1

Introduction

1.1 Motivation

Structural formation through self-assembly or external assembly is omnipresent in a large number of natural and technological systems. At the smallest atomistic scales of individual (macro-)molecules (few to thousands of atoms connected by covalent bonds) examples such as polymers, carbon nano tubes [1, 2], and polyoxometalates (large polyatomic ion structures) [3, 4] exist in technology. These macromolecules provide for example material building blocks or catalysts for oxidization of organic compounds in the case of polyoxometalates. Similarly and likely even more important for all living organisms, virtually all proteins require a specific three dimensional structure, also called conformation or secondary/tertiary structure, to enable their function. Issues with regard to their conformation directly impact function e.g. leading to various diseases such as in the context of allergies [5]. This high impact has lead to significant scientific interest to understand protein folding as shown in the 'critical assessment of protein structure prediction' (CASP) [6], a biennial competition of protein folding prediction algorithms, which has notably been won in 2018 and 2020 by the deep-learning algorithm AlphaFold [7].

Similarly, these structural formation mechanisms extend hierarchically to larger assemblies of multiple macromolecules, which is the focus of this work. These *macromolecular assemblies* are defined by both composition and structure to enable their function. Depending on the field other terms such as *supramolecular assembly* in (supramolecular) chemistry and nanotechnology or *quaternary structure* in biology are also common [8, 9]. The occurring structural formation is caused by non-covalent intermolecular interactions, which is the distinguishing element from individual macromolecules (see IUPAC definition [8]). However, some structures such as chemically

cross-linked gels [10], which are connected by covalent bonds, might also be considered to fall in the same category of macromolecular assemblies while not adhering to the previous definition. Macromolecular assemblies in general may vary with regard to their function, size, selectivity of intermolecular binding sites, regularity of structural organization, assembly mechanisms, and other properties. In the following, some examples of macromolecular structures will be provided with special regard to their function.

In the biological context, a variety of biopolymers (polypeptides [polymers of amino acids, e.g. collagen; protein when sufficiently large with biological functionality], polynucleotides [e.g. RNA, DNA], polysaccharides [e.g. alginate] [8]) and non-polymeric biomolecules (e.g. lipids) build larger structures through self-assembly of multiple copies either of the same biomolecule or different types to enable their function. One example is the field of viruses, which often contain structural proteins with the ability to assemble into regular structures, called virus capsids or virus-like particles (VLP). These structures are critical for the function of the overall virus during infection and reproduction, as well as for the immune system recognition [11]. Examples are the hepatitis B virus [12], adenoviruses [13], and coronaviruses [14].

Another example is the field of multi-enzymatic biocatalysis [15], where different enzymes (proteins with biocatalytic function) catalyze a cascade chemical reaction. Many times such systems achieve their high activity through structural formation leading to effects like metabolic channeling [16, 17]. Examples for this are the pyruvate dehydrogenase complex (PDC) [18], fatty acid synthase [19], glutamine synthetase [20], and others. Adaptations and possibly *de novo* creations of multi-enzymatic biosynthetic reactions are consequently also of high interest and being developed for industrial applications [21–23].

Further examples exist in the context of material science, e.g. in regard to colloids or gels [8]. For a variety of dispersed and continuous phases the molecular assembly is critical to ensure its function, e.g. with regard to mechanical stability. Examples are hydrogels and aerogels [24], which rely on their cross-linked polymer network structure to enable mechanical stability. Underlying polymers can be a variety of natural polymers, such as alginate [25], as well as synthetic polymers, such as polyethylene glycol [26]. Other examples from supramolecular chemistry and nanotechnology include self-assembled monolayers [27] and host-guest chemistry [28, 29].

In summary, the large body of examples, literature, and features highlights the great interest of both the scientific and industrial community in understanding, modifying, and possibly *de novo* creating such macromolecular structures. In order to gain this state of the art understanding, a variety of experimental and numerical techniques have already been developed. Nonetheless, limitations apply steering from the challenging multiscale

nature of such phenomena, as well as high dynamics and partially disordered structural elements. Focus of this work will be placed on numerical simulation of these systems to improve mechanistic understanding. For this, a novel physics-based and data-driven methodology will be presented capable of reaching the micro-meter and milli-second scales using bottom-up parameterization, thus advancing capabilities in the field. In the following, the state of the art with regard to numerical simulation approaches will be depicted.

