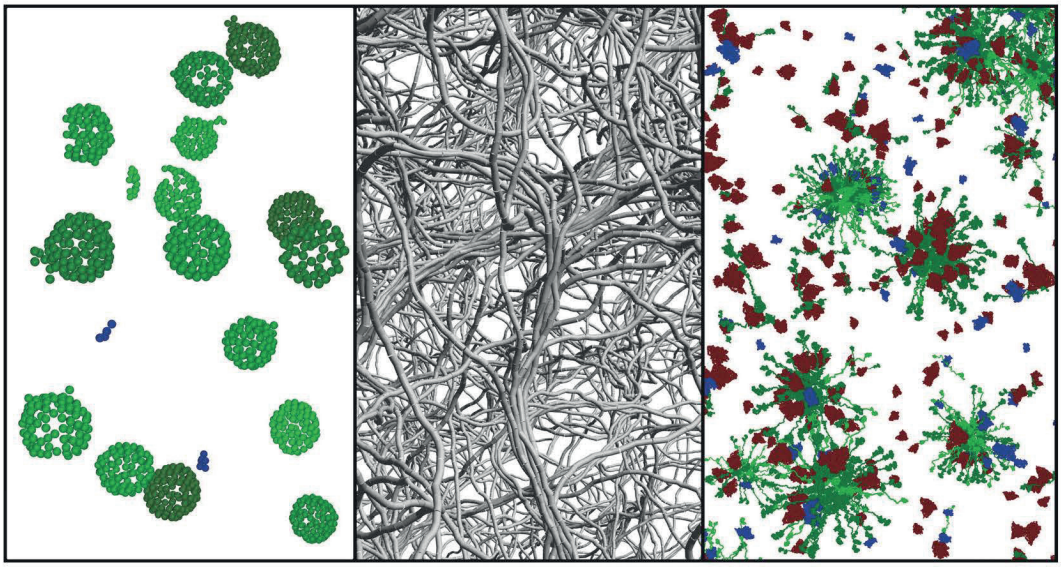
INSTITUTE OF SOLIDS PROCESS ENGINEERING & PARTICLE TECHNOLOGY
TUHH

SPE-Schriftenreihe 25
Herausgegeben von Prof. Dr.-Ing. habil. Prof. E.h. Dr. h.c. Stefan Heinrich

Philipp Nicolas Depta

Physics-Based and Data-Driven Multiscale Modeling of the Structural Formation in Macromolecular Systems



Cuvillier Verlag Göttingen
Internationaler wissenschaftlicher Fachverlag

Physics-Based and Data-Driven Multiscale

Modeling of the Structural Formation in Macromolecular Systems

Physics-Based and Data-Driven Multiscale Modeling of the Structural Formation in Macromolecular Systems

Vom Promotionsausschuss der
Technischen Universität Hamburg
zur Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

genehmigte Dissertation

von

Philipp Nicolas Depta

aus

Frankfurt am Main

2024

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliographische Daten sind im Internet über http://dnb.d-nb.de abrufbar.
1. Aufl. - Göttingen: Cuvillier, 2024
Zugl.: (TU) Hamburg, Univ., Diss., 2024

Gutachter:
1. Prof. Dr.-Ing. habil. Prof. E.h. Dr. h.c. Stefan Heinrich
2. Prof. Dr. Pavel Gurikov
3. Prof. Dr.-Ing. Carsten Schilde

Tag der mündlichen Prüfung: 19. Januar 2024

CUVILLIER VERLAG, Göttingen 2024
Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0 Telefax:
0551-54724-21
www.cuvillier.de

1. Auflage, 2024
Gedruckt auf umweltfreundlichem, säurefreiem Papier aus nachhaltiger Forstwirtschaft.

ISBN 978-3-7369-7972-7
eISBN 978-3-7369-6972-8

Acknowledgments

I would like to express my gratitude to Prof. Stefan Heinrich for the opportunity of doing my doctorate in his institute, as well as providing the invaluable guidance, unwavering support, and encouragement throughout my research. This environment and possibility for exchange at numerous national and international conferences really enabled me to grow as a researcher.

Special thanks go to former Jun. Prof. Maksym Dosta for his relentless support, inspiration, insightful feedback, and collaborative approach shaping my perspectives as both a researcher and a person.

Similarly, I would also like to express my thanks to Prof. Pavel Gurikov and Prof. Carsten Schilde for their commitment as examiners of my dissertation, as well as Prof. Andreas Liese as committee chair.

Furthermore, I would like to thank all my collaborators, without whom this research would not have been possible. I am grateful to Dr. Uwe Jandt, Prof. Pavel Gurikov, Dr. Baldur Schroeter, Dr. Attila Forgács, Prof. József Kalmár, Dr. Geo Paul, Prof. Leonardo Marchese, Dr. Mariana Kozlowska, and Prof. Wolfgang Wenzel (in chronological order). Particularly, I would like to thank Dr. Uwe Jandt for his collaboration and guidance during my initial stages as a researcher.

Moreover, I would like to express my deepest gratitude to the SPE team and its support throughout the years – you are the most remarkable group and research family anyone could wish for. Particularly I would like to thank Dr. Britta Buck for the warm welcome; my office mates Olga Ochkin-König, Abdullah Sadeq, Jess – Chan Tsz Tung, and Marian Schmitt for providing an amazing office atmosphere through all ups and downs; as well as Vasyl Skorych for pushing my programming skills and entertaining all my IT projects and ideas.

Additionally, I am also very thankful to the Deutsche Forschungsgemeinschaft (DFG) for funding my research in the context of SPP 1934 DiSPBiotech and the High-Performance Computing Center Stuttgart (HLRS, Acid 44178) for providing the necessary computational resources. In this regard, I would also like to thank all participants of the SPP DiSPBiotech for their fruitful discussions and wonderful exchanges.

Finally, I am immensely grateful to my fantastic family and Tajda for their relentless support throughout all phases of this dissertation – you made all this possible.

Abstract

Macromolecular structural formation and hierarchical self-assembly is crucial for a variety of systems in both nature and technology. Such systems may retain a remarkable structural organization from the atomistic up to the macroscopic scale enabling crucial features for their function. Many of these systems achieve this through self-assembly and consequently do not rely on external assembly mechanisms. In the field of material science one example is hydrogels, which achieve significant mechanical strength through polymer network formation on the molecular level. In the field of biology examples are abundant including most enzymes and viruses. One example is the hepatitis B virus, which contains a structural protein that assembles into regular spherical structures to transport the genetic material of the virus. Another example is the pyruvate dehydrogenase complex, which is crucial for cellular respiration and achieves its high biocatalytic activity through structural formation, thereby enabling features such as metabolic channeling. While there is an abundant amount of examples, investigation is challenging both experimentally and numerically. The phenomena involved in such structural assembly spread over vast scales in length and time and contain not only regular structures, but often also disordered elements. Consequently, capturing the mechanisms of formation, especially their kinetics, is inherently difficult.

In order to improve understanding of these phenomena, this work proposes a physics-based and data-driven multiscale modeling framework capable of describing structural formation on the micro-meter and milli-second scale, while retaining large amounts of molecular detail. The framework achieves this by abstracting the elementary macromolecules of a system as anisotropic unit objects and describes the interaction between units as well as the environment through data-driven models, e.g. 6D interaction potential fields. The models are parameterized in a bottom-up fashion and validated top-down. The framework is applied to and validated on three model systems: the gelation of alginate in $CaCl_2$ solution, the self-assembly of the hepatitis B core antigen into virus-like particles, and the assembly and agglomeration of the pyruvate dehydrogenase complex. Results are validated using literature data and experimental data provided by collaborators, which show good agreement with measurable characteristics. Consequently, the developed framework enables novel scales to be investigated using numerical simulations and proposes a streamlined bottom-up parameterization, thus paving the way towards physically-mechanistic modeling of such structural assembly processes.

Contents

Symbols

Abbreviations

acetyl-CoA	Acetyl Coenzym A
AMR	Adaptive Mesh Refinement
AN1 / AN2	Annealing Procedures 1 and 2 for Alginate (Sec. 6.1.2.1)
ANN	Artificial Neural Network
AUC	Analytical Ultracentrifugation
AS	Active Site
BAC	Bond-Angle Correlation
BD	Brownian Dynamics
BJH	Barrett-Joyner-Halendia
BLI	Bio-Layer Interferometry
BLUE	Best Linear Unbiased Estimate
BPM	Bonded Particle Model
CASP	Critical Assessment of Protein Structure Prediction
CFD	Computational Fluid Dynamics
CG	Coarse-Grained
CG-MD	Coarse-Grained Molecular Dynamics
COM	Center of Mass
CRW	Conditional Reversible Work
cryo-EM	Cryoelectron Microscopy
DEM	Discrete Element Method
DFT	Density Functional Theory
DLS	Dynamic Light Scattering
DMD	Discrete Molecular Dynamics
DOF	Degree of Freedom
DPD	Dissipative Particle Dynamics
EFCG	Effective Force Coarse-Grained
E1	PDC enzyme pyruvate dehydrogenase
E2	PDC enzyme dihydrolinpoamide acetyltransferase
E3	PDC enzyme dihydrolipoamide dehydrogenase
E3BP	PDC enzyme E3 binding protein
FCS	Fluorescence Correlation Spectroscopy

FEM	Finite Element Method
FM	Force Matching
G	Guluronic Acid (α-L-guluronate)
GG	α-L-guluronate dimer
GLF	Generalized Logistic Function
GM	α-L-guluronate and β-D-mannuronate dimer
GPR	Gaussian Process Regression
GPU	Graphics Processing Unit
H	High G content (0.63)
HBcAg	Hepatitis B Core Antigen
$HBcAg_2$	Hepatitis B Core Antigen dimer
HIV	Human Immunodeficiency Virus
IBI	Iterative Boltzmann Inversion
IM 1 / IM 2	Ion Model for Alginate (Sec. 4.2.2)
ICF	Instantaneous Collective Forces
IMC	Inverse Monte Carlo
ITC	Isothermal Titration Calorimetry
L	Low G content (0.33)
LD	Langevin Dynamics
LJ	Lennard-Jones
M	Mannuronic Acid (β-D-mannuronate)
MB	Maxwell-Boltzmann
MC	Monte Carlo
MD	Molecular Dynamics
MDEM	Molecular Discrete Element Method
ML	Machine Learning
MM	Molecular Mechanics
MM	β-D-mannuronate dimer
MPC	Multi-Particle Collision Dynamics
MSCG	Multi-Scale Coarse Graining
MSD	Mean Square Displacement
NMR	Nuclear Magnetic Resonance
NPT	Isothermal-Isobaric Ensemble
NVT	Canonical Ensemble
PBC	Periodic Boundary Condition
PCA	Principle Component Analysis
PDB	Protein Database
PDC	Pyruvate Dehydrogenase Complex
PDF	Probability Density Function
PME	Particle Mesh Ewald
PMF	Potential of Mean Force
PSD	Pore Size Distribution
PW	Polarizable Water in MD
QM/MM	Hybrid Quantum Mechanical / Molecular Mechanics

RBF	Radial Basis Function
REM	Replica Exchange Method
RDF	Radial Distribution Function
RMS	Root Mean Square
RMSD	Root Mean Square Deviation
RPY	Rotne-Prager-Yamakawa
SANS	Small Angle Neutron Scattering
SAS	Self-Assembled Structure
SAXS	Small Angle X-Ray Scattering
SP-PDC-	Simulation Procedure for PDC (Sec. 8.2)
SP-VLP-	Simulation Procedure for HBcAg / VLP (Sec. 7.2)
SRD	Stochastic Rotational Dynamics
STD	Standard Deviation
SVD	Singular Value Decomposition
SVR	Support Vector Regression
TCA	Tricarboxylic Acid
TEA	Truncated Expansion Ansatz
TPP	Thiamine Pyrophosphate
VLP	Virus-Like Particle
WHAM	Weighted Histogram Analysis Method

Greek symbols

α	first angle in Eulerian orientation representation	rad
β	second angle in Eulerian orientation representation	rad
γ	third angle in Eulerian orientation representation	rad
γ	friction coefficient in LD	s^{-1}
γ_Y	residual variogram in Kriging	kJ^2/mol^2
δ	distance	m
δ_m	minimum distance between BB of molecules	m
δ_r	RMSD between two molecular conformations	m
Δt	time step	s
ϵ	potential scaling factor	J
ζ	normally distributed random number with zero mean and unit variance in LD	-
η	dynamic viscosity	$kg\,s^{-1}\,m^{-1}$
η	random force in LD	N
θ	angle	rad
κ	estimation location in Kriging	-
λ	Lagrange multiplier in Kriging	-
μ	systematic trend in Kriging	J
μ_I	effective / reduced mass moment of inertia	kgm^2

μ_m	effective / reduced mass	kg
ξ_{i-j}	normalized contact number between i-j	-
ξ_t / ξ_r	normally distributed random number with zero mean and unit variance in diffusion model	-
ρ	density	$\mathrm{kg\,m^{-3}}$
σ	standard deviation	-
τ	time constant	s
ϕ	fraction	-
ϕ_i	angular grid spacing	rad
Ψ	force constant in LD	$\mathrm{N^2}$
ω	rotational velocity	$\mathrm{rad\,s^{-1}}$
ω_0	natural frequency	$\mathrm{rad\,s^{-1}}$

Latin symbols

$\vec{b}$	bond vector	m
BAC	bond-angle correlation	-
c	concentration	$\mathrm{kg/m^3}$
d	diameter or distance	m
d_{b-b}	distance between centers of ion volumes in IM1 and IM2	m
d_m	location of potential minimum in LJ potential	m
D	diffusion coefficient	$\mathrm{m^2/s}$ or $\mathrm{rad^2/s}$
f	ion fraction	-
f_m	basic fitting functions	-
f_{stab}	objective function for stability	-
$\vec{F}$	force vector	N
G	Gibb's free energy	$\mathrm{kJ\,mol^{-1}}$
h	cartesian grid spacing	m
h	heaviside function	-
H	enthalpy	$\mathrm{kJ\,mol^{-1}}$
I	mass moment of inertia	$\mathrm{kgm^2}$
k	coordination number	-
k_{B}	Boltzmann constant (1.38×10^{-23})	$\mathrm{J\,K^{-1}}$
K_D	equilibrium dissociation constant / affinity constant	M
L	Lagrangian energy	J
m	mass	kg
$\vec{M}$	torque	N m
$\mathbf{M}$	relative rotation matrix	-
M_w	molecular weight	kDa
n	nugget in Kriging variogram	-
N	number of respective index	-
$N_{\mathrm{bonded},i}$	number of bonds for particle / molecule i	-

N_p	number of particles / molecules	-
p	probability	-
$\vec{p}$	position vector	m
P	potential components in Kriging	-
$\mathbf{q}$	quaternion (prescript indicates element)	-
r	radius or diffusion distance or variogram range	m
s	sill in Kriging variogram	-
t	time	s
T	temperature	K
T_0	oscillation period	s
U	potential energy	J
v	velocity	$\mathrm{m\,s^{-1}}$
V	volume	$\mathrm{m^3}$
w	weight in Kriging estimation	-
x	position in x-axis	m
$\dot{x}$	velocity in x-axis	$\mathrm{m\,s^{-1}}$
$\ddot{x}$	acceleration x-axis	$\mathrm{m/s^2}$
y	position in x-axis	m
Y	random component in Kriging	J
z	position in z-axis	m

Indices

A	molecule A (interaction partner)
AB	compound of molecule A and B
an	annealing
asymp	asymptotical value
avail	available (ions)
ave	average
B	molecule B (interaction partner)
base	base case
bb	back-bone reference structure
bcp	bond contact point
bind	binding
body	body frame of reference
Ca2+	calcium ion
cavity	binding cavity for ion
coll	collision of molecule atoms / back-bone
COM	center of mass
const	constant
cont	contact of abstracted molecules
cool	cooling of annealing procedure

cor	correction
crit	critical
cubic	cubic function
cur	current (time)
cut	cutoff
des	desired (ions)
dif	diffusion
dis	dissipative drag
el	(volume) element
eq	equilibrium
exp	exponential function or in reference to experimental data
F	force
fib	fiber
fluc	fluctuation
full	full atom reference structure
gauss	Gaussian function
gel	gel / hydrogel
GLF	Generalized Logistic Function
glob	global reference
gyr	gyration
i	particle / molecule i (counter)
$i-j$	between i and j
int	interaction
j	particle / molecule j (counter)
K	Kriging
kind	molecule kind
life	life (time) of a structure
lin	linear
M	torque or counter
MD	relative to MD sample
min	minimum
max	maximum
n	normal
N	counter
outer	outer area / diameter
part	particle
pp	particle-particle (molecule-molecule)
r	rotational axis
ref	reference
rel	relative
rep	repulsive
rot	rotation
S	Stokes
SAS	self-assembled structure

sim	simulation
sph	spherical
struc	structured (contact)
t	tangential or translational axis
tot	total
unstruc	unstructured (contact)
used	used (ions)
V	volume
x	x-axis
y	y-axis
z	z-axis
0.9 nm	contact distance of coordination number
95	95 % of asymptotic value

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]

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

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

lower-scale methods such as atomistic MD, while top-down approaches aim at the reproduction of measurable properties on scales of the coarse-grained model through optimization of force-field parameters. In many cases bottom-up approaches are either not sufficiently accurate or prohibitively expensive leading to hybrid approaches [30]. In the following, some of these approaches are summarized in more detail and examples presented. Note that alternatively coarse-grained force-fields can e.g. be characterized as being derived vs. parameterized [107] or physics based vs. knowledge based [109]. In this work slight focus is placed on a bottom-up/top-down point of view, but some flexibility maintained as a clear differentiation is not always possible. For reviews see e.g. refs. [107–111, 132, 133].

Approaches with a bottom-up and derived focus attempt a direct determination of (pairwise) potentials between coarse-grained beads based on the atomistic and possibly quantum mechanical level (*ab initio* MD). Traditional approaches in the aforementioned formalism aim to reproduce the thermodynamic properties on the coarse-grained scales as accurately as possible through the coarse-grained potential U^{CG}, also called *potential of mean force* (PMF) [30]. One classical method for derivation is through *thermodynamic integration*, which applies constraints on the desired DOF in atomistic simulations to determine the forces for different configurations [30]. In the following, the overall coarse-grained potential can be determined through numerical integration of the average constraint forces. Similarly, *umbrella sampling* [134] employs restrain potentials instead of constraints (e.g. of harmonic shape), which can then be used either via their forces similarly to thermodynamic integration or via the distribution of configurations. Commonly this is achieved through the weighted histogram analysis method (WHAM) [135] (see [136] for a further generalization). Alternatively to using forces (via constraints/restrains), the *free energy perturbation* method [137] (also called *thermodynamic perturbation* [30]) determines the potential difference between nearby configurations using the Boltzmann factor of a perturbation (i.e. slight change in configuration). The individual potential differences between nearby configurations are then used to reconstruct the overall coarse-grained potential. In a similar direction, *steered molecular dynamics* (also referred to as *pulling* simulations) [30, 138] applies an external force on the desired DOF (either force or velocity controlled) to change the configuration over time. In the limit of zero pulling rates (i.e. without friction forces), the coarse-grained potential can be calculated from the work of the pulling force. For finitely larger, but still small pulling rates, the reversible work can be estimated from a set of simulations using the Jarzynsky equation [139, 140]. Further examples include the *effective force coarse-grained (EFCG)* [141] and *conditional reversible work (CRW)* method [142].

Approaches with a top-down and parameterized focus attempt to derive force-field parameters through optimization towards macroscopic properties on the scales of the

coarse-grained model and above. These properties are often structure-based, but can also be force-based. Consequently, a distinct differentiation to bottom-up approaches is not always possible, leading to hybrid approaches. One of the most widely used structure-based method is the *iterative Boltzmann inversion (IBI)* method [143], which performs an iterative refinement of the interaction potential based on the radial distribution function (RDF) between the beads, e.g. determined from all-atom MD simulations. Similarly *inverse Monte Carlo (IMC)* [144] iteratively refines based on RDFs, but performs corrections based on statistical mechanical arguments. More generally, the *relative entropy formalism* [145] aims to minimize the relative entropy difference between the atomistic and coarse-grained system (i.e. loss of information through coarse-graining) captured through relative probability density distributions for force-field parameterization. Both IBI and IMC are special cases [107]. Alternatively, in a non-iterative fashion the *generalized-Yvon-Born-Green theory (GYBG)* [146] employs coupled linear integration equations based on CG bead structural correlation functions to determine CG potentials. Force-based methods attempt to match the force-distribution of the coarse-grained model to that of the atomistic [107] - thus being similar to the previously mentioned bottom-up approaches. For example the *force matching (FM)* [147] method does so by matching to atomic forces and trajectories, which was later extended e.g. in the *multiscale coarse graining (MSCG)* method [148] to enable increased number of parameters.

In addition to these more traditional approaches, *machine learning* has gained increasing interest in the context of coarse-grained modeling generally, as well as with regard to force-field development. This includes specifically the usage of *neural networks* to describe (components of the) coarse-grained model. Examples include *CGnets* [130] with a reformulation of FM as a supervised learning problem and the *deep coarse-grained potential (DeePCG)* method [131] employing neural networks trained with all-atom data. Further details on machine learning approaches will be provided in the following section. For reviews see e.g. ref. [149].

A variety of coarse-grained models and force-fields have been proposed in literature and some examples of the most widely established (transferable) ones will be presented in the following. One of the most popular examples is the MARTINI force-field [53, 125, 126], which has initially been developed for lipids [53] and extended to a variety of other applications such as biological systems (e.g. proteins) [126] and material science [150]. Further examples in the context of proteins include SIRAH [151, 152], UNICORN [153], AWSEM [154], and OPEP [155]. Specific developments for nucleic acids include e.g. HiRE-RNA [156] and oxDNA/oxRNA [157, 158]. Further fields include e.g. polymers [132], metals [159], and clay [160]. Most of these transferable coarse-grained models combine a few atoms into a bead (e.g. four non-hydrogen atoms in the case of MARTINI [53]), thus

enabling moderate speedup while maintaining applicability to multiple systems. Higher (ultra-)coarse-grained levels are inherently more specialized and less transferable. Examples of such higher levels can be found in refs. [161, 162] and include simplified anisotropic bead shapes [163–166] and potentials [163, 166, 167].

Thus far, focus was placed on coarse-grained approaches in space. While these do also enable larger time scales through increased time steps (resulting from increased mass to force stiffness ratio) and accelerated dynamics (e.g. when neglecting random and friction forces from implicit DOF), dynamic detail of the relevant DOF is largely preserved. For systems with systematic forces changing slower than time scales of γ^{-1} with regard to the velocity distribution, Langevin's equation can be averaged over these time scales and thus the acceleration/inertia term in eq. 1.7 be neglected ($m\ddot{x} \approx 0$) [30]. This approximation leads to inertia-free **Brownian Dynamics (BD)** describing the velocity as

$$\dot{x} = -\frac{D}{k_\mathrm{B}T}\nabla U^\mathrm{CG} + \sqrt{2D}\zeta, \tag{1.9}$$

where D is the diffusion coefficient according to the Einstein relation [168]

$$D = \frac{k_\mathrm{B}T}{m\gamma} \tag{1.10}$$

with k_B being the Boltzmann constant and T temperature. At the cost of loosing dynamic detail, this approximation increases addressable time scales and is explored in a variety of examples [169–176].

1.2.3 Other Derivatives and Related Methods: From Monte Carlo to Machine Learning

In addition to the mentioned coarse-graining strategies in space and time, there is a variety of further derivatives from MD and related methods. In the following, some examples related to this work are provided. For reviews on related methods see e.g. refs. [30, 32].

With regard to the **structural formation of colloidal systems**, a variety of methods have been developed in which particles form rigid structures upon contact with a defined probability. Depending on their propagation and contact scheme, these are usually called **ballistic, diffusion-limited, or reaction-limited cluster-cluster aggregation (BLCA, DLCA, RLCA)** [177–180].

With regard to **static properties of molecular systems**, these approaches employ for example statistical sampling, such as Monte Carlo [32, 181–183], or more recently data-driven methods, such as in machine learning [149, 184, 185]. **Monte Carlo (MC)**

methods are widely employed in the context of molecular modeling with regard to generating representative statistical ensembles of molecular system configurations at specific thermodynamic conditions [30, 186]. This is achieved by applying random pertubations to those system configurations according to their (thermodynamic) probability [187]. Thus, the conformational space is sampled instead of analyzing trajectories in time [30, 187]. Additionally, hybrid MC/MD have been developed to combine dynamic information from MD with improved conformational sampling through MC [188]. Similarly, event driven methods such as **discrete molecular dynamics (DMD)** [189–191] provided improved sampling and increased time scales. In addition, data-driven methods have gained increasing interest in recent years through the rapid development of **machine learning (ML)** supported by increasing data sets and computational resources (e.g. graphics processing units, GPUs). On small scales, this has lead e.g. to machine learning descriptions of potential energy surfaces from quantum mechanics [56, 57] and coarser atomistic systems [130, 131]. On larger scales, ML has been employed to predict protein complexes and their quaternary structure [192], design materials and molecules [193, 194], as well as pharmaceuticals [195]. Similarly, these trends show in increasing efforts for benchmarking and comparison of methods [6, 196]. Nonetheless, the majority of research in the field of molecular modeling remains in the context of discrete methods around molecular dynamics, as these methods enable detailed *dynamic* insights and physics-based descriptions. For reviews on machine learning in molecular modeling see e.g. refs. [149, 184, 185].

With regard to **larger macroscopic scales**, further methods including both discrete and continuum approaches have been developed. For macroscopic objects in the diffusion-free limit, the **discrete element method (DEM)** [197–199] provides a framework to study particle and granular systems in engineering applications. Methodologically this is equivalent to MD, i.e. Newton's equation of motion is solved for discrete objects, but requires different interaction models (i.e. force-fields) and no thermostats or barostats. Similarly, bonds can be employed, which are especially useful to study mechanical breakage of structures [200–203]. Furthermore, DEM has been applied to nano- and microscopic particles and agglomerates for insights in formation (e.g. using spray-drying [204]) or mechanical characterization [205] - particularly of battery materials [206, 207]. In addition, continuum methods through finite volumes or finite differences enables investigation of fluids (**computational fluid dynamics (CFD)** [208]) and solids (**finite element method (FEM)** [209]). Coupling of these methods is also frequently performed: CFD-DEM [210–212] and FEM-CFD [213] enable for example the investigation of interacting solids and fluids.

Additionally, there are methods on the intermediate scale to describe fluid systems using discrete approaches, while attempting to conserve macroscopic properties such

as mass, momentum, and energy. For example **multi-particle collision dynamics (MPC)** [214–216], also called **stochastic rotational dynamics (SRD)**, uses a particle description for (complex) fluids. It performs alternating streaming/collision steps for particles and has gained interest for a variety of fluid systems including binary mixtures and micro-emulsions [215, 217]. Another example is **dissipative particle dynamics (DPD)** [218, 219], which employs a method similarly to LD by applying friction and noise terms to describe e.g. the rheology of polymers and other complex fluids [220–222].

1.2.4 Scope of this Work

In the scope of this work, a novel abstraction will be proposed for Langevin Dynamics using data-driven methods. In contrast to state of the art methods typically employing simple 1D interaction potentials between coarse-grained beads, this abstraction describes entire macromolecules as anisotropic unit objects and interactions through data-driven potential fields in 6D space of relative position and orientation. As a result, a high level of detail can be retained within these potential fields, while exploiting larger time scales through increased coarse-graining. Furthermore, an implicit model for anisotropic diffusion and enforcement of the desired canonical ensemble will be proposed. The method will be demonstrated on the structural formation of the hepatitis B core antigen and pyruvate dehydrogenase complex, including a bottom-up parameterization strategy. Additionally, a specialized form will be presented for modeling the structural formation of alginate during calcium mediated gelation. In the following, these model systems will be presented in detail.

1.3 Model Systems

This chapter is based on the following publications:

P. N. Depta, U. Jandt, M. Dosta, A.-P. Zeng, and S. Heinrich. Toward Multiscale Modeling of Proteins and Bioagglomerates: An Orientation-Sensitive Diffusion Model for the Integration of Molecular Dynamics and the Discrete Element Method. *J. Chem. Inf. Model.*, 59(1):386–398, 2019

P. N. Depta, P. Gurikov, B. Schroeter, A. Forgács, J. Kalmár, G. Paul, L. Marchese, S. Heinrich, and M. Dosta. DEM-Based Approach for the Modeling of Gelation and Its Application to Alginate. *J. Chem. Inf. Model.*, 62(1):49–70, 2022

P. N. Depta, M. Dosta, W. Wenzel, M. Kozlowska, and S. Heinrich. Hierarchical Coarse-Grained Strategy for Macromolecular Self-Assembly: Application to Hepatitis B Virus-Like Particles. *Int. J. Mol. Sci.*, 23(23):14699, 2022

P. N. Depta, M. Dosta, and S. Heinrich. Data-Driven Multiscale Modeling of Self-Assembly and Hierarchical Structural Formation in Biological Macro-Molecular Systems. In W. E. Nagel, D. H. Kröner, and M. M. Resch (editors), *High Performance Computing in Science and Engineering '21*. Springer International Publishing, Cham, 2023

P. N. Depta, M. Dosta, and S. Heinrich. Data-Driven Multiscale Modeling of Self-Assembly and Hierarchical Structural Formation in Biological Macro-Molecular Systems: Pyruvate Dehydrogenase Complex. In W. E. Nagel, D. H. Kröner, and M. M. Resch (editors), *High Performance Computing in Science and Engineering '22*. Springer International Publishing, Cham, 2024 (in print)

P. N. Depta, M. Dosta, and S. Heinrich. Multiscale Model-Based Investigation of Functional Macromolecular Agglomerates for Biotechnological Applications. In A. Kwade and I. Kampen (editors), *Dispersity, Structure and Phase Changes of Proteins and Bio Agglomerates in Biotechnological Processes*. Springer International Publishing, Cham, 2024 (in print)

In the context of this work, three model systems from a material science and biological background will be studied and used for validation of the proposed model framework. While macromolecular structural formation occurs in a variety of other fields, these fields exhibit some of the most interesting phenomena with regard to human life and technological development.

1.3.1 Alginate Gelation

Cross-linked hydrophilic polymers, also called hydrogels in solution, find a variety of applications in material science and biomedicine such as for example soft contact lenses. Additionally, in recent decades the further processing into lightweight nano-porous materials called aerogels through supercritical drying has gained increasing interest [24, 229, 230]. In this context, biopolymers such as alginate are important base materials and chosen as one model system for this work based on the respective publication Depta *et al.* [224].

Alginate is an umbrella term of polysaccharides naturally occurring in brown algae with weight-averaged molecular weights around 200 kDa and molar-mass dispersities between 1.5 and 3 [231]. It is a binary copolymer of mannuronic (M) and guluronic (G) acid with linearly 1-4-linked residues as it can be seen in the atomistic structure visualized in Fig. 1.1. As shown, the polymer composition is structured in block-wise patterns of homopolymeric regions (G- and M-blocks), as well as regions of alternating residues [231]. While G- and M-units are of identical molar mass (175 g/mol) and overall similar, their configuration of the C-5 atom leads to different conformations of the pyranose ring. Based on this, homopolymeric M and alternating GM regions exhibit the shown flat ribbon-like structure with increased flexibility, while homopolymeric G regions exhibit a buckled structure with increased rigidity.

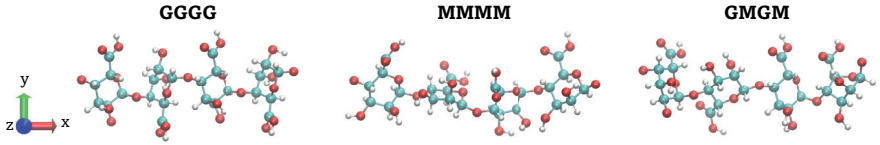


FIGURE 1.1: Visualization of the alginate reference structure from data provided by Hecht *et al.* [232]. The sequences GGGG (left), MMMM (center), and GMGM (right) are shown and colored according to atom type (carbon in cyan, oxygen in red, hydrogen in white).

Alginates are characterized by their molecular weight and polymer composition. This includes at the fundamental level the fraction of G and M units ϕ_G and ϕ_M, respectively, but also more complete statistical descriptions through fractions of diads, triads, tetrads, and higher multads [233]. Due to their biosynthetic production, composition varies for example depending on season and part of the brown algae plant [231].

Gelation of alginate is achieved through addition and binding of di- and trivalent cations (except Mg^{2+}) [234] leading to the polymer network formation considered a (physical, i.e. non-covalently bonded) hydrogel on the macroscopic scale. In this regard, gelation through calcium cross-linking is the most widely studied alginate system due to its

straightforward production via pouring of aqueous sodium alginate into a solution of dissolved calcium carbonate salt. While the exact mechanisms of metal binding remain a subject of scientific debate, qualitatively cross-linking is established to be driven through the binding of M^{2+} ions to two adjacent G units favored by the nest-like structure with carboxylate and hydroxyl groups of guluronic acid. With increasing ion concentrations this leads to a dimerization of G-M^{2+}-G sites on different polymer fibers resulting in cross-linking of zigzag-shaped junction zones and an auto-cooperative zipping. This mechanism termed 'egg-box' model has been initially proposed and described by Grant *et al.* [235] and revised by numerous authors in the following decades [236–238]. The mechanism is established to be driven by the nest-like conformation of poly-G regions and hindered for the flatter conformation of poly-M and alternating GM regions. The cross-linking subsequently leads to inter-cluster associated multimers and the overall gel network [239], as visualized schematically in Fig. 1.2.

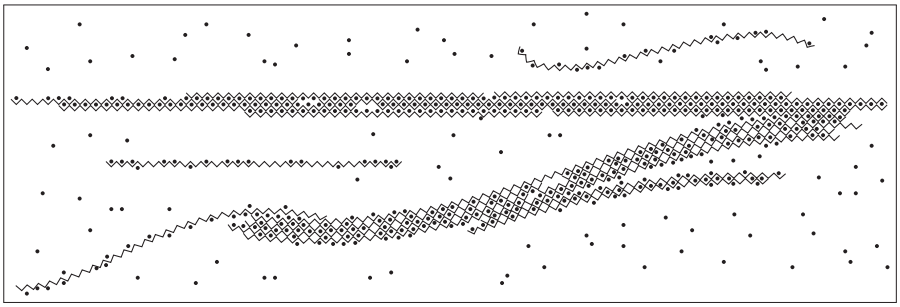


FIGURE 1.2: Schematic drawing of the calcium mediated gelation of alginate. The zigzag lines represent the alginate polymer fiber inspired by the 'egg-box' model and circles represent calcium ions.

During gelation, various parameters such as ion concentration, alginate molecular weight, and composition have a direct impact on the gel structure and its mechanical properties [232]. For example, alginates with high G fraction of $\phi_G > 0.7$ and homopolymeric G regions of tens of residues exhibit increased mechanical strength in comparison to those with high ϕ_M. Based on this, industrial alginates are typically characterized as either high-G or low-G alginates [233]. Further effects will be discussed in more detail in the respective results section.

With regard to modeling, the majority of approaches in literature were conducted around MD and DFT, thus providing primarily insight on small scales. From these insights, for example the 'egg-box' model [235] has been modified concerning its molecular structure and the importance of electrostatic effects [236, 237]. Similarly, the specifics of ion acceptance with regard to free energy have been investigated [240]. Furthermore, on slightly larger scales of multiple poly-G chains cross-linkage was investigated elucidating the association mechanisms [241, 242]. In the following, similar works were performed

extended to alternating G and M compositions of fibers and their influence on association [232, 243]. With regards to larger scales of modeling, e.g. BD and MC were used to improve the understanding of polymer stiffness and persistence lengths [170]. The subsequently proposed model and framework builds upon these previous works for parameterization and validation. The framework attempts to advance the field by providing a meso-scale model for the overall structural formation during calcium mediated gelation of alginate.

1.3.2 Hepatitis B Core Antigen (HBcAg)

A variety of viruses, such as the hepatitis B virus, contain structural proteins with the ability to assemble into regular structures, called virus-like particles (VLP) or virus capsids. These structures are critical for the overall function of the virus during infection and reproduction, as well as for the immune system recognition [11]. Due to this importance, this work investigates the self-assembly of the hepatitis B core antigen (HBcAg) into its VLP. The HBcAg VLP, shortly referred as VLP, is composed of either 90 or 120 dimer units, subsequently termed $HBcAg_2$, in an icosahedral capsid structure [12, 244, 245]. For the fully expressed HBcAg, the 120 dimer capsid was found to account for more than 95 % of the population [245]. A visualization of this capsid including its atomistic secondary structure, as well as coarse-grained structure, can be found in Fig. 1.3.

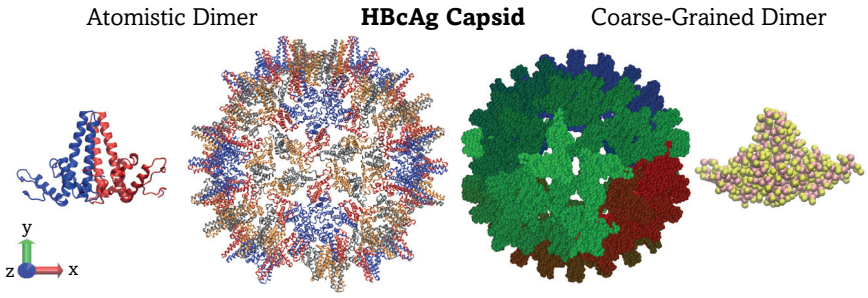


FIGURE 1.3: Visualization of the HBcAg virus capsid assembly for the coarse-grained reference structure (right, coloring by index to improve contrast) and atomistic structure (left, coloring by chain to improve contrast) based on data provided by Mariana Kozlowska and processed by Uwe Jandt based on PDB 6HTX [246] and PDB 1QGT [244] as outlined in Depta *et al.* [225].

Within the capsid, two slightly different dimer conformations (AB and CD) are typically distinguished build from two copies of the HBcAg monomer (root-mean-square deviations between 0.5 - 1.2 Å in α-carbons) [244]. While the majority of the conformation is the same, larger differences in conformation of AB and CD are in the inter-dimer interaction regions for the capsid assembly, particularly residues 128-136 [244]. Note that in

the context of this work, only one reference structure of the dimer is modeled. This is sufficient, as the molecule's structure is flexible in itself during MD for parameterization of all models.

In literature, a variety of attempts have been conducted to better understand the VLPs, specifically those of HBcAg and the human immunodeficiency virus (HIV). Atomic structure and organization of the HBcAg monomer, dimer, and resulting VLPs are well understood through electron cryomicroscopy (cryo-EM) [12, 244]. Similarly, the role of specific sections of the composing amino acid chain with respect to capsid self-assembly and binding of genetic material is well understood through experimental testing [247, 248]. Furthermore, such systems are regularly being adapted for vaccines [249, 250] and drug delivery [251, 252]. However, understanding the self-assembly mechanics remains difficult, as the process is nucleation-limited and trapping intermediates is nearly impossible [225, 253, 254]. Similarly to experimental insights, numerical methods such as atomistic and coarse-grained MD have improved understanding of subunits and capsid stability [255–258], but are unable to capture time and length scales necessary for VLP assembly. For understanding these processes, hundreds to thousands of copies of the structural proteins need to be simulated over milli-seconds and longer. In this context, so far only very simplified and specialized models have been proposed based for example on trapezoidal/triangular shapes, patchy-spheres, or hard pseudoatoms [163–166]. Consequently, transferability to different systems and process aspects remains difficult. In this regard, the subsequently proposed model and framework targets to provide a generic meso-scale model for structural formation including a data-driven bottom-up parameterization approach building upon aforementioned works for validation.

1.3.3 Pyruvate Dehydrogenase Complex (PDC)

The pyruvate dehydrogenase complex (PDC) links the anaerobic glycolytic energy pathway to the aerobic tricarboxylic acid (TCA) cycle by catalyzing the conversion of pyruvate into acetyl coenzyme A (acetyl-CoA) [18], thus enabling cellular respiration. As a result, PDCs exist in various living organisms from bacteria (e.g. *Escherichia coli*) to mammals such as humans [18], in which form it is the focus of this work. For enabling this biocatalytic functionality at sufficient rates, PDC relies on structural assembly. It is composed of multiple copies of four different proteins shown in Fig. 1.4 for mammalian and specifically human PDC: pyruvate dehydrogenase (E1), dihydrolinpoamide acetyltransferase (E2), dihydrolipoamide dehydrogenase (E3), and the E3 binding protein (E3BP) [259, 260]. As it can be seen, E2 and E3BP contain a flexible linker arm, often called swinging arm, connecting catalytic, binding, and lipoyl domains [261]. In

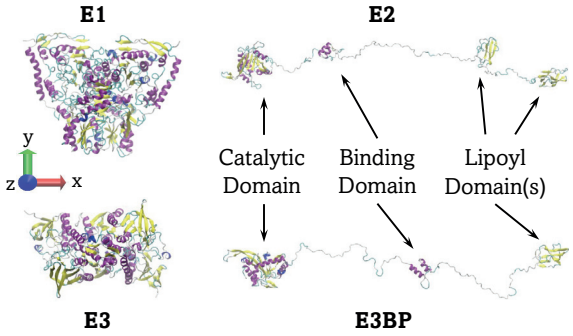


FIGURE 1.4: Visualization of the secondary structure of all PDC components based on their atomistic structure. Adapted with permission from Depta *et al.* [223]. Copyright 2019 American Chemical Society.

contrast, the heterotetramer E1 and homodimer E3 are structurally more globular. Note that some organisms can lack the E3BP protein [262, 263].

Human PDC forms a comparatively large 60-mer core of E2 and E3BP in pentagonal dodecahedral shape [264, 265] with studies indicating composition from two trimer populations of either three E2 or two E2 with one E3BP [259, 266]. The 60-mer core stoichiometry is still a matter of some debate with earlier studies finding a 48×E2 + 12×E3BP stoichiometry more likely [264, 267] including stability indications from modeling [268], while recent studies are rather supporting a 40×E2 + 20×E3BP stoichiometry [259, 265, 266]. As a visual example, the 60-mer core based on a composition of 48×E2 + 12×E3BP is shown in Fig. 1.5. The E1 and E3 proteins then bind as shown to the

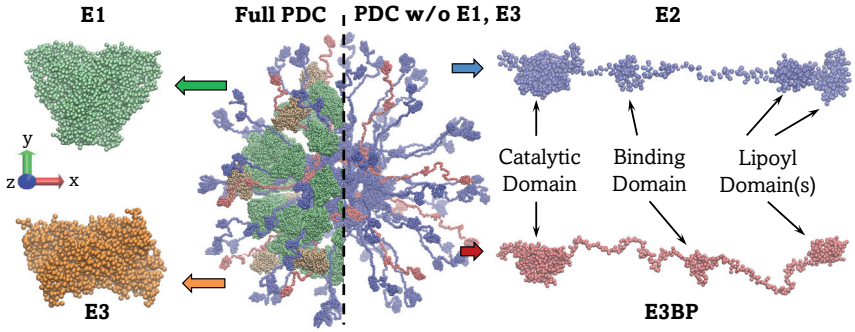


FIGURE 1.5: Visualization of the PDC assembly (48×E2 + 12×E3BP) and components based on their coarse-grained structure. Note that this is the structure after representative clustering for E2 and is consequently different from that in Fig. 1.4. Adapted with permission from Depta *et al.* [223]. Copyright 2019 American Chemical Society.

binding domains of E2 and E3BP, respectively assembling the overall complex. This self-assembled structure of PDC aids its function of biocatalysis through mechanisms such as metabolic channeling [261] by avoiding free diffusion of metabolites [16]. Nonetheless, activity of PDC has also been observed for smaller structural assemblies than the 60-mer core as well as larger ones [269]. Consequently, understanding and predicting the structural assembly of macromolecular systems such as PDC is of great interest.

As discussed before, the investigation of such systems is challenging both experimentally and specifically numerically. While the majority of previously mentioned findings with regard to PDC has been developed experimentally, numerical investigations to the same extend are lacking. Nonetheless, an increasing number of studies have investigated the PDC system typically employing MD to gain insight to structural mechanisms on the atomistic scale.

With regard to the modeling of (human) PDC, only a limited number of publications are available primarily with a focus on MD leading to atomistic insights. In this regard, atomistic MD has been used to further establish the molecular structure of PDC subunits and their dimer and trimer interaction to aid detailed understanding of the core assembly [106]. Similarly, atomistic simulations have been extended to the 60-mer core without linker arms providing similar results [270]. As such atomistic MD models are strongly limited with regard to reachable length and time scales, coarse-grained MD models of PDC have been developed and validated by the underlying atomistic models [268]. With these models it has been possible to investigate the 60-mer core stability including linker arms with respect to stoichiometry (suggesting a higher stability of the 48×E2 + 12×E3BP composition [268] in contrast to recent experimental works pointing towards 40×E2 + 20×E3BP [265]), as well as binding of E1 and E3 to the 60-mer core improving understanding of binding and assembly stoichiometry (suggesting approximately 30×E1 per 60-mer core) [268, 271]. Furthermore, MD simulations have been performed with regard to the binding of thiamine pyrophosphate (TPP) to E1 for various mutations caused by disease to identify the underlying mechanisms [272]. Similarly, substrate binding and inhibition of E1 has been investigated by other authors [273–276]. Note that while these models provide detailed insight on the atomistic scale, they are unable to capture length scales for structures beyond the 60-mer as well as time scales required for assembly. In order to improve understanding on such scales, the same generic data-driven multiscale model as for the VLP system will be applied to study the more complex PDC system.

1.4 Outline

The main focus of this thesis is on method development in the field of structural formation of macromolecular systems. In Ch. 2, the proposed model framework will be presented and all components outlined. In total, there are three components, which will be described in the following chapters:

The first model component, the diffusion and thermodynamic model, will be presented in Ch. 3. An anisotropic diffusion model for enforcing the appropriate thermodynamics (canonical ensemble, i.e. constant temperature) will be derived based on Langevin dynamics. The model will be at the foundation of the diffusion driven structural formation.

The second model component, the intermolecular interaction, will be presented in Ch. 4. Two different approaches for describing intermolecular interaction will be provided. The first approach will be a specialized model for ion mediated gelation of alginate based on literature and theoretical considerations. The second approach will be a generic data-driven approach for deriving the potential fields for interactions between macromolecules from MD, which can then be used for meso-scale simulations.

The third model component, the bonded interaction model, will be presented in Ch. 5. Two different approaches will be provided. The first approach will be on a pairwise elastic bond model including orientation, which is useful for complex macromolecular bonds. The second approach will be a simplified fiber bond model used in gelation modeling.

After providing the methods, the results will be presented for each model system including alginate (Ch. 6), HBcAg virus-like particles (Ch. 7), and PDC (Ch. 8). Lastly, in Ch. 9 the main results will be summarized and an outlook provided.

2

Model Framework

This chapter is based on the following publications:

P. N. Depta, M. Dosta, W. Wenzel, M. Kozlowska, and S. Heinrich. Hierarchical Coarse-Grained Strategy for Macromolecular Self-Assembly: Application to Hepatitis B Virus-Like Particles. *Int. J. Mol. Sci.*, 23(23):14699, 2022

P. N. Depta, M. Dosta, and S. Heinrich. Data-Driven Multiscale Modeling of Self-Assembly and Hierarchical Structural Formation in Biological Macro-Molecular Systems. In W. E. Nagel, D. H. Kröner, and M. M. Resch (editors), *High Performance Computing in Science and Engineering '21*. Springer International Publishing, Cham, 2023

P. N. Depta, U. Jandt, M. Dosta, A.-P. Zeng, and S. Heinrich. Toward Multiscale Modeling of Proteins and Bioagglomerates: An Orientation-Sensitive Diffusion Model for the Integration of Molecular Dynamics and the Discrete Element Method. *J. Chem. Inf. Model.*, 59(1):386–398, 2019

2.1 Introduction and Framework Overview

As outlined in the introduction, macromolecular structural formation is at the foundation of many processes in material science and technology, as well as virtually all processes in biology and biotechnology. At small scales, molecular dynamics (MD) related methods are typically applied in order to capture these structural formation, self-assembly, and agglomeration processes. MD related methods have the advantage of describing systems at atomistic resolution and thus provide a wealth of knowledge. However, while coarse-grained methods exist, their applicability to the scales necessary

for describing structural formations on the micro-meter and milli-second scale remains difficult.

In order to provide insight into these processes, this work aims at providing a generically applicable model framework capable of describing structural formation phenomena on the micro-meter and milli-second scale, building upon MD and literature data for parameterization. For this, **an entire macromolecule is abstracted as an anisotropic object with a certain position, orientation, and spatial extent**. As the proposed abstraction level is in between MD and DEM, the methodology is termed the **Molecular Discrete Element Method (MDEM)**. Three model components, which can be seen in Fig. 2.1, are then used to describe the interaction between macromolecules and with the environment, e.g. solvent and salts. While the model components are chosen similarly to the typically employed strategies in MD, model formulations and parameters are derived specific to the abstraction and system. At the basis of the framework stands the molecular reference structure, which is available for many systems from the protein database (PDB). Details on the origin of reference structures for the model systems in this work are provided in Sec. 2.3. Similarly, often information on structural assembly and binding locations are available, which are optional information for this framework.

As the first model component, a Langevin Dynamics (LD) and implicit solvent based **diffusion model** provides the foundation for describing the anisotropic diffusion kinetics of molecular movement in translation and rotation, as well as enforcing the desired canonical ensemble [223]. The proposed model provides a full methodology for deriving anisotropic 6D diffusion coefficients from MD based on the molecular reference structure. Furthermore, effects of solvent temperature and viscosity can be readily modeled and do not require a re-parameterization as long as no significant conformational changes of the reference structure occur. Alternatively, diffusion coefficients can also be estimated from the molecular reference structure based on theoretical considerations. Hydrodynamic interaction between units, i.e. forces resulting from their relative velocities coupled through the solvent, is neglected in line with literature for anisotropic biomolecules [277]. More details will be provided in Ch. 3.

As a second model component, a **pairwise intermolecular interaction model** provides a model for the interaction of macromolecules in 6D space of all relative configurations. Two different approaches were explored within this work. One approach provided a probabilistic functional model to describe alginate gelation based on literature data and theoretical considerations [224], which is consequently specific to the system. The other approach explores a generic data-driven methodology for deriving intermolecular interaction potentials from MD [225, 226] for wide applicability. More details will be provided in Ch. 4.

As a third model component, **bonded interaction** of strong intermolecular assemblies can be captured separately from the pairwise intermolecular interaction. This optional component can be advantageous from a numerical point of view and provides a more effective description of bonded structures, which are very stable over the course of a simulation. Such bonded interaction can either be parameterized by decoupling from the intermolecular interaction potential or through information on the structural assembly. In the context of this work, bonded interaction is avoided as far as possible to capture assembly from the most fundamental monomer or dimer units. More details will be provided in Ch. 5.

The three model components provide the force and torque contributions on each macromolecule resulting from interaction with the environment and other molecules. Time

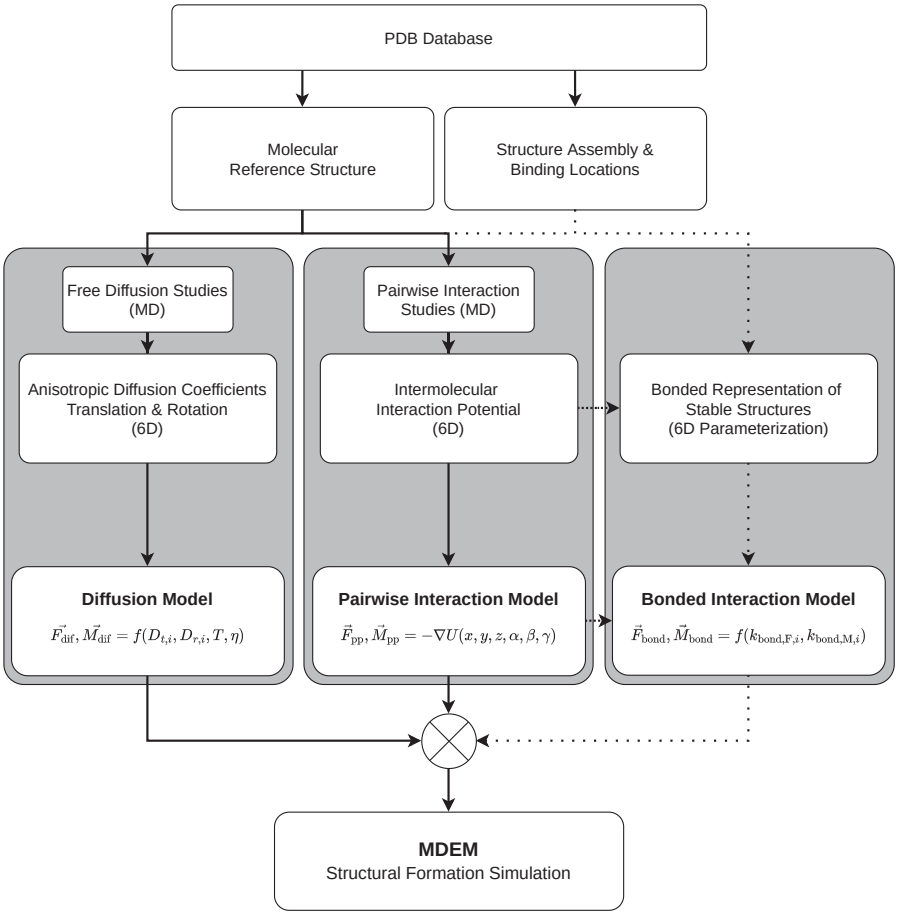


FIGURE 2.1: Overview of the physics-based and data-driven framework for macromolecular structural formation.

integration can then be performed in the context of Newton's equation of motion and will be discussed in more detail in the following section. Note that during parameterization of model parameters in MD, the molecular reference structure is flexible. Consequently, structural variability is at least partially sampled during parameterization. For a more detailed overview including sub-components the interested reader is referred to App. A.2.

2.2 MDEM Implementation

The proposed framework was implemented in a custom C++ and CUDA based software suite building upon the open-source code *GROMACS* [41, 278] for MD and a heavily modified version of the open-source DEM code *MUSEN* [279] for the structural formation simulations of MDEM. *GROMACS* was used for investigating the molecules of interest on an atomistic and coarse-grained MD scale and no development was performed. The code for pre-/post-processing, especially for data-driven interaction potential extraction, was written in C++ with a hybrid MPI+OpenMP parallelization and will be described in more detail in Ch. 4. The focus of this section is on the modifications of the DEM code *MUSEN* [279]. For more detail on *MUSEN* the interested reader is referred to ref. [279]. In the following, the term 'particles' will typically be used consistent to DEM, which refers in the context of this work to the abstracted macromolecular unit structures.

Equation of Motion and Numerical Time Integration In order to model the structural assembly over time, numerical time integration has to be performed. For all time steps, the total force $\vec{F}_{\mathrm{tot},i}$ and torque $\vec{M}_{\mathrm{tot},i}$ of each particle i has to be calculated from all contributions as

$$\vec{F}_{\mathrm{tot},i} = \vec{F}_{\mathrm{dif},i} + \sum_{j \in N_p} \vec{F}_{\mathrm{pp},i-j} + \sum_{k \in N_{\mathrm{bonded},i}} \vec{F}_{\mathrm{bond},i-k}, \tag{2.1}$$

$$\vec{M}_{\mathrm{tot},i} = \vec{M}_{\mathrm{dif},i} + \sum_{j \in N_p} \vec{M}_{\mathrm{pp},i-j} + \sum_{k \in N_{\mathrm{bonded},i}} \vec{M}_{\mathrm{bond},i-k}, \tag{2.2}$$

where N_p is the number of particles in the system and $N_{\mathrm{bonded},i}$ is the number of bonded interactions of particle i. As the particles, i.e. macromolecules, are highly anisotropic, their mass moment of inertia is a tensor. The reference structures of all macromolecules are orientated along their principle component axes in ascending order of x, y, z (body frame of reference in center of mass). Consequently, time-integration of rotation has to be performed in this body frame of reference and the forces and torques from the global reference frame rotated accordingly for each particle. This is different from the typical DEM formulation, which usually works with spherical particles. Time integration is

subsequently performed using the leap-frog algorithm and quaternion integration from ref. [198]. Normalization of unit quaternions was performed after each time step. Note that the convergence of explicit time integration algorithms is highly dependent on the time step size. Consequently, for each model component a critical time step was determined.

Inter-particle/molecule Contact and Detection As inter-particle contact detection for structurally anisotropic macromolecules is challenging, two adaptations were implemented. First, in order to build upon the normal contact detection algorithm implemented in *MUSEN* [279], an extended (spherical) contact radius was implemented covering all orientations for which a contact can occur including long-range interactions. Second, in order to reduce the computational demand of the pairwise interaction model, a boolean field was pre-calculated for each interaction pair, which saves whether a relative position and orientation of two objects requires a calculation of the interaction model, thus reducing the number of contacts.

Boundary Conditions In the context of this work only periodic boundary conditions (PBC) were used in order to avoid boundary effects. If required in the future, e.g. for interface phenomena, other boundaries could be implemented.

Initialization In order to initialize the system, all particles / macromolecules were randomly placed and oriented [280] inside the simulation domain with zero translation and rotation velocity, as well as disregarding potential overlaps. This includes structures such as alginate polymer fibers. Equilibration took place at the beginning of each simulation study by the respective model. Velocity equilibration was performed using the diffusion model and took place typically within 2-3 critical time steps (≈ 1 ps). Overlaps of initially placed molecules were resolved by the repulsive component of the interaction model (Ch. 4). Alginate polymer fibers were initially placed straight in the simulation domain, followed by an equilibration procedure (see 6.2).

Numerical Precision Similarly to MD, all simulations were performed using single point floating precision for accelerated runtime and the *MUSEN* source code adapted accordingly. While higher precision is advantageous to ensure reproducible results in deterministic DEM simulations, in the context of diffusion and numerical thermostats this is not required and would only lead to unnecessarily longer runtimes.

2.3 Molecular Reference Structures

The molecular reference structures provide the foundation for parameterization of the proposed framework shown in Fig. 2.1. For most macromolecules, especially in the biological context, structural information is available from the protein data bank (PDB). Furthermore, as the question of a molecular reference structure is directly equivalent to the widely studied topic of protein folding, there is an increasing number of predictive models including e.g. AlphaFold [7]. Based on this structural information, the proposed framework can then be employed. With regard to the studied model systems, the following reference structures were used. Note that all of these reference structures were centered in their center of mass and oriented along their principle component axes (x first, then y, then z), which is considered their body frame of reference.

Alginate The atomistic reference structures of mannuronic and guluronic acid including permutations along alginate polymer chains were provided by Hecht *et al.* [232] as shown in the introduction Sec. 1.3.1 Fig. 1.1 [224]. No modifications were performed.

Hepatitis B Core Antigen (HBcAg) The atomistic reference structure of $HBcAg_2$ for AB and CD shown in the introduction Sec. 1.3.2 Fig. 1.3 (left) [225, 226] was provided by Mariana Kozlowska based on a modified version of PDB 6HTX [246] and PDB 1QGT [244]. Two residues (74 and 97) of 6HTX were changed back to the original ones and reconstructed by *ROSETTA* using 1QGT as a template. Residues from the C-terminal of chain D were used to reconstruct missing residues on chain A. Furthermore, the missing three residues of the terminal of chain C were added by loop homology modeling in *Modeller* 9.21. In order to derive the coarse-grained reference structure primarily used within this work (see Fig. 1.3 right), representative clustering of the AB dimer was performed by Uwe Jandt using the linkage method as implemented in *GROMACS* [42] on the *Martini* coarse-grained structure. For this, conformations of a 10 ns MD run with 10 ps savings interval were used (process conditions 293 K and 150 mM NaCl in addition to charge neutralization, see further description in models section). Root-mean-square deviation (RMSD) of the determined reference structure in comparison to the original structure was 0.39 nm. Note that in the context of this work, only one reference structure of the $HBcAg_2$ dimer is modeled. This is sufficient, as the molecule's structure is flexible in itself during MD for parameterization of all models.

Pyruvate Dehydrogenase Complex (PDC) The atomistic and coarse-grained reference structures of the four proteins E1, E2, E3, and E3BP belonging to PDC were

provided by Uwe Jandt based on Hezaveh *et al.* [106, 268, 271] as shown in the introduction Sec. 1.3.3 Fig. 1.4 and 1.5 [223, 227, 228]. Due to the more volatile linker arm of E2, additionally representative clustering was performed using the linkage method as implemented in *GROMACS* [42] based on 61 *Martini* coarse-grained pairwise interactions between E2-E2 for 0.5 ns (process conditions 300 K and no additional ions, only charge neutralized, see further description in models section). Root-mean-square deviation (RMSD) of the determined E2 reference structure in comparison to the original structure was 3.27 nm highlighting the effects of the linker arm. The results published in Depta *et al.* [223] are based on the non-clustered E2 structure, while the updated structure based on representative clustering is used for this work.

3

Diffusion and Thermodynamics

This chapter is based on the following publication:

P. N. Depta, U. Jandt, M. Dosta, A.-P. Zeng, and S. Heinrich. Toward Multiscale Modeling of Proteins and Bioagglomerates: An Orientation-Sensitive Diffusion Model for the Integration of Molecular Dynamics and the Discrete Element Method. *J. Chem. Inf. Model.*, 59(1):386–398, 2019

3.1 Introduction

Molecular phenomena are generally influenced by their system's temperature, resulting in a specific velocity distribution and subsequent motion of all objects known as diffusion or Brownian motion. The velocity distribution increases with temperature from absolute zero and follows a Maxwell-Boltzmann distribution for each object and degree of freedom (DOF). As discussed in detail in Sec. 1.2, this is modeled for the individual atoms in MD through thermostat models typically distinguishing between stochastic methods, strong-coupling methods, weak-coupling methods, and extended system dynamics [30]. Furthermore, these aspects are strongly related to the representation of the solvent in coarse-graining approaches leading e.g. to an implicit solvent representation and a stochastic thermostat as in Langevin dynamics (see Sec. 1.2.2).

In contrast to the point masses modeled in atomistic and coarse-grained MD, the abstracted macromolecules in the proposed methodology further require treatment of rotational diffusion and directional anisotropy resulting from their complex molecular structure, e.g. the linker arm of PDC's E2 enzyme inducing diffusion anisotropy (Fig. 1.4). In order to address this, an orientation-sensitive diffusion model was formulated [223] based on the MD thermostat by Goga *et al.* [281], which will be presented subsequently.

3.2 Model Description

3.2.1 Overview

Consider as outlined in Ch. 2 an anisotropic macromolecule of mass m oriented along its principle component axes in ascending order x, y, z (body frame of reference in center of mass, rotational axes α, β, γ) with moment of inertia I_i (i indicating the axis). The molecule possesses a translatory velocity $v_{\text{global},i}$ and angular velocity $\omega_{\text{global},i}$ in the global frame of reference, which can be transformed into the body frame of reference ($v_{\text{body},i}$ and $\omega_{\text{body},i}$) through the unit quaternion $\mathbf{q}$ describing the orientation. The molecule is placed in a dilute Newtonian fluid with dynamic viscosity η and temperature T, which is to be modeled implicitly (i.e. no fluid particles) through stochastic exciting and systematic damping forces, thus leading to Langevin dynamics (LD) as shown in Sec. 1.2.2 eq. 1.6. Note that this assumes that the molecule has sufficiently small m and I_i for diffusion effects to be relevant regarding their average thermokinetic energy $\frac{\mathrm{k_B}T}{2}$ per DOF, where $\mathrm{k_B}$ is the Boltzmann constant and T the system's temperature [114].

The diffusive movement of the molecule resulting from the atomistic structure and its interaction with the solvent environment is described by three translatory diffusion coefficients $D_{\mathrm{t},i}$ and three rotatory diffusion coefficients $D_{\mathrm{r},i}$ in the principle component axes, which will be determined subsequently in Sec. 3.3. Note that this formulation assumes that the diffusion tensor collapses onto its diagonal components in the body frame of reference, which will be checked in Sec. 3.5 by comparison with the fully resolved movement in MD. The diffusion coefficients are determined at a reference viscosity η_{ref} and temperature T_{ref}, which can be different from the system viscosity η and temperature T as long as the molecular structure and its interaction with the solvent environment remain stable. In order to model the diffusive movement and enforce the desired canonical ensemble for these abstracted molecules in the MDEM framework (see Fig. 2.1 and eqs. 2.2ff) with explicit time integration in increments Δt, the diffusive forces and torques in each time step and for each DOF i (x, y, z for translation; α, β, γ for rotation) can be calculated as [223]

$$F_{\text{dif,body},i} = -c_{\text{dis,t},i}\, v_{\text{body},i} + F_{\text{fluct},i}\, \xi_{\mathrm{t},i}, \tag{3.1}$$

$$M_{\text{dif,body},i} = -c_{\text{dis,r},i}\, \omega_{\text{body},i} + M_{\text{fluct},i}\, \xi_{\mathrm{r},i}, \tag{3.2}$$

where $\xi_{\mathrm{t},i}$ and $\xi_{\mathrm{r},i}$ are six independently drawn random numbers with zero mean and unit variance, while c_{dis} and F_{fluct} / M_{fluct} are the dissipative drag and fluctuating force / torque coefficients calculated as

$$c_{\mathrm{dis,t},i} = \frac{m}{\Delta t}\left(1 - e^{-\frac{\eta}{\eta_{\mathrm{ref}}}\frac{\mathrm{k_B}T_{\mathrm{ref}}}{mD_{\mathrm{t},i}(T_{\mathrm{ref}},\eta_{\mathrm{ref}})}\Delta t}\right), \tag{3.3}$$

$$F_{\mathrm{fluct},i} = \frac{\sqrt{m\mathrm{k_B}T\left(1 - e^{-2\frac{\eta}{\eta_{\mathrm{ref}}}\frac{\mathrm{k_B}T_{\mathrm{ref}}}{mD_{\mathrm{t},i}(T_{\mathrm{ref}},\eta_{\mathrm{ref}})}\Delta t}\right)}}{\Delta t}, \tag{3.4}$$

$$c_{\mathrm{dis,r},i} = \frac{I_i}{\Delta t}\left(1 - e^{-\frac{\eta}{\eta_{\mathrm{ref}}}\frac{\mathrm{k_B}T_{\mathrm{ref}}}{I_i D_{\mathrm{r},i}(T_{\mathrm{ref}},\eta_{\mathrm{ref}})}\Delta t}\right), \tag{3.5}$$

$$M_{\mathrm{fluct},i} = \frac{\sqrt{I_i\mathrm{k_B}T\left(1 - e^{-2\frac{\eta}{\eta_{\mathrm{ref}}}\frac{\mathrm{k_B}T_{\mathrm{ref}}}{I_i D_{r,i}(T_{\mathrm{ref}},\eta_{\mathrm{ref}})}\Delta t}\right)}}{\Delta t}. \tag{3.6}$$

Additionally, the model can be used as a direct coupling to computational fluid dynamics by replacing the particle velocity with the differential to the fluid velocity. Thus, enabling for example the investigation of shear flow on molecular assemblies.

However, note that the proposed model neglects hydrodynamic interaction, i.e. forces acting through the solvent by relative movement of molecules in proximity, and instead only accounts for molecular interaction as a function of relative position and orientation in the overall MDEM framework (Ch. 2). This is done due to the significant complexity of solving the relative friction tensor for arbitrarily aniostropic molecules and in line with literature. Instead, a reduction of the effective viscosity is performed as an approximation, which is discussed in detail in App. A.3.

3.2.2 Background

The proposed model is based on Langevin dynamics and is specifically an extension of the "impulsive Langevin leap-frog algorithm for systems without constraints" by Goga *et al.* [281] implemented in *GROMACS* [278, 282, 283]. While the original model serves as a thermostat for atoms in molecular systems to enforce a canonical ensemble in an impulse-based formulation [281], it was extensively modified to reproduce the anisotropic diffusion kinetics of abstracted molecules in an implicit solvent. For this, the formulation was written in a force-based solution and friction coefficients were derived to reproduce the anisotropic diffusion of the abstracted molecules including rotational diffusion. Thus, an adequate modeling of the movement and thermodynamics of complex molecules on the

abstraction level necessary for supramolecular assemblies was achieved. Furthermore, the formulation enables modification of the system temperature and solvent viscosity without requiring a re-parameterization of the diffusion coefficients. Derivation details are provided in the supplementary of Depta *et al.* [223] and accuracy of the model can easily be verified by inserting the dissipative drag (c_{dis}) and fluctuating force / torque coefficients (F_{fluct} / M_{fluct}) into the fluctuation-dissipation theorem in eq. 1.8 as will be done subsequently in Sec. 3.4.1 for derivation of the critical time step.

3.2.3 Simplification for Isotropic Diffusion

While the proposed model is specifically designed to address complex molecules with anisotropic diffusion, it can also be used for spherical molecules or nanoparticles by using equal diffusion coefficients for x, y, z and α, β, γ, respectively. Furthermore, the diffusion coefficient is related to Stoke's radius through Einstein's relation [168] as

$$D_{\mathrm{t},i}\left(T_{\mathrm{ref}}, \eta_{\mathrm{ref}}\right) = \frac{\mathrm{k_B} T_{\mathrm{ref}}}{6\pi\eta_{\mathrm{ref}} r_{\mathrm{s,t},i}}, \tag{3.7}$$

$$D_{\mathrm{r},i}\left(T_{\mathrm{ref}}, \eta_{\mathrm{ref}}\right) = \frac{\mathrm{k_B} T_{\mathrm{ref}}}{8\pi\eta_{\mathrm{ref}} r_{\mathrm{s,r},i}^3}. \tag{3.8}$$

As a result, the radius of a nanoparticle can be used directly to parameterize the diffusion model. However, note that potentially a shell of hydration has to be accounted for. This model has for example been used to model the diffusive component of industrial zeolite production including a fluid flow coupling for describing the effects of shear flow [284].

3.3 Parameterization

3.3.1 Approaches

In order to employ the proposed diffusion model for macromolecules, knowledge of the diffusion coefficients is necessary. Generally speaking, these diffusion coefficients can be determined either through experimental investigation, analytical considerations, or numerical investigation. With regard to **experimental** investigation, the determined diffusion coefficients are typically isotropic due to limited resolution and differentiation on the required scales of length and time. Hence, limiting usefulness for complex molecules such as PDC. Regarding rotational diffusion, methods include for example nuclear magnetic resonance (NMR) [285], dynamic magnetic susceptibility measurements

[286], and fluorescence correlation spectroscopy (FCS) [123]. With regard to **analytical** considerations, theoretical works on the fundamentals of diffusion by Einstein [119], Debye [120], and Perrin [121, 122] enable the direct calculation of diffusion coefficients from radii using eq. 3.7. However, these are limited to simple objects such as spheres or ellipsoids. Consequently, while this can be used to estimate anisotropic diffusion coefficients - even for elongated molecules such as the E2 enzyme of PDC - more detailed insight is preferential. Lastly, **numerical** investigation through MD enables detailed insight including anisotropy by fully resolving the molecules and solvent environment. Consequently, this approach was chosen for parametrization and the setup will be described next. However, note that for the specialized model of the alginate system analytical estimation was employed in line with the overall approach of literature-based modeling for this model system.

3.3.2 Parameterization through Molecular Dynamics

The open-source software package *GROMACS* [41, 278] version 5.1.1 was used to perform all MD simulations. As a force-field, the coarse-grained (CG) *Martini* force-field version 2.2P [125, 287] was used with polarizable water (PW) [287] and the particle mesh Ewald (PME; fourth-order, 1.2 nm real space cutoff, 0.16 nm Fourier mesh spacing) method [54] for electrostatics to improve accuracy in comparison to the standard *Martini* water. The CG model was previously validated structurally for PDC [106, 268, 271] and is employed in this context as atomistic MD simulations are slower by 1-2 orders of magnitude and consequently infeasible for this purpose [223]. However, validation simulations with the atomistic OPLS-AA force-field were performed for the E1 enzyme of PDC yielding comparable results with regard to the isotropic diffusion coefficients (isotropic translation: 53.1 ± 3.63 µm^2 s^{-1} for CG vs. 46 ± 29.4 µm^2 s^{-1} for atomistic; isotropic rotation: 2.70 ± 0.18 Mrad^2 s^{-1} for CG vs. 2.73 ± 0.05 Mrad^2 s^{-1} for atomistic) [223]. Note that an anisotropic investigation was not possible due to limited statistics resulting from computational requirements.

Each molecule was placed in a cubic (E1 and E3) or triclinic box (all other) with periodic boundary conditions (PBC, x-y-z) and a minimum distance of 23.5 nm to its mirror image, which was determined through a convergence study (see supplementary data in Depta *et al.* [223]). All systems were charge-neutralized and for the HBcAg system an additional 150 mM of sodium chloride ions added, none for the PDC system. Temperature was maintained at 293 K for the HBcAg system and 300 K for the PDC system using the velocity-rescaling algorithm [288] (coupling time constant $\tau = 1$ ps).

In a first step, each system was solvated and energy minimization performed using the steepest descent algorithm [289] to a tolerance of 10'000 kJ/mol/nm for up to 100'000 steps. Next, equilibration was performed for 20 ps in an NVT ensemble using a time step of 5 fs with position restraints on the back-bone atoms using a force constant of 1000 kJ/mol nm^2. Subsequently, another equilibration was performed for 40 ps in an NPT ensemble using the Parrinello-Rahman barostat [290, 291] (coupling constant $p = 12$ ps, isothermal compressibility 3×10^{-4} bar^{-1}) ensuring stable density. Lastly, production MD was run on a NPT ensemble with PW, PME, and without any constraints using a time step of 20 fs and savings step of 1 ps for 2 ns, which is significantly above the minimum time for the correlation between mean-square-displacement (MSD) and diffusion coefficient to hold [119] (minimum for investigated molecules is 3.1 ps) and was validated through a convergence study.

GROMACS utilities were used to calculate positions and orientations over time by fitting the current structure to the reference structure, as well as analyze kinetic energies. The initial 10 ps of each simulation were discarded allowing for additional equilibration. *Matlab* version 2018a was subsequently used to calculate the MSD displacement using inverse time integration in the body frame of reference and compensate for possibly drift of the center of mass (COM) resulting, e.g., from flexibility of E2 and E3BP linker arms. A convergence study concerning the number of replicates was carried out by incrementally increasing the number of replicates by 50 until relative changes remained below 5 % for two steps, which was fulfilled for all molecules by 600 replicates. Diffusion coefficients were subsequently determined by least-square fitting the MSD displacements of each DOF in the body frame of reference according to [119, 168]

$$\langle r_i^2 \rangle = 2D_{\mathrm{t},i}t, \tag{3.9}$$

$$\langle \theta_i^2 \rangle = 2D_{\mathrm{r},i}t. \tag{3.10}$$

The determined diffusion coefficients will be provided in the results chapters of the respective model system, while the remainder of this chapter will discuss the various convergence aspects of the diffusion model. Credit for the MD setup and simulations is given to Uwe Jandt.

3.4 Convergence

In order to ensure numerical reliability of the diffusion model, a convergence study was performed analyzing the influence of the simulation time step on reproduction of the

desired diffusion coefficient (determined through MSD in eq. 3.10) and kinetic energy. For this, in a first step the critical time step was determined based on theory subsequently followed by the numerical investigation. As the numerical investigation is additionally dependent on the statistical sample size, a parallel study on the system size n (number of molecule copies) was carried out. All convergence evaluations were carried out on the molecules of the PDC system (see Ch. 8 for MD parameterization including diffusion constants), specifically the E2 enzyme[1], as it contains the most complex and anisotropic structure. All MDEM simulations were performed with 100 replicates (multiplied by system size n) in dilute state without any molecular interaction for 5 ns with a 1 ps saving step (if time step was larger, then time step was used for saving). Unless otherwise stated, a time step of $\Delta t = 10^{-13}$ s and system size of $n = 5 \times 10^5$ was used.

3.4.1 Critical Time Step

As outlined in Sec. 1.2.2, the diffusion model based on Langevin dynamics has to fulfill the fluctuation-dissipation theorem (eq. 1.8) to ensure a stationary process with time independent velocity correlations of a canonical ensemble. Thus, leading to [223]

$$e^{-\frac{\mathrm{k_B}T_{\mathrm{ref}}\eta}{mD_{\mathrm{t},i}(T_{\mathrm{ref}},\eta_{\mathrm{ref}})\eta_{\mathrm{ref}}}\Delta t} \simeq 1, \tag{3.11}$$

$$e^{-\frac{\mathrm{k_B}T_{\mathrm{ref}}\eta}{I_i D_{\mathrm{r},i}(T_{\mathrm{ref}},\eta_{\mathrm{ref}})\eta_{\mathrm{ref}}}\Delta t} \simeq 1, \tag{3.12}$$

which is only fulfilled for time steps Δt

$$\Delta t \ll \frac{mD_{\mathrm{t},i}\left(T_{\mathrm{ref}},\eta_{\mathrm{ref}}\right)\eta_{\mathrm{ref}}}{\mathrm{k_B}T_{\mathrm{ref}}\eta} \tag{3.13}$$

$$\Delta t \ll \frac{I_i D_{\mathrm{r},i}\left(T_{\mathrm{ref}},\eta_{\mathrm{ref}}\right)\eta_{\mathrm{ref}}}{\mathrm{k_B}T_{\mathrm{ref}}\eta}. \tag{3.14}$$

As a result, the critical time step τ_{crit} of the diffusion model can be defined as

$$\tau_{\mathrm{crit}} = \min\left(\frac{mD_{\mathrm{t},i}\left(T_{\mathrm{ref}},\eta_{\mathrm{ref}}\right)\eta_{\mathrm{ref}}}{\mathrm{k_B}T_{\mathrm{ref}}\eta}, \frac{I_i D_{\mathrm{r},i}\left(T_{\mathrm{ref}},\eta_{\mathrm{ref}}\right)\eta_{\mathrm{ref}}}{\mathrm{k_B}T_{\mathrm{ref}}\eta}\right) \tag{3.15}$$

with $i \in \{x, y, z, \alpha, \beta, \gamma\}$. Relative to this critical time step the error of discrete time integration is investigated subsequently.

[1] Note that for this work a slightly updated reference structure of E2 was used in contrast to ref. [223].

3.4.2 Thermal Equilibration Speed

First, the speed of thermal equilibration, i.e. how fast the system reaches its steady state temperature from a different previous temperature, was investigated. It was found that independent of previous temperature and temperature difference, the system equilibrates to its average kinetic energy with less than a 1 % deviation in $2-3\times\tau_{\mathrm{crit}}$. This result is in agreement with theory as the critical time constant is also the thermal rate constant [281]. Consequently, no equilibration procedure is needed before MDEM simulations.

3.4.3 Diffusion Coefficient

In order to evaluate convergence of the diffusion coefficient, firstly the system size was varied between $n = 10^3$ and 10^7 for the model molecule E2 to determine the statistic limits resulting from finite sample size at a constant time step of $\Delta t = 10^{-13}$ s, which is considerably below the critical time steps of E2 shown in Tab. 3.1. As it can be seen in Fig. 3.1A, the root-mean-square (RMS) error decreases linearly in the double logarithmic plot with system size. As expected, DOFs with large critical time constants (specifically x, β, γ) follow a linear trend throughout the investigated system sizes, while DOFs with small critical time constants (specifically α and partially y) exhibit an asymptotic error trend towards large system sizes. This is attributed to increasing errors from time integration due to the small critical time step relative to the simulation time step of 10^{-13} s. In order to further study the convergence of the simulation time step Δt, based on these results a system size of $n = 5\times10^5$ was chosen as its error limit of 0.2 % through sample size is below the expected accuracy from parameterization of the diffusion coefficient.

Table 3.1: Critical time constants for each DOF of the model enzyme E2. Reprinted with permission from Depta *et al.* [223]. Copyright 2019 American Chemical Society. Note that in contrast to Depta *et al.* [223] an updated reference structure of E2 was subsequently used in this work, specifically Ch. 8.

Direction	D_{t} [$\mu\mathrm{m}^2\,\mathrm{s}^{-1}$]	τ_{crit}[ps]	D_{r} [$\mathrm{Mrad}^2\,\mathrm{s}^{-1}$]	τ_{crit}[ps]
x/α	48.0	1.136	8.74	0.588
y/β	37.7	0.894	0.33	1.250
z/γ	41.2	0.976	0.36	1.371

Subsequently, the time step Δt was varied between 2×10^{-14} s and 5×10^{-12} s for all molecules of PDC as shown in Fig. 3.1B. As it can be seen, the RMS error decreases linearly in the double logarithmic plot with decreasing time step until reaching the asymptotic limit of approximately 0.2 % imposed by the finite system size. Independent of model molecule and DOF, all data points follow the same trend when normalized by

their respective critical time constant underlining the importance of τ_{crit}. Consequently, the relationship between RMS and Δt can be used to estimate the error of the apparent diffusion coefficient *a priori*, thus, giving guidance for choosing the time step. For this, the data between an RMS of 0.3 % and 100 % was used to find the correlation

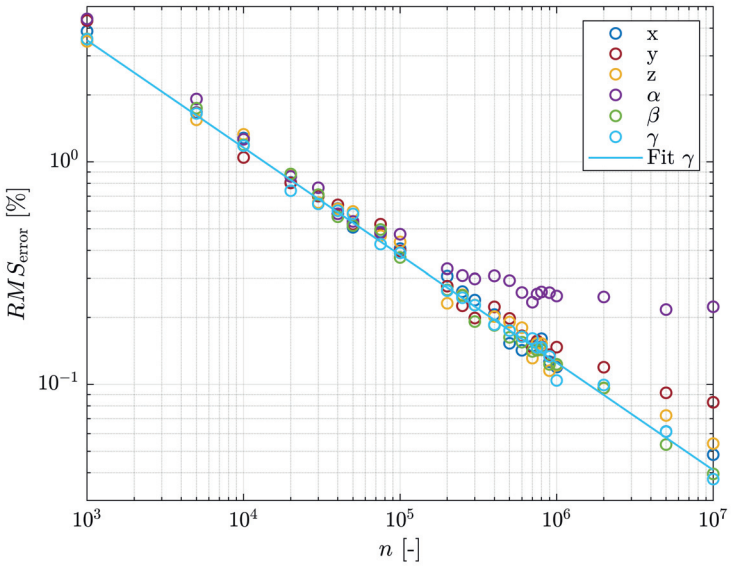
$$\Delta t = \tau_{\text{crit}} \cdot 10^{\left(\frac{\log(RMS_{\text{error}}[\%]/1\%)-0.90}{1.97}\right)}. \tag{3.16}$$

by least square fitting. Note that validity of the correlation is limited to an RMS above 0.3 %. However, the correlation likely remains accurate beyond this and additionally accuracy of diffusion parameters is typically less.

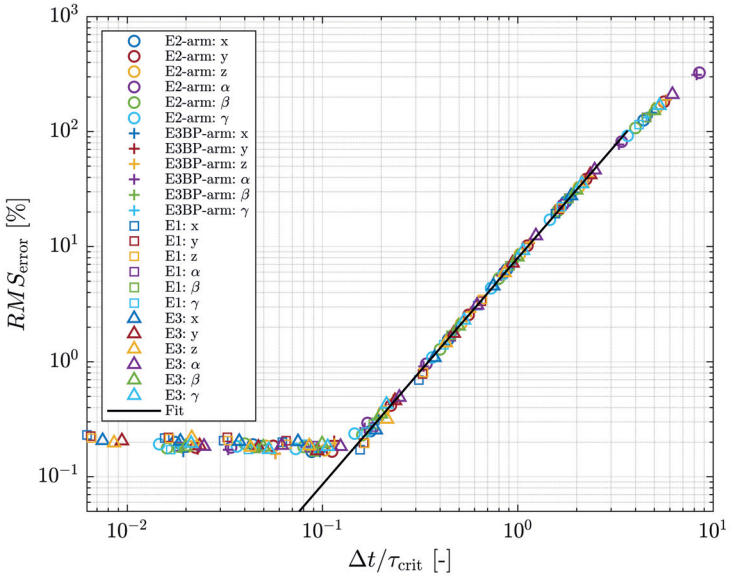
3.4.4 Kinetic Energy

Lastly, convergence of the apparent kinetic energy was investigated as shown in Fig. 3.2. For this, statistical analysis was performed over all molecule copies and times for each DOF in its respective body frame of reference and compared to the expected average thermokinetic energy of $\frac{k_B T}{2}$ per DOF [114]. As it can be seen, trends are qualitatively similar to those for the convergence of the diffusion coefficient, but with errors one to two order of magnitude lower. Consequently, errors of the apparent diffusion coefficient dominate and thus the previous discussion holds overall.

Nonetheless, as a difference it should be noted that DOF with higher critical time constants lead to slightly higher errors visible at sufficiently large sample sizes provided by large n. This opposite trend compared to the apparent diffusion coefficient is attributed to the critical time constant also being the thermal rate constant [281] (see Sec. 3.4.2). Thus, leading to longer restoration times to thermal equilibrium and thereby larger errors. However, as the error in kinetic energy always remains significantly below the error in the apparent diffusion coefficient, these differences are not relevant for the overall MDEM framework.

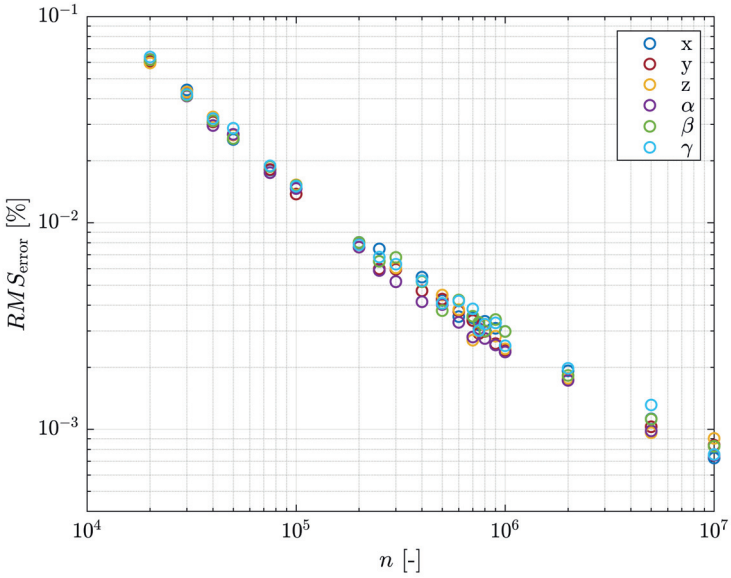


(A) Diffusion coefficient: System size for E2.

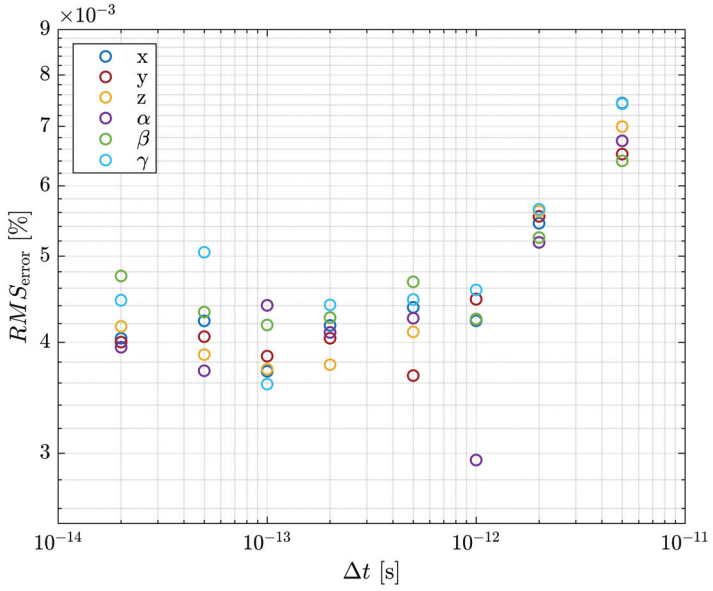


(B) Diffusion coefficient: Normalized time step for all PDC enzymes.

FIGURE 3.1: Convergence of diffusion coefficient for varying system size (A) and normalized time step (B) for individual DOF. The root-mean-squared (RMS) error relative to the specified diffusion coefficient is displayed. When varying the system size a time step of 10^{-13} s was specified and when varying the time step a system size of 5×10^5 was specified. In (A) a fit is provided for the DOF γ and in (B) a fit is provided for errors between 0.3 % and 100 %. Reprinted with permission from Depta *et al.* [223]. Copyright 2019 American Chemical Society.



(A) Energy: System size for E2.

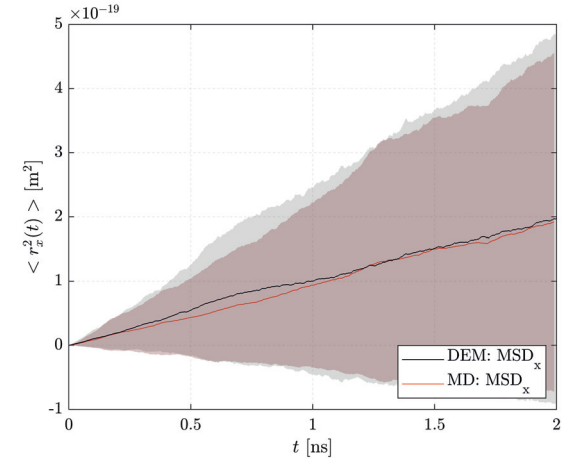


(B) Energy: Time step for E2.

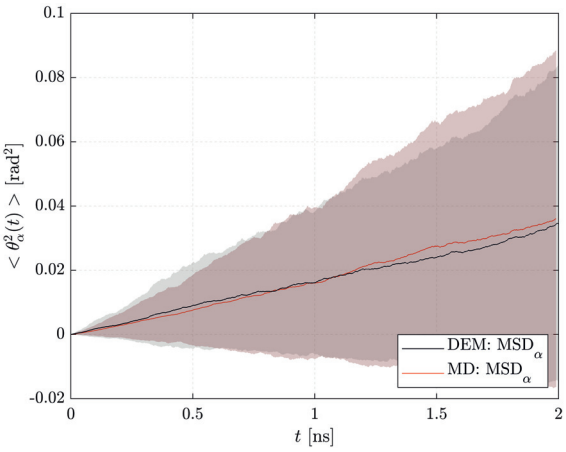
FIGURE 3.2: Convergence of kinetic energy for varying system size (A) and time step (B) for individual DOF. The root-mean-squared (RMS) error relative to the desired energy per DOF ($\frac{k_B T}{2}$) is displayed. When varying the system size a time step of 10^{-13} s was specified and when varying the time step a system size of 5×10^5 was specified. Reprinted with permission from Depta *et al.* [223]. Copyright 2019 American Chemical Society.

3.5 Comparison with Molecular Dynamics Data

Furthermore, the anisotropic diffusive movement and thermokinetic energy were compared between the proposed model and the MD model (Sec. 3.3.2) for all molecules of PDC. In order to ensure comparable stastistics, 600 replicates over 2 ns were used for the MDEM model in agreement with the number of MD replicates. Based on the



(A) Translation - X

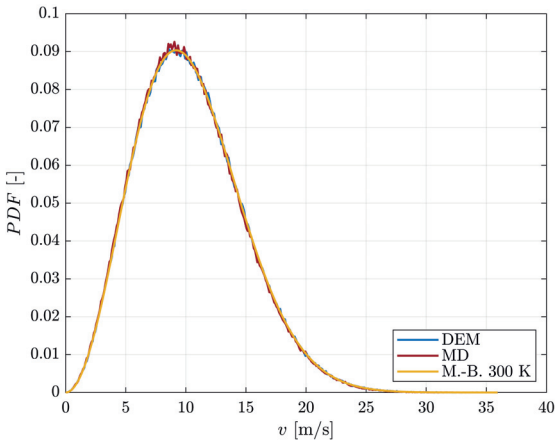


(B) Rotation - α

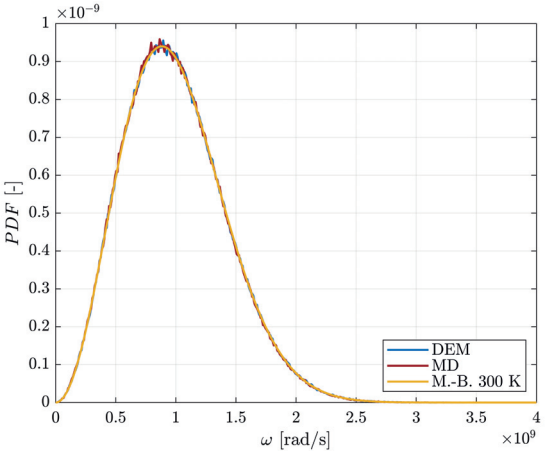
FIGURE 3.3: Comparison of MD data and DEM diffusion model data using the mean-squared-displacement (MSD) plots for x and α DOF of the model enzyme E2. Solid lines indicate mean values (MSD) and shaded regions the standard deviation. To ensure a good comparability, both MD and DEM statistic consisted of 600 enzyme repetitions. Reprinted with permission from Depta *et al.* [223]. Copyright 2019 American Chemical Society.

convergence study in Sec. 3.4, a time step of 10^{-13} s was used for the MDEM model leading to an RMS error below 0.3 % for the apparent diffusion coefficient.

Concerning anisotropic diffusive movement, the mean-square displacement (MSD) is shown exemplary for the x and α DOF of the E2 enzyme in Fig. 3.3. As it can be seen, for translation and rotation both the average (solid lines) and distribution (transparent regions) of the MSD are in good agreement with the MD model, thereby accurately



(A) Translation



(B) Rotation

FIGURE 3.4: Comparison of the velocity probability-density-functions (PDF) for the DEM diffusion model, MD results, and a Maxwell-Boltzmann distribution at 300 K. The enzyme E2 is used as a model enzyme. For the DEM diffusion model the rotational velocity was determined by weighting body frame velocity components with their respective moment of inertia and for MD by using the rotational energy and the average moment of inertia. Reprinted with permission from Depta *et al.* [223]. Copyright 2019 American Chemical Society.

reproducing the diffusive movement even for structurally complex molecules such as E2. Similarly, all other PDC components and their respective DOF are accurately reproduced as shown in App. B Fig. B.1. Likewise, the velocity distribution agrees well for both translation and rotation with the MD model and theoretical predictions of a Maxwell-Boltzmann velocity distribution as shown in Fig. 3.4 for E2. The proposed diffusion model therefore accurately reproduces the diffusion behavior of anisotropic molecules for the MDEM framework.

3.6 Enhanced Sampling of the Conformation Space through Simulated Annealing

Molecular simulations are typically performed at constant temperature employing one of the thermostats outlined in Sec. 1.2 similarly to the proposed diffusion model in this section. Starting from an initial condition of the molecular system, also called conformation, simulations are performed over a finite period of time constrained by computational resources with the goal of reaching equilibrium. However, many times various potential minima exist separated by high potential barriers from the initial conformation [30]. Thus, sampling of the conformation space is restricted to a small subset and naturally occurring conformations may be missing.

In order to remedy this limitation, various methods have been proposed to enhance sampling of the conformation space. One example is *simulated annealing* [75], which temporarily increases and then again lowers temperature according to a protocol to enhance crossing of potential barriers and takes its name from the related metallurgy process. Another example is the *replica exchange method* (REM), which simulates replicas of the system at various temperature and performs probabilistic temperature exchanges between replicas. Further examples include *expanded ensembles* [77] and *simulated tempering* [292].

In order to be able to address these aspects of enhancing conformation sampling, **simulated annealing** has been implemented for the proposed diffusion model. For this, the temperature of the simulation domain can be varied over the course of the simulation by updating the fluctuating force coefficient $F_{\text{fluct},i}$ and fluctuating torque coefficient $M_{\text{fluct},i}$, which is implemented through a scaling factor of $\sqrt{T_{\text{cur}}(t)/T_{\text{eq}}}$, where T_{cur} is the desired current temperature at time t and T_{eq} is the original equilibrium temperature. The viscosity of the fluid was kept constant. As a temperature protocol, periodic step increases (time period $\tau_{\text{an,period}}$, repeated until final time $t_{\text{an,finished}}$) to the maximum temperature $T_{\text{an,max}}$ followed by a linear decay over a time of $\tau_{\text{an,cool}}$ back to equilibrium were implemented subsequent to testing.

4

Intermolecular Interaction

4.1 Introduction

During the assembly of supramolecular structures the intermolecular interaction is of critical importance. As shown in Ch. 1, this is commonly defined by non-covalent interaction between macromolecules. In traditional all-atom and coarse-grained MD (see Sec. 1.2), this is modeled by the underlying interaction between individual atoms defined through force-fields. However, while providing great detail, an investigation on the scales in length and time necessary for larger molecular assemblies is not possible with current computational resources. In order to address these scales, this work abstracts each entire macromolecule as an object with a position and orientation resembling an ultra-coarse-grained approach (see Ch. 2). As a result, the intermolecular interaction model is (up to) six dimensional (6D) capturing the relative position (3D) and orientation (3D). Furthermore, this interaction has to be uniquely formulated for each pairwise interaction of two macromolecules, which cannot directly be transferred to other molecules including for significant changes in secondary and tertiary structure. As a result, it becomes a trade-off whether the work deriving such a unique force-field between two macromolecules is reasonable in comparison to the speed-up the model provides in understanding the system of interest. In the method development focus of this work, two approaches are explored for deriving such intermolecular interaction models: one specialized model based on literature for alginate gelation, as well as a generic data-driven approach to derive interaction potentials based on MD.

The task of formulating such intermolecular interaction models can be embedded in the broader field of *surrogate modeling*, which has received increasing attention in recent decades with the rise of machine learning (ML) methods. The goal of surrogate modeling

is to derive an effective model with fast evaluation times for a complex process by mapping an arbitrarily dimensional input space to an arbitrarily dimensional output space. Specific to the task of interaction models this means a description of the potentials/ forces/torques (output) for relative positions and orientations (input). In the following, a few of the most commonly applied surrogate modeling techniques will be highlighted. Traditionally the mapping of input-output relationships for surrogate modeling has been performed using fitting of *functional descriptions* such as polynomials. Optimization of parameters is then performed using e.g. least squares or *support vector regression (SVR)* [293]. Traditional MD force-fields widely use such functional 1D descriptions. However, this approach becomes significantly difficult for highly dimensional and complex relationships. In this regard, one of the most popular approaches without a function and parametric description is *Kriging* [294–297], also called *Wiener–Kolmogorov prediction* or *Gaussian process regression (GPR)* in ML context [298, 299], as well as being further embedded in *Bayesian statistics* as a form of regression/inference. Kriging provides the best linear unbiased estimator (BLUE) [294] for the value of a variable in an arbitrarily dimensional space based a set of observations (data) and their correlation, i.e. a statistical description of the data. Not only the estimate, but also an error estimate is achieved, which is especially useful in understanding and improving the non-parametric surrogate model. Kriging is used in this work for the data-driven approach to estimate the interaction potential with further details provided in Sec. 4.3.4. In a related direction, *radial basis functions (RBF)* [300, 301] provide a description based on the linearly weighted combination of the distance to specific center points in space to formulate the surrogate model. Furthermore, artificial neural networks (ANN) [298, 302] provide more general and flexible descriptors by employing complex networks of neurons between the input and output space. ANNs have gained significant interest in ML and are applied to a variety of problems such as natural language processing [303, 304]. For further examples and reviews on surrogate modeling see e.g. refs. [305–308].

With regard to the parameterization of such surrogate models for the desired intermolecular interaction, similar approaches to those discussed in Sec. 1.2.2 for coarse-grained MD can be employed. As shown, these approaches are often classified as bottom-up, top-down, and hybrid - thus incorporating a variety of information as purely first-principle based methods are frequently not sufficient [30]. Related bottom-up methodologies include e.g. thermodynamic integration [30], free energy perturbation [137], umbrella sampling [134], and steered molecular dynamics. In addition to these traditional methods, Gaussian progress regression (GPR) has previously employed for example by Bartók *et al.* [309] to derive potentials based on quantum models (for broader review see ref. [310]). Following works have employed GPR in combination with umbrella sampling, constraints, or instantaneous collective forces (ICF) to derive potential energies up to

4D for very small molecules such as alanine tripeptide [311, 312] and methanol/benzene [129]. In similar regards and with similar limitations, Poisson equation formalism's have been applied [313]. Additionally, ANN have been employed increasingly on similar scales of coarse-graining [130, 131, 314] in addition to the atomistic [56, 315]. However, while providing detailed insight, these methods are largely restricted to few degrees of freedom by computational requirements - thus they are not directly applicable to the 6D interaction space between molecules.

The subsequently presented models attempt to remedy some of these limitations. The first model presented in Sec. 4.2 is specialized for alginate interaction to implicitly capture ions through a probabilistic binding and unbinding model, therefore alleviating the need to explicitly model solvent or ions. It is fully based on theoretical considerations and literature knowledge, thus employing a top-down knowledge-based parameterization. The second model presented in Sec. 4.3 is a generic formulation for the full 6D case and employs a bottom-up focus, while also permitting incorporation of empirical data. It employs a Kriging-based approach with a similar methodology as free energy perturbation, while sacrificing some thermodynamic accuracy to gain practicality in computation for the 6D interaction space. The determined interaction potential is then captured in a homogeneously gridded 6D field on which the gradient operation to derive forces and torques is carried out for each contact between macromolecules during the higher-level MDEM simulation. This approach has the benefit of being flexible with regard to the shape of the potential and maintains a constant runtime independent of its complexity, which is a significant advantage over e.g. ANNs or functional descriptions. However, due to the memory requirements to store the field, the number of different molecules in the system is limited to a few. This trade-off was found to be acceptable for the investigated systems and is expected to become less relevant with increasing computer memory over time.

4.2 Probabilistic Interaction Model for Calcium Mediated Alginate Gelation Based on Literature and Theory

This chapter is based on the following publication:

P. N. Depta, P. Gurikov, B. Schroeter, A. Forgács, J. Kalmár, G. Paul, L. Marchese, S. Heinrich, and M. Dosta. DEM-Based Approach for the Modeling of Gelation and Its Application to Alginate. *J. Chem. Inf. Model.*, 62(1):49–70, 2022

During the gelation of alginate, the presence of ions is crucial to mediate the binding process of polymer chains and resulting network formation. This network of polymer

fibers is the structural foundation of the macroscopic gel. In the context of this work, calcium is used as the mediating ion as is visualized in the introductory Fig. 1.1 and 1.2. In the following, a probabilistic interaction model for the calcium mediated pairwise interaction of alginate polymer fiber (dimer) units will be proposed. The model will take a functional form and will be derived and parameterized based purely on literature, theoretic considerations, and approximations. As previously stated, the smallest units of description are the dimer units of alginate, i.e. GG, MM, and GM/MG. In the following, the basic functional description followed by the probabilistic ion binding and unbinding model will be presented.

4.2.1 Interaction Model

A subset of alginate components, namely the GG - GG dimers and to some extend MM - MM, possess an attractive interaction during gelation. This interaction of two GG dimers with one calcium ion bound in the center is characteristic for the alginate system and has originally been described by Grant *et al.* as the 'egg-box' model [235]. At the same time, other polymer chain components, namely ones with alternating G and M units (GM), are known to not form connections and consequently posses only a repulsive interaction. As a result, for the functional description of the forces between dimers units, a basic function capable of capturing these trends is necessary: repulsive behavior at small distances and a defined potential minimum / well, i.e. attractive behavior, at specific distances for a subset of interactions.

In order to capture this, the Lennard-Jones (LJ) potential was chosen as an effective potential to describe the potential shape of unbonded pairwise interaction. The LJ potential is widely used in the field of molecular modeling to describe atomic interaction and is especially useful in this context due to its defined potential well inducing an attractive interaction at large distances, which effectively models the 'egg-box' interaction, as well as the possibility to be repulsive-only for other interaction pairs. While other potential shapes are certainly conceivable, the LJ potential is a reasonable first estimation. The basic LJ potential and forces can be calculated as

$$U_{\mathrm{LJ,base}} = \epsilon((\frac{d_m}{d})^{12} - 2(\frac{d_m}{d})^{6}) + U_{\mathrm{cor}}, \tag{4.1}$$

$$F_{\mathrm{pp}} = -\nabla U_{\mathrm{LJ,base}} = 12\frac{\epsilon}{d_m}((\frac{d_m}{d})^{13} - (\frac{d_m}{d})^{7}), \tag{4.2}$$

where d_m is the location of the potential well (minimum), ϵ is the potential scaling factor, and U_{cor} is the correction potential to ensure zero potential at the employed cutoff d_{cut}. Additionally, two corrections have been implemented: First, a maximum force F_{max} was implemented to avoid numerical instabilities resulting from the singularity of the

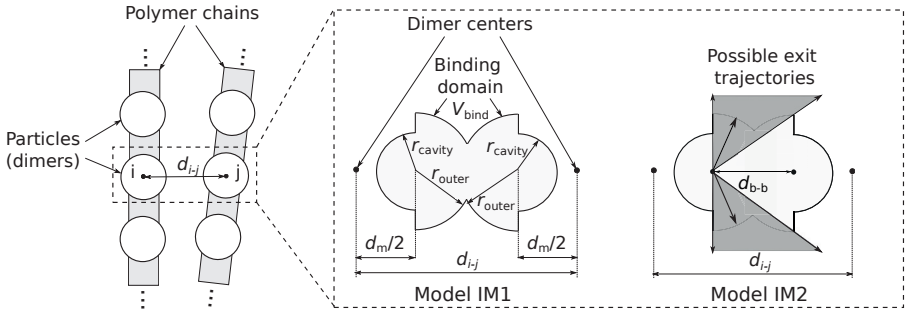


FIGURE 4.1: Visualization of ion models IM1 and IM2 for alginate gelation. Adapted with permission from Depta *et al.* [224] under CC-BY 4.0 license.

LJ potential at zero. Consequently, once the force exceeded this threshold and remains constant, the potential increases linearly. Second, in cases when no calcium is present to mediate the interaction or the interaction partners only possess a repulsive interaction, the potential was set to zero above d_m and shifted accordingly. The forces were adapted accordingly and the remaining interaction is repulsive only for such cases.

4.2.2 Ion Model

As previously noted, the availability of ions is critical in the gelation of alginate. Depending on the presence of a (calcium) ion, a pairwise interaction between two GG dimers can be either repulsive or attractive - mediated by the ion. Consequently, to describe gelation, modeling the presence, binding, and unbinding of ions is of crucial importance. The limit case of an unlimited supply of ions at any location in the polymer solution consequently represents an idealized case: It leads to very homogeneous and densely connected gels, which overestimate the experimental reality. Preliminary simulations with such a model (termed IM0 for ion model 0) were carried out and exhibited the described behavior. In order to capture the physics of ion mediated gelation more realistically, two probabilistic models were derived based on theoretical considerations and the geometric structure of the binding zone based on the 'egg-box' model [235].

IM1 The first ion model (IM1) describes the availability and consequently **binding probability of ions as a random process** dependent on the calcium concentration in the system. The model is visualized in Fig. 4.1. A limited number of ions is available in the simulation domain and decreases with every bound ion at a pairwise interaction. The probability of an ion being at the binding site of an interaction (capable of accepting

an ion) in each time step is modeled as

$$p_{\text{bind}} = \frac{V_{\text{bind}}}{V_{\text{tot}}} \cdot N_{\text{avail,ions}}, \tag{4.3}$$

where V_{bind} is the volume of the binding site, V_{tot} is the volume of the system domain, and $N_{\text{avail,ions}}$ is the number of currently available ions in the system. Once an ion binds to a pairwise interaction, the ion is bound until the connection breaks, i.e. their cutoff distance is exceeded. Consequently, unbinding is neglected, which will be considered in the next model IM2.

The main parameter defining the binding probability besides ion concentration thus becomes the volume of the binding site V_{bind}. Generally speaking, V_{bind} is a function of the conformation of interacting molecules, their distance, resulting internal charge distribution, affinity of the (calcium) ion to the interaction (i.e. binding energy), as well as thermodynamic effects (i.e. diffusion, leading to a time step dependency during simulation). To understand this in its entirety, DFT simulations of various configurations would be necessary similar to ref. [240], which goes beyond the scope of this work. In order to estimate V_{bind} for the purpose of an effective and simplified model, the following approach was chosen.

As it can be seen in Fig. 4.1, each molecule of a pairwise interaction possesses its own binding volume. The binding volume of each molecule is centered at half the equilibrium distance $d_m/2$, i.e. location of the ion in the 'egg-box' cavity and location of potential well in the pairwise interaction. The binding volumes can overlap when being in proximity, but are independent from each other when sufficiently spatially separated. The shape is approximated as spherical with contributions from an electrostatic r_{ion} and a diffusive/thermodynamic component r_{dif}. The part of the binding volume away from the molecule is unconstrained by the conformation and described by radius r_{outer}, which is the sum of r_{ion} and r_{dif}. In contrast, the part of the binding volume towards the molecule is constrained by the molecule's conformation and described by radius r_{cavity} for which r_{ion} is a sufficient approximation as there is no diffusive component. Furthermore, it is modeled that the binding volume can be closed off by the interacting molecules to account e.g. for the enclosure of the 'egg-box'. For this, the binding volume is modeled as closed off / inaccessible from the outside if the distance between binding volume $(d_{i-j} - d_m)$ centers does not exceed a minimum gap $d_{\text{gap,min}}$.

As previously noted, the binding model IM1 assumes that an ion will remain bound until the interaction breaks, i.e. cutoff is exceeded. As a result, the overall binding probability is a function of the cutoff distance.

IM2 In order to address this, the second ion model (IM2) includes the **unbinding probability of calcium ions** in addition to the previous binding probability, leading to a dynamic equilibrium of binding and unbinding. The model is visualized in Fig. 4.1. Inclusion of unbinding comes at the cost of significantly increased computational requirements primarily resulting from the decreased gelation kinetics. Generally speaking, the unbinding probability depends on the binding energy, temperature, solvent effects, and conformational constraints. These aspects will be discussed and modeled next.

As noted earlier, calculation of the exact binding energy $U_{\text{ion,bind}}$ of an ion at a molecule requires detailed methods such as DFT and is beyond the scope of this work. Thus, to estimate $U_{\text{ion,bind}}$ efficiently, it is assumed to be half the interaction potential ϵ. This is reasonable as the ion is primarily bound to one molecule and effects of the interaction partner are captured in the context of geometric constraints. Using the Maxwell-Boltzmann distribution of the kinetic energy of the ion at temperature T, the probability of the kinetic energy exceeding the binding energy can then be estimated. This probability is further lowered by geometric constraints to determine the unbinding probability. Specifically, the ion can only unbind in a direction away from both interacting molecules, as it is shown in Fig. 4.1. In accordance with the previously described binding volumes, the exit ratio over all directions in 3D space associated with unbinding can be calculated as

$$c_s = \frac{1}{2 \cdot \sqrt{(\frac{r_{\text{outer}}}{d_{b-b}})^2 + 1}}, \tag{4.4}$$

where $d_{b-b} = r - r_m$ is the distance between centers of ion volumes and r_{outer} is the previously defined outer radius of the binding volume. c_s is always less than 0.5 and the overall probability of unbinding can be calculated from the Maxwell-Boltzmann distribution as

$$p_{\text{unbind}} = c_s \cdot \left(1 - \text{erf}\left(\frac{v}{\sqrt{2}a}\right) - \sqrt{\frac{2}{\pi}} \cdot \frac{v}{a} \cdot e^{\frac{-v^2}{2a^2}}\right), \tag{4.5}$$

$$v = \sqrt{\frac{2U_{\text{ion,bind}}}{m_{\text{ion}}}}, \tag{4.6}$$

$$a = \sqrt{\frac{\text{k}_\text{B}T}{m_{\text{ion}}}}, \tag{4.7}$$

where v is the velocity corresponding to $U_{\text{ion,bind}}$, erf is the error function, k_B is the Boltzmann constant, and m_{ion} is the ions mass. Similar to IM1, below the minimum gap $d_{\text{gap,min}}$ no unbinding is assumed possible, i.e. $p_{\text{unbind}} = 0$. Note that for the used time step of 10^{-13}s there are no inertia effects of the calcium ion and consequently p_{unbind} is uncorrelated in time.

Generally speaking, with an increasing overlap of binding regions, the binding energy $U_{\text{ion,bind}}$ increases and unbinding probability p_{unbind} should decrease. At the same time, the exit ratio c_s increases with decreasing overlap, leading to opposite effects. Some of these inaccuracies are at the same time mitigated by the minimum gap $d_{\text{gap,min}}$. Overall, as exact modeling is non-trivial, the proposed simplified model was chosen.

All model parameters including binding volumes and probabilities as a function of distance will be presented in Sec. 6.1. Note that the presented models can be extended easily in the future to model ion heterogeneities by replacing the global concentration of ions by a local concentration derived from a coupled CFD simulation of ion diffusion.

4.2.3 Critical Time Step

In order to estimate the critical time step $\tau_{\text{crit,pp,gel}}$ of the intermolecular interaction model for the gelation of alginate, the oscillation period $T_{0,\text{pp,gel}}$ of the corresponding two-mass spring system of particles i and j can be used. The derivation based on Lagrangian mechanics can be found in most mechanics textbooks and is omitted here as it presents a simplified case of the one shown in the following section. The critical time step for the interaction model can be estimated as

$$\tau_{\text{crit,pp,gel}} = 2\pi \cdot \min\left(\sqrt{\frac{\mu_{m,i-j}}{k_{\text{pp,max},i-j}}}\right) \tag{4.8}$$

over all particle interactions $i - j$, where $k_{\text{pp,max},i-j}$ is the maximum stiffness of the interaction pair at any difference and $\mu_{m,i-j}$ represents the reduced mass defined as

$$\mu_{m,i-j} = \frac{m_i m_j}{m_i + m_j}. \tag{4.9}$$

Note that this is an approximation of the oscillation period. Due to the fact that the used time step is typically lower by a factor of five [42], this approximation is sufficient.

4.3 Data-Driven Interaction Potential Fields Based on MD

This chapter is based on the following publications:

P. N. Depta, M. Dosta, W. Wenzel, M. Kozlowska, and S. Heinrich. Hierarchical Coarse-Grained Strategy for Macromolecular Self-Assembly: Application to Hepatitis B Virus-Like Particles. *Int. J. Mol. Sci.*, 23(23):14699, 2022

P. N. Depta, M. Dosta, and S. Heinrich. Data-Driven Multiscale Modeling of Self-Assembly and Hierarchical Structural Formation in Biological Macro-Molecular Systems. In W. E. Nagel, D. H. Kröner, and M. M. Resch (editors), *High Performance Computing in Science and Engineering '21*. Springer International Publishing, Cham, 2023

As previously outlined, modeling intermolecular interaction is crucial for understanding the behavior and structural assembly of macromolecules. This section presents a data-driven approach for determining interaction potential fields based on MD, which can subsequently be used to describe interactions of the abstracted macromolecules in MDEM. Conceptually speaking, this approach transfers the detailed information necessary to describe intermolecular interaction from many 1D atom interactions into a single gradient operation on a 6D potential field (see Fig. 4.2). As a result, high levels of detail are maintained in the complex potential field, while computational requirements are drastically reduced. Thus, an investigation of larger system sizes and times is possible.

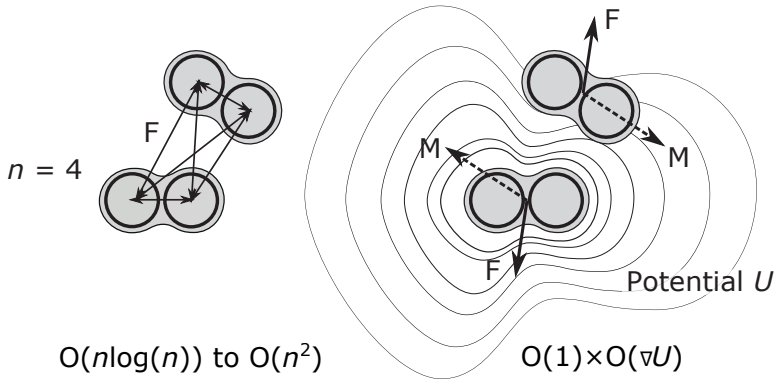


FIGURE 4.2: Comparison between atomistic representation of intermolecular interaction (left) and abstraction as anisotropic beads with interaction potential (right) on computational complexity (n is number of atoms, neglecting solvent and ions). Note that for example a single interaction of $HBcAg_2$ is equivalent to $n = 9432$ and further increased by the solvent atoms ($n \approx 10^5$). Reprinted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

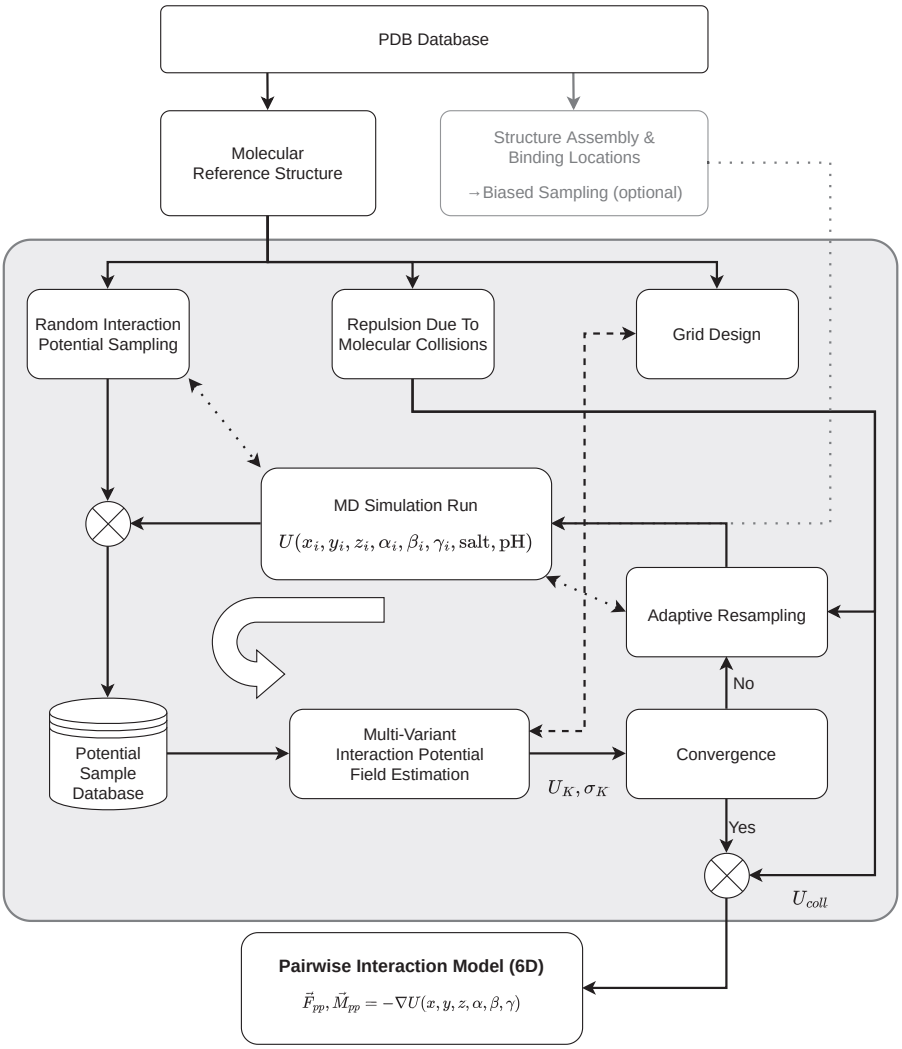


FIGURE 4.3: Visualization of the method for interaction potential determination based on molecular dynamics. Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

In order to derive the necessary intermolecular potential fields, a generic approach was formulated based on MD, which is visualized in Fig. 4.3 and will be explained in detail in the following sections. At the basis of the method lies the molecular reference structure of the macromolecules, which is used within the MD model to study and parameterize interaction including effects of solvent molecules and ions (i.e. salts). The molecular reference structures of many macromolecules are typically available from the PDB based on a variety of previous works in literature, which is also what the reference structures in this work are largely based upon. The derived interaction potential fields from the method can then be used in the MDEM structural formation simulations by employing a numerical gradient operation.

Note that in the following section the **conformation** refers to the 3D structure of a molecule, i.e. in the case of a protein its folded structure, while the **configuration** refers to the relative position and orientation of two molecules interacting with each other.

4.3.1 Molecular Dynamics Setup and Potential Groups

The open-source software package *GROMACS* [41, 278] version 2020.1 was used to perform all MD simulations and the respective protocols were slightly adjusted from the previous diffusion parameterization (Sec. 3.3.2) as follows. As a force-field, the coarse-grained *Martini* force-field version 2.2P [125, 287] was used with polarizable water (PW) [287] and the particle mesh Ewald (PME) method [54] for electrostatics to improve accuracy in comparison to the standard *Martini* water. The coarse-grained model was previously validated structurally for PDC [106, 268, 271] and is employed in this context as atomistic MD simulations are slower by 1-2 orders of magnitude and consequently infeasible for this purpose [223]. Full credit for this MD setup and portion of this work is given to Uwe Jandt and will be summarized in the following. Out of this, the MD setup for E3BP – E3 was provided by Cornelius Jacobi. In the context of the remaining work, the MD model is considered a *black-box* model for determining interaction potential contributions at a relative position and orientation of two macromolecules, which is then used to estimate an overall potential by spatial structure and overall relation. The remaining methodology for estimating the overall interaction potential based on these samples using Universal Kriging and employing it for structural formation simulations is decoupled from the underlying MD model, which is interchangeable depending on improvements on the MD side.

The 'new' parameter set for the *Martini* force-field was used as a basis for the MD setup and PW and PME employed unless otherwise stated. Each simulation consisted of two

molecules A and B (depending on types investigated) at a specified relative position and orientation, which could flexibly change during the simulation. The molecules were centered in a triclinic box with a minimum of 5.5 nm to all periodic boundary conditions (PBC, x-y-z). A larger distance of 8 nm to all PBC was tested in a convergence study for $HBcAg_2$ and found largely similar potential trends. All systems were charge-neutralized and for the HBcAg system an additional 150 mM of sodium chloride ions added, none for the PDC system. Temperature was maintained at 293 K for the HBcAg system and 300 K for the PDC system using the velocity-rescaling algorithm [288].

In a first step, each system was solvated using normal *Martini* water (no PW and no PME) and energy minimization performed using the steepest descent algorithm [289] to a tolerance of 10'000 kJ/mol/nm for up to 100'000 steps. Afterwards, the solvent was replaced by polarizable water (and PME enabled) and a second energy minimization performed using the steepest descent algorithm to a tolerance of 10'000 kJ/mol/nm for up to 50'000 steps. Next, equilibration was performed for 50 ps using a time step of 5 fs with position restraints on the back-bone atoms using a force constant of 1000 kJ/mol nm^2. The Berendsen barostat [72] with a compressibility of 3×10^{-4} bar^{-1} was used to avoid oscillations with position restraints and a coupling constant of 4 ps used. Lastly, production MD was run on a NPT ensemble with PW, PME, and without any constraints for 0.6 ns using a time step of 20 fs. As no constraints were used, the Parrinello-Rahman barostat [290, 291] with the previous compressibility and coupling constant of 12 ps was employed. Energies between all groups (A, B, PW, ions) were calculated every 20 steps and saved along with trajectories every 500 steps.

GROMACS utilities were used to perform postprocessing of each MD simulation. For each saving step, relative positions and orientations were calculated by fitting the current structure to the reference structure. The initial 0.5 ns of each production run were discarded to allow for further equilibration to take place and (small) conformational changes to occur, e.g. during binding events. Between 0.5 and 0.6 ns all energies, positions, and orientations were then averaged. Potential components were grouped and Lennard-Jones and Coulomb potentials were added for each potential group when applicable leading to the following potential components[1]: A-B, A-A + B-B, A-PW + B-PW, PW-PW, A-ions + B-ions, PW-ions, ions-ions, bonds, G96-angles, improper dihedral angles, Coulomb reciprocal. Note the inclusion of effects due to water, ions, bonds, and long-range electrostatics. Furthermore, MD runs were checked for errors and quality criteria enforced concerning e.g. distances to mirror images, which are documented in App. C.4. The remainder of this section will focus on deriving a simplified overall description of the intermolecular interaction based on this information using Universal Kriging.

[1] Dash '-' indicates the potential between, plus '+' an addition of potentials.

Note that carrying out such MD simulations is only possible when no collisions of atoms/molecules occur. Consequently, no information is available for colliding conformations. This limitation including a proposed solution is addressed in Sec. 4.3.6. However, in a first step a definition of the collision distance is required in order to avoid simulations of entangled molecules. For this, a collision distance $d_{\text{coll,full}} = 0.4$ nm is defined between any atom / bead combination of A and B using their full reference structure including side-chains. No MD simulation can be and is performed if such a collision is present for a starting configuration and the number of such collisions between A and B is termed $N_{\text{coll,full}}$. This limit is motivated by the fact that during all interaction sampling, no distance less than 0.39 nm between any atom / bead combination of A and B was detected, which is attributed to the force-field. Additionally, a collision distance of $d_{\text{coll,bb}} = 0.305$ nm is defined between the back-bone atoms and the number of such collisions between A and B is termed $N_{\text{coll,bb}}$. As the stiff back-bone structure is assumed to be largely responsible for the repulsive potential during molecular collision in comparison to the flexible side-chains, this definition will become important when accounting for the collision potential in Sec. 4.3.6. The limit is motivated by the equilibrium distance of carbon-carbon bonds in the studied structures (determined from back-bone structure to be approximately 0.34 nm), which leads to a repulsive distance of 0.305 nm under the assumption of a Lennard-Jones potential.

4.3.2 Spatial Descriptors

The interaction space (configuration space) of two macromolecules is described by their relative position and orientation, leading to a six dimensional space of coordinates $x, y, z, \alpha, \beta, \gamma$ (see App. A.1 for definition of Euler angles). These relative coordinates are coordinates of B with respect to A, which is located at the origin. Additionally to these coordinates, we will introduce spatial descriptors to enable a lower-dimensional, ideally 1D, description. Two groups of spatial descriptors are distinguished: A-B spatial descriptors, which provide a lower-dimensional coordinate for B with respect to A; and B-B spatial descriptors, which provide a distance measure between two configurations of B (A is always located at the origin) and are required to investigate spatial continuity. While a well-established solution for the B-B spatial descriptor exists, the A-B spatial descriptor is more complicated.

A-B: The spatial descriptors between A-B are motivated by the goal of trend modeling in a lower dimensional space. Universal Kriging, which will be introduced in the following sections, assumes data to be composed of a deterministic trend and random component with a spatial continuity. As trend modeling of the interaction potential in 6D

space is challenging (and also not required with Universal Kriging), a lower-dimensional description is favorable. Generally for molecular interaction, the emphasis is on small distances between molecules, as well as their contact / proximity intensity (i.e. if two small side-chains are in proximity or if a substrate is at a tailored binding zone of an enzyme forming an extensive 'contact surface'). Motivated by this, three different descriptors were tested with the following definition, advantages, and disadvantages. Let $N_{bb,A}$ and $N_{bb,B}$ be the number of back-bone atoms of A and B, respectively, and δ_{i-j} the distance between the center of atom i of A and j of B.

- δ_m is defined as the minimum distance between the back-bone (bb) atoms of A and B corrected by $d_{\text{coll,bb}}$. Advantages of this descriptor are that it is referenced to zero (distinct position for binding) and provides a simple measure. A disadvantage is that it does not capture proximity intensity.

$$\delta_m = \max\{\min_{j \in N_{bb,B}} \min_{i \in N_{bb,A}} \delta_{i-j} - d_{\text{coll,bb}}, 0\} \tag{4.10}$$

- $\overline{\delta_m}$ is defined as the average minimum distance of all back-bone atoms of B to any back-bone atom of A corrected by $d_{\text{coll,bb}}$. An advantage of this descriptor is that it captures proximity intensity well. A disadvantage is that it starts at a non-zero value determined by the molecular conformations and complex trends with multiple minima/maxima are possible in case of more than one binding site.

$$\overline{\delta_m} = \sum_{j=1}^{N_{bb,B}} \max\{\min_{i \in N_{bb,A}} \delta_{i-j} - d_{\text{coll,bb}}, 0\}/N_{bb,B} \tag{4.11}$$

- $\overline{\delta}$ is defined as the average distance between all A and B back-bone atoms. An advantage of this descriptor is that it captures proximity intensity well. A disadvantage is that it starts at a non-zero value determined by the molecular conformations and complex trends with multiple minima/maxima are possible in case of more than one binding sites.

$$\overline{\delta} = \sum_{i=1}^{N_{bb,A}} \sum_{j=1}^{N_{bb,B}} \frac{\delta_{i-j}}{N_{bb,A} N_{bb,B}} \tag{4.12}$$

All three spatial descriptors were investigated and an overview will be provided in App. D.1 for HBcAg. Overall, δ_m was chosen for trend modeling as it provides a simple overall descriptor for all ranges, including the collision region, which is especially useful and will be discussed in Sec. 4.3.6. Alternatively, a wide variety of other spatial descriptors could be derived, e.g. simple functions between strategic back-bone atoms similar

to radial basis functions. However, this is beyond the scope of this work and could be addressed in the future.

B-B: The spatial descriptor between B-B is motivated by providing a distance measure between two conformations, i.e. two different relative positions and orientations of molecule B. In this context, A is always located at the origin. Such a measure is for example necessary to describe spatial continuity between two MD simulations in a lower-dimensional space, which captures both translatory and rotatory differences. Such a measure is provided by the well established root-mean-squared deviation (RMSD) of the reference structure of B as

$$\delta_r = \sqrt{\frac{1}{N_B} \sum_{i=1}^{N_B} \delta_{i,B1-B2}^2} \,, \tag{4.13}$$

where $N_{bb,B}$ is the number of back-bone atoms of molecule B and $\delta_{i,B1-B2}$ is the distance between atom i in the two conformations of B, i.e. B1 and B2. As it can be seen, the back-bone structure of B is used as it provides sufficient detail and enables a faster computation in comparison to using the full reference structure. Overall, it provides a meaningful 1D descriptor capturing translatory and rotatory differences.

4.3.3 Basic Functions for Trend and Variogram Modeling

For trend and variogram modeling, a set of basic functions is required. In the context of trend modeling, it is desired for these functions to be continuous and asymptotic towards a constant value. In the context of variogram modeling, beyond similar considerations more stringent mathematical requirements exist (e.g. conditional definiteness) [294]. As fulfillment of these requirements is difficult for arbitrary functions, a set of basic functions is typically used in literature [294, 295, 297]. Such a subset of functions is used in the context of this work and shown next. The nomenclature follows literature, where n is termed nugget[2], s sill, and r range. As basic functions, the constant, linear, exponential, spherical, Gaussian, cubic, and generalized logistic function (GLF) are used. All functions are used for trend modeling (linear typically only for box size compensation) and all except for the constant function and GLF used for variogram modeling. The

[2]Note that the nugget(-effect) describes the discontinuity at zero and is inspired by the geological background of Kriging in the context of finding a gold nugget. Typically, this discontinuity is required for Kriging to honor the observed measurements and for most models desired. In the context of this work, the discontinuity at zero is omitted in order to avoid the introduction of noise into the overall interaction potential.

equations are

$$\xi = x/r, \qquad \chi = x - r \,, \tag{4.14}$$

$$f_{\text{const}}(x) = c \,, \tag{4.15}$$

$$f_{\text{lin}}(x) = a + bx \,, \tag{4.16}$$

$$f_{\text{exp}}(x) = n + (s-n)(1 - e^{-3\xi}) \,, \tag{4.17}$$

$$f_{\text{sph}}(x) = n + (s-n)(h(-\chi) \times (1.5\xi - 0.5\xi^3) + h(\chi)) \,, \tag{4.18}$$

$$f_{\text{gauss}}(x) = n + (s-n)(1 - e^{-x^2 r^{-2/3}}) \,, \tag{4.19}$$

$$f_{\text{cubic}}(x) = n + (s-n)(h(-\chi) \times (7\xi^2 - 8.75\xi^3 + 3.5\xi^5 - 0.75\xi^7) + h(\chi)) \,, \tag{4.20}$$

$$f_{\text{GLF}}(x) = n + (s-n)(1/(1 + e^{-B\chi})) \,, \tag{4.21}$$

where h is the heaviside function and B is an additional fitting variable for GLF. A visualization of all base functions can be found in Fig. 4.4. Fitting is performed using weighted-least-squares and the trust region reflection algorithm as implemented in *SciPy* [316] (version 1.3.3) and *Matplotlib* [317] (version 3.3.4) in *Python*. For all functions, reasonable start conditions were specified and the range constrained between the minimum and maximum x value. For variogram fitting, n and s were constrained to positive values as required to be meaningful. For each data set to be fitted, all function fits were performed and the best considering its R^2 selected.

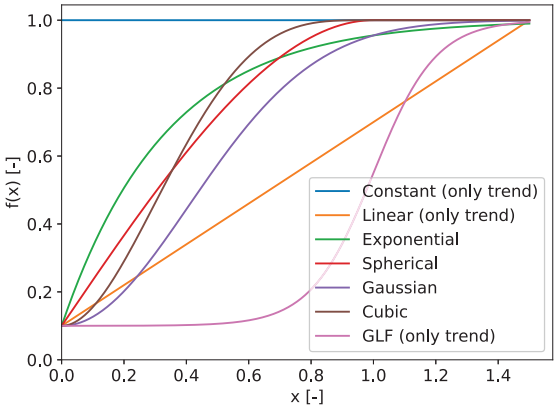


FIGURE 4.4: Visualization of all basic functions with the following parameters: $r = 1, n = 0.1, s = 1, c = 1, a = 0.1, b = 0.6, B = 10$

Concerning validity and usage in the overall Kriging algorithm, a variogram fit is considered valid if $n < s$, $r < 5$ nm, $\sigma(r) < 0.2r$, $\sigma(s) < 0.2|n - s|$, or $\sigma(n) < 0.2|n - s|$, where σ indicates the estimation variance of a parameter. A trend fit is considered valid

if $r_{95} < 4$ nm, $r_{95} > 0.2$ nm, $\sigma(r) < 0.2r$, $\sigma(s) < 0.2|n-s|$, $\sigma(n) < 0.2|n-s|$, or for GLF $\sigma(B) < 0.25B$, where r_{95} is the range when $f = n + 0.95(s-n)$ for GLF.

4.3.4 Multi-Variant Field Interpolation using Univeral Kriging

In order to estimate the interaction potential field based on a set of MD samples, a Universal Kriging approach was implemented [225, 226]. Kriging is most widely applied in the field of geostatistics, as it provides the best linear unbiased estimation for arbitrarily dimensional problems and is consequently especially useful when sampling is expensive (e.g. in geological exploration). Similarly, it provides a useful tool for the intermolecular field interpolation as the dimensionality of the problem is comparatively large with regards to the sampling cost. In the following, the essential equations and background on Universal Kriging will be provided following literature and the terminology therein [294–297].

4.3.4.1 Method

The overall aim of Kriging is to estimate the value of a spatially distributed random variable at any location in an arbitrary dimensional space based on a set of observations, also called data set in the following. Kriging assumes that the estimation can be inferred by a linearly weighted combination of the observations and provides the *best linear unbiased estimation (BLUE)* upon the fulfillment of certain mathematical assumptions (e.g. intrinsic stationarity) [294]. In this regard, optimality (*best*) refers to minimum estimation variance. Consequently, the main question to be answered by Kriging is the determination of optimal weights for estimation. While a variety of Kriging algorithms exist, typically, three kinds of Kriging are distinguished: Simple Kriging, Ordinary Kriging, and Universal Kriging. Of the three, Universal Kriging presents the most general one and will be presented in the following, as it is used within this work. For details on Simple and Ordinary Kriging the interested reader is referred to literature [294–297]. In brief, Simple Kriging requires a zero mean of the investigated field / process, while Ordinary Kriging requires a constant mean (at least locally), which can be unknown. Due to the strong trend over the distance between molecules, neither one is applicable.

In this work, the variable to be estimated is the interaction potential $U_K(\vec{x}, \mathbf{q})$, which is a super-position of the potential components P (see Sec. 4.3.1), in the space of relative position $\vec{x}$ and orientation $\mathbf{q}$. The estimation is performed as a linear combination of

N_K observations $U_{p,i}$ as

$$U_K(\vec{x}, \mathbf{q}) = \sum_{p=1}^{P} U_{K,p}(\vec{x}, \mathbf{q}) = \sum_{p=1}^{P} \sum_{i=1}^{N_K} w_{p,i} U_{p,i}(\vec{x}_i, \mathbf{q}_i) \ . \tag{4.22}$$

Note that in most cases, only a subset of observations $N_K \subseteq N_{tot}$ (local neighborhood) out of the total set of observations is used for estimation at a given location. This is motivated by computational feasibility, as a linear system of equation has to be solved for the determination of weights[3], and also leads to an improved estimate, as Universal Kriging estimates the local mean based on this observation set (local neighborhood). A convergence study of the required size of N_K along with the search algorithm is provided in App. C.2.

The underlying process, in this case potential U (each potential component separately, index p dropped for simplicity), is assumed by Univeral Kriging to be decomposable into a systematic trend $\mu(\vec{x}, \mathbf{q})$ and random component $Y(\vec{x}, \mathbf{q})$ as

$$U(\vec{x}, \mathbf{q}) = \mu(\vec{x}, \mathbf{q}) + Y(\vec{x}, \mathbf{q}) \ , \tag{4.23}$$

where the systematic trend $\mu(\vec{x}, \mathbf{q})$ can be modeled by the linear combination of M deterministic basic functions f_m

$$\mu(\vec{x}, \mathbf{q}) = \sum_{m=0}^{M} b_m f_m(\vec{x}, \mathbf{q}) = \sum_{m=0}^{M} b_m f_m(\delta_m) \ . \tag{4.24}$$

The basic functions f_m can either be modeled in the same space as the original process or a fitting derived space. Due to the significant complexity of generally highly anisotropic macromolecular interaction, a lower-dimensional space of the minimum distance δ_{min} between back-bone atoms of molecules A and B was chosen (see Sec. 4.3.3 for more details). This approach has the advantage of general applicability and negative effects due to the simplification are mitigated by the local estimation of the mean in the neighborhood of the estimate. To address bias as a result of sampling heterogeneity, the measurement data set is weighted using an inverse Gaussian weighting scheme using δ_r as a distance measure and a Gaussian width of 2 nm. Based on these weights for all measurement observation, the trend fitting approach in Sec. 4.3.3 is employed and the best resulting fit in combination with the constant function for local mean estimation used to model the trend. Note that for potential components, which are dependent on the MD system size (box size), trend fitting was performed on the residuum after

[3]For a data set of 100'000 points the memory requirements of the linear system of equation alone would exceed 37 GB at single point floating precision.

box size compensation (linear fit through either the number of water molecules or ions depending on potential component).

After subtraction of the trend, the remaining random component Y has approximately a zero mean in δ_m space for all tested systems as required by Universal Kriging. In addition, Universal Kriging requires Y to be intrinsically stationary and optimality of the estimation relies on Y being Gaussian [294]. This poses a challenge in the case of molecular interaction as the width of the (approximately) Gaussian distribution changes from a finite value at small δ_m towards zero for large δ_m, i.e. a delta distribution. To solve this, the problem and data set is split into sections in δ_m space. Within each section, the requirement of intrinsic stationarity is approximately fulfilled. At the same time, the requirement of a Gaussian distribution for estimation optimality is fulfilled in the important close (binding) region and only violated at large distances, which tend towards zero anyway.

Over the range of the trend function, five sections are used and the data within each section, as well as the upper and lower neighboring section, used. By inclusion of data points from the neighboring upper and lower section issues with jumps of the estimate (not estimation variance) at the boundaries and spatial paradoxes are avoided. The spatial continuity of Y is then described for each section by a separate residual variogram γ_Y using the root-mean-square distance of back-bone atoms δ_r as a distance measure[4] between two relative locations:

$$\gamma_Y(\delta_r) = \frac{1}{2}\mathrm{Var}(Y((\vec{x},\mathbf{q}) + \delta_r) - Y(\vec{x},\mathbf{q}))\ . \tag{4.25}$$

γ_Y is estimated based on the correlations between all data point for each section. As the number of correlations is in the order of 10^{10} and consequently direct curve fitting is unfeasible, an automatic iterative binning strategy was derived (see App. C.1 including discretized form of eq. 4.25) followed by the fitting procedure described in Sec. 4.3.3 using the standard deviation in each bin as its uncertainty. Note that in the context of variogram modeling, the nugget value n (see Sec. 4.3.3) can be interpreted as the intrinsic measurement error (e.g. due to the thermodynamic ensemble), while the sill can be interpreted as the overall variance of the process. Any failed sectional Variograms (see Sec. 4.3.3) are replaced by a valid neighboring sectional Variogram, preferably from a lower section.

Based on the trend and variogram models, which are separately derived for each potential component, the optimal weights providing the unbiased estimate with minimum estimation variance at a location $\vec{x}$ and $\mathbf{q}$ denoted as κ can then be determined solving

[4] Note that generally, anisotropic direction-dependent variograms models are possible. Due to the dimensionality of the problem and sampling cost, this is not feasible in this context.

the Universal Kriging system

$$\begin{bmatrix} \gamma_Y(\delta_{r,1-1}) & \dots & \gamma_Y(\delta_{r,1-N}) & 1 & f_1(\delta_{m,1}) & \dots & f_M(\delta_{m,1}) \\ \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ \gamma_Y(\delta_{r,N-1}) & \dots & \gamma_Y(\delta_{r,N-N}) & 1 & f_1(\delta_{m,N}) & \dots & f_M(\delta_{m,N}) \\ 1 & \dots & 1 & 0 & 0 & \dots & 0 \\ f_1(\delta_{m,1}) & \dots & f_1(\delta_{m,N}) & 0 & 0 & \dots & 0 \\ \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ f_M(\delta_{m,1}) & \dots & f_M(\delta_{m,N}) & 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_N \\ \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_M \end{bmatrix} = \begin{bmatrix} \gamma_Y(\delta_{r,\kappa-1}) \\ \vdots \\ \gamma_Y(\delta_{r,\kappa-N}) \\ 1 \\ f_1(\delta_{m,\kappa}) \\ \vdots \\ f_M(\delta_{m,\kappa}) \end{bmatrix}$$

As all components have to be on the same order [318], a normalization was performed for both variogram and trend functions. For details on the background, derivations, and proofs the interested reader is referred to refs. [294–297]. As it can be seen, the system of equations contains information on the spatial correlation between the estimate location and data points, the spatial correlation within the data points, and trend. Furthermore, row $N + 1$ constrains the sum of weights to one and leads to a local estimation of the mean. Based on the resulting weights for each potential component, the estimate can be calculated using eq. 4.22. Furthermore, the estimate variance for each potential component P can be calculated as

$$\sigma_K^2(\vec{x}_\kappa, \mathbf{q}_\kappa) = \sum_{i=1}^{N} w_i \gamma_Y(\delta_{r,\kappa-i}) + \sum_{j=0}^{M} \lambda_j f_j(\vec{x}_\kappa, \mathbf{q}_\kappa) \, . \tag{4.26}$$

The estimation variance can be and is later used to iteratively resample and improve the estimation of the potential field. Note in this context that solving the Universal Kriging system requires no knowledge of the actual value of the data points, but only their location, as well as a trend and variogram model.

When solving the system of equation, one of the most frequent issues is that the matrix can become ill-conditioned, especially for a variogram model based on a Gaussian function due to the zero slope for small x. In order to avoid such issues [319], the system is solved using the bidiagonal divide and conquer singular value decomposition (SVD) using double precision and a maximum factor between smallest and largest eigenvalue of 10^6. The overall algorithm was implemented in C++ with hybrid MPI+OpenMP parallelization using the library *Eigen* (revision 14db78c53) for solving the linear system of equations using SVD [226]. For function fitting, the code was coupled to *SciPy* [316] (version 1.3.3) and *Matplotlib* [317] (version 3.3.4) in *Python*. As previously mentioned, the described procedure is carried out for all potential components separately, as trend and spatial correlation can differ. The overall interaction potential can be calculated as a super-position of all potential components as shown in eq. 4.22. It was found that out of

all potential components only the A-B potential contained sufficient spatial correlation and signal-to-noise ratio to perform Kriging. For all remaining potential components, only the trend was used.

4.3.4.2 Grid Design

The presented Universal Kriging approach provides a method to estimate the respective variable of interest, in this case interaction potential, at one location based on nearby samples. In order to use this knowledge during a MDEM simulation, a more general representation is needed, as Kriging cannot be performed for all contacts during a simulation. For this, the interaction potential is determined on a grid. On this grid, a numerical gradient operation can then be carried out to determine forces and torques of two macromolecules in contact. At this point, the most important aspects for homogeneous grid design of translatory and rotatory dimensions will be discussed. Alternative representations will be discussed later, including methods for refinement.

In order to design the interaction potential grid, three aspects are important: the structure of both molecules, the interaction cutoff $d_{\text{inter,cut}}$[5], and the available memory. Molecule A is located at the origin, while molecule B can have any configuration nearby, leading to a 6D representation of all relative positions x, y, z and orientations α, β, γ (see App. A.1). In this context, the cartesian space of relative positions has to incorporate all configurations with an orientation inside the cutoff. In order to calculate the extend of the cartesian space $\mathbf{C}$, the maximum distance of any atom of B from its center of mass (COM) has to be added to the bounding box of molecule A at the origin extended by $d_{\text{inter,cut}}$ in all directions. After $\mathbf{C}$ is known, the equivalent rotational equivalent $\mathbf{R}$ has to be determined. While the extend is known for all systems to be $-\pi$ to π for α and γ and $-\pi/2$ to $\pi/2$ for β, an equivalent angular resolution to a cartesian resolution has to be determined. For this, the maximum distance of any atom of B from each axis in its reference frame at the COM can be used to calculate the angular equivalent for which the segment of a circle with that radius is equal to the cartesian distance. Note that this methodology employs the strictest conditions on the angular resolution for cartesian equivalence. Alternatively, the RMS distance from the axis could be used. Based on $\mathbf{C}$, $\mathbf{R}$, and the angular equivalents, the finest grid resolution for a specified field size (memory size available) can be determined. Unless otherwise indicated, any grid definition refers to this methodology and is consequently fully described by the cartesian grid resolution. In addition, a more advanced multi-grid hierarchical field approach was developed, which will be presented in the following.

[5]The interaction cutoff $d_{\text{inter,cut}}$ is defined as the maximum of the trend cutoff (largest spatial descriptor, δ_m , after which the absolut trend sum is less than 1 kJ/mol) and Kriging cutoff (largest spatial descriptor, δ_m , after which the absolut trend residual of any data point is less than 1 kJ/mol).

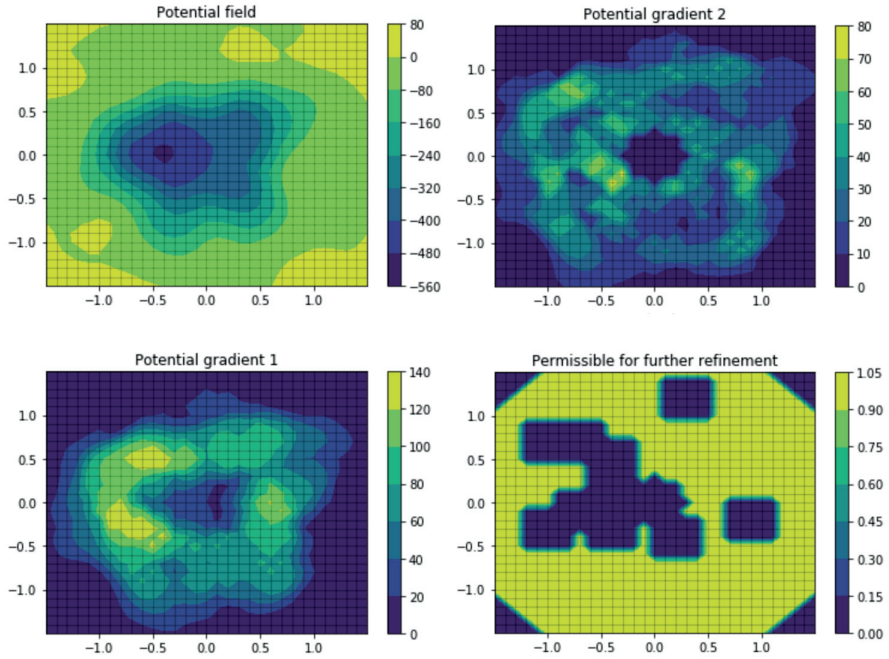


FIGURE 4.5: Visualization of the multi-grid hierarchical field approach in 2D.

Multi-Grid Hierarchical Fields While the presented grid design covers the entire interaction space, it is limited in local resolution due to constraints in memory. In order to resolve this problem, an approach was developed to automatically create higher-resolution grids in regions of interest, which are then hierarchically stacked and searched during calculation of forces and torques. This approach enables an efficient calculation of forces and torques due to the maintained homogeneous grid design. Alternative approaches will be discussed in Sec. 4.4.2.

The main idea of this approach is to locally improve grid resolution based on four possible criteria in descending order of priority: potential minima, potential maxima, gradient maxima (estimated as absolute maximum difference to a neighboring grid point), and second gradient maxima (similarly from first gradient). In the following, the approach will be described and a 2D visualization can be found in Fig. 4.5.

1. Specification of the number of refinement grids for each criteria, which are in the previously stated order of priority.

2. Load overall potential field and mark all grid locations outside of cutoff and with collisions (conditional on being further than 0.6 nm from the closest MD point) as not permissible for refinement.

3. Identify grid location matching the extrema of the current criteria.
4. Create a new grid centered in found location with 10 steps at a resolution of 0.15 nm (angular equivalent according to previous methodology) using the Universal Kriging model. Account for wrap-around of angular dimensions.
5. Mark all grid locations of overall grid within new grid as not permissible for further refinement.
6. Return to step 3 until the specified number of grids for each criteria are generated or no more permissible points are available for refinement.

As it can be seen in Fig. 4.5, the proposed approach provides a good methodology to refine the grid resolution in regions of interest, while keeping the memory requirements low.

4.3.4.3 Initial Sampling and Iterative Refinement

One of the main advantages of Universal Kriging is that it provides an error estimate and consequently a powerful way to improve the overall estimate. Nonetheless, Kriging requires a sufficient initial data set to perform statistical analysis concerning trends and spatial correlation. Consequently, a two-step process was employed to first perform systematic random sampling over the interaction region, followed by iterative resampling. This procedure will be discussed in the following.

Initial Sampling In the field of sampling statistics, a variety of techniques exist. Examples are purely random sampling, cluster sampling, and latin hypercube sampling, see e.g. ref. [320] for further. In the context of this work, three aspects are most important: Firstly, interaction occurs within a certain interaction distance of a 6D space. The space extends towards larger distances, while smaller distances are especially crucial for binding. Consequently, simple random sampling is not favorable. Secondly, the interaction region is a highly complex subregion of the original 6D space. In combination with the dimensionality, advanced placement is challenging. Thirdly, for spatial correlation analysis at least a subset of samples have to be in proximity to provide information on spatial correlations at small distances. As a result, a systematic random approach consisting of two components was employed to generate the initial sample set.

- **Systematic random sampling** at different distance classes was performed to control sampling density over the interaction distances. Distance was measured as the minimum distance between the center of mass of atoms of A and B. Classes

were defined as following: I) 0.4 - 0.5 nm for an emphasis on binding locations II) 0.5 - 2.5 nm in 0.2 nm increments for an emphasis on binding behavior III) 2.5 - 5.0 nm in 0.5 nm increments for an ephasis on determining cutoff. Typically 25'000 samples were carried out in region I and 5'000 for each increment in regions II and III. Each sample was created by randomly inserting (position and orientation) both molecules in a 100 nm box until their distance was within the specified region. Credit for the implementation is given to Uwe Jandt.

- **Proximity resampling** was performed for the sample set to contain samples at small distances between each other and improve Variogram estimation. For this, a set of 50 samples in each increment of region II was replicated 50 times and the dynamic nature of the MD simulation used, which caused all final sampling locations to deviate slightly from each other. This sample set was only used for Variogram analysis and not for estimation of the field to avoid any bias. This was only employed for the PDC system and not necessary for the HBcAg system.

Iterative Refinement In order to improve the potential field estimate, an iterative resampling strategy was derived. Resampling was performed based on four criteria: (normalized) estimation variance maxima, potential maxima, potential minima, and gradient maxima. The derived strategy will be outlined in the following and results presented for each model system in the results section. In each case, a set of resampling points is determined per iteration for which MD simulations are run.

First, the special case of estimation variance maxima will be discussed. Normalized and absolute estimation variance are distinguished, as per Kriging section the absolute values differ and the overall estimation variance is meant to be improved, which also includes outer points for which the absolute variance is lower. One of the main advantages of Kriging is that not only an estimation variance is known along with the field estimate, but also that no knowledge of the actual data point values is required to determine the variance. During variance resampling, this advantage was exploited to iteratively determine the resampling locations as:

1. Find maximum (normalized) estimation variance on the grid under the constraint that no collision of molecules is present (see Sec. 4.3.1).
2. Add virtual data point to data set at this location and recalculate the variance of affected nearby grid points.
3. Return to step 1 until the desired number of resampling points is found.

Although this placement procedure of resampling points is not necessarily optimal as *combined* placement of *multiple* resampling points can provide a better result, the procedure provides very good results and exploits knowledge of the model in a straightforward and computationally efficient manner. Furthermore, due to the dynamic nature of the MD simulations, no exact placement of resampling points is possible either way. This also motivates the iterative resampling strategy. For each iteration of variance minimization, a set of 5'000 resampling points was run. Variance minimization resampling was always performed before extrema resampling to first improve the overall estimate and avoid false-negatives concerning e.g. binding location identification.

Second, the more general cases of extrema resampling will be discussed, i.e. potential minima, potential maxima, gradient maxima. Extrema resampling is motivated by improving the estimation of location and potential value of e.g. binding locations and repulsive locations. For this, an extrema search algorithm was implemented:

1. Find the current (global) extrema on the grid under the constraint that no collision of molecules is present (see Sec. 4.3.1).
2. Determine all neighboring points of the current extrema for which a continuous increase/decrease (minima/maxima) is present. This is called the neighborhood of the extrema. Direct neighbors are called first-level neighbors.
3. Remove extrema and neighborhoods from eligibility for current extrema and return to step 1 until the desired number of resampling points is found.

The gradient field is calculated beforehand as the maximum absolute potential difference to a neighboring grid point. For each iteration of extrema resampling, a set of extrema points and random first-level neighborhood points relative to the extrema's neighborhood size (after limiting the maximum to 25 %) was run. Results of iterative resampling will be provided for each model system in the results section. In the following, a simplified 2D example will be shown for validation and visualization.

4.3.4.4 2D Example

In order to validate the algorithm and provide a visual understanding, a two dimensional example was created. The example consists of a random scalar field (no units) between two objects of radius 0.15 and can be found in Fig. 4.6 and Fig. 4.7. The random truth field (Fig. 4.6 top left) has similar statistical properties as typical MD data and was created in the following fashion: In a first step, a random field was generated using sequential Gaussian simulation [321] using a Gaussian variogram model with a range of

0.7, nugget of 3000, and sill of 10000. In a second step, a scaling of the random field to zero was performed between a minimum distance of 0.4 and 1.2 using a Gaussian function. Lastly, a trend of -400 at object contact and zero at range one with Gaussian shape was super-positioned.

The simplified example shows how the Kriging algorithm performs resampling based on normalized variance minimization. Initially 20 samples are randomly selected and then iteratively 10 samples are added for each of 18 iterations (200 samples in total). Note that in order to ensure sufficient quality of the variogram, the full field was provided for statistical analysis. Selected iterations in Fig. 4.7 show how the algorithm automatically chooses resampling locations to reduce the overall estimation variance and 'learn' the entire interaction potential. As Fig. 4.6 shows, the overall trends and extrema (e.g. minima, meaning binding locations) are identified and the final estimation error consists largely of noise and small-scale discontinuities. In addition to this simplified example, the overall sampling algorithm employs even more advanced resampling techniques based on extrema resampling to localize and quantitatively evaluate e.g. binding, leading to even better results.

4.3.5 Biased MD and Insertion of Empirical Data

In many cases a bottom-up parameterization of an effective surrogate model is not sufficient. Due to limitations of the lower-scale model, e.g. model simplifications or sampling resolution during parameterization of the effective model, not all effects of the actual system can often be captured. Similar limitations were found to apply to the parameterization using MD. These include, but are not limited to the following:

As computational resources are always limited, only finite sampling can be performed for parameterization. Especially in the context of the highly complex conformational space during binding, similarly to the problem of protein folding, this leads to a significant noise ratio in the data and capturing binding events, especially at short MD durations, can be challenging due to its low probability. Furthermore, to the knowledge of the author, no MD force-field is parameterized with the motivation of accurately reproducing intermolecular interaction potentials, especially for generic systems. While deriving more appropriate and generally applicable MD force-fields will likely remain a research topic for decades, two approaches were attempted to overcome some of the limitations in the context of sampling.

First, the impact of simulation time in the context of binding was evaluated by performing **biased MD simulations at the binding locations**. For this, the binding locations were first extracted based on literature knowledge of the structural assembly,

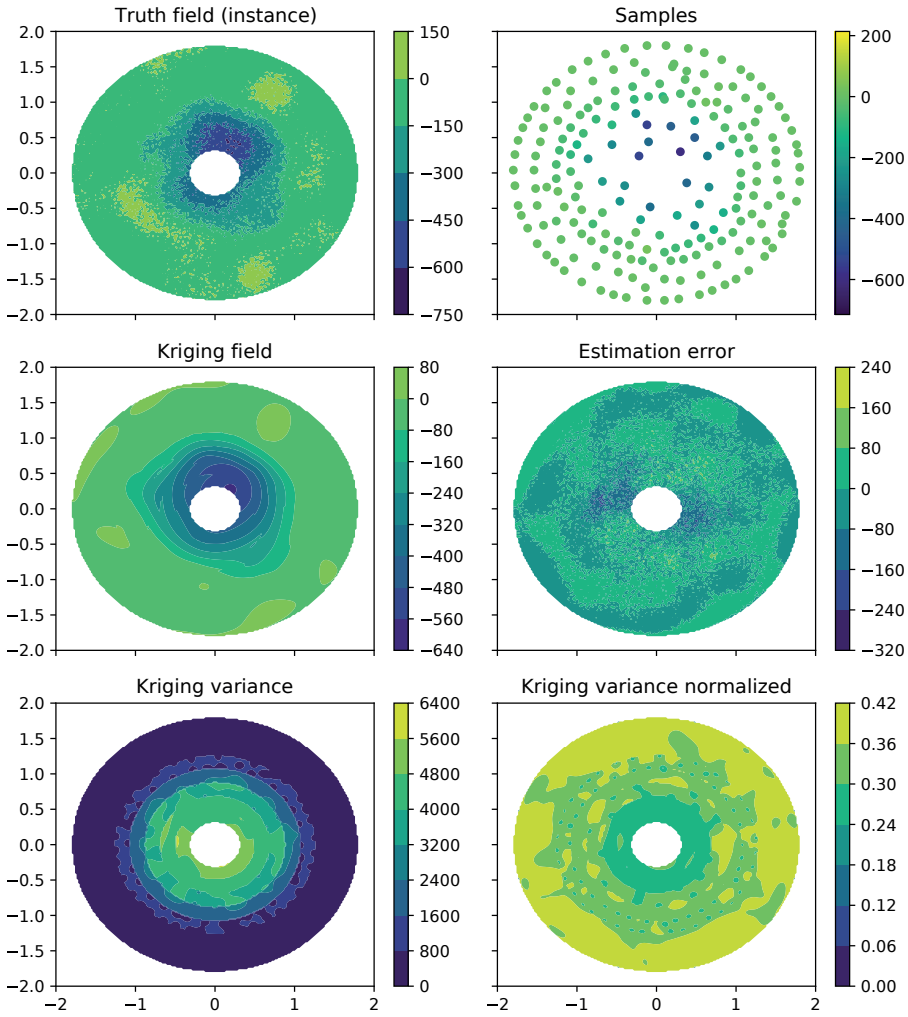


FIGURE 4.6: 2D example of Universal Kriging algorithm starting from 20 samples after 18 iterations with 10 samples per iteration (200 samples total). For variogram determination the entire truth field was provided to ensure sufficient statistics.

e.g. for $HBcAg_2$ four binding locations based on PDB 1QGT capsid. As the exact binding positions typically contain overlapping atoms of the reference structures, in a first step an algorithm was written to find the closest (w.r.t. δ_r , root-mean-square distance of back-bone atoms) collision-free location. The algorithm was based on a grid search with five steps in all directions and dimensions with a resolution of 0.2 nm or 10°. For each binding location, 1'016 MD simulations of 10 ns duration were performed with the same settings as described in Sec. 4.3.1. These biased simulations of extended time were performed to check whether this might improve identification of binding in the overall

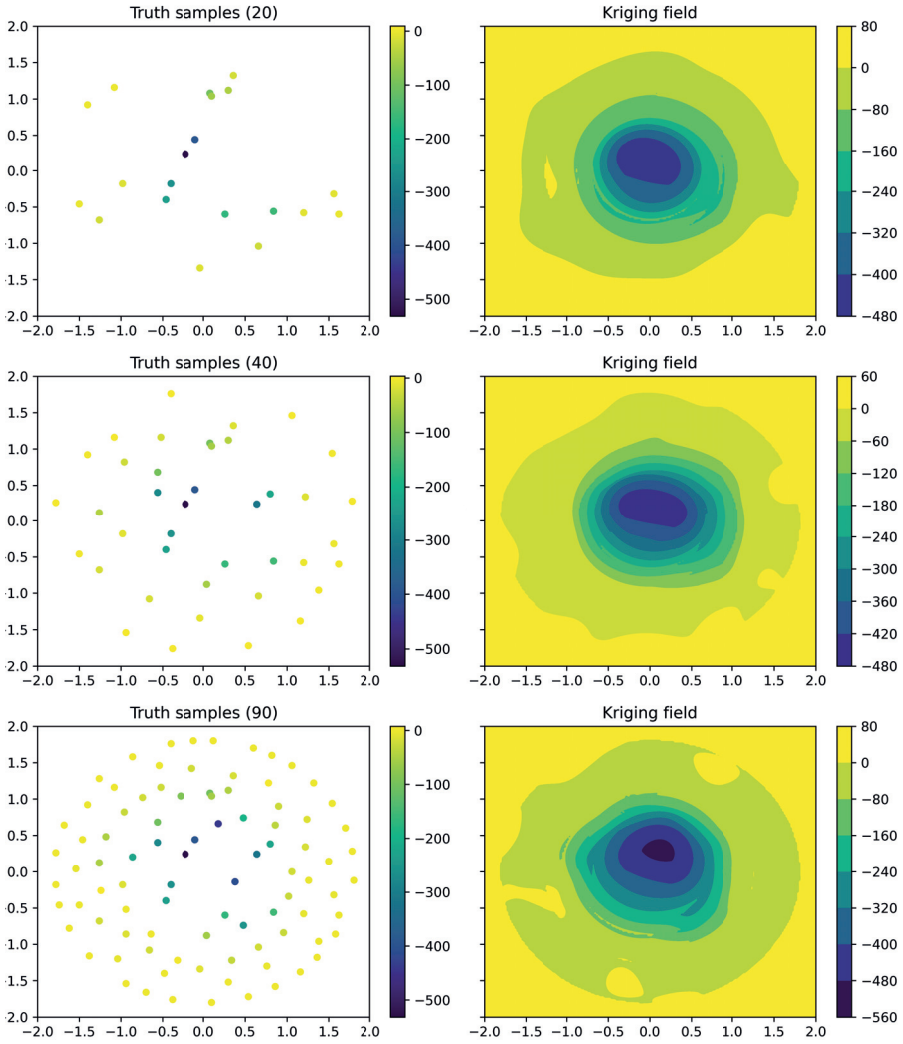


FIGURE 4.7: 2D example of Universal Kriging algorithm after initial random sampling (20 samples total), two iterations of 10 samples each (40 samples total), and seven iterations of 10 samples each (90 samples total).

potential.

Second, various approaches to **insert additional knowledge through empirical data points** were evaluated and the most effective one found will be presented in this work. For this, two sets of empirical data points are introduced into the data set for potential A-B. These points are considered virtual data points and have no impact on the trend or variogram. The first set of points has a constant potential $U_{\text{emp,bind,center}}$ and is located on a regular grid centered at the binding location (not corrected for avoiding atom collisions) with one step $d_{\text{step,center}}$ (e.g. 0.1 nm equivalent to -0.1, 0.0, and 0.1 nm) in all directions including rotational equivalents. This replication is necessary due to the possible existence of data points from MD sampling in the proximity of the binding site, causing an insufficient impact of a single empirical data point. The second set of points is also located at the binding location, but has an increasing potential with increasing distance from the binding location representing the shape of the potential minima. The best solution found has a Gaussian shape (see eq. 4.19) with a range of $r_{\text{emp,bind}}$ in δ_r space and an asymptotic value of $U_{\text{emp,bind,outer}}$ located on a regular grid with two steps $d_{\text{step,outer}}$ in all directions (e.g. 0.2 nm equivalent to -0.4, -0.2, 0.0, 0.2, 0.4 nm), including rotational equivalents. Empirical points are added up to the range $r_{\text{emp,bind}}$ and up to an increase of $N_{\text{emp,coll,inc}}$ in back-bone atom collisions relative to that at the binding location to limit overlapping configurations.

In order to quantitatively evaluate the impact on structural stability and derive appropriate parameters for the insertion of empirical data points, an objective function for structural stability was used (see App. C.3). A variety of parameters were evaluated and the results including biased MD simulations will be presented in the results.

Additional approaches to overcome the sampling problem worth exploring might be in similar contexts as protein folding, e.g. the replica exchange method. However, in the context of this work a further investigation into this exceeds the computational resources.

4.3.6 Molecular Collisions

The presented methodology is limited by the fact that configurations containing molecular collisions cannot be investigated in the same manner. While it might be feasible to use steered MD to parameterize collision configurations, the complexity would become similar to the non-colliding potential. As the primary interest and consequent deployment of computational resources of this work lies on the non-overlapping potential, a simplified model for capturing collisions was implemented inspired by the mechanism of repulsion: overlapping atoms.

When two molecules are overlapping, the interaction potential is modeled to increase as a function of the number of atom collisions and proximity of the configuration (relative position and orientation) to MD data. The motivation for including the proximity to MD data comes from the fact that conformation changes can occur during interaction / binding. When such conformation changes are present in the MD data, nearby configurations are considered possible, as the conformation changes avoid molecular collisions. Consequently, a configuration between two molecules is described by the number of back-bone collisions $N_{\text{coll,bb}}$ (definition see Sec. 4.3.1), the number of side-chain atom collisions $N_{\text{coll,side}} = N_{\text{coll,full}} - N_{\text{coll,bb}}$ (definition see Sec. 4.3.1), the distance to the next MD point d_{MD} (δ_r as distance measure). The respective values are calculated for each point on the interaction potential grid.

The proposed model increases the interaction potential as a function of these measures. The model will be described qualitatively first to emphasize the background and motivation, followed by the mathematical formulation. As the collisions field ($N_{\text{coll,bb}}$ and $N_{\text{coll,side}}$) contains jumps due to its discrete nature, in a first step smoothing of the collision fields is performed using a Gaussian kernel G with a width w_{coll}. In order for a collision to increase the potential, the following criteria need to be fulfilled:

- the smoothed number of collisions exceed a threshold value of $N_{\text{min,bb}}$ or $N_{\text{min,side}}$, respectively; and
- d_{MD} is larger than $d_{\text{MD,min}} + w_{\text{MD,min}}$. If it is between $d_{\text{MD,min}}$ and $d_{\text{MD,min}} + w_{\text{MD,min}}$, it is linearly scaled from zero to one to achieve a smooth increase in repulsion potential.

Based on the fulfillment of these conditions, the potential is increased by a repulsion coefficient $c_{\text{rep,bb}}$ or $c_{\text{rep,side}}$, respectively. Overall, this leads to

$$\begin{aligned} U &= U_{\text{krig}} + f(d_{\text{MD}}) \times \\ &\quad (c_{\text{rep,bb}} \times h(\zeta_{\text{bb}}) \times \zeta_{\text{bb}} + c_{\text{rep,side}} \times h(\zeta_{\text{side}}) \times \zeta_{\text{side}}) \,, && (4.27) \\ \zeta_{\text{bb}} &= G(N_{\text{coll,bb}}, w_{\text{coll}}) - N_{\text{min,bb}} \,, && (4.28) \\ \zeta_{\text{side}} &= G(N_{\text{coll,side}}, w_{\text{coll}}) - N_{\text{min,side}} \,, && (4.29) \\ f(d_{\text{MD}}) &= \begin{cases} 1 & \text{if } d_{\text{MD}} > d_{\text{MD,min}} + w_{\text{MD,min}} \\ 0 & \text{if } d_{\text{MD}} < d_{\text{MD,min}} \\ \frac{d_{\text{MD}} - d_{\text{MD,min}}}{w_{\text{MD,min}}} & \text{else} \end{cases} && (4.30) \end{aligned}$$

where h is the heaviside function. Optionally, the base potential U_{krig} can also be smoothed with a Gaussian kernel G of width w_{krig} (none as default). More complex

approaches were investigated, but did not lead to improved results. In order to determine the parameters, a combination of optimization and parameter studies based on the stability of a HBcAg capsid and PDC 60-mer (see App. C.3 for objective function) was carried out, which will be presented in the results sections.

4.3.7 Method Summary and Uncertainties

Overall, an extensive framework for deriving data-driven interaction potential fields from MD for the configuration space of two macromolecules has been presented. A visualization of the method procedure can be found in Fig. 4.3. Focus of this work is the multi-variant estimation of interaction potential including a sampling procedure based on a set of MD potential data. The presented approach provides a methodology for deriving the interaction potential field based on a set of observations using Universal Kriging as a *best linear unbiased estimator*, including an estimation variance. The various components of the methodology include spatial descriptors (Sec. 4.3.2), basic functions for trend and variogram modeling (spatial correlation, Sec. 4.3.3), Universal Kriging (Sec. 4.3.4), and treatment of molecular collisions (Sec. 4.3.6). Furthermore, an approach for insertion of empirical knowledge has been presented to enable the often needed correction due to limitations of the lower-scale model (Sec. 4.3.5). The methodology has been presented in detail providing parameters, background, and a 2D example. Overall, the presented method procedure is summarized:

1. **Trend fitting** of all potential components using basic model functions (Sec. 4.3.3) in a lower-dimensional space using spatial descriptors (Sec. 4.3.2).
2. **Spatial correlation** analysis and **Variogram fitting** of all potential components for trend-removed residuals (Sec. 4.3.3 and 4.3.4).
3. **Additional data (optional):** Insertion of empirical knowledge, e.g. from experimental findings (Sec. 4.3.5).
4. **Universal Kriging** of qualifying potential components to determine **best linear unbiased estimate** (Sec. 4.3.4).
5. **Summation** of all potential components for overall potential (Sec. 4.3.4).
6. **Molecular collision** analysis for increasing overall potential (Sec. 4.3.6).
7. **Iterative resampling** for improvement of estimation (Sec. 4.3.4.3).

Lastly, the main uncertainties of the method should be named. The first main uncertainty comes from the underlying MD model. In the context of this work, this model

is considered a black-box model, which was provided by collaborators, and it is only noted that this model comes with limitations due to the force-field, general extraction of potentials, and more. Note that to the knowledge of the author, no force-field is parameterized with the goal of providing accurate (relative) interaction potentials. This work, however, focuses on how such data can be post-processed to a meso-scale model, as well as providing a sampling approach.

The second main uncertainty comes from the estimation at each grid location of the potential field. The presented Universal Kriging approach provides the *best linear unbiased estimator* including an estimation variance based on the surrounding data including its spatial structure, agreement with trend, and local estimation of the mean. Consequently, a good estimation of the uncertainty of the Kriging model is provided. The main limitation of the approach lies in the fact that some potential components possess such significant noise that Kriging cannot be reasonably performed, in which case only the potential trend is used. While this is the most reasonable approach at this point, it also presents the biggest opportunity for future improvement: More elaborate trend models could provide a more detailed and improved overall potential. However, special care would have to be taken to avoid over-fitting and such a further improvement is beyond the scope of this work.

The third main uncertainty comes from the dimensionality of the interaction space. Due the highly complex 6D space, sampling is - at least with current computational resources - always limited. In this context, the presented approach for insertion of empirical knowledge (Sec. 4.3.5) is especially useful, as it provides means to make up at least partially for this sampling limitation.

In the following section, the approach for deriving forces and torques from the interaction potential will be described. This can then be used to model the interaction of two macromolecules in conjunction with the diffusion model presented in Ch. 3.

4.4 Derivation of Interaction Forces and Torques From Potential Fields

In this section, the background and equations for deriving interaction forces and torques from the previously determined potential fields will be discussed. Theoretically, the forces and torques could be pre-calculated before the simulation and access performed using linear interpolation. However, it was found that the penalty on the grid resolution in the context of memory consumption due to the additional dimension was too

significant. Additionally, alternative representations and possible simplifications will be provided.

4.4.1 Direct Usage of Potential Field

In order to calculate the forces and torques on molecule A resulting from a pairwise interaction with molecule B (position $\vec{x}_{body,A\to B}$ in cartesian space and orientation $\vec{\theta}_{body,A\to B}$ in Eulerian space), the gradient operations

$$\vec{F}_{pp,body,A} = -\nabla_{\mathrm{t}} U(\vec{x}_{body,A\to B}, \vec{\theta}_{body,A\to B}) \,, \tag{4.31}$$

$$\vec{M}_{pp,body,A} = -\nabla_{\mathrm{r}} U(\vec{x}_{body,A\to B}, \vec{\theta}_{body,A\to B}) \tag{4.32}$$

have to be carried out. This can be done numerically using finite differences and has to be carried out in the local coordinate system of molecule A (body frame of reference). For conciseness, the index 'body, $A \to B$' is omitted in the following. In the context of this work, a central difference form was chosen to calculate the gradient as

$$-\nabla_{\mathrm{t},i} U = -\frac{U(\vec{x} + \hat{i}h_i, \vec{\theta}) - U(\vec{x} - \hat{i}h_i, \vec{\theta})}{2h_i} + O(h_i^2) \tag{4.33}$$

in second-order approximation or

$$-\nabla_{\mathrm{t},i} U = -\frac{-U(\vec{x} + 2\hat{i}h_i, \vec{\theta}) + 8U(\vec{x} + \hat{i}h_i, \vec{\theta}) - 8U(\vec{x} - \hat{i}h_i, \vec{\theta}) + U(\vec{x} - 2\hat{i}h_i, \vec{\theta})}{12h_i} + O(h_i^4) \tag{4.34}$$

in fourth-order approximation in cartesian space dimension $i \in \{x, y, z\}$, where h_i is the grid spacing. When the point of gradient evaluation is not located directly on the grid, two options of evaluating U were implemented: First, a linear interpolation in 6D space, which is computationally expensive due to having to access all $2^6 = 64$ grid points around the evaluation point of U, but provides a more accurate result. And second, a nearest-neighbor search, which is computationally efficient, but provides a less accurate and non-continuous result of the gradient. Tests on energy conservation during time integration showed that the usage of a fourth-order gradient scheme and linear interpolation is advantageous for coarse grids to improve energy conservation in the absence of thermodynamic effects over longer time periods. However, in the presence of the diffusion model and resulting thermodynamically induced randomness, a second-order scheme and nearest-neighbor search was found to be sufficient.

While carrying out the finite difference scheme in cartesian space is straight-forward, calculation in Eulerian space is not as simple. Reason for this is that the performed orientation step has to be distributed over A and B as both are free in space, which

consequently leads to a change in relative position of B with respect to A, i.e. rotation and translation are coupled[6]. Note that this *cannot* be neglected. Furthermore, the gradient in Euler space has to be transformed into a gradient in $\{x, y, z\}$ space using the respective Jacobian. As a result, the calculation of $\nabla_{\mathrm{r}} U$ has to be adapted accordingly as

$$-\overrightarrow{\nabla_{\mathrm{r}} U} = -\begin{bmatrix} \cos(\beta)\cos(\gamma) & -\sin(\gamma) & 0 \\ \cos(\beta)\sin(\gamma) & \cos(\gamma) & 0 \\ -\sin(\beta) & 0 & 1 \end{bmatrix} \begin{bmatrix} \nabla_{\mathrm{r},\alpha} U \\ \nabla_{\mathrm{r},\beta} U \\ \nabla_{\mathrm{r},\gamma} U \end{bmatrix}, \tag{4.35}$$

$$\nabla_{\mathrm{r},i} U = \frac{U(\mathbf{M}^{-1}\vec{x}, \vec{\theta} + \hat{\imath}\phi_i/2) - U(\mathbf{M}\vec{x}, \vec{\theta} - \hat{\imath}\phi_i/2)}{2\phi_i} + O(\phi_i^2), \tag{4.36}$$

$$\tag{4.37}$$

with grid spacing ϕ_i in Eulerian space $i \in \{\alpha, \beta, \gamma\}$, where $\mathbf{M}$ is the relative rotation matrix from $\vec{\theta}$ to $\vec{\theta} + \hat{\imath}\phi_i/2$. The fourth-order approximation of the gradient can be calculated similarly.

Note that previously stated tests concerning order of gradient operation and search scheme were carried out including orientation. For dimensions with periodicity (α and γ) a wrap around was implemented, while for non-periodic dimension a mapping onto the boundary was implemented when leaving the grid during gradient operation.

Note further that significant effort was placed on optimizing the implementation of this gradient operation for Nvidia GPUs, as this operation is the most time-consuming of the entire MDEM simulation.

4.4.2 Alternative Representations and Simplifications

This section serves to provide some alternative representations and simplifications, which might be worth exploring in future works. While the chosen representation of homogeneous multi-grid fields provides many advantages, such as being one of the most flexible representations of arbitrary interaction potentials, it also comes with disadvantages. Most importantly, the grid resolution is limited by memory. While hardware memory has improved very significantly and is expected to improve in the future, some limitations will always apply. Additionally, the numerical gradient operation to be performed for every contact during the simulation comes at a computational cost, which could be reduced by an alternative representation.

[6]The step in cartesian space also has to be distributed over both A and B. However, a translatory movement of neither A nor B results in a changing relative orientation of B, i.e. in this case translation and rotation are effectively decoupled.

The first alternative representation to be named is **adaptive mesh refinement (AMR)**, which is more commonly used in CFD to improve accuracy in solution sensitive regions. AMR might provide a more powerful alternative to multi-grid fields, but will come at the cost of more complexity for the numerical gradient operation and was therefore not used within this work. Furthermore, a **principle component analysis (PCA)** might provide more insight into degrees-of-freedom, which contain little information and can therefore be neglected. Consequently, the interaction potential fields could be simplified. Additionally, **function representations** of the 6D field might provide an alternative for simple molecules. While generic formulations in a 6D space are challenging, for certain cases sufficient solutions might exist, possibly aided by a PCA resulting in a lower-dimensional representation. In a slightly more detailed fashion, the interaction potential could be saved in a very **coarse-grained representation** of the molecules structures with simple functional potentials between beads, similar to a very coarse-grained force-field in MD. Lastly, the interaction potential could also be represented in an **ANN** building upon the Kriging model to provide post-processed data. Detailed investigation of the neural network design would be crucial to ensure an appropriate representation, avoid over-fitting, and manage computational requirements during usage.

4.5 Critical Time Step

In order to estimate the critical time step $\tau_{\mathrm{crit,pp,field}}$ of the intermolecular interaction model using data-driven potential fields, the oscillation period $T_{0,\mathrm{pp,field}}$ of the corresponding two-mass spring system of particles i and j can be used. The critical time step for the interaction model can be estimated as

$$\tau_{\mathrm{crit,pp,field}} = 2\pi \cdot \min\left(\sqrt{\frac{\mu_{m,i-j}}{\max(\|\overrightarrow{{}^2\nabla_{\mathrm{t}} U}\|)}}, \sqrt{\frac{\min(\mu_{I,i-j,\mathrm{comp}})}{\max(\|\overrightarrow{{}^2\nabla_{\mathrm{r}} U}\|)}}\right), \tag{4.38}$$

over the potential fields and structural features of all particle interactions $i-j$, where ${}^2\nabla$ is the second partial derivative of the potential field U defined as ${}^2\nabla_{\mathrm{t}} = \left(\frac{\partial^2}{\partial x^2}, \frac{\partial^2}{\partial y^2}, \frac{\partial^2}{\partial z^2}\right)$ for translation and ${}^2\nabla_{\mathrm{r}} = \left(\frac{\partial^2}{\partial \alpha^2}, \frac{\partial^2}{\partial \beta^2}, \frac{\partial^2}{\partial \gamma^2}\right)$ for rotation (calculated equivalently to Sec. 4.4.1 using central finite differences with second-order accuracy and nearest-neighbor approximation). Further, $\mu_{m,i-j}$ represents the reduced mass defined as

$$\mu_{m,i-j} = \frac{m_i m_j}{m_i + m_j}. \tag{4.39}$$

and $\mu_{I,i-j,\text{comp}}$ the (minimum) reduced mass moment of inertia over all permutations of particle interaction $i-j$ and degrees-of-freedom $k \in \{\alpha, \beta, \gamma\}$ as

$$\mu_{I,i-j,\text{comp}} = \frac{I_{i,k} I_{j,k}}{I_{i,k} + I_{j,k}}. \tag{4.40}$$

As a result, the chosen approximation represents the worst-case approximation with regard to the time step. Alternatively, one could calculate at significant computational cost more accurately the reduced mass moment of inertia relative to $\overrightarrow{{}^2\nabla_\text{r} U}$ for the entire field. However, this more conservative approximation was found to be sufficient as other model components were dominating the critical time step for the investigated systems. Further note that the used time step is typically lower by a factor of five [42].

5

Bonded Interaction

5.1 Introduction

The main purpose of bonded interactions in the context of molecular modeling is to describe strong interactions between the primary units. In the context of MD, bonded interactions are typically used to represent covalent interaction between atoms or coarse-grained beads and are consequently an integral part of each force-field [30]. A variety of bonded interactions have been proposed in literature not just for pairwise interactions, but also three and four body interactions. Applications of multi-body bonded interactions are e.g. dihedral angles. Concerning the potential shape of bonded interactions, beyond harmonic potentials a variety of more complex potentials such as the Morse and Ryckaert-Bellemans potential have been used [30] and are e.g. implemented in *GROMACS* [42]. For bonded interactions, oscillation frequencies can become a limiting factor concerning time integration. This is especially true for light particles / atoms, such as the hydrogen atom [42]. In order to address this, constraint algorithms have been developed in literature. Examples are the algorithms SHAKE [322], RATTLE [323], SETTLE [324] (for water molecules), LINCS [325], and its parallel version P-LINCS [326]. For an in-depth discussion of (hononomic) constraint algorithms the interested reader is referred to refs. [30, 42]. Constraint parameters are often addressed directly in the force-field parameterization, e.g. for the Martini force-field [53, 125, 126].

In the context of macro-scale methods such as DEM, bonded interaction is typically used to model structures and their breakage in a similar concept as the Finite Element Method (FEM). This approach is typically called the Bonded Particle Model (BPM) and applications range from models for rock mechanics [201, 202], over concrete [203], to bio-polymers [200]. The main strength of such an approach in contrast to e.g. FEM is that breakage can easily be modeled. However, parameterization of micro-scale bond

parameters is especially challenging. While methods for parameterization and calibration have been proposed [327], parameterization based on macro-scale properties is not yet possible. From a methodological point of view, BPM is equivalent to the bonded interaction in MD and the main difference lies in application and scale.

In the context of this work, many of the previously developed approaches in MD and DEM are build upon. However, in contrast to previously developed methods, the chosen abstraction of an entire macromolecule as the smallest unit in this work does not only contain positions for each object, but also its orientation. This is especially important in the context of macromolecular bonds, as these are primarily meant to describe structures such as dimers, trimer, and multimers for which the relative orientation is crucial. Consequently, this feature will be incorporated when formulating the bond models in the following sections. Overall, two bond models will be provided in the scope of this work. The first model will address pairwise bonding including normal, shear, torsion, and bending using an orientation-defined approach. The second model will address polymer fibers for which a three body bond model will be formulated to capture normal and bending deformation of fibers.

5.2 Pairwise Elastic Bond Model (incl. Orientation)

5.2.1 Model Description

In order to model stable structures of multiple macromolecules, the following pairwise linear elastic bond model was formulated based on relative position and orientation. As it can be seen in Fig. 5.1, it contains four stiffness parameters for normal $k_{\mathrm{bond},Fn}$, shear $k_{\mathrm{bond},Ft}$, torsion $k_{\mathrm{bond},Mn}$, and bending $k_{\mathrm{bond},Mt}$ load. Shear and bending are uniform within their load plane. Terminology is based on that of classical mechanics and represents different components of a harmonic oscillator, which will be simplified as being independent of each other. Stiffness parameters can either be determined through analytical considerations, MD simulations, or optimization using empirical stability information. As the model is specifically meant to describe interactions between macromolecules, which are stable over the entire simulation time, no breakage was implemented. However, if a model system requires, breakage conditions based on a critical bond energy, strain, or stress can be added.

Let a bond connect two particles i and j with the initial positions $\vec{p}_{0,i}$ and $\vec{p}_{0,j}$ and current positions $\vec{p}_i$ and $\vec{p}_j$, as well as the initial orientations $\mathbf{q}_{0,i}$ and $\mathbf{q}_{0,j}$ and current orientations $\mathbf{q}_i$ and $\mathbf{q}_j$. The average relative orientation of the bond at the current point

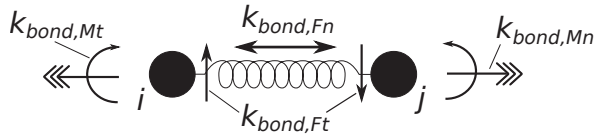


FIGURE 5.1: Visualization of the pairwise bond model.

in time can be calculated as

$$\mathbf{q}_{\text{bond,rel},i-j} = \overline{\{\mathbf{q}_i\mathbf{q}_{0,i}^{-1}, \mathbf{q}_j\mathbf{q}_{0,j}^{-1}\}}, \tag{5.1}$$

being the quaternion average [328] of both particles' relative orientations with respect to their initial orientation. Based on this, the relative rotation for i to reach equilibrium ($\mathbf{q}_{\text{bond,rel},i-j}$) is

$$\Delta\mathbf{q}_i = \mathbf{q}_{\text{bond,rel},i-j}\mathbf{q}_{0,i}\mathbf{q}_i^{-1}. \tag{5.2}$$

For the typically small angles of $\Delta\mathbf{q}_i$, the corresponding displacement angles can be calculated from the quaternion elements as

$$\Delta\theta_i = 2\arccos({}^0\Delta q_i), \tag{5.3}$$

$$\Delta\vec{\theta}_i = \Delta\theta_i \begin{pmatrix} {}^1\Delta q_i \\ {}^2\Delta q_i \\ {}^3\Delta q_i \end{pmatrix}. \tag{5.4}$$

Based on the normalized bond vector

$$\hat{b}_{i-j} = \frac{\vec{p}_j - \vec{p}_i}{\|\vec{p}_j - \vec{p}_i\|}, \tag{5.5}$$

the displacement angle can be split into a normal (Mn, torsion) and perpendicular (Mt, bending) component

$$\Delta\vec{\theta}_{i,Mn} = \hat{b}_{i-j}(\hat{b}_{i-j} \bullet \Delta\vec{\theta}_i), \tag{5.6}$$

$$\Delta\vec{\theta}_{i,Mt} = \Delta\vec{\theta}_i - \Delta\vec{\theta}_{i,Mn}. \tag{5.7}$$

Similarly, using $\mathbf{q}_{\text{bond,rel},i-j}$ and the original bond vector

$$\vec{b}_{0,i-j} = \vec{p}_{0,j} - \vec{p}_{0,i}, \tag{5.8}$$

the desired bond vector $\vec{b}_{\mathrm{des},i-j}$ (equilibrium) and consequently normal and shear displacements can be calculated as

$$\vec{b}_{\mathrm{des},i-j} = \mathbf{q}_{\mathrm{bond,rel},i-j}\vec{b}_{0,i-j}\mathbf{q}^{-1}_{\mathrm{bond,rel},i-j}, \tag{5.9}$$

$$\vec{\delta}_{i-j} = \vec{p}_j - \vec{p}_i - \vec{b}_{\mathrm{des},i-j}, \tag{5.10}$$

$$\vec{\delta}_{i,Fn} = \hat{b}_{i-j}(\hat{b}_{i-j} \bullet \vec{\delta}_{i-j}), \tag{5.11}$$

$$\vec{\delta}_{i,Ft} = \vec{\delta}_{i-j} - \vec{\delta}_{i,Fn}. \tag{5.12}$$

Consequently, the bond force and torque on particle i and j can be calculated as

$$\vec{F}_{\mathrm{bond},i\leftarrow j} = k_{\mathrm{bond},Fn}\vec{\delta}_{i,Fn} + k_{\mathrm{bond},Ft}\vec{\delta}_{i,Ft}, \tag{5.13}$$

$$\vec{M}_{\mathrm{bond},i\leftarrow j} = k_{\mathrm{bond},Mn}\Delta\vec{\theta}_{i,Mn} + k_{\mathrm{bond},Mt}\Delta\vec{\theta}_{i,Mt} + 0.5(\vec{p}_j - \vec{p}_i) \times \vec{F}_{i\leftarrow j}, \tag{5.14}$$

$$\vec{F}_{\mathrm{bond},j\leftarrow i} = -k_{\mathrm{bond},Fn}\vec{\delta}_{i,Fn} - k_{\mathrm{bond},Ft}\vec{\delta}_{i,Ft}, \tag{5.15}$$

$$\vec{M}_{\mathrm{bond},j\leftarrow i} = -k_{\mathrm{bond},Mn}\Delta\vec{\theta}_{i,Mn} - k_{\mathrm{bond},Mt}\Delta\vec{\theta}_{i,Mt} + 0.5(\vec{p}_i - \vec{p}_j) \times \vec{F}_{j\leftarrow i}, \tag{5.16}$$

where $\vec{F}_{j\leftarrow i} = -\vec{F}_{i\leftarrow j}$ and while also accounting for the (asymmetric) torque resulting from shear and ensuring a zero moment balance over both particles. The bond stiffnesses can be assigned either universally for all bonds or unique for each individual bond. Alternatively, as many interactions are softening with increasing distance, e.g. in the PDC 60-mer core the intra-trimer bonds are shorter and stiffer in comparison to inter-trimer bonds. In order to model this with as few as possible parameters, the bond stiffnesses can be formulated as

$$k_{\mathrm{bond},Fn,i-j} = \frac{\kappa_{\mathrm{bond},Fn}}{b_{0,i-j}}, \tag{5.17}$$

$$k_{\mathrm{bond},Ft,i-j} = \frac{\kappa_{\mathrm{bond},Ft}}{b_{0,i-j}}, \tag{5.18}$$

$$k_{\mathrm{bond},Mn,i-j} = \frac{\kappa_{\mathrm{bond},Mn}}{b_{0,i-j}}, \tag{5.19}$$

$$k_{\mathrm{bond},Mt,i-j} = \frac{\kappa_{\mathrm{bond},Mt}}{b_{0,i-j}}, \tag{5.20}$$

where $b_{0,i-j}$ is the initial bond length between i and j.

Note that this bond model was not employed for any of the systems studied in this work and is provided as an extension for future model systems requiring treatment of interaction portions through numerically more stable bonds.

5.2.2 Bond Contact Point

The previously formulated bond model is valid for all pairwise bonds connecting the center of mass (COM) of particles. However, when contact points are not coinciding with the COM, the model needs to be modified. An example of a model system requiring this is the E2 component of the PDC complex. While the COM is along the linker arm of the enzyme, bonded interaction when forming trimer and 60mer structures occurs at the catalytic domain of E2. In order to address such cases, the following modifications can be applied.

Let $\vec{p}_{bcp,i}$ be the bond contact point for a particle i in the local coordinate frame of the particle. When calculating bond displacements all occurrences of particle positions $\vec{p}$ for i and j have to be replaced by the position of the bond contact point as

$$\vec{p}_{bcp,i} = \vec{p}_i + \mathbf{q}_i \vec{p}_{bcp,i} \mathbf{q}_i^{-1}. \tag{5.21}$$

Additionally, the resulting torque has to be adapted as

$$\vec{M}_{\text{bond},i\leftarrow j} = k_{\text{bond},Mn}\Delta\vec{\theta}_{i,Mn} + k_{\text{bond},Mt}\Delta\vec{\theta}_{i,Mt} + (\vec{p}_{bcp,i} - \vec{p}_i + 0.5(\vec{p}_{bcp,j} - \vec{p}_{bcp,i})) \times \vec{F}_{i\leftarrow j}, \tag{5.22}$$

$$\vec{M}_{\text{bond},j\leftarrow i} = -k_{\text{bond},Mn}\Delta\vec{\theta}_{i,Mn} - k_{\text{bond},Mt}\Delta\vec{\theta}_{i,Mt} + (\vec{p}_{bcp,j} - \vec{p}_j + 0.5(\vec{p}_{bcp,i} - \vec{p}_{bcp,j})) \times \vec{F}_{j\leftarrow i}, \tag{5.23}$$

to account for the modified contact point and lever arm.

5.2.3 Critical Time Step

In order to estimate the critical time step $\tau_{\text{crit,bond,pair}}$ of this bond model, the oscillation period $T_{0,\text{bond,pair}}$ of the corresponding two-mass spring system of particles i and j can be used. The derivation based on Lagrangian mechanics can be found in most mechanics textbooks and is omitted here as it presents a simplified case of the one shown in the following section. The critical time step for this model can be estimated as[1]

$$\tau_{\text{crit,bond,pair}} = 2\pi \cdot \min\left(\sqrt{\frac{\mu_{m,i-j}}{k_{\text{bond},Fn,i-j}}}, \sqrt{\frac{\mu_{m,i-j}}{k_{\text{bond},Ft,i-j}}}, \sqrt{\frac{\mu_{I,n,i-j}}{k_{\text{bond},Mn,i-j}}}, \sqrt{\frac{\mu_{I,t,i-j}}{k_{\text{bond},Mt,i-j}}}\right) \tag{5.24}$$

[1] Note that the critical time steps of all degrees of freedom are estimated independently.

over all bonds $i-j$, where $\mu_{m,i-j}$ represents the reduced mass defined as

$$\mu_{m,i-j} = \frac{m_i m_j}{m_i + m_j}. \tag{5.25}$$

Equivalently, the remaining effective mass moment of inertia in normal $\mu_{I,n,i-j}$ and bending $\mu_{I,t,i-j}$ direction can be calculated relative to the bond center. Note that this is an approximation of the oscillation period. Due to the fact that the used time step is typically lower by a factor of five [42], this approximation is sufficient.

5.3 Fiber Bond Model

This chapter is based on the following publication:

P. N. Depta, P. Gurikov, B. Schroeter, A. Forgács, J. Kalmár, G. Paul, L. Marchese, S. Heinrich, and M. Dosta. DEM-Based Approach for the Modeling of Gelation and Its Application to Alginate. *J. Chem. Inf. Model.*, 62(1):49–70, 2022

5.3.1 Model Description

In order to model the behavior of polymer fibers and the interaction between neighboring objects within the fiber, a linear elastic bond model (i.e. harmonic potential) was implemented. Note that in this context, no information on the orientation of the units making up the fibers has to be available. As it can be seen in Fig. 5.2, the model acts on two degrees of freedom: the distance between neighboring particles of the fiber with the stiffness $k_{\mathrm{bond},Fn}$, as well as the fiber curvature described by the bond angle between three consecutive particles of the fiber with the stiffness $k_{\mathrm{bond},Mt}$. The remaining degrees of freedom are free. While the model acts between three consecutive particles connected by bonds, it captures in its essence both structural stability of the fiber chain as well

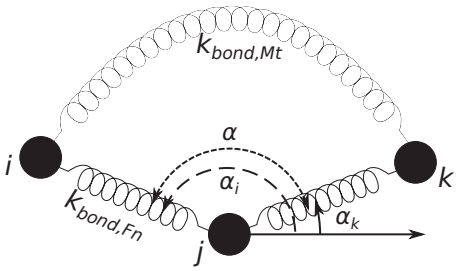


FIGURE 5.2: Visualization of the fiber bond model.

as long range interaction between macromolecules/particles in proximity, such as electrostatic interaction (e.g. repulsion in the case of alginate). As the model is specifically meant to describe interactions between macromolecules based on covalent bonds (e.g. in the case of alginate a C-O-C bond), which are stable over the entire simulation time, no breakage was implemented. However, if a model system requires, breakage conditions based on a critical bond energy, strain, or stress can be added.

Let a bond connect three particles i, j, and k in consecutive order within a fiber with the initial positions $\vec{p}_{0,i}$, $\vec{p}_{0,j}$, and $\vec{p}_{0,k}$, as well as current positions $\vec{p}_i$, $\vec{p}_j$, and $\vec{p}_k$. $b_{0,i-j}$ denotes the initial bond length, $\vec{b}_{i-j}$ the bond vector, and $\hat{b}_{i-j}$ between i and j defined as

$$\vec{b}_{i-j} = \vec{p}_j - \vec{p}_i, \tag{5.26}$$

$$\hat{b}_{i-j} = \frac{\vec{b}_{i-j}}{\|\vec{b}_{i-j}\|}. \tag{5.27}$$

Definitions between j and k are equivalent. Based on this, the bond angle α can be calculated as

$$\alpha = \arccos\left(\hat{b}_{j-i} \bullet \hat{b}_{j-k}\right). \tag{5.28}$$

Based on the equilibrium angle α_{eq}, the relative bond angle α_{rel} can be calculated as

$$\alpha_{\mathrm{rel}} = \alpha - \alpha_{\mathrm{eq}}. \tag{5.29}$$

Consequently, the bond torque becomes

$$\vec{M}_{\mathrm{bond}} = k_{\mathrm{bond},Mt}|\alpha_{\mathrm{rel}}|(\hat{b}_{j-k} \times \hat{b}_{j-i}) \tag{5.30}$$

and is set to zero for $\alpha_{\mathrm{rel}} < 0.01\,\mathrm{rad}$ to avoid problems with numerical singularities. The forces acting on the three particles are then calculated as

$$\vec{F}_{\mathrm{bond},i\leftarrow ijk} = \frac{(\vec{M}_{\mathrm{bond}} \times \hat{b}_{j-i})}{\|\vec{b}_{j-i}\|} + k_{\mathrm{bond},Fn}(\|\vec{b}_{i-j}\| - b_{0,i-j})\hat{b}_{i-j}, \tag{5.31}$$

$$\vec{F}_{\mathrm{bond},k\leftarrow ijk} = \frac{(\hat{b}_{j-k} \times \vec{M}_{\mathrm{bond}})}{\|\vec{b}_{j-k}\|} + k_{\mathrm{bond},Fn}(\|\vec{b}_{k-j}\| - b_{0,k-j})\hat{b}_{k-j}, \tag{5.32}$$

$$\vec{F}_{\mathrm{bond},j\leftarrow ijk} = -\vec{F}_{\mathrm{bond},i\leftarrow ijk} - \vec{F}_{\mathrm{bond},k\leftarrow ijk}, \tag{5.33}$$

based on the current length of the bonds as a lever arm and including normal forces. As a result of this formulation, the angular stiffness decreases with increasing bond length, which is in agreement with physical expectation e.g. of the alginate model system. The bond stiffnesses can be assigned either universally for all bonds or unique for each

individual bond. Due to the nature of the alginate system for which this model was used, universal bond stiffnesses were implemented. Note that the model could alternatively be formulated by Taylor series expansion to avoid the inverse trigonometric function [35]. However, this form was found to be sufficient for the studied systems.

5.3.2 Critical Time Step

In order to estimate the critical time step $\tau_{\text{crit,bond,fib}}$ of the fiber bond model, the oscillation period $T_{0,\text{bond,fib}}$ of the corresponding three-mass spring system of particles i, j, and k was determined. To simplify the derivation, the planar case with the center particle j being stationary was assumed, as well as all masses being equal to m. For the derivation using Lagrangian mechanics, a polar coordinate system with its origin located at j is used. Consequently i is located at (b_{j-i}, α_i) and k at (b_{j-k}, α_k). The Lagrangian energy can be calculated as

$$\begin{aligned} L = &\frac{1}{2}m((b_{j-i}\dot{\alpha}_i)^2 + \dot{b}_{j-i}^2) + \frac{1}{2}m((b_{j-k}\dot{\alpha}_k)^2 + \dot{b}_{j-k}^2) \\ &- \frac{1}{2}k_{\text{bond},Mt}(\alpha_i - \alpha_k - \alpha_{\text{eq}})^2 - \frac{1}{2}k_{\text{bond},Fn}((b_{j-i} - b_{0,j-i})^2 + (b_{j-k} - b_{0,j-k})^2). \end{aligned} \tag{5.34}$$

Using the relative angle $\alpha = \alpha_{\text{rel}} = \alpha_i - \alpha_k - \alpha_{\text{eq}}$ and normal displacements $\delta_{j-i} = b_{j-i} - b_{0,j-i}$ (similarly for δ_{j-k}), the expression can be simplified as

$$\begin{aligned} L = &m((b_{j-i}\dot{\alpha}_i)^2 + \dot{\delta}_{j-i}^2) + m((b_{j-k}\dot{\alpha}_k)^2 + \dot{\delta}_{j-k}^2) \\ &- k_{\text{bond},Mt}\alpha^2 - k_{\text{bond},Fn}(\delta_{j-i}^2 + \delta_{j-k}^2). \end{aligned} \tag{5.35}$$

Assuming that with regard to the angular movement, the changes in b_{j-i} and b_{j-k} are negligible and both are equal to b_0, it follows that

$$\begin{aligned} L = &mb_0(\dot{\alpha}^2 + 2\dot{\alpha}_k(\dot{\alpha} + \dot{\alpha}_k)) + m(\dot{\delta}_{j-i}^2 + \dot{\delta}_{j-k}^2) \\ &- k_{\text{bond},Mt}\alpha^2 - k_{\text{bond},Fn}(\delta_{j-i}^2 + \delta_{j-k}^2). \end{aligned} \tag{5.36}$$

Using Lagrange's equation without constraints for the degree of freedom $x \in \{\alpha, \alpha_k, \delta_{j-i}, \delta_{j-k}\}$

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = 0, \tag{5.37}$$

the equation of motion can be determined. The solution leads to

$$\ddot{\alpha} = -\frac{2k_{\mathrm{bond},Mt}}{mb_0^2}\alpha, \tag{5.38}$$

$$\ddot{\delta}_{j-k} = -\frac{k_{\mathrm{bond},Fn}}{m}\delta_{j-k}, \tag{5.39}$$

$$\ddot{\delta}_{j-i} = -\frac{k_{\mathrm{bond},Fn}}{m}\delta_{j-i}. \tag{5.40}$$

Consequently, the oscillation period and critical time step can be calculated according to

$$\tau_{\mathrm{crit,bond,fib}} = 2\pi \min\left(\sqrt{\frac{mb_0^2}{2k_{\mathrm{bond},Mt}}}, \sqrt{\frac{m}{k_{\mathrm{bond},Fn}}}\right). \tag{5.41}$$

Note that this is an approximation of the oscillation period. Due to the fact that the used time step is typically lower by a factor of five [42], this approximation is sufficient.

6

Results: Alginate System

This chapter is based on the following publication:

P. N. Depta, P. Gurikov, B. Schroeter, A. Forgács, J. Kalmár, G. Paul, L. Marchese, S. Heinrich, and M. Dosta. DEM-Based Approach for the Modeling of Gelation and Its Application to Alginate. *J. Chem. Inf. Model.*, 62(1):49–70, 2022

6.1 Model Parameters

6.1.1 Structural Model

In order to study the underlying mechanisms of cross-linking and network formation during gelation, the polymer chains of the alginate model system were abstracted using their composing dimer units as primary building blocks, also called units or particles. Consecutive dimer units along the polymer chain were then connected by elastic bonds (see Sec. 5.3) representing the overall chain. As introduced in more detail in Sec. 1.3.1, the calcium mediated gelation of alginate is highly dependent on the distribution of monomers (G and M units), leading to structural features such as the 'egg-box' conformation (two consecutive G units) [235]. Consequently, in the case of alginate the choice of dimers as primary building blocks is reasonable.[1] This abstraction leads to three types of dimer units: GG, MM, and GM. For simplicity, the MG dimer is considered equal to GM. The mass of each dimer unit is $350\,\mathrm{Da} = 5.812 \times 10^{-25}\,\mathrm{kg}$ and no orientation is modeled.

[1] Note that for other polymers abstractions of monomers or larger multimers as primary building blocks might be more appropriate.

TABLE 6.1: Overview of alginate gelation model parameters. Non-specified permutations are described as involving XX. Table adapted with permission from Depta *et al.* [224] under CC-BY 4.0 license.

Parameter	Value	Literature
Structural Model		
$d_{\mathrm{GG-GG}}$	0.92 nm	[232]
$d_{\mathrm{GG-MM}}$, d_{GG-GM}	0.97 nm	[232]
$d_{\mathrm{MM-MM}}$	1.01 nm	[232]
$d_{\mathrm{GM-GM}}$, $d_{\mathrm{MM-GM}}$	1.02 nm	[232]
Dimer mass m	350 Da	
Fiber molar mass	100 - 200 kDa	
Functional Model		
Diffusion and Thermodynamics		
Temperature T	300 K	
Dynamic viscosity η	$8.54 \times 10^{-4}\,\mathrm{Pa\,s}$	[329]
Stokes radius r_S	0.35 nm	[232]
Intermolecular Interaction		
Shape	Lennard-Jones	
Cutoff d_{cut}	1.5 nm	
Gradient limit ∇U_{max}	±500 kJ/mol/nm	
GG-GG d_m	0.6 nm	[243]
GG-GG ϵ	22 kJ/mol	attr.+rep., [239]
MM-MM d_m	0.7 nm	[243]
MM-MM ϵ	3 kJ/mol	attr.+rep., [239]
XX-XX d_m	0.7 nm	[243]
XX-XX ϵ	1 kJ/mol	rep. only
Ion Model		
$m_{\mathrm{Ca^{2+}}}$	40 u	
r_{ion}	1.2 Å	[330]
r_{dif}	0.3 Å	[331]
$d_{\mathrm{gap,min}}$	1 Å	
GG-GG r_{cavity}	$r_{\mathrm{ion}} = 1.2$ Å	
MM-MM r_{cavity}	0 Å	
XX-XX r_{cavity}	0 Å	
GG-GG r_{outer}	$r_{\mathrm{ion}} + r_{\mathrm{dif}} = 1.5$ Å	
MM-MM r_{outer}	$r_{\mathrm{ion}} + r_{\mathrm{dif}} = 1.5$ Å	
XX-XX	0 Å	
Bonded Interaction		
Normal stiffness $k_{\mathrm{bond,Fn}}$	$2.5\,\mathrm{N\,m^{-1}}$	
Bending stiffness $k_{\mathrm{bond,Mt}}$	$2.5 \times 10^{-19}\,\mathrm{N\,m\,rad^{-1}}$	[232, 332]

TABLE 6.2: Second-order Markov chain properties of alginate (fractions of classes) based on refs. [224, 234].

Main Fractions								
	ϕ_{GGG}	ϕ_{GGM}	ϕ_{GMG}	ϕ_{GMM}	ϕ_{MGG}	ϕ_{MGM}	ϕ_{MMG}	ϕ_{MMM}
High G (H)	0.475	0.050	0.055	0.050	0.050	0.055	0.050	0.215
Low G (L)	0.180	0.020	0.045	0.085	0.020	0.110	0.085	0.455
Resulting Fractions								
	ϕ_G	ϕ_M	ϕ_{GG}	ϕ_{GM}	ϕ_{MM}			
High G (H)	0.630	0.370	0.525	0.105	0.265			
Low G (L)	0.330	0.670	0.200	0.130	0.540			

In order to generate a polymer solution, random instances of polymer chains are necessary. For this, knowledge of the chain conformation (i.e. distance between dimer units), composition (i.e. sequence probabilities), and fiber molar mass (i.e. length) is necessary. The chain conformation was provided by Hecht *et al.* [232] and from this atomistic structure the equilibrium distance between dimer units was extracted (center of mass distance d_{i-j}), which can be found in the overview Tab. 6.1. These equilibrium distances and the resulting bending stiffnesses are in agreement with literature as poly-G > poly-M > alternating GM [232] (see also Sec. 1.3.1). Concerning fiber composition and molar mass, information is typically available or can be determined for most industrially used materials. For this work, the alginate properties of Agulhon *et al.* [234] in the corrected form of Depta *et al.* [224] were used, which can be found in Tab. 6.2. The composition is modeled by a second-order Markov process, which describes the fractions ϕ_{ijk} of ternary sequences i, j, k along the chain. Concerning the fiber molar mass, different masses between 100 - 200 kDa (285 - 571 dimers per polymer chain) were investigated.

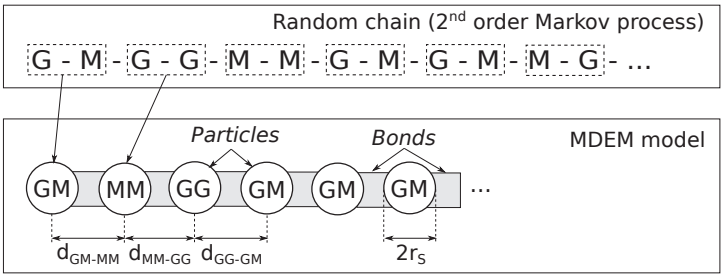


FIGURE 6.1: Structural model of the alginate polymer chain. Figure adapted with permission from Depta *et al.* [224] under CC-BY 4.0 license.

Based on this knowledge, a two-step process was employed for the generation of random polymer chains, which is visualized in Fig. 6.1. In a first step, random sequences of G and M monomers were generated as a function of the second-order Markov process and length of the polymer chain. In a second step, the generated monomer units were combined pairwise to the dimer units GG, MM, and GM/MG. The units were then

connected consecutively by elastic bonds thereby making up the polymer chain as shown in Fig. 6.1. Note that this two-step process enables the direct usage of typically available material parameters, but comes at the expense of an error when combining into dimer units. The resulting overestimation of GM units over GG and MM units is considered acceptable, as longer poly-G sections are crucial for stable gelation.

The gelation process of alginate is mediated in the model system by calcium ions. While it would be possible to model those calcium ions explicitly, i.e. in the structural model, it would significantly increase the computational requirements and decrease the required time step approximately by a factor of 5. Consequently, an implicit calcium model was formulated in Sec. 4.2, which will be parameterized in the following as part of the functional model components.

6.1.2 Functional Model

6.1.2.1 Diffusion and Thermodynamics

Diffusion and the respective thermodynamics of the desired canonical ensemble were modeled using the isotropic translational version of the presented diffusion model in Sec. 3.2.3. The model captures the interaction of the dimer units with the solvent environment and requires the objects masses and diffusion coefficients (or equivalent Stokes radii), as well as the solvent temperature and dynamic viscosity as parameters.

TABLE 6.3: Stokes radii of alginate system for all dimer units in x, y, z. Table adapted with permission from Depta *et al.* [224] under CC-BY 4.0 license. For the model, the average Stokes radius of 0.35 nm was used.

	$r_{S,x}$	$r_{S,y}$	$r_{S,z}$
GG	0.29 nm	0.34 nm	0.40 nm
MM	0.29 nm	0.41 nm	0.42 nm
GM	0.29 nm	0.35 nm	0.39 nm

In agreement with normal experimental conditions, simulation studies were carried out at an equilibrium temperature of 300 K and a resulting dynamic viscosity for water of $8.54 \times 10^{-4}\,\mathrm{Pa\,s}$ [329]. Furthermore, the dimer unit masses are readily available from their atomic structure as 350 Da. As shown in Sec. 3, the diffusion coefficients can for example be determined in detail using molecular dynamics simulations. These diffusion coefficients were found to be largely in agreement with structural predictions even for the highly anisotropic molecules of PDC. Consequently, for the purpose of this model, the diffusion coefficients / Stokes radii were estimated from the atomistic reference structure [232] by the following methodology: The unit dimers were oriented along their principle axes and the radius of gyration r_{gyr} calculated for all axes. Stokes

radii r_S were then estimated by accounting for the shell of hydration using the relation $r_S = r_{gyr}/0.77$ in each direction, which is widely employed for globular proteins. As it can be seen in Tab. 6.3, the resulting values vary between 0.29 - 0.42 nm indicating a low degree of anisotropy. Motivated by this, as well as model simplicity and computational performance, the average Stokes radius of 0.35 nm was used for all dimer units.

Simulated Annealing Simulated annealing procedures have been employed to improve the probability of reaching the global potential minimum through enhanced sampling and reduce overall simulation times to reach equilibrium. For details on simulated annealing please refer to Sec. 3.6. Different annealing procedures have been developed depending on the simulation stage (system generation, relaxation, equilibration, and production, see Sec. 6.2). Parameters of annealing procedures provided in Sec. 3.6 and used in the context of alginate gelation are:

- **AN1:** $\tau_{an,cool} = 5\,\mu s$, $\tau_{an,period} = 5\,\mu s$, $t_{an,finished} = 5\,\mu s$, linear temperature decay, $T_{an,max} = 2000\,K$ (used for relaxation step)
- **AN2:** $\tau_{an,cool} = 2\,\mu s$, $\tau_{an,period} = 2\,\mu s$, $t_{an,finished} = 10\,\mu s$, linear temperature decay, $T_{an,max} = 2000\,K$ (partially used for production step)

Validation of annealing procedure AN1 against standard conditions found equivalent system properties concerning fiber structures (see Sec. 6.1.2.3). Results of annealing procedure AN2 will be presented in direct comparison to runs without annealing in the results section. Note that unless specifically indicated, no annealing is performed.

6.1.2.2 Intermolecular Interaction

The pairwise intermolecular interaction of alginate dimers was captured using the model presented in Sec. 4.2. The intermolecular interaction depends on the dimer type, as well as presence of a calcium ion, which is modeled through an implicit binding (IM1) and unbinding (IM2) process. A schematic representation is reprinted in Fig. 6.2 and parameters will be derived in the following. All parameters are summarized in Tab. 6.1 and are based on literature and estimations. Future studies might determine some of the parameters in more detail, e.g. through steered MD. However, this is beyond the intended scope of this work.

Interaction Potential The gelation process of calcium mediated alginate is dominated largely by the dimerization of poly-G regions through an 'egg-box' association

between two GG dimers on interacting chains mediated by a central calcium ion, as established initially by Grant *et al.* [235]. Subsequently, a lateral association and zipping mechanism of the polymer chains occurs as described e.g. by Hecht *et al.* [232, 243]. This mechanism is attributed to the 'egg-box' structure of the binding zone encapsulating the ion, while other areas such as poly-M and and alternating GM aggregate less readily due to their flatter structure [232, 235, 243]. The binding enthalpy of this association mechanism has been experimentally measured by Fang *et al.* [239] using isothermal titration calorimetry. The binding enthalpy was found to be $\Delta H = -11.6$ kJ/mol for calcium alginate with a G fraction of 46% and $\Delta H = -15.0$ kJ/mol for a system with a G content of 64% (Gibb's free energy $\Delta G = -23.0$ kJ/mol in both cases) [239]. Modeling works of the 'egg-box' binding mechanism have found similar, but slightly lower Gibb's free energies in the range of $\Delta G = -35$ kJ/mol to -60 kJ/mol [237, 240].

These mechanisms are replicated in the intermolecular interaction model through the binding energy and equilibrium distance dependent on the type of interaction and ion availability. In order to estimate the binding energies, the experimental measurements by Fang *et al.* [239] were used. Extrapolating the measured G content to 0 and 1 was used to estimate the binding enthalpy for pure poly-G as -22 kJ/mol and for pure poly-M as -3 kJ/mol, which was used for the GG-GG and MM-MM interaction, respectively. These values are reasonable in regard to the previously discussed association mechanisms. As other interaction permutations are understood to aggregate significantly less readily, these interactions were set to be repulsive only with a scaling factor ϵ = -1 kJ/mol. With regard to estimating the equilibrium distance d_m during association, the work by Hecht [243] on investigating the composition dependent associating of alginate using MD was used. The estimated equilibrium distances based on this were found to be $d_m = 6$ Å for GG-GG and $d_m = 7$ Å for MM-MM and all other permutations.

Based on the interaction potential (eq. 4.2) with these parameters, the cutoff distance was chosen in agreement with literature [333] as 1.5 nm ($2.1 \times d_m$ to $2.5 \times d_m$). Furthermore, the gradient limit $\nabla U_{\max}$ was chosen as ± 500 kJ/mol/nm to avoid numerical instabilities resulting from the singularity at zero distance. $\nabla U_{\max}$ was reached only very rarely at the beginning of equilibration.

Ion Model In order for a GG-GG or MM-MM interaction to be attractive, a calcium ion needs to be present. Consequently, the ion model is crucial for describing the ion availability, binding, and unbinding. Thus, controlling both dynamics and steady-state of gel formation and breakage. Two separate models have been proposed in Sec. 4.2.2 to model ion binding (IM1) and unbinding (IM2). Concerning parameterization, the ion binding model IM1 requires parameters with regard to the ion type (Ca^{2+}), conformation

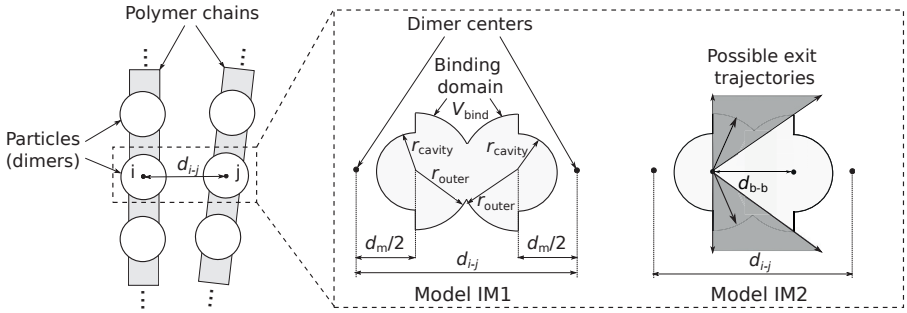


FIGURE 6.2: Visualization of intermolecular interaction and ion models IM1 and IM2 for alginate gelation (reprint of Fig. 4.1 with permission from Depta *et al.* [224] under CC-BY 4.0 license).

of interaction partners, and simulation time step to estimate the binding volume. The binding site volume is considered to contain an electrostatic and diffusive component. The electrostatic component is estimated using the ionic radius $r_{\text{ion}} \approx 1.2\,\text{Å}$ for calcium, which is experimentally known to be in the region $1.0 - 1.4\,\text{Å}$ according to Shannon *et al.* [330].[2] In addition, the time step dependent diffusive component is estimated as the average of the diffusive length based on the diffusion coefficient according to ref. [331] ($0.16\,\text{Å}$ for $\Delta t = 10^{-13}$s) and the characteristic length of mean velocity ($0.43\,\text{Å}$ for $\Delta t = 10^{-13}$s), leading to $r_{\text{dif}} = 0.3\,\text{Å}$ for $\Delta t = 10^{-13}$s. This diffusive component was added to the outer half-sphere of the interaction region in addition to the electrostatic component.

Depending on the conformation of the interaction zone, the binding domain can extend into a cavity of the interacting molecules (see r_{cavity}). For the 'egg-box' conformation of the GG-GG interaction, the cavity is approximately as large as the ionic radius leading to the estimation $r_{\text{cavity}} \approx r_{\text{ion}} = 1.2\,\text{Å}$. For the flatter conformation of the MM-MM interaction no such cavity exists and only the outer half-sphere is considered as the sum of the ionic and diffusive radius $r_{\text{outer}} = r_{\text{ion}} + r_{\text{dif}} = 1.5\,\text{Å}$.

Additionally, the binding model describes the encapsulation of the ion through the conformation of the interacting molecules for small distances. This is achieved by a required minimum gap of $d_{\text{gap,min}} = 1\,\text{Å}$ beyond the equilibrium distance to permit entry and exit of the ion. The proposed value of $d_{\text{gap,min}} = 1\,\text{Å}$ is slightly lower than $r_{\text{ion}} = 1.2\,\text{Å}$ in order to account for slight conformational changes of the interacting molecules.

Overall, this leads to a binding domain volume $V_{\text{bind},i-j}$ as a function of center of mass distance as shown in Fig. 6.3 for IM1. As an additional parameter for the unbinding

[2]Note that hydration effects of the calcium ion were neglected. Future studies might investigate this influence by increasing the effective ionic radius in regard of strong dissociation energies of water molecules from calcium ions($\sim$ 21.9 kcal/mol for first water molecule and $\sim$ 26.3 kcal/mol for second according to ref. [334]).

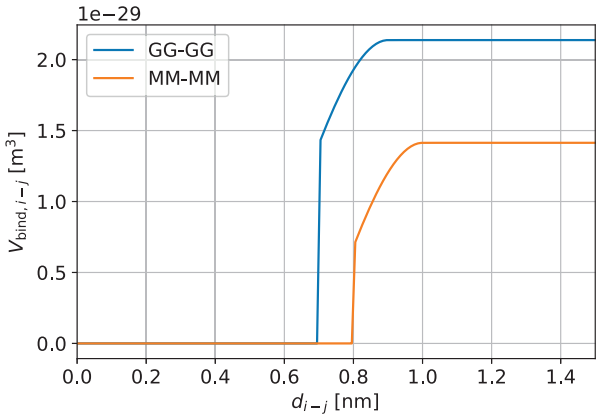


FIGURE 6.3: Volume of binding domain for pairwise interaction at a temperature of 300 K with a time step of 10^{-13} s using ion model IM1.

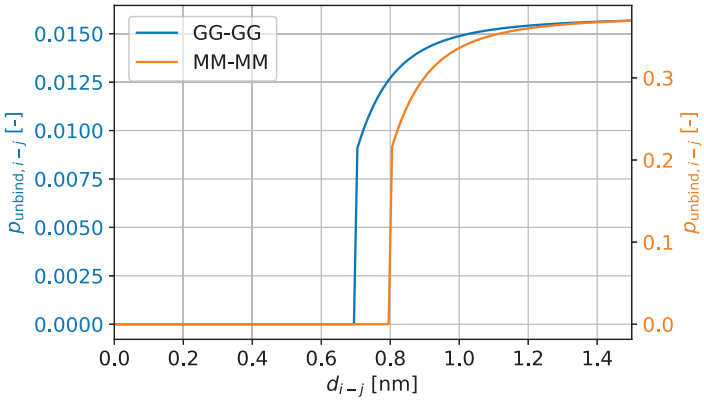


FIGURE 6.4: Ion unbinding probability for pairwise interaction at a temperature of 300 K with a time step of 10^{-13} s using ion model IM2.

model IM2 only the ion mass of $m_{\mathrm{Ca}^{2+}} = 40\,\mathrm{u} = 6.64 \times 10^{-26}\,\mathrm{kg}$ is required. The unbinding probability as a function of distance d_{i-j} is shown in Fig. 6.4. During simulation, all interaction forces were tabulated with a resolution of 0.01 Å and all binding/unbinding probabilities with a resolution of 0.1 Å. An overview of all parameters is provided in Tab. 6.1.

6.1.2.3 Bonded Interaction

The elastic bonds connecting the dimer units along the polymer chains were described using the bonded interaction model presented in Sec. 5.3. The model contains the normal stiffness $k_{\mathrm{bond},Fn}$ and bending stiffness $k_{\mathrm{bond},Mt}$ as parameters to be specified. These parameters were chosen to be universal for connections of all dimer unit types, leading implicitly to stiffnesses in order poly-G > poly-M > alternating GM, which is consistent with literature [232] (see also Sec. 1.3.1). These parameters could be determined through a variety of approaches, e.g. steered MD or umbrella sampling. For this work, an optimization approach using experimental data and the employed diffusion model was chosen. The bond-angle correlation (BAC) [332] and resulting persistence length l_p of the polymer chain's shape was chosen as an optimization target for the **bending stiffness** and tuned to the experimental data of ref. [232]. The BAC is defined as

$$BAC(i) = \left\langle \frac{\vec{b}_j \cdot \vec{b}_{j+i}}{\|\vec{b}_j\| \cdot \|\vec{b}_{j+i}\|} \right\rangle = \langle \cos(\theta) \rangle, \tag{6.1}$$

$$BAC(x) = \exp\left(\frac{-x\sigma}{l_p}\right), \tag{6.2}$$

being the scalar product of normalized bond vectors $\vec{b}$ at increasing distances along the chain [232, 332]. It exhibits an exponential decay at increasing distances and is characterized through the persistence length l_p. Experimental results from ref. [232] predict a l_p between 20 - 30 nm for poly-M and poly-G fibers at saturated calcium conditions, as well as a $l_p < 5$ - 10 nm for alternating GM. Consequently, the optimization target was set between l_p = 15 - 20 nm for the primary model system high-G (H). Concerning the **normal stiffness**, the bond strain was used as an optimization target. However, no literature data was available concerning polymer fiber stretching of alginate, which is attributed to the stable structure and inherently strong C-O-C bond between monomers. Consequently, the optimization target was defined more flexibly with a reasonably stable bond of near zero mean strain and maximum strain < 0.25. In order to narrow the stiffness parameter range and initial estimation, energy estimations were performed based on the thermodynamic energy of the dimer units. Afterwards, free diffusion experiments (dilute solution with no interaction between fibers) were conducted with 100 polymer fibers of 200 kDa for 50 µs using a time step of 10^{-13} s (additionally 5×10^{-14} s convergence study). As the best stiffness parameters the values of $k_{\mathrm{bond,Fn}} = 2.5\,\mathrm{N\,m^{-1}}$ and $k_{\mathrm{bond,Mt}} = 2.5 \times 10^{-19}\,\mathrm{N\,m\,rad^{-1}}$ were found. For this parameterization the persistence length l_p is 16 nm and the normal bond strain has an average value of 0.004 (minimum -0.16, maximum 0.22).

6.1.2.4 Critical Time Step

In order to estimate the critical time step τ_{crit} necessary for a convergent and numerically stable solution, the previously discussed methods for the individual model components were used, see Sec. 3.4.1, 4.2.3, and 5.3.2. Based on this, the critical time step can be estimated as 1.0×10^{-13} s for the diffusion model, 7.3×10^{-13} s for the intermolecular interaction model, 3.0×10^{-12} s for the normal direction of the bond model, and 6.2×10^{-12} s for the bending direction of the bond model. For bonded interactions in discrete simulations with explicit time-stepping typically a simulation time step of $0.2 \times \tau_{\text{crit}}$ is recommended [42], which is fulfilled by a simulation time step of 10^{-13} s. Concerning diffusion and the related thermodynamics, this is also at the advisable limit with an error of 8.0 % regarding RMS displacement (see Sec. 3.4.1). Thus, the simulation time step was chosen as 10^{-13} s and an additional convergence study at 5×10^{-14} s showed no notable differences.

6.2 Simulation Setup and Procedure

A variety of simulation setups with varying polymer and ion conditions was investigated in cubic simulation domains with periodic boundary conditions to study the gelation process. Details on the MDEM implementation can be found in Sec. 2.2 and all model parameters are provided in Tab. 6.1. The simulation procedure is closely related to the widely employed workflows in molecular dynamics and consists of the following four steps:

1. **System generation:** A system of defined box size, polymer concentration, fiber composition, and fiber molar mass is generated. Each fiber is generated with a random composition resulting from the respective second-order Markov process as specified in Sec. 6.1.1. The fibers are positioned and oriented randomly in the simulation box and may cross any periodic boundary condition. All polymer fibers are initially perfectly linear.

2. **Relaxation:** The initially perfectly linear polymer fibers are relaxed using the employed diffusion model without interacting with each other (similar to a perfectly dilute system). The fibers transition from their artificial state of generation to their natural state with respective to curvature / BAC. Annealing procedure AN1 (see Sec. 6.1.2.1) was validated and employed, which enabled a reduction of required simulation time from approximately 50 µs without annealing to 15 µs with annealing.

3. **Equilibration:** The system is equilibrated for 5 µs without any annealing to correct possible overlaps between fibers and reach its natural state without calcium ions in the system. Accordingly, all intermolecular interaction models are set to be repulsive only, i.e. without any ions present.

4. **Production (Gelation):** The desired concentration of calcium ions is added to the system and the full intermolecular interaction model employed (IM1 or IM2) representing the calcium mediated gelation. As the gelation process is more prone to run into local potential minima during network formation, both normal simulation (no annealing) and annealing procedure AN2 (see Sec. 6.1.2.1) have been performed and will be compared in the results section.

Note that all simulation times were chosen such as to ensure equilibration of the respective stage, i.e. steady-state conditions with respect to the fast modes of relaxation of various global properties (e.g. global interaction potential, bond potential, kinetic energy, BAC; see following section).

6.3 Analysis and Postprocessing

Simulation results were analyzed both online and in postprocessing to study a variety of energetic and structural features. The analyzed system properties include the following and are taken from Depta *et al.* [224]:

- global system energy U in components of the potential energy $U_{\text{glob.int.}}$ due to pairwise interaction (normalized by number of dimers/particles N_{part}), bonded potential energy $U_{\text{glob.bond.}}$, and kinetic energy $U_{\text{glob.kin.}}$;
- ion ratio f of used f_{used}, available $f_{\text{avail.}}$, or desired ions $f_{\text{des.}}$. Desired ions constitute unbonded contacts capable of receiving an ion;
- normalized contact number $\xi_{i-j} = 2N_{\text{cont},i-j}/\sqrt{N_i N_j}$, where $N_{\text{cont},i-j}$ is the number of contacts between type i and j based on the interaction cutoff of 1.5 nm and N_i (N_j) is the number of particles of type i (j);
- coordination number with regard to its average $\overline{\text{k}}_{0.9\text{nm}}$ and histogram fractions $\phi(\text{k}_{0.9\text{nm}})$ based on a reduced contact distance of 0.9 nm to capture only contacts in close proximity. Note that the average coordination number is inversely proportional to the specific surface of the polymer;
- volume fractions ϕ_V of spatially-resolved dimer/particle concentrations. A discretization of the simulation domain into 5 nm cubes on a regular grid is applied.

The concentration in each cell is calculated as the number of dimers/particles $N_{di.}$ divided by the cell volume $V_{el.}$. From this discretization, volume fractions of pores (empty cells) and fiber bundles (cells with high concentration) were calculated;

- pore size distribution based on chord length distribution for comparison with experimental data. The *Python* library *Porespy* was used to calculate the chord length distribution based on the previous 5 nm discretization applying various limits as to whether a cell is occupied or not for sensitivity analysis;
- thickness of fiber bundles through the number of fibers and bundle diameter (maximum distance between dimers/particles in bundle cross-section) for comparison with experimental data. A contact network search algorithm determined the dimers/particles in the bundle cross-section for each particle in the system, while only considering the closest contact (maximum contact distance 0.9 nm as for coordination number) between dimers/particles of two fibers and only passing each fiber once.

Based on this, the results will be provided in the following. For a selected number of simulations the full transient data is provided. For the full transient data of all simulations the interested reader is referred to the supplementary data of the respective publication Depta *et al.* [224].

6.4 Results

A total of 33 case studies were simulated to study the mechanisms of network formation during gelation and validate the modeling framework. An overview of all conditions can be found in Tab. 6.4. All studies were performed in 0.5 µm cubic boxes with periodic boundary conditions in all directions. Primary focus of the case studies was placed on investigating the influence of the calcium concentration ($f = 0.1 - 1.0$). Additionally, the polymer concentration ($c = 0.5 - 1.0$ wt.%), polymer composition (high and low G content), and molecular weight ($M_w = 100 - 200$ kDa) were varied. Lastly, the influence of unbinding of ions in addition to binding (IM2 vs. IM1), as well as possible improvements through annealing procedures (AN2) were investigated. Note that in addition to the gelled conditions, the equilibrated calcium-free conditions provide additional insights and data.

TABLE 6.4: Case study overview with base case marked gray. All cases are in a 500 nm cubic box with periodic boundary conditions. