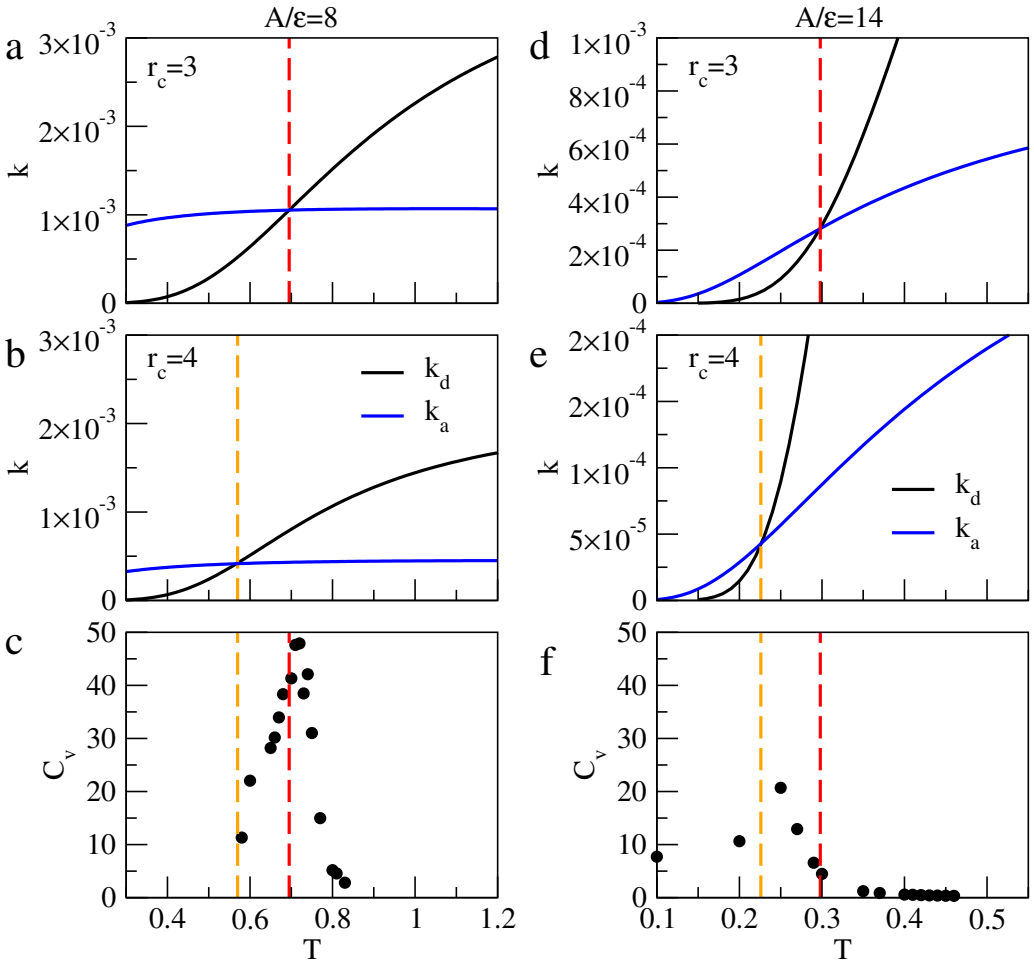

Figure 10.3: Illustration of the semi-analytical method proposed to compute T_m. (a-c) refers to $A/\epsilon = 8$, while (d-f) to $A/\epsilon = 14$. In (a) and (d), the rates are computed with $r_c = 3$, while in (b) and (e), with $r_c = 4$. Continuous black and blue curves denote the temperature-dependent dissociation $k_d(T)$ and association $k_a(T)$ rates, respectively. Dashed red and orange vertical lines refer to the location of T_m according to the criteria $k_d(T_m) = k_a(T_m)$ for $r_c = 3$ and $r_c = 4$, respectively. Panels (c) and (e) represent the specific heats $C_V(T)$ also as a function of T computed from Brownian simulations of the many-body systems with $N = 100$ particles and $\rho = 6.25 \times 10^{-3}$.

lines, validating the theoretical approach introduced here. We also show representative configurations of each region of the phase diagram. Particularly, we show configurations for $A/\epsilon = 10$ at two distinct temperatures, as indicated by the blue and red crosses. It can be seen that at low temperatures multiple small aggregates are formed in conformity with the expected behavior observed in microphase separation, while at high temperatures the system is almost all diluted. We note that for small values of A/ϵ, deviations are expected, since in that case the potential does not present a barrier, and one cannot interpret a particle subject to that potential as a problem of two states (associated and disassociated); thus, the association and dissociation rates defined by the inverse of the MFPTs might not be meaningful. In those situations, the association and dissociation dynamics is closer to a diffusion-limited process instead of a reaction-limited one, implicitly assumed in our

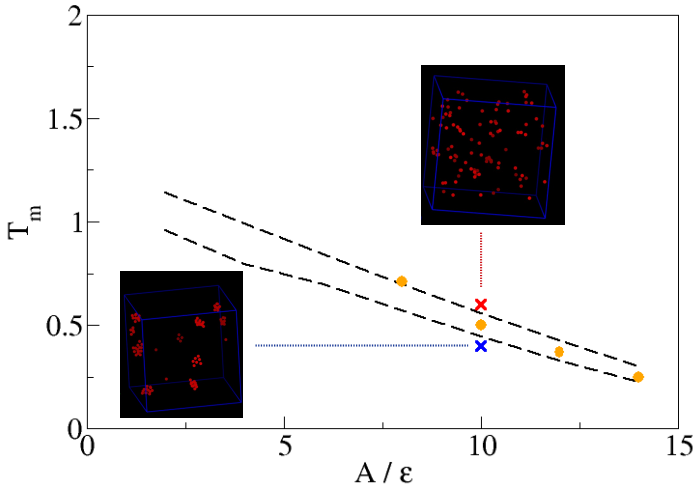

Figure 10.4: Phase diagram where the dashed lines represent the theoretical estimates of the temperature T_m that delimits the onset of microphase separation as a function of A/ϵ for $r_c = 3$ (top line) and $r_c = 4$ (bottom line). The orange points indicate the location of the peak of the specific heats obtained by the many-body simulations. The configurations on the top and bottom are typical ones of each region of the phase diagram; here we used particularly $A/\epsilon = 10$, while $T = 0.4$ and $T = 0.6$.

approach.

In conclusion, we expect that the approach that is being proposed here may be valuable for first estimates of the regions of the phase diagram where microphase separation, i.e., the appearance of the equilibrium cluster phase, takes place in systems with competing interactions. It is worth noting that we have deliberately avoided calling T_m a transition temperature since it is known that, for such systems, the microphase separation may preempt the usual first-order phase transition of simple colloidal particles with attractive plus excluded volume interactions [114]. Indeed, by increasing the strength of the repulsion, e.g., by increasing the ratio A/ϵ, the system passes through the so-called Lifshitz point, where the first-order macrophase separation is replaced by microphase separation [120]. The latter, which is also known as a crossover transition, does not present signatures of a first-order phase transition, and it is not even clear if it can be accommodated in a usual thermostatistical phase transition definition. Indeed, from the simulations presented in this chapter, i.e., for $A/\epsilon \geq 8$, we did not find signatures of first-order transitions. In particular, the PDFs of the RCs, such as energy E and the largest cluster size n, do not display any bimodal behavior, which is characteristic of first-order phase transitions in finite systems. Thus, we conclude that these points of the phase diagram are above the Lifshitz point. This region was indeed the one we were interested in, since below the Lifshitz point the system behaves like a simple aggregating colloidal system [120], so that usual approaches for computing phase diagrams, such as the law of correspondent states, should work better.

We also note that the study presented in this chapter is still in progress and we aim to perform complementary analyses in the future. For example, we used the peak of the specific heats as an indicator of cooperative behavior associated with the formation of the clusters. However, other approaches seem to rely on the structure factor, which should present a peak at a finite value of the wavevector, indicating a characteristic size directly related to the cluster sizes. We might perform such kind of additional analysis in the future to see how it compares to our results. Although the structure factor analysis provides more direct evidence of cluster formation, it also has a drawback because one usually needs to use heuristic criterion to define a threshold for the height of the peak related to the formation of finite-sized clusters [120]. Moreover, it will be interesting to see how our theoretical computation of T_m performs at different concentrations, since we were able to include some indirect information about the concentration via $\tau_{c,N}$ and r_0, albeit the analysis is almost strictly based on the properties of the pairwise potential. In this way it should be possible to build a 3D phase diagram (T, A/ϵ, ρ). Certainly, though, the approach will fail for sufficiently higher concentrations, where many-body interactions are important and other interesting phenomena, such as gelation, occur.

Part V

Conclusions

Chapter 11

Concluding remarks

In this work we were mainly interested in the kinetic description of many-body aggregating systems and also in how the kinetic properties are related with quantities extracted from usual equilibrium analysis. This problem turns out to be an issue of reduction of dimensionality. As one can not usually track the full system's dynamics in configurational space because of the huge dimensionality of such space, it is necessary to perform projections of the configurational space onto low-dimensional RCs. Although one can nowadays numerically integrate the configurational space by means of simulations, low-dimensional RCs are still needed for interpretation of the results. Indeed, it would be almost impossible even to visualize a space consisting of all positions of a many-body system, let alone interpret it physically and build effective theories looking directly at that high-dimensional space. This closely relates to the founding of thermodynamics, in which it was necessary to derive a description of the equilibrium properties of systems with many particles looking only at a limited set of macroscopic observables, such as energy, volume, and others. The computation of order parameters, specifically, requires such reduction of dimensionality. Here, we were interested in computing not only good order parameters, which describe well the equilibrium states of the system, but also in a steep forward, that is, reliable good reaction coordinates, which also describe the system's kinetics.

In Part II, we trackled the problem of computing an exact or optimal RC, which allows for a comprehensive, quantitative, and simple kinetic description. Although any projected RC can, in principle, be described by the Mori-Zwanzig projector operator formalism, the projection over usual RCs is expected to present strong non-Markovian effects. In the Mori-Zwanzig formalism, such non-Markovian effects can be described with generalized Langevin equations with a memory kernel, which is challenging to compute in practical situations. Moreover, it complicates the simple interpretation of diffusion on a potential landscape, which holds for the usual Langevin dynamics. By approximating the description of the systems with a diffusive model, our results indicate that one usually gets inaccurate descriptions with faster kinetics than the real one. One alternative is to optimize the RC in order to mitigate the non-Markovian effects. In Part II we introduced

a methodology to compute the optimal committor RC for the aggregation transition in many-body systems [23]. By combining a systematic way of computing invariant collective coordinates with a nonparametric variational approach, we were able to determine such optimal RC for both 2D lattice systems evolving through Monte Carlo stochastic dynamics, as well as 3D systems with LJ interacting particles simulated with Brownian dynamics. We confirmed that the projected dynamics along the committor can be accurately described as simple diffusion on the associated free-energy landscape.

In Part III, we turn our focus to obtaining reliable kinetic information from suboptimal reaction coordinates. We showed that energy can be used as an RC under specific conditions, i.e., with a suitable separation of timescales. By doing that, we demonstrated that our proposed closed expressions for the aggregation and dilution rates can accurately describe the system's kinetics [24]. This achievement allows one to make a more direct connection of kinetic aspects of phase transformation with usual equilibrium descriptors, since energy is the usual variable in thermostatistical approaches.

As a final contribution, we also demonstrated, in Part IV, that valuable information about the phase diagram of systems with competing interactions can be obtained solely by considering the pairwise interaction potential through the MFPT-based stochastic approach introduced here. By analyzing the association and dissociation of a single particle subject to the potential, we could infer the phase behavior of many-body systems, i.e., the location of the regions where microphase separation occurs.

Bibliography

[1] REIF, F. *Fundamentals of statistical and thermal physics.* Waveland Press, 1965.

[2] LABASTIE, P.; WHETTEN, R. L. Statistical thermodynamics of the cluster solid-liquid transition. *Phys. Rev. Lett.*, v. 65, p. 1567–1570, 1990.

[3] CHO, W. J.; KIM, J.; LEE, J.; KEYES, T.; STRAUB, J. E.; KIM, K. S. Limit of metastability for liquid and vapor phases of water. *Phys. Rev. Lett.*, v. 112, p. 157802, 2014.

[4] AUER, S.; KASHCHIEV, D. Phase diagram of α-helical and β-sheet forming peptides. *Phys. Rev. Lett.*, v. 104, p. 168105, 2010.

[5] WANG, Y.; BUNCE, S. J.; RADFORD, S. E.; WILSON, A. J.; AUER, S.; HALL, C. K. Thermodynamic phase diagram of amyloid-β(16-22) peptide. *Proc. Natl. Acad. Sci. U.S.A.*, v. 116, p. 2091, 2019.

[6] DUAN, C.; WANG, R. Electrostatics-induced nucleated conformational transition of protein aggregation. *Phys. Rev. Lett.*, v. 130, p. 158401, 2023.

[7] SOSSO, G. C.; CHEN, J.; COX, S. J.; FITZNER, M.; PEDEVILLA, P.; ZEN, A.; MICHAELIDES, A. Crystal nucleation in liquids: Open questions and future challenges in molecular dynamics simulations. *Chem. Rev.*, v. 116, p. 7078, 2016.

[8] HAO, M.; SCHERAGA, H. A. Monte carlo simulation of a first-order transition for protein folding. *J. Phys. Chem.*, v. 98, p. 4940–4948, 1994.

[9] LIU, Y.; KELLOGG, E.; LIANG, H. Canonical and micro-canonical analysis of folding of trpzip2: An all-atom replica exchange Monte Carlo simulation study. *J. Chem. Phys.*, v. 137, p. 045103, 2012.

[10] HAGAN, M. F.; GRASON, G. M. Equilibrium mechanisms of self-limiting assembly. *Rev. Mod. Phys.*, v. 93, p. 025008, 2021.

[11] KNOWLES, T. P. J.; VENDRUSCOLO, M.; DOBSON, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell. Biol.*, v. 15, p. 384–396, 2014.

[12] EVES, B. J.; DOUTCH, J. J.; TERRY, A. E.; YIN, H.; MOULIN, M.; HAERTLEIN, M.; FORSYTH, V. T.; FLAGMEIER, P.; KNOWLES, T. P. J.; DIAS, D. M.; LOTZE, G.; SEDDON, A. M.; SQUIRES, A. M. Elongation rate and average length of amyloid fibrils in solution using isotope-labelled small-angle neutron scattering. *RSC Chem. Biol.*, v. 2, p. 1232–1238, 2021.

[13] NADLER, W.; HANSMANN, U. H. E. Generalized ensemble and tempering simulations: A unified view. *Phys. Rev. E*, v. 75, p. 026109, 2007.

[14] BALAKRISHNAN, V. *Elements of nonequilibrium statistical mechanics*. Springer, 2021.

[15] KUHNHOLD, A.; MEYER, H.; AMATI, G.; PELAGEJCEV, P.; SCHILLING, T. Derivation of an exact, nonequilibrium framework for nucleation: Nucleation is a priori neither diffusive nor markovian. *Phys. Rev. E*, v. 100, p. 052140, 2019.

[16] TEPPER, L.; DALTON, B.; NETZ, R. R. Accurate memory kernel extraction from discretized time-series data. *J. Chem. Theory Comput.*, v. 20, p. 3061–3068, 2024.

[17] KRIVOV, S. V. On reaction coordinate optimality. *J. Chem. Theory Comput.*, v. 9, p. 135–146, 2013.

[18] BEREZHKOVSKII, A. M.; SZABO, A. Diffusion along the splitting/commitment probability reaction coordinate. *J. Phys. Chem. B*, v. 117, p. 13115–13119, 2013.

[19] LU, J.; VANDEN-EIJNDEN, E. Exact dynamical coarse-graining without time-scale separation. *J. Chem. Phys.*, v. 141, p. 044109, 2014.

[20] PETERS, B. Reaction coordinates and mechanistic hypothesis tests. *Annu. Rev. Phys. Chem.*, v. 67, p. 669–690, 2016.

[21] DU, R.; PANDE, V. S.; GROSBERG, A. Y.; TANAKA, T.; SHAKHNOVICH, E. S. On the transition coordinate for protein folding. *J. Chem. Phys.*, v. 108, p. 334–350, 1998.

[22] BANUSHKINA, P. V.; KRIVOV, S. V. Nonparametric variational optimization of reaction coordinates. *J. Chem. Phys.*, v. 143, p. 184108, 2015.

[23] TRUGILHO, L. F.; AUER, S.; RIZZI, L. G.; KRIVOV, S. V. Nonparametric determination of the committor in multimolecular systems. *J. Chem. Theory Comput.*, v. 21, p. 10080–10085, 2025.

[24] TRUGILHO, L. F.; AUER, S.; RIZZI, L. G. A density of states-based approach to determine temperature-dependent aggregation rates. *J. Chem. Phys.*, v. 161, p. 051101, 2024.

[25] RIZZI, L. G. Kinetics of first-order phase transitions from microcanonical thermostatistics. *J. Stat Mech.*, v. 2020, p. 083204, 2020.

[26] TRUGILHO, L. F.; RIZZI, L. G. Shape-free theory for the self-assembly kinetics in macromolecular systems. *EPL*, v. 137, p. 57001, 2022.

[27] GROSS, D. H. E. *Microcanonical termodynamics: Phase transitions in "small" systems.* World Scientific, 2001.

[28] SCIORTINO, F.; TARTAGLIA, P.; ZACCARELLI, E. One-dimensional cluster growth and branching gels in colloidal systems with short-range depletion attraction and screened electrostatic repulsion. *J. Phys. Chem. B*, v. 109, p. 21942–21953, 2005.

[29] BELLAC, M. L. *Quantum and statistical field theory.* Clarendon Press, 1991.

[30] GILLESPIE, D. T. Fluctuation and dissipation in brownian motion. *Am. J. Phys.*, v. 61, p. 1077, 1993.

[31] ZHOU, H. Rate theories for biologists. *Q. Rev. Biophys.*, v. 43, p. 219–293, 2010.

[32] JANSSEN, J.; MANCA, O.; MANCA, R. *Applied diffusion processes from engineering to finance.* Wiley, 2013.

[33] LANGEVIN, P. Sur la théorie du mouvement brownien. *C. R. Acad. Sci. (Paris)*, v. 146, p. 530–533, 1908.

[34] LEMONS, D. S.; GYTHIEL, A. Paul langevin's 1908 paper "on the theory of brownian motion" ["sur la théorie du mouvement brownien", c. r. acad. sci. (paris) 146, 530–533 (1908)]. *Am. J. Phys.*, v. 65, p. 1079, 1997.

[35] KAMPEN, N. G. V. *Sochastic process in physics and chemistry (third edition).* North-Holland Personal Library. Elsevier, 2007.

[36] GARDINER, C. W. *Handbook of stochastic methods.* Springer, 2004.

[37] BEREZHKOVSKII, A. M.; SZABO, A. Committors, first-passage times, fluxes, markov states, milestones, and all that. *J. Phys. Chem.*, v. 150, p. 054106, 2019.

[38] TREBST, S.; HUSE, D. A.; TROYER, M. Optimizing the ensemble for equilibration in broad-histogram monte carlo simulations. *Phys. Rev. E*, v. 70, p. 046701, 2004.

[39] METZNER, P.; SCHÜTTE, C.; VANDEN-EIJNDEN, E. Illustration of transition path theory on a collection of simple examples. *J. Chem. Phys.*, v. 125, p. 084110, 2006.

[40] SZABO, A.; SCHULTEN, K.; SCHULTEN, Z. First passage time approach to diffusion controled reactions. *J. Chem. Phys.*, v. 72, p. 4350, 1980.

[41] BEREZHKOVSKII, A. M.; PUSTOVOIT, M. A.; BEZRUKOV, S. M. Channel-facilitated membrane transport: Average lifetimes in the channel. *J. Chem. Phys.*, v. 119, p. 3943–3951, 2003.

[42] BEREZHKOVSKII, A. M.; DAGDUG, L.; BEZRUKOV, S. M. Mean direct-transit and looping times as functions of the potential shape. *J. Phys. Chem. B*, v. 121, p. 5455–5460, 2017.

[43] BANUSHKINA, P. V.; KRIVOV, S. V. Optimal reaction coordinates. *WIREs Comput. Mol. Sci.*, v. 6, p. 748–763, 2016.

[44] CHAVANIS, P. H. Hamiltonian and brownian systems with long-range interactions: Ii kinetic equations and stability analysis. *Physica A*, Amsterdam, v. 361, p. 81–123, 2006.

[45] E, W.; VANDEN-EIJNDEN, E. Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.*, v. 61, p. 391–420, 2010.

[46] KHOO, Y.; LU, J.; YING, L. Solving for high-dimensional committor functions using artificial neural networks. *Res. Math. Sci.*, v. 6, p. 1, 2019.

[47] KRIVOV, S. V. Numerical construction of the p-fold (committor) reaction coordinate for a markov process. *J. Phys. Chem. B*, v. 115, p. 11382–11388, 2011.

[48] LI, Q.; LIN, B.; REN, W. Computing committor functions for the study of rare events using deep learning. *J. Chem. Phys.*, v. 151, p. 054112, 2019.

[49] KANG, P.; TRIZIO, E.; PARRINELLO, M. Computing the committor with the committor to study the transition state ensemble. *Nat. Comput. Sci.*, v. 4, p. 451–460, 2024.

[50] CHEN, Y.; HOSKINS, J.; KHOO, Y.; LINDSEY, M. Committor functions via tensor networks. *J. Comput. Phys.*, v. 472, p. 111646, 2023.

[51] CHAHINE, J.; OLIVEIRA, R. J.; LEITE, V. B. P.; WANG, J. Configuration-dependent diffusion can shift the kinetic transition state and barrier height of protein folding. *Proc. Natl. Acad. Sci. U.S.A.*, v. 104, p. 14646–14651, 2007.

[52] KASHCHIEV, D. *Nucleation: Basic theory with applications.* Butterworth-Heinemann, 2000.

[53] SOSSO, G. C.; CHEN, J.; COX, S. J.; FITZNER, M.; PEDEVILLA, P.; ZEN, A.; MICHAELIDES, A. Crystal nucleation in liquids: Open questions and future challenges in molecular dynamics simulations. *Chem. Rev.*, v. 116, p. 7078–7116, 2016.

[54] ZIERENBERG, J.; SCHIERZ, P.; JANKE, W. Canonical free-energy barrier of particle and polymer cluster formation. *Nat. Commun.*, v. 8, p. 14546, 2017.

[55] VROYLANDT, H. On the derivation of the generalized langevin equation and the fluctuation-dissipation theorem. *EPL*, v. 140, p. 62003, 2022.

[56] VANDEN-EIJNDEN, E.; VENTUROLI, M.; CICCOTTI, G.; ELBER, R. On the assumptions underlying milestoning. *J. Chem. Phys.*, v. 129, p. 174102, 2008.

[57] VANDEN-EIJNDEN, E.; VENTUROLI, M. Exact rate calculations by trajectory parallelization and tilting. *J. Chem. Phys.*, v. 131, p. 044120, 2009.

[58] KRIVOV, S. V. Protein folding free energy landscape along the committor - the optimal folding coordinate. *J. Chem. Theory Comput.*, v. 14, p. 3418–3427, 2018.

[59] KRIVOV, S. V.; KARPLUS, M. Diffusive reaction dynamics on invariant free energy profiles. *Proc. Natl. Acad. Sci. U.S.A.*, v. 105, p. 13841–13846, 2008.

[60] RHEE, Y. M.; PANDE, V. S. One-dimensional reaction coordinate and the corresponding potential of mean force from commitment probability distribution. *J. Chem. Phys.*, v. 109, p. 6780–6786, 2005.

[61] HUSIC, B. E.; PANDE, V. S. Markov state models: From an art to a science. *J. Am. Chem. Soc.*, v. 140, p. 2386–2396, 2018.

[62] KRIVOV, S. V. Nonparametric analysis of nonequilibrium simulations. *J. Chem. Theory Comput.*, v. 17, p. 5466–5481, 2021.

[63] APAYDIN, M. S.; BRUTLAG, D. L.; GUESTRIN, C.; HSU, D.; LATOMBE, J.; VARMA, C. Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion. *Journal of Computational Biology*, v. 10, n. 3-4, p. 257–281, 2003.

[64] SINGHAL, N.; SNOW, C. D.; PANDE, V. S. Using path sampling to build better markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.*, v. 121, p. 415, 2004.

[65] GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[66] BANUSHKINA, P. V.; KRIVOV, S. V. Nonparametric reaction coordinate optimization with histories: A framework for rare event dynamics, 2025.

[67] KRIVOV, S. V. Blind analysis of molecular dynamics click to copy article link. *J. Chem. Theory Comput.*, v. 17, p. 2725–2736, 2021.

[68] HERRINGER, N. S. M.; DASETTY, S.; GANDHI, D.; LEE, J.; FERGUSON, A. L. Permutationally invariant networks for enhanced sampling (pines): Discovery of multimolecular and solvent-inclusive collective variables. *J. Chem. Theory Comput.*, v. 20, p. 178–198, 2024.

[69] LANDAU, D. P.; BINDER, K. *A guide to monte carlo simulations in statistical physics*. Cambridge University Press, 2014.

[70] SHERIFF, K.; FREITAS, R.; TREWARTHA, A.; TORRISI, S. Simultaneous discovery of reaction coordinates and committor functions using equivariant graph neural networks. In: . c2024.

[71] ARREDONDO, S. C.; TANG, C.; TALMAZAN, R. A.; MEGÍAS, A.; CHEN, C. G.; CHIPOT, C. From atoms to dynamics: Learning the committor without collective variables, 2025.

[72] SANZ, E.; MARENDUZZO, D. Dynamic Monte Carlo versus Brownian dynamics: A comparison for self-diffusion and crystallization in colloidal fluids. *J. Chem. Phys.*, v. 132, p. 194102, 2010.

[73] MÜLLER, F.; CHRISTIANSEN, H.; JANKE, W. Phase-separation kinetics in the two-dimensional long-range ising model. *Phys. Rev. Lett.*, v. 129, p. 240601, 2022.

[74] KIKUCHI, K.; YOSHIDA, M.; MAEKAWA, T.; WATANABE, H. Metropolis monte carlo method as a numerical technique to solve the fokker-planck equation. *Chem. Phys. Lett.*, v. 185, p. 335–338, 1991.

[75] NEWMAN, M. E. J.; BARKEMA, G. T. *Monte Carlo methods in statistical physics*. Claredon Press Oxford, 1999.

[76] TRUGILHO, L. F.; RIZZI, L. G. Microcanonical thermostatistics of aggregation transition in a system with anisotropically interacting molecules. *J. Phys. Conf. Ser.*, v. 1483, p. 012011, 2020.

[77] METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equation of state calculations by fast computing machines. *J. Chm. Phys.*, v. 21, p. 1087–1092, 1953.

[78] CABRIOLU, R.; KASHCHIEV, D.; AUER, S. Breakdown of nucleation theory for crystals with strongly anisotropic interactions between molecules. *J. Chem. Phys.*, v. 137, p. 204903, 2012.

[79] BINGHAM, R. J.; RIZZI, L. G.; CABRIOLU, R.; AUER, S. Communication: Non-monotonic supersaturation dependence of the nucleus size of crystals with anisotropically interacting molecules. *J. Chem. Phys.*, v. 139, p. 241101, 2013.

[80] RIZZI, L. G.; HEAD, D. A.; AUER, S. Universality in the morphology and mechanics of coarsening amyloid fibril networks. *Phys. Rev. Lett.*, v. 114, p. 078102, 2015.

[81] RIZZI, L. G.; AUER, S. Amyloid fibril solubility. *J. Phys. Chem. B*, v. 119, p. 14631–14636, 2015.

[82] KNOWLES, T. P.; FITZPATRICK, A. W.; MEEHAN, S.; MOTT, H. R.; VENDRUSCOLO, M.; DOBSON, C. M.; WELLAND, M. E. Role of intermolecular forces in defining material properties of protein nanofibrils. *Science*, Washington, v. 318, p. 1900, 2007.

[83] BANUSHKINA, P. V.; KRIVOV, S. V. Fep1d: a script for the analysis of reaction coordinates. *J. Comput. Chem.*, v. 36, p. 878–882, 2015.

[84] MIE, G. Zur kinetischen theorie der einatomigen körper. *Annalen der Physik*, v. 316, n. 8, p. 657–697, 1903.

[85] HASEGAWA, M.; OHNO, K. The dependence of the phase diagram on the range of the attractive intermolecular forces. *J. Phys.: Condens. Matter*, v. 9, p. 3361, 1997.

[86] VLIEGENTHART, G. A.; LODGE, J. F. M.; LEKKERKERKER, H. N. W. Strong weak and metastable liquids structural and dynamical aspects of the liquid state. *Physica A*, Amsterdam, v. 263, p. 378–388, 1999.

[87] CAMPBELL, A. I.; ANDERSON, V. J.; VAN DUIJNEVELDT, J. S.; BARTLETT, P. Dynamical arrest in attractive colloids: The effect of long-range repulsion. *Phys. Rev. Lett.*, v. 94, p. 208301, 2004.

[88] DUARTE, L. K. R.; TEIXEIRA, A. V. N. C.; RIZZI, L. G. Microrheology of semiflexible filament solutions based on relaxation simulations. *Soft Matter*, v. 17, p. 2920, 2021.

[89] CALLEN, H. B. *Thermodynamics: an introduction to the physical theories of equilibrium thermostatics and irreversible thermodynamics.* John Wiley & Sons, 1960.

[90] SCHNABEL, S.; SEATON, D. T.; LANDAU, D. P.; BACHMANN, M. Microcanonical entropy inflection points: Key to systematic understanding of transitions in finite systems. *Phys. Rev. E*, v. 84, p. 011127, 2011.

[91] QI, K.; BACHMANN, M. Classification of phase transitions by microcanonical inflection-point analysis. *Phys. Rev. Lett.*, v. 120, p. 180601, 2018.

[92] SAUER, T. A look back at the ehrenfest classification. *Eur. Phys. J. Special Topics*, v. 226, p. 539, 2017.

[93] TRUGILHO, L. F.; RIZZI, L. G. Microcanonical characterization of first-order phase transitions in a generalized model for aggregation. *J. Stat. Phys.*, v. 186, p. 40, 2022.

[94] FERRENBERG, A. M.; SWENDSEN, R. H. Optimized monte carlo data analysis. *Phys. Rev Lett.*, v. 63, p. 1195–1198, 1989.

[95] BERG, B. A.; NEUHAUS, T. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev Lett.*, v. 68, p. 9–12, 1992.

[96] LEE, J. New monte carlo algorithm: Entropic sampling. *Phys. Rev Lett.*, v. 71, p. 211–214, 1993.

[97] DE OLIVEIRA, P. M. C.; PENNA, T. J. P.; HERRMANN, H. J. Broad histogram method. *Braz. J. Phys.*, v. 26, p. 677–683, 1996.

[98] WANG, F.; LANDAU, D. P. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, v. 86, p. 2050, 2001.

[99] KIM, J.; STRAUB, J. E.; KEYES, T. Statistical-temperature monte carlo and molecular dynamics algorithms. *Phys. Rev. Lett.*, v. 97, p. 050601, 2006.

[100] KIM, J.; KEYES, T.; STRAUB, J. E. Communication: Iteration-free, weighted histogram analysis method in terms of intensive variables. *J. Chem. Phys.*, v. 135, p. 061103, 2011.

[101] ZIERENBERG, J.; JANKE, W. Exploring different regimes in finite-size scaling of the droplet condensation-evaporation transition. *Phys. Rev. E*, v. 92, p. 012134, 2015.

[102] MÖDDEL, M.; JANKE, W.; BACHMANN, M. Systematic microcanonical analyses of polymer adsorption transitions. *Phys. Chem. Chem. Phys*, v. 12, p. 11548–11554, 2010.

[103] FRIGORI, R. B.; RIZZI, L. G.; ALVES, N. A. Microcanonical thermostatistics of coarse-grained proteins with amyloidogenic propensity. *J. Chem. Phys.*, v. 138, p. 015102, 2013.

[104] JUNGHANS, C.; BACHMANN, M.; JANKE, W. Microcanonical analyses of peptide aggregation processes. *Phys. Rev. Lett.*, v. 97, p. 218103, 2006.

[105] JUNGHANS, C.; BACHMANN, M.; JANKE, W. Thermodynamics of peptide aggregation processes: An analysis from perspectives of three statistical ensembles. *J. Chem. Phys*, v. 128, p. 085103, 2008.

[106] JANKE, W. Canonical versus microcanonical analysis of first-order phase transitions. *Nucl. Phys. B*, v. 63, p. 631–633, 1998.

[107] SCHIERZ, P.; ZIERENBERG, J.; JANKE, W. First-order phase transitions in the real microcanonical ensemble. *Phys. Rev. E*, v. 94, p. 021301, 2016.

[108] SCHIERZ, P.; ZIERENBERG, J.; JANKE, W. Molecular dynamics and monte carlo simulations in the microcanonical ensemble: Quantitative comparison and reweighting techniques. *J. Chem. Phys*, v. 143, p. 134114, 2015.

[109] HÄNGGI, P.; TALKNER, P.; BORKOVEC, M. Reaction-rate theory: fifty years after kramers. *Rev. Mod. Phys.*, v. 62, p. 251, 1990.

[110] KRAMERS, H. A. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, v. 7, p. 284, 1940.

[111] BACHMANN, M. Efficient microcanonical histogram analysis and application to peptide aggregation. *Phys. Rev. Lett.*, v. 135, p. 138401, 2025.

[112] DANIELSSON, J.; MU, X.; LANG, L.; WANG, H.; BINOLFI, A.; THEILLET, F.-X.; BEKEI, B.; LOGAN, D. T.; SELENKO, P.; WENNERSTRÖM, H.; OLIVEBERG, M. Thermodynamics of protein destabilization in live cells. *Proc. Natl. Acad. Sci. U.S.A.*, v. 112, n. 40, p. 12402–12407, 2015.

[113] VAN DER WAALS, J. D. *On the continuity of the gaseous and liquid states of matter*. North Holland. Edited by J. S. Rowlinson, 1988.

[114] RUIZ-FRANCO, J.; ZACCARELLI, E. On the role of competing interactions in charged colloids with short-range attraction. *Annu. Rev. Condens. Matter Phys.*, v. 12, p. 51–70, 2021.

[115] MUTHUKUMAR, M. *Physics of charged macromolecules*. Cambridge University Press, 2023.

[116] CAMPA, A.; DAUXOIS, T.; RUFFO, S. Statistical mechanics and dynamics of solvable models with long-range interactions. *Phys. Rep.*, v. 480, p. 57–159, 2009.

[117] BOMONT, J.; PASTORE, G.; COSTA, D.; MUNAÒ, G.; MALESCIO, G.; PRESTIPINO, S. Arrested states in colloidal fluids with competing interactions: A static replica study. *J. Chem. Phys.*, v. 160, p. 214504, 2024.

[118] MONTERO, A. M.; YUSTE, S. B.; SANTOS, A.; DE HARO, M. L. Discontinuous structural transitions in fluids with competing interactions. *Entropy*, v. 27, p. 95, 2025.

[119] SEAR, R. P.; GELBART, W. M. Microphase separation versus the vapor-liquid transition in systems of spherical particles. *J. Chem. Phys.*, v. 110, p. 4582–4588, 1999.

[120] COSTA, D.; MUNAÒ, G.; BOMONT, J.; MALESCIO, G.; PALATELLA, A.; PRESTIPINO, S. Microphase versus macrophase separation in the square-well-linear fluid: A theoretical and computational study. *Phys. Rev. E*, v. 108, p. 034602, 2023.

[121] NORO, M. G.; FRENKEL, D. Extended corresponding-states behavior for particles with variable range attractions. *J. Chem. Phys.*, v. 113, p. 2941, 2000.

[122] GODFRIN, P. D.; VALADEZ-PÉREZ, N. E.; CASTAÑEDA-PRIEGO, R.; WAGNER, N. J.; LIU, Y. Generalized phase behavior of cluster formation in colloidal dispersions with competing interactions. *Soft Matter*, v. 10, p. 5061–5071, 2014.

Appendix A

Transition path segments

The transition path segments (TPS) are segments of the trajectory beginning and ending at the boundary states A and B. Pathways returning to the same node (e.g., from A and back to A) are included so that the states are still sampled with the equilibrium probability and, in particular, detailed balance is still satisfied [17]. After partitioning the trajectory into these segments, each of them is continued indefinitely by repeating the first and last nodes infinitely in each direction. In fact, for practical purposes like computing the profiles, it is sufficient to repeat both first and last nodes by Δt, the sampling interval used to compute the profile. By selecting the segments in such a way that all transition paths between the boundaries are accounted for by construction, there are no undetectable transitions even for large Δt. As a consequence, $Z_{C,1}(q, \Delta t)$ is always independent of q and Δt and equals the fixed value, N_{AB}. This can be viewed alternatively as follows. The mathematical proof of the validation criteria Eq. 3.33 is based on the committor definition Eq. 3.47. If Eq.3.47 holds for q, then $Z_{C,1}(q, \Delta t)$ is independent of q and Δt [17]. However, the boundary states satisfy Eq. 3.48 rather than Eq. 3.47. In this way, $Z_{C,1}(q, \Delta t)$ can only vary for nodes i connected with the boundaries ($n_{iA} \neq 0$ or $n_{iB} \neq 0$). For small Δt, this corresponds only to a negligible fraction of the nodes near each boundary, but for large enough Δt, the whole $Z_{C,1}(q, \Delta t)$ curve is dragged down. The transition path segments simply recover the property of Eq. 3.47 for the boundaries so that $Z_{C,1}(q, \Delta t)$ is constant everywhere [58].

An illustration of the construction of the TPS is given in Fig. A.1, and we consider a simple system with only three states to exemplify how the total squared displacement (TSD) and the cut-based profiles can be computed with the TPS, in a construction adapted from Ref. [17].

The system has two boundary states, A and B, and an intermediate state I, and is defined by the transition probabilities (in the MSM notation) $P_{AI} = P_{BI} = 1/2$ and $P_{IA} = P_{IB} = 1$. Recall that P_{ij} is the probability of going from j to i. There is a set of four possible transition paths: AIA, AIB, BIB, and BIA. Note also that all four paths are equiprobable due to the rules we imposed on P_{ij}. Because of that, the equilibrium number

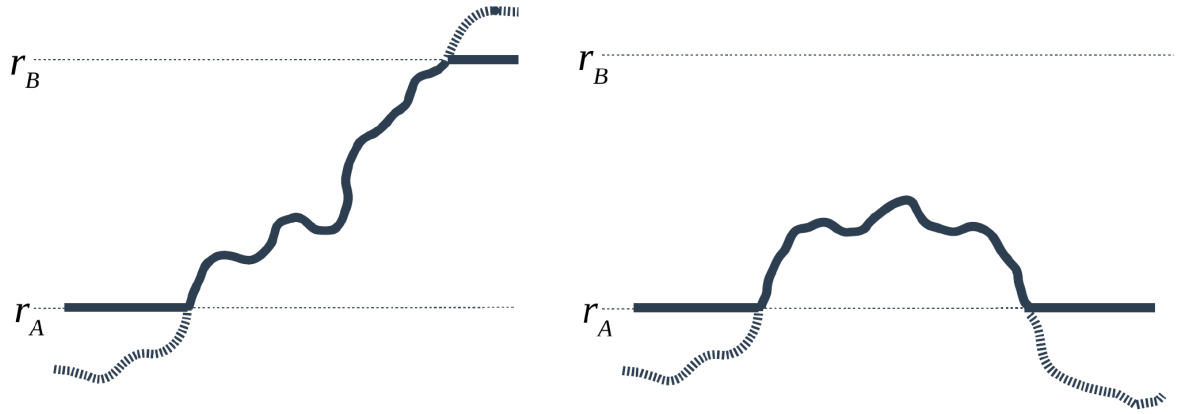

Figure A.1: Illustration of the construction of transition path segments. The dashed plus continuous lines represent part of the full trajectory. The transition paths are identified as pathways returning to the same state (left panel) or going from one state to the other (right panel). Then, the segments are extended by repeating the first and last nodes, which are represented here by the continuous horizontal lines.

of segments of each type, for a trajectory of length $\mathcal{T}$ (in time units) sampled with Δt_0, is $\mathcal{T}/(8\Delta t_0)$. This is because $\mathcal{T}/\Delta t_0$ is the total number of trajectory points, which is divided by four to account for the four types of transition paths and by two because each segment has two state changes.

Now we can calculate the TSD, $\Delta' r^2(\Delta t)$ (we use a slightly different notation for the TSD with a prime here, compared with Sec. 3.3.4 for a reason that will be clear soon), of a general reaction coordinate r_i and for different sampling intervals $\Delta t = k\Delta t_0$ for this simple system using the TPS convention. We first consider segment AIB, which is extended as ...AAAAIBBBB.... The contribution of this segment to the TSD for $\Delta t = \Delta t_0$ $(k = 1)$ is $\Delta' r^2(\Delta t_0) = \sum_i [r(i\Delta t_0) - r(i\Delta t_0 + \Delta t_0)]^2 = (r_A - r_I)^2 + (r_I - r_B)^2$ where i runs over the segment. Now, for $\Delta t = 2\Delta t_0$ $(k = 2)$, one has $\Delta' r^2(2\Delta t_0) = \sum_i [r(i\Delta t_0) - r(i\Delta t_0 + 2\Delta t_0)]^2 = (r_A - r_I)^2 + (r_I - r_B)^2 + (r_A - r_B)^2$. Note that the computation is performed with different starting snapshots here because $\Delta t > \Delta t_0$ so that if one starts at the second A before I, we have a transition from A to I and then from I to B because $k = 2$. Conversely, starting from the first A before I, we have a transition directly from A to B. Note also the reason why it is only necessary to extend the segment by $\pm\Delta t$: for extended nodes further away from $|\Delta t_0|$, the transitions are just inside each boundary. Repeating this procedure for $\Delta t = k\Delta t_0$, considering k starting nodes, one gets $\Delta r'^2(k\Delta t_0) = \sum_i [r(i\Delta t_0) - r(i\Delta t_0 + k\Delta t_0)]^2 = (r_A - r_I)^2 + (r_I - r_B)^2 + (k-1)(r_A - r_B)^2$. For the segment BIA extended as ...BBBBIAAAA..., one gets exactly the same result since it is just the same sequence reversed. Now, for the segment AIA extended as ...AAAAIAAAA... the contribution to the TSD for $\Delta t = \Delta t_0$ is $\Delta' r^2(\Delta t_0) = (r_A - r_I)^2 + (r_I - r_A)^2$. Moreover, in this case increasing Δt does not change

the contribution of this segment since for most starting nodes the displacement is null, so we have $\Delta' r^2(k\Delta t_0) = (r_A - r_I)^2 + (r_I - r_A)^2$. In an analogous way, one has, for the segment BIB extended as ...BBBBIBBBB..., $\Delta' r^2(k\Delta t_0) = (r_B - r_I)^2 + (r_I - r_B)^2$. The TSD is then computed by summing it all up and multiplying by the number of segments of each type, $(\mathcal{T}/(8\Delta t_0))$, which gives

$$\Delta' r^2(k\Delta t_0) = (\mathcal{T}/8\Delta t_0)[4(r_A - r_I)^2 + 4(r_B - r_I)^2 + 2(k-1)(r_A - r_B)^2]. \tag{A.1}$$

We can then define the mean squared displacement (MSD), dividing the TSD by the number of trajectory snapshots $(\mathcal{T}/\Delta t_0)$. Note that such defined MSD grows indefinitely with Δt (or k) even for a system with finite configuration space because it is computed with segments growing in size while being divided by the original trajectory length $(\mathcal{T}/\Delta t_0)$, which is constant. It will not saturate at high Δt, as can be seen, e.g., in Fig. 8.2. Conversely, the MSD computed in the conventional way, directly from a trajectory time series, will always saturate to an equilibrium value for systems with limited configuration space. Both ways to compute the MSD will be in agreement for small times, when the boundary states can be neglected [17]. Moreover, in the more usual way, one usually computes the MSD by selecting a single state as a starting point, while when using the TSD, an average over all states as starting points (with equilibrium probabilities) is considered.

Specifying Eq. A.1 for the committor q_i, where for this simple system with three states one simply has $q_A = 0$, $q_I = 1/2$ and $q_B = 1$, one gets

$$\Delta' q^2(k\Delta t_0) = \frac{\mathcal{T}}{4\Delta t_0} k, \tag{A.2}$$

while the MSD is simply $k/4$. Albeit the different numerical factors, both increase linearly with Δt (k) for all times, as it is for simple diffusion. Now, recall that in Sec. 3.3.4 we stated that, for the committor, the TSD attains a minimum given by $\Delta q^2 = 2N_{AB}$, which is independent of Δt (see Eq. 3.52). The matter is just that in Sec. 3.3.4 we defined the TSD in a slightly different manner, without summing over different starting points of the TPS (or the trajectory in general) for $\Delta t > \Delta t_0$, like we did in this section. That is why we used the different notation with a prime $\Delta' r^2$ here. Since for $\Delta t = k\Delta t_0$, $k > 1$, there are k different starting points available, both definitions are simply related by $\Delta r^2 = \Delta' r^2/k$. In other words, for bigger sampling intervals, one can average over the starting point, i.e., sum over the k starting points and then divide by k. In this way, for the system with three states considered here, we have $\Delta q^2(k\Delta t_0) = \Delta' q^2(k\Delta t_0)/k = \mathcal{T}/(4\Delta t_0)$ so that $N_{AB} = \mathcal{T}/(8\Delta t_0)$ by virtue of Eq. 3.52. This is consistent since $\mathcal{T}/(8\Delta t_0)$ is the equilibrium number of segments, AIB and BIA, which in this case is exactly the number of transitions between the boundaries, which in equilibrium is equal in both

directions $N_{AB} = N_{BA}$. The interpretation of a constant $\Delta q^2(k\Delta t_0)$ is that while the total squared displacement grows linearly with k, the number of trajectory points considered for computing $\Delta q^2(k\Delta t_0)$ is reduced like k^{-1}. For example, for $k = 2$, only half the trajectory points are considered. Equivalently, one can sum the two intercalating trajectories but then take the mean, dividing by two.

Now we exemplify the computation of the cut-based profile, $Z_{C,1}(r, \Delta t)$, using the TPS. $Z_{C,1}(r, \Delta t)$ is defined by Eq. 3.31 with $a = 1$. But as it equals half the sum of times the trajectory passes through point r after successive time displacements Δt, weighted by the length of each coordinate displacement, it can be written more compactly as

$$Z_{C,1}(r, k\Delta t_0) = \frac{1}{2}\sum_i{}' |r(ik\Delta t_0) - r(ik\Delta t_0 + k\Delta t_0)| \tag{A.3}$$

where the prime indicates that the sum is over all trajectory (or TPS) snapshots i such that r is between $r(ik\Delta t_0)$ and $r(ik\Delta t_0 + k\Delta t_0)$. Similar to the TSD, it can be averaged over the initial points for $\Delta t > \Delta t_0$

$$Z_{C,1}(r, k\Delta t_0) = \frac{1}{2k}\sum_i{}' |r(i\Delta t_0) - r(i\Delta t_0 + k\Delta t_0)|. \tag{A.4}$$

Now we can compute it for the system with three states. As a difference compared to the computation of the TSD, here one needs to discriminate what transitions contribute to each coordinate interval. We assume $r_A < r_I < r_B$, which will be valid for the committor. Segment ...AAAAIAAAA..., for example, only contributes to the interval $r_A < r < r_I$ since there are only transitions between A and I, so these transitions do not pass through B. For $\Delta t = \Delta t_0$, it contributes to the profile in that interval by $Z_{C,1}(r, \Delta t_0) = 1/2(|r_A - r_I| + |r_I - r_A|) = |r_A - r_I|$. For $\Delta t = 2\Delta t_0$, the same transitions occur, but one needs to divide by two to average, so $Z_{C,1}(r, 2\Delta t_0) = 1/4(2|r_A - r_I|) = 1/2|r_A - r_I|$. Thus one has for $\Delta t = k\Delta t_0 = 1/k|r_A - r_I|$. Analogously, segment ...BBBBIBBBB... contributes with $Z_{C,1}(r, k\Delta t_0) = 1/k|r_B - r_I|$, but now for the interval $r_I < r < r_B$. Now, segment ...AAAAIBBBB... contributes to both intervals. In relation to the interval $r_A < r < r_I$, one has for $\Delta t = \Delta t_0$, $Z_{C,1}(r, \Delta t_0) = 1/2|r_A - r_I|$. For $\Delta t = 2\Delta t_0$, $Z_{C,1}(r, 2\Delta t_0) = 1/4(|r_A - r_I| + |r_A - r_B|)$, so for $\Delta t = k\Delta t_0$, $Z_{C,1}(r, k\Delta t_0) = 1/(2k)[|r_A - r_I| + (k-1)|r_A - r_B|]$. Similarly, in the interval $r_I < r < r_B$, $Z_{C,1}(r, k\Delta t_0) = 1/(2k)[|r_I - r_B| + (k-1)|r_A - r_B|]$. Note that transitions from A to B contribute to both intervals. Finally, the segment ...BBBBIAAAA... contributes identically compared to the ...AAAAIBBBB... segment. Summing it all up and multiplying by the equilibrium number of each segment, $\mathcal{T}/(8\Delta t_0)$, one gets for each interval

$$Z_{C,1}(r, k\Delta t_0) = \frac{\mathcal{T}}{8\Delta t_0}\left[\frac{2}{k}|r_A - r_I| + \frac{k-1}{k}|r_A - r_B|\right], \qquad r_A < r < r_I \tag{A.5}$$

and

$$Z_{C,1}(r, k\Delta t_0) = \frac{\mathcal{T}}{8\Delta t_0}\left[\frac{2}{k}|r_B - r_I| + \frac{k-1}{k}|r_A - r_B|\right], \qquad r_I < r < r_B \tag{A.6}$$

Note that $Z_{C,1}(r, k\Delta t_0)$ does not diverge with k since we have performed the average by construction when dividing by k in Eq. A.4; instead, it becomes independent of k for $k \to \infty$. In this limit $Z_{C,1}(r, k\Delta t)$ tends to $N_{AB}\Delta r^\dagger$ for both intervals since $N_{AB} = \mathcal{T}/(8\Delta t_0)$ and we defined $\Delta r^\dagger = |r_B - r_A|$. Specifying for the committor, Eq. A.5 and A.6 give $Z_{C,1}(q, k\Delta t) = N_{AB}$ not only in the high k limit, but for all values of sampling interval, which illustrate the committor validation test, Eq. 3.33, for this simple model. That is to say that for any coordinate, the TSD will become linear with k for high enough k, while it is always in this diffusive regime for the committor. This is expected since for any coordinate the memory effects will be negligible at some (unspecified) high enough timescale. In contrast, for the committor, the diffusive model holds at all timescales. These conclusions were illustrated using the simple three-state model but are in fact rather general. In particular, $Z_{C,1}$ computed with the TPS will always converge for large Δt as

$$Z_{C,1}|_{k\gg 1} \approx \frac{1}{2}\left(N_{AA}|r_A - r_A| + N_{BB}|r_B - r_B| + N_{AB}|r_A - r_B| + N_{BA}|r_B - r_A|\right) = \\ = N_{AB}|r_A - r_B| \equiv N_{AB}\Delta r^\dagger, \tag{A.7}$$

because most of the transitions will be between the boundaries.

Now, using the identity [17]

$$\Delta' r^2 = 2k\bar{Z}_{C,1}(k\Delta t_0), \tag{A.8}$$

where $\bar{Z}_{C,1}(k\Delta t_0)$ is the integral of $Z_{C,1}(r, k\Delta t_0)$, which can be computed in the high k limit using Eq. A.7,

$$\bar{Z}_{C,1}|_{k\gg 1} = \int_{r_A}^{r_B} Z_{C,1}|_{k\gg 1}\mathrm{d}r = N_{AB}\Delta r^\dagger \int_{r_A}^{r_B} \mathrm{d}r = N_{AB}(\Delta r^\dagger)^2, \tag{A.9}$$

so that for large k, the TSD is given by

$$\Delta' r^2|_{k\gg 1} \approx 2N_{AB}(\Delta r^\dagger)^2 k. \tag{A.10}$$

Note that this is consistent with the leading term for high k (proportional to k) that we obtained for the three-state model (Eq. A.1). Finally, the limiting value of the MSD for an arbitrary coordinate at high temporal scales, when using the TPS construction, is

obtained by dividing by the trajectory length $\mathcal{T}/\Delta t_0$

$$\frac{\Delta' r^2|_{k\gg 1}}{\mathcal{T}/\Delta t_0} = 2J_{AB}(\Delta r^\dagger)^2 k\Delta t_0, \tag{A.11}$$

where we used the definition of the flux $J_{AB} = N_{AB}/\mathcal{T}$, the number of transitions per unit time. Eq. A.11 specified for the energy gives the expected behavior for high k that we used in Sec. 8.2, i.e., the dashed line in Fig. 8.2(b).

Also in Sec. 8.2, we considered an energy-dependent diffusion coefficient obtained via Eq. 3.34, i.e.,

$$D(E, k\Delta t_0) = \frac{Z_{C,1}(E, k\Delta t_0)}{k\Delta t_0 Z_H(E, k\Delta t_0)}, \tag{A.12}$$

but as we assumed an effective coordinate-independent D^* to derive the rate equations in that section, we implemented the equilibrium average in Eq. 8.17, i.e.,

$$D(k\Delta t_0) = \int p_{\mathrm{eq}}(E) D(E, k\Delta t_0) \mathrm{d}E. \tag{A.13}$$

$D(k\Delta t_0)$ varies with sampling interval $\Delta t = k\Delta t_0$ because energy is not an optimal coordinate. In this way D^* was defined by taking the value of $D(k\Delta t_0)$ in the limit of high Δt or k, where the memory effects become negligible. The exact value of D^* can be computed as follows. First, note that $p_{\mathrm{eq}}(E) = Z_H(E, \Delta t_0)/\mathcal{Z}$ (Eq. 3.28), where the normalization constant is just $\mathcal{Z} = \mathcal{T}/\Delta t_0$ for a trajectory sampled with Δt_0. In this way Eq. A.13 can be rewritten as

$$D(k\Delta t_0) = \frac{\Delta t_0}{\mathcal{T}} \int k Z_H(E, k\Delta t_0) D(E, k\Delta t_0) \mathrm{d}E, \tag{A.14}$$

where we used $Z_H(E, \Delta t_0) = k Z_H(E, k\Delta t_0)$. Now, using Eq. A.12

$$D(k\Delta t_0) = \frac{k\Delta t_0}{\mathcal{T}} \int \frac{Z_{C,1}(E, k\Delta t_0)\mathrm{d}E}{k\Delta t_0} = \frac{\bar{Z}_{C,1}(k\Delta t_0)}{\mathcal{T}} \tag{A.15}$$

Finally, the integral $\bar{Z}_{C,1}(k\Delta t_0)$ can be computed for $k \gg 1$ and is given in Eq. A.9 so that

$$D^* = \lim_{k\to\infty} D(k\Delta t_0) = \frac{N_{AB}(\Delta E^\dagger)^2}{\mathcal{T}} = J_{AB}(\Delta E^\dagger)^2. \tag{A.16}$$

Alternatively, one can consider the MSD as defined from the TSD to obtain another estimate of the diffusion coefficient, say $\bar{D}(k\Delta t_0)$. It can be obtained in a usual way as half the time derivative of the MSD:

$$\bar{D}(k\Delta t_0) = \frac{1}{2}\frac{\Delta t_0}{\mathcal{T}}\frac{\mathrm{d}\Delta' E^2}{\mathrm{d}\Delta t}, \tag{A.17}$$

note that $\bar{D}(k\Delta t_0)$ is independent of the coordinate E because the TSD is already com-

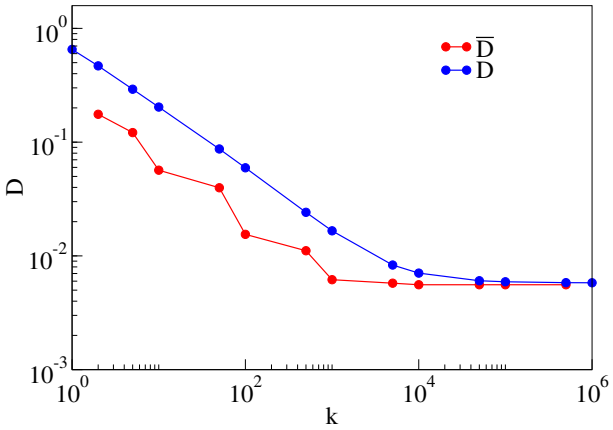

Figure A.2: Estimates of $D(k\Delta t_0)$ obtained via Eq. A.12-A.13 (blue circles) and of $\bar{D}(k\Delta t_0)$ obtained via Eq. A.17 (red circles) for the isotropic lattice system considered in Sec. 8.2. Continuous lines are just guides to the eyes.

puted as an average over the initial states, as discussed above. In contrast, $\bar{D}(k\Delta t_0)$ does depend on Δt since the MSD is not linear with time for low k, see Fig. 8.2(b). Nevertheless, for high k, the MSD approaches the linear behavior. Then $\bar{D}^*$ can be obtained by performing the derivative in this limit. The estimate is just half the derivative of the right-hand side of Eq. A.11 with respect to $\Delta t = k\Delta t_0$, so that $\bar{D}^*$ is equivalent to the previously defined D^*.

In Fig. A.2 we show both $D(k\Delta t_0)$ obtained through Eq.A.12-A.13 and $\bar{D}(k\Delta t_0)$ from Eq. A.17 as functions of k for the isotropic lattice system that we considered in Sec. 8.2. In the first case, the blue symbols represent the selected values of k in which we computed $Z_{C,1}(E, k\Delta_0)$. While in the second case, the red symbols represent the three-point numerical derivatives that we performed from the data of the MSD displayed in Fig. 8.2(b). In both cases the continuous lines are just guides to the eyes. It can be seen that while both estimates deviate a bit for low k, they converge to the same value in the high k limit, as expected.

Appendix B

Permutationally invariant CVs

Here we report a detailed comparison of the committor estimates obtained using the two more common ways of computing permutationally invariant CVS with our proposed ordering strategy employing the full distance matrix d_{ij}.

The first usual way of obtaining invariant CVs is to sum all features (distances in our case) over all identical molecules (pairs of molecules for distances) [68]. We tried summing functions of the pairwise distances d_{ij}, such as exponential and Gaussian functions, with free parameters that varied, allowing for the construction of a set of CVs. For example, in Fig. B.1(a) we consider the isotropic lattice system with $N = 400$ and show the $F_{C,1}$ validation criteria for a putative committor RC obtained by performing the optimization scheme with a set of CVs with $x_m = \sum_{i<j} \exp(-\alpha_m d_{ij})$, where $\{\alpha_m\}$ represents a set of four hundred logarithmically spaced numbers between 0.01 and 15. As can be seen in the figure, albeit being more optimal than n (compare with Fig. 5.4), the $F_{C,1}$ profiles still vary considerably with Δt when compared with the same plots for our final proposed committor, which we show again on Fig. B.1(d) at the same axis scale. Importantly, the $F_{C,1}$ profiles change not only with Δt but also with q, indicating that the optimization is performed in a non-uniform way. Particularly, the computed $N_{AB} = 146$ from the diffusive model along the committor obtained with the sum-based CVs still deviates from the real value computed directly from the trajectories $N_{AB} = 112$. Other choices of functions gave similar results.

A second possible strategy is to perform a permutationally invariant ordering. This is usually done by sorting all pairwise distances between the N molecules from the smallest to the highest, as is usually done. The time series of each of those $N(N-1)/2$ distances is a permutationally invariant CV. But the large amount of such CVs led to difficulties when dealing with them: one cannot use all of these time series together for optimization due to memory limitations. We tried using a representative fraction of those CVs. In Fig. B.1(b) we show the validation criteria for the resulting committor obtained from the optimization scheme with x_m consisting of the j_m-th smallest distances with $j_m = \text{int}(2^{8+i_m/50})$, $i_m = 0, 1, ..., 399$, where the function int() converts the argument to

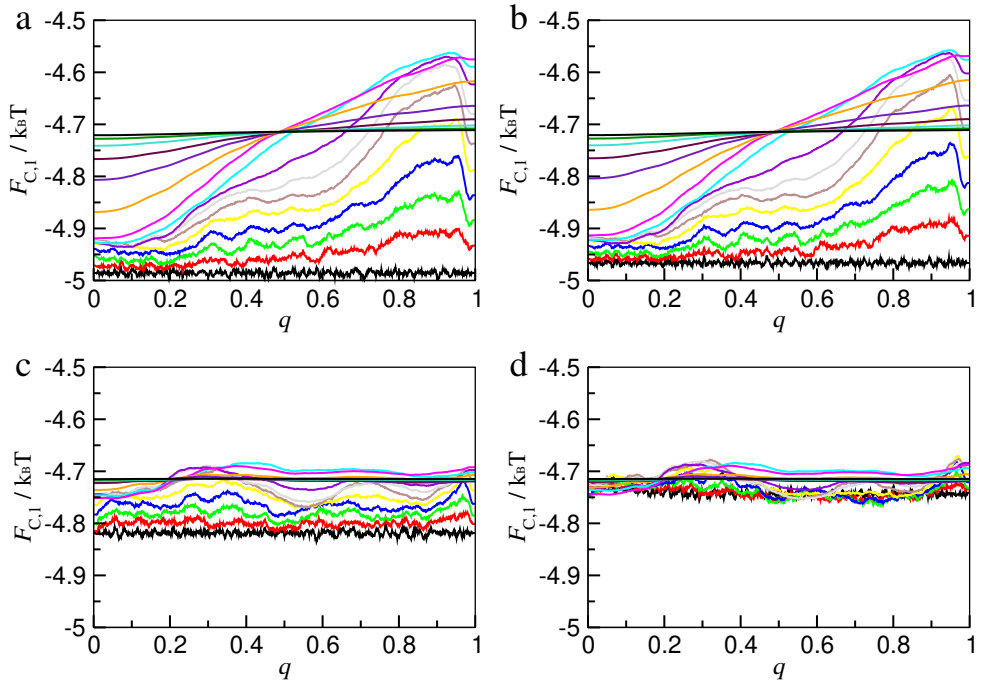

Figure B.1: Generalized cut-based free-energy profiles $F_{C,1}(q, \Delta t)/k_B T$ for different estimates of the committor q. (a) Optimization performed with $\{x_m\}$ obtained by summing up exponential functions of the pairwise distances. (b) $\{x_m\}$ consisting of selected sorted distances in the usual way. (c) $\{x_m\}$ consisting of the entries of only one axes $\bar{d}_{i,7}$ of the sorted matrix. (d) Final estimate obtained employing eight axis of the sorted matrix. In all plots, different colors correspond to different $\Delta t = 2^i \Delta t_0$, with $i = 0, 1, 2, ..., 15$ (from the bottom to the top). Figure adapted from Ref. [23].

the nearest smallest integer. The results are very similar compared to the sum-based CVs, slightly more optimal for $\Delta t = \Delta t_0$ (black curves). Both cases are worse even when compared with the committor obtained considering only one axis of the sorted matrix, of which we show the validation test in Fig. B.1(c). Not only do the $F_{C,1}$ profiles fluctuate less in Fig. B.1(c), but they are optimized in a more uniform way, both in temporal (Δt) and "spatial" (q) domains. The choice of the m-th distances out of $N(N-1)/2$ linearly spaced entries illustrates the difficulty of selecting the representative set of distances.

In Fig. B.1(c) we consider the putative RC obtained by taking as $\{x_m\}$ only the axis $\bar{d}_{i,7}$ (a total of 400 CVs since $N = 400$) of the sorted matrix, and the RC is already very close to the committor. Particularly, this putative RC is much more optimal than the ones obtained by more usual strategies described above and displayed in Fig. B.1(a) and (b) with the same number of CVs (400). Moreover, one can continue optimization by subsequently considering other axes in order to improve the putative RC. To avoid memory issues, we implemented that by performing the ordering at each time step or a batch of time steps (which require the full matrix) but then only storing one axis time series at a time. Depending on the memory available, this can be readily adapted to saving more axes together if desired. In Fig. B.1(d) we show the validation test considering eight

axes uniformly distributed between $\bar{d}_{i,7}$ and $\bar{d}_{i,42}$, and the resulting RC approaches almost exactly, i.e., within statistical error, the true committor function. Importantly, simply increasing the number of CVs employed in the other two strategies described before does not seem to significantly improve the achieved results, since no substantial improvements were detected going from 65 to the reported 400 number of CVs. It seems that when dealing with the sorted distances in the usual way, one would need to employ a clever way of selecting the representative subset of CVs, which seems to be much simpler using the sorted matrix that we employ.

Appendix C

Additional results for larger systems

Here we show some additional results obtained for larger isotropic systems. All constructed committors were obtained with the same procedure described in Sec. 5.1 using the nonparametric approach Eq. 4.1-4.3 supplied with a few axes extracted from the sorted matrix $\bar{\mathbf{d}}$.

In Table C.1 we show the number of transitions N_{AB} calculated from Eq. 3.52 and directly from the time series for both $q(n)$ and q, as well as $\tau_{B\to A}$ calculated via Eq. 3.39 and directly from the RC time series also for both RCs, for different system sizes. We consider isotropic ($\xi = 1$) systems of size $L = 200$ containing $N = 400$ molecules (same as in Sec. 5.1.2), $L = 260$ containing $N = 676$ molecules, and $L = 320$ containing $N = 1024$ molecules, so that the concentration is fixed at $\rho = N/L^2 = 10^{-2}$. We note that the discrepancies observed for the diffusive models along the n-based coordinate are persistent with increasing system size and that the diffusive model over our proposed optimized coordinate q can be used to describe the kinetics of all system sizes considered. For the system with $N = 400$ molecules, six simulations of length 10^8 MCs were considered, while for the system with $N = 676$, we considered four simulations of the same length, and for the system with $N = 1024$, we considered four simulations but with doubled length. We note that for the largest system we used a slightly different temporal resolution $\Delta t_0' = 2.5\Delta t_0 = 10^3$MCs due to hardware limitations.

Table C.1: Estimates for the number of transitions N_{AB} and the MFPT $\tau_{B\to A}$ obtained from both the n-based RC $q(n)$ and the commitor q for systems with different numbers of molecules N with the same concentration. All times are given in units of $10^3 \times \Delta t_0$.

	$N = 400$	$N = 676$	$N = 1024$
$(T^*)^{-1}$	2.73087	2.69033	2.66521
N_{AB} from Eq. 3.52 with $q(n)$	336	108	63
N_{AB} from time series of $q(n)$	122	30	24
$\tau_{B\to A}$ from Eq. 3.39 with $q(n)$	1.8	3.1	9.9
$\tau_{B\to A}$ from time series of $q(n)$	4.7	8.7	23
N_{AB} from Eq. 3.52 with q	113	29	21
N_{AB} from time series of q	112	27	24
$\tau_{B\to A}$ from Eq. 3.39 with q	5.4	13	31
$\tau_{B\to A}$ from time series of q	5.3	11	26

Appendix D

Evaluation of the Gaussian integrals for the rate equations

In Sec. 8.1 we arrived at the following Gaussian integrals for $\tau_{a\to d}$:

$$\tau_{a\to d} =\approx \frac{1}{D^*}\int_{E_a}^{E_d}\frac{\mathrm{d}E}{p^*(E;\beta)}\int_{-\infty}^{E^*}p_a(E;\beta)\mathrm{d}E \tag{D.1}$$

with $p^*(E;\beta)$ and $p_a(E;\beta)$ given by the Gaussian expressions, Eq. 8.6 and 8.8, respectively. Given that $p^*(E;\beta)$ and $p_a(E,\beta)$ are considerably different from zero only near E^* and E_a, respectively, one can extend the integration limits to $\pm\infty$ in both integrals, which are then computed as

$$\int_{-\infty}^{\infty}\frac{\mathrm{d}E}{p^*(E;\beta)} = \frac{\mathcal{Z}(\beta)}{\Gamma^*(\beta)}\sqrt{\frac{2\pi}{\gamma^*}} \qquad \text{and} \qquad \int_{-\infty}^{\infty}p_a(E;\beta)\mathrm{d}E = \frac{\Gamma_a(\beta)}{\mathcal{Z}(\beta)}\sqrt{\frac{2\pi}{\gamma_a}} \tag{D.2}$$

so that

$$\tau_{a\to d} = \frac{1}{k_d} \approx \frac{1}{D^*}\frac{\Gamma_a(\beta)}{\Gamma^*(\beta)}\frac{2\pi}{\sqrt{\gamma^*\gamma_a}}, \tag{D.3}$$

and using the expressions 8.7 and 8.9 for $\Gamma^*(\beta)$ and $\Gamma_a(\beta)$, one gets Eq. 8.13 for the dilution rate.

Similarly, for $\tau_{d\to a}$, we had

$$\tau_{d\to a} =\approx \frac{1}{D^*}\int_{E_a}^{E_d}\frac{\mathrm{d}E}{p^*(E;\beta)}\int_{E^*}^{\infty}p_d(E;\beta)\mathrm{d}E \tag{D.4}$$

so that an analogous argument leads to

$$\tau_{d\to a} = \frac{1}{k_a} \approx \frac{1}{D^*}\frac{\Gamma_d(\beta)}{\Gamma^*(\beta)}\frac{2\pi}{\sqrt{\gamma^*\gamma_d}} \tag{D.5}$$

and using Eq. 8.7 and 8.9 leads to Eq. 8.12 for the aggregation rate.

In the expressions above, we used the Kramers-like approximations (see, e.g., Chap-

ter 5 of Ref. [36]), which are expected to hold for high enough free-energy barriers. In this way, by performing the algebra that leads from Eq. D.5 and D.3 to Eq. 8.12 and 8.13, one arrives at the pre-factors given by Eq. 8.16.

When dealing with lower free-energy barriers, it turned out that a slightly different approximation is more suitable. For $\tau_{a\to d}$ we start again from Eq. D.1. For the first integral we made the same approximation of extending the integration limits to $\pm\infty$, but in the second integral we maintain the limits finite in the region where p_a is appreciable so that it can be related to error functions,

$$\int_{E_a}^{E^*} p_a(E;\beta)\mathrm{d}E = \frac{\Gamma_a(\beta)}{\mathcal{Z}(\beta)}\sqrt{\frac{2}{\gamma_a}}\left[\mathrm{erf}\left(\sqrt{\frac{\gamma_a}{2}}(E^*-E_0)\right) - \mathrm{erf}\left(\sqrt{\frac{\gamma_a}{2}}(E_a-E_0)\right)\right], \quad \text{(D.6)}$$

where $E_0 = E_a - (\beta-\beta^*)/\gamma_a$ (see Eq. 8.8) and $\mathrm{erf}(x) = \int_0^x e^{-y^2}\mathrm{d}y$. As a final approximation, we consider $\mathrm{erf}(x) \approx x$ so that the second integral in Eq.D.1 is given simply by

$$\int_{-\infty}^{E^*} p_a(E;\beta)\mathrm{d}E \approx \frac{\Gamma_a(\beta)}{\mathcal{Z}(\beta)}(E^*-E_a) = \frac{\Gamma_a(\beta)}{\mathcal{Z}(\beta)}\Delta E_a^{\ddagger}, \quad \text{(D.7)}$$

while the first integral is still approximated by the first expression in Eq. D.2, so that one has

$$\tau_{a\to d} = \frac{1}{k_d} \approx \frac{1}{D^*}\frac{\Gamma_a(\beta)}{\Gamma^*(\beta)}\sqrt{\frac{2\pi}{\gamma^*}}\Delta E_a^{\ddagger}. \quad \text{(D.8)}$$

Substituting the expressions for $\Gamma^*(\beta)$ and $\Gamma_a(\beta)$, one gets again Eq. 8.13 for the dilution rate, but now with the pre-factor given by Eq. 8.15. And an analogous argument leads to starting from Eq. D.4

$$\tau_{d\to a} = \frac{1}{k_a} \approx \frac{1}{D^*}\frac{\Gamma_d(\beta)}{\Gamma^*(\beta)}\sqrt{\frac{2\pi}{\gamma^*}}|\Delta E_d^{\ddagger}| \quad \text{(D.9)}$$

Substituting the expressions for $\Gamma^*(\beta)$ and $\Gamma_a(\beta)$, one gets Eq. 8.12 for the aggregation rate with the pre-factor given by Eq. 8.15.