1.2 Theory and State of the Art in Molecular Mechanics

Modeling and simulation of real-world systems is required to be both accurate and efficient - thus the chosen model description is dependent on the system and properties of interest.[1] In the context of molecular modeling and specifically molecular mechanics of the aforementioned systems, neither the treatment of quantum dynamical or relativistic effects, nor abstractions as a continuum are accurate and efficient. As a result, most developed simulation methods (under the assumption of interest in the *dynamics* of the system) describe such systems using discrete modeling approaches in the context of molecular dynamics (MD) related methods. As such, they assume the atomistic objects of the system to behave non-relativistic (i.e. velocities are much smaller than the speed of light), the Born-Oppenheimer approximation to hold (i.e. electrons move much faster than nuclei), and atomic motion to be following classical mechanics including inertia effects. By abstractions of the discrete units from individual atoms to coarse-grained (CG) beads the level of detail and thus numerically reachable scales of length and time can be controlled. In the following, the fundamentals of MD and related methods will be discussed. For textbooks and reviews see e.g. refs. [30–35]. Before going into the details, it should be noted that other modeling approaches exist when one is primarily interested in the *static* properties of such systems, e.g. regarding molecular assembly, but not formation mechanisms and dynamics. These approaches will be discussed briefly in Sec. 1.2.3.

1.2.1 Molecular Dynamics (MD)

As just highlighted, molecular dynamics (MD) describes the dynamic motion of atomic nuclei (referred to simply as atoms in the following) using classical mechanics to study molecular systems in the fields of physical chemistry, biochemistry, and others. For these systems, the Born-Oppenheimer approximation holds and electrons are typically

[1] This overview is conceptually based on Berendsen [30] and begins on his level 4 abstraction.

assumed to be in their ground state. Thus, atom interaction is fundamentally captured by the time-independent Schrödinger equation depending on nuclei positions and electrons. Effective interactions are subsequently modeled through so called 'force-fields' enabling a simple description, thus not requiring the explicit treatment of electron distributions. Starting from classical mechanics, the **motion of an atom** i with mass m_i in a system of N atoms is described using Newton's equation of motion as

$$m_i \ddot{\vec{x}}_i = \vec{F}_i = -\nabla U_i, \tag{1.1}$$

where the acceleration $\ddot{\vec{x}}_i$ (second time-derivative of position $\vec{x}_i$) corresponding to the force $\vec{F}_i$ results from the interaction potential U_i with other atoms. Assume the interaction potential U to be known at this point. The resulting velocity and trajectory in time can be calculated based on an initial condition of coordinates and velocities using **numerical time integration** with time steps Δt. In MD, numerical time integrations is typically performed using explicit time stepping using e.g. Verlet [36] or leap-frog [37] algorithms. In order to enable a numerically stable solution, a sufficiently small time step has to be chosen. As can be seen directly in eq. 1.1, this is primarily influenced by light atoms leading to time step requirements in the 1 fs (10^{-15} s) range required by hydrogen bonds [38].

Due to the high complexity and small time steps, only small system sizes on scales up to tens or hundreds of nano-meters can currently be modeled. The shape and size of the **simulation domain**, as well as its **boundary conditions** has to be chosen suitable to avoid boundary effects. While open boundaries are generally possible, the represented molecules in dilute gas phases or vacuum are of limited interest. Thus, periodic boundary conditions are most widely employed and to a lesser extend continuum boundary conditions (e.g. for surface absorption) or restrained-shell boundary conditions [30, 39]. Additionally, note that the shape of simulation domains, also with periodic boundary conditions, is not restricted to cubic domains, but may also include e.g. triclinic shapes, hexagons, and more [40] - as might be advantageous e.g. for studying crystals.

Having provided the simulation domain for atoms to be placed in and time integration algorithms to determine trajectories based on the forces an atom experiences, the main question becomes a descriptor for the forces $\vec{F}_i$ (see eq. 1.1) between interacting atoms, which are equivalent to the negative gradient of the potential energy $-\nabla U_i$. As previously noted, **force-fields** provide the effective description for the interaction of all atoms or groups of atoms (beads, see coarse-graining in next section) in a system. Consequently, such force-fields provide an effective model of the electron distribution in their ground state, i.e. no chemical reactions, resulting from the time-independent Schrödinger equation. As such, they have to be sufficiently simple to solve large atomistic systems

over reasonably long times, while also providing sufficient accuracy. For this, force-fields typically take the covalent structure of molecules into account[2] and separate the energy contributions as

$$U = U_{\text{covalent}} + U_{\text{non-covalent}}, \tag{1.2}$$

$$U_{\text{covalent}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihedral}} + U_{\text{improp. dihedral}}, \tag{1.3}$$

$$U_{\text{non-covalent}} = U_{\text{electrostatic}} + U_{\text{van der Waals}}, \tag{1.4}$$

where U_{covalent} are energy terms of covalent bonds[3] including U_{bond} describing bond stretching, U_{angle} describing bond angles formed by three atoms (e.g. O-C-O in CO_2), and U_{dihedral} / $U_{\text{improp. dihedral}}$ describing dihedral angles between four atoms in different planes (improper dihedral to keep planar groups like aromatic rings planar); $U_{\text{non-covalent}}$ are non-bonded interactions (bonded interaction pairs excluded/modified) that are pairwise additive including $U_{\text{electrostatic}}$ describing electrostatic interaction (Coulomb potential) and $U_{\text{van der Waals}}$ describing van der Waals interaction (typically modeled as a Lennard-Jones potential). Typically, established functional descriptions are used for the respective energy contributions and tabulation is employed for computational efficiency [41, 42]. More elaborate force-fields might incorporate additional features such as polarizability, virtual interaction sites, dummy particles, coupling terms, flexible constraints, charge distributions, multipoles, reactive components, and others [30]. A variety of force-fields have been developed with the motivation of providing an as widely applicable atom interaction parameterization as possible. However, research has shown that such (simple) force-fields are largely only applicable to a class of systems and less transferable as they would ideally be[4]. Details on parameterization of force-fields is beyond the scope of this work, but approaches include e.g. *ab initio* quantum calculations and adjustments according to empirical observations [43]. Examples of important classical force-fields are AMBER [44], CHARMM [45], GROMOS [46], and OPLS [47]. Examples of polarizable force-fields are further developments of AMBER [48] and CHARMM [49, 50]. An example for reactive force-fields (i.e. incorporating chemical reactions) is ReaxFF [51]. For more details on force-fields see e.g. with regard to protein simulation ref. [52].

In order for pairwise contacts of non-covalent contributions to be calculated efficiently for reasonably large systems (i.e. not scaling with a computational complexity of $O(N^2)$ for the number of atoms), **cutoff distances** are employed. Each force-field comes

[2] Thus no chemical reactions, changes in covalent structure, redox states, or protonation may take place [30]. Such systems have to be treated differently, e.g. using quantum-chemical methods.

[3] Note that covalent bonds are sometimes also represented through constraints. Such approaches will be discussed in more detail in Ch. 5.

[4] Limitations in force-field transferability might e.g. result from non-additivity of constituent terms, neglect of contributions, or adjustments to empirical observations [30]

with specific cutoff distances integral for the overall energy balance and reproduction of desired properties, see e.g. ref. [53]. In order to account for the discontinuity at cutoff, a variety of switching and shifting methods have been developed, see e.g. in refs. [30, 42]. In order for **long-range interactions** (specifically electrostatic interactions and especially with polarization in medium) to be modeled more accurately, coupled field approaches such as the (smooth) Particle-Mesh-Ewald method [54, 55] have been developed.

Note that in addition to classical functional descriptions a variety of machine learned force-fields have been developed recently using (deep) neural networks on quantum mechanical data [56–58]. Alternatively to an effective force-field, *ab initio* molecular dynamics, initially proposed by Car and Parrinello [59], solves first principle quantum mechanical methods (such as the density functional theory (DFT) and approximations like 'divide-and-conquer' DFT [60] or 'tight-binding' DFT [61]) to gain more detailed information on the electron distribution and potential energy at the cost of significantly higher computational demand [59, 62, 63]. Such approaches can further capture electrons in excited states, e.g. for chemical reactions. Additionally, a variety of methods for treating subsystems at the quantum mechanical scale while maintaining effective MD force-fields in the remaining have been developed in the context of hybrid quantum mechanical / molecular mechanics (QM/MM) methods, initially proposed by Warshel and Levitt [64], see also refs. [65, 66]. More details with regard to force-fields and intermolecular interaction will be provided in Ch. 4.

Until this point, systems in molecular dynamics were considered as a simulation domain filled with atoms that evolve in time from a given initial condition. However, such a system is merely one form of a **thermodynamic ensemble** in statistical mechanics - specifically a *microcanonical* ensemble of constant number of particles N, volume V, and energy E. Alternatively, instead of constraining the number of particles N one can constrain the chemical potential μ; instead of the volume V one can constrain the pressure p; and instead of energy E one can constrain the temperature T (or enthalpy H). Some of the most widely employed ensembles and their names are listed in Tab. 1.1. In the context of MD and specifically this work, the *canonical* NVT and *isothermal-isobaric* NPT ensemble are most crucial. Thus, the main question consequently becomes how pressure and temperature control can be achieved.

Table 1.1: Most widely used thermodynamic ensembles.

Abbreviation	Name
NVE	Microcanonical
NVT	Canonical
μVT	Grand canonical
NPT	Isothermal-isobaric
NPH	Isoenthalpic-isobaric

In order to enforce the desired ensemble or perform *non-equilibrium* simulations, a variety of **temperature and pressure coupling** methods (also called **thermostats** and **barostats**) have been developed and can typically be classified as stochastic methods, strong-coupling methods, weak-coupling methods, and extended system dynamics [30]: Stochastic methods apply either stochastic exciting forces in combination with friction forces or randomly reassign certain variables (e.g. velocities for temperature control). They are particularly used to control temperature and enforce a canonical ensemble. Examples are the Anderson thermostat [67] (work also contains a barostat) and more generally Langevin dynamics (see following section). Strong-coupling methods constrain a certain variable (e.g. velocities for temperature control) employing e.g. a scaling in every time step. Examples are the isokinetic Gauss thermostat [68, 69] and related barostats by Evans *et al.* [70, 71]. Weak-coupling methods apply non-stochastic perturbations to enable a first-order decay of deviations from the desired controlled quantities (temperature via velocity scaling or pressure via coordinate scaling). An example is the Berendsen thermostat [72]. Extended system dynamics add additional degrees of freedom to control quantities and an example is the Nosé-Hover thermostat [69, 73, 74]. A more detailed discussion can be found in ref. [30]. Additional details with regard to temperature control and diffusion will be provided in Ch. 3.

Note that while the majority of MD simulations are performed at constant temperature, a variety of additional methods exist, which are advantageous for **enhanced sampling**, e.g. conformation sampling through thermodynamic state changes. Examples are simulated annealing [75], replica exchange MD [76], and expanded ensembles [77]. More details will be provided in Sec. 3.6.

Furthermore, as most molecular systems exist in solution, **solvent modeling** is given special attention in MD. This is especially true for modeling **water**, which is the most common solvent in nature and also many technical systems - thus the focus in the following. For many of such systems the computational load resulting from the modeling of the solvent is quite significant - often exceeding that of the actual molecules investigated. In this regard, it is crucial which properties of the solvent one wants to reproduce, e.g. phase changes and dielectric constants. With regard to water, a large variety of *explicit models* have been developed employing at least up to six sites for parameterization [78, 79]. Widely used examples are the SPC [80], SPC/E [81], TIP3 [82], TIP2P/TIP3P/TIP4P [83], and MCDHO [84]. The high number of water models indicates the challenge in reproducing all properties accurately, especially with regard to varying conditions. In this context, special attention has to be paid e.g. on polarizability and induced dipols [30]. For reviews see e.g. ref. [85]. In addition, *implicit water models* have gained interest in recent decades to reduce the overall computational requirements, while describing the molecules of interest with atomistic resolution [86–94]. Examples are semi-heuristic

methods like ASP [95], Generalized-Born models [96], or more generally ones based on Poisson-Boltzmann theory [97]. For reviews see e.g. refs. [98–102].

In summary, molecular dynamics provides an established and still heavily researched framework for studying molecular phenomena on atomistic scales under the assumption of non-relativity and applicability of the Born-Oppenheimer approximation. Applications extend from crystal cracks / defects [103] to protein folding [104], protein-ligand binding [105], and protein-protein interaction [106]. Various molecular dynamic codes are available, both free and commercial, in order to investigate chemical, biological, and other systems. Examples are the codes AMBER, CHARMM, GROMACS, TINKER, OpenMM, NAMD, and LAMMPS. For MD simulations in this work the code GROMACS was used, as it provides a free and open-source platform.

1.2.2 Coarse-Graining in Space and Time

In order to investigate systems on larger scales of length and time, various methods have been developed beyond atomistic MD [30]. These methods employ the same ideas based on classical mechanics, but perform coarse-graining with regard to **space**, i.e. reduction of the degrees of freedom by combining multiple atoms to a unit/bead, as well as **time**, e.g. neglecting inertia terms. In the following, the most widely employed approaches will be presented. For reviews see e.g. refs. [30, 107–111].

Generally speaking, every coarse-graining approach consists of a structural and a functional model. The **structural model** separates the system into *relevant (explicit)* and *omitted (implicit)* degrees of freedom (DOF) / particles, and provides a **mapping** methodology to combine multiple relevant atoms / particles into a coarse-grained bead. The **functional model** provides a **coarse-grained force-field** describing the interaction between coarse-grained beads and possibly implicit aspects of the omitted DOF. As a result, coarse-grained approaches are inherently more specialized and less transferable than all-atom descriptions. Existing coarse-grained models thus employ a variety of approaches for definition and parameterization of the structural and functional model. Before going into their detail, a more general formalism based on a bottom-up abstraction will be provided following Berendsen [30]. With regard to mathematical formulations, this will be restricted to the cartesian degrees of freedom in their center of mass. For generalized coordinates the reader is e.g. referred to ref. [30].

The Mori-Zwanzig projection-operator formalism [112–114] presents a systematic (bottom-up) approach to derive the evolution of a subsystem in phase space. Based on this, the equation of motion for the *relevant (explicit)* particles i (*omitted/implicit* j)

becomes [30]

$$m_i\vec{\ddot{x}}_i = \underbrace{-\nabla U_i^{\mathrm{CG}}}_{\substack{\text{systematic forces}\\ \text{between explicit DOF}\\ \text{(CG beads)}}} \underbrace{-\sum_j \int_0^t m_i\gamma_{ij}(\tau)\vec{\dot{x}}_i(t-\tau)d\tau}_{\substack{\text{friction forces}\\ \text{from implicit DOF}}} + \underbrace{\vec{\eta}_i(t)}_{\substack{\text{random forces}\\ \text{from implicit DOF}}} \quad , \tag{1.5}$$

where m_i is the mass of the bead, $\vec{\ddot{x}}_i$ its acceleration, and U_i^{CG} the coarse-grained potential describing the systematic forces between beads (to be defined later, often called *potential of mean force*). Effects of the *omitted (implicit)* DOF are captured through the frictional forces resulting from the friction kernel γ_{ij} (including its time dependence), as well as the random forces $\vec{\eta}_i$. Note that this formulation assumes the systematic force (gradient of potential of mean force) to be curl free, frictional forces to be linearly dependent on velocity (i.e. laminar flow with a Reynolds number of less than one for macroscopic systems), and omitted (implicit) DOF to equilibrate much faster than relevant (explicit) DOF. This formulation is equivalent to the *generalized Langevin equation*[5] [116] and one arrives at the following **Langevin Dynamics (LD)** formulation in the memory-free Markovian limit[6] applicable for the time scales of coarse-grained simulations as

$$m_i\vec{\ddot{x}}_i = -\nabla U_i^{\mathrm{CG}} - m_i\gamma_i\vec{\dot{x}}_i + \vec{\eta}_i, \tag{1.6}$$

or more commonly written in 1D as

$$m\ddot{x} = -\nabla U^{\mathrm{CG}} - m\gamma\dot{x} + \sqrt{\Psi}\zeta, \tag{1.7}$$

where the random force is decomposed into a constant Ψ and a normally distributed random number ζ with zero mean, unit variance, and no correlation in time[7]. Resulting from the assumption of a stationary process with time-independent velocity correlations

[5] Langevin's equation was introduced in 1908 by Paul Langevin [115] as a stochastic differential equation to describe Brownian motion of particles in a fluid. Newton's equation of motion is extended by a random exciting force and systematic damping force, which represent the collision with high-velocity fluid molecules and the fluid drag, respectively. The equation is discretized in Langevin Dynamics (LD) and frequently employed in coarse-graining approaches to represent the forces of neglected degrees of freedom through friction and noise, see e.g. ref. [30]. As a result, it acts essentially as a thermostat and enforces a canonical ensemble, while accounting for the solvent (similar to an implicit solvent model, but not accounting e.g. for electrostatic screening) and neglected degrees of freedom implicitly.

[6] For works in the non-Markovian limit see e.g. ref. [117].

[7] As previously noted we assume the memory-free Markovian limit and omitted (implicit) DOF to equilibrate much faster than relevant (explicit) DOF, which is reasonably fulfilled for most coarse-graining applications. See e.g. discussion in ref. [30].

of a canonical ensemble, the friction and random force are related by the *fluctuation-dissipation theorem* [30]

$$\Psi = 2m\gamma k_B T. \tag{1.8}$$

Note that while this is only generally valid without systematic forces, it yields consistent dynamics with proper equilibrium fluctuations under the chosen assumptions of a memory-free Markovian process [30]. For more details the reader is referred to refs. [114, 118].

For practical purposes the question remains how to derive the friction coefficients. A variety of approaches based on theory (e.g. Einstein [119], Debye [120], and Perrin [121, 122]), experiments (e.g. FCS [123]), and detailed MD simulations (e.g. ref. [124]) have been explored.

Further note that many coarse-grained approaches do not employ this formalism and the resulting LD-related formulation, thus do not include additional friction and random forces resulting from the neglected degrees of freedom (e.g. MARTINI [53, 125, 126]). Other approaches scale down these contributions by a factor between 10 - 1000 through an effective viscosity [110, 127, 128]. While this is not expected to influence equilibrium properties, their dynamics are likely to be accelerated [30]. In the context of this work, this formalism and resulting Langevin Dynamics will be employed later in the limit of representing entire macromolecules as (ultra-)coarse-grained beads in implicit solution. Thus, the friction and random forces represent specifically the solvent and their parameterization is performed using detailed MD simulations. In the same context, it should be noted that for the simulation of non-dilute solutions in addition to the systematic force between coarse-grained beads accounting for *hydrodynamic interaction*, i.e. forces resulting from their relative velocities coupled through the solvent, can become important. More detail on hydrodynamic interaction is provided in App. A.3.

Having accounted for the omitted DOF through implicit random and friction forces, the remainder of this section will focus on the derivation and parameterization of **coarse-grained force-fields** describing the systematic forces between beads. Note that these forces are not necessarily pairwise additive, but may also include multi-body contributions. In their functional description they are often closely related to atomistic force-fields (see eq. 1.3, also termed neoclassical [110]), but alternative descriptions through e.g. neural networks have found increasing interest in recent years [129–131]. In order for the respective force-fields to be parameterized, a variety of approaches exist in literature, which are often classified as *bottom-up*, *top-down*, or *hybrid* approaches [108, 110]. Bottom-up approaches parameterize the coarse-grained force-fields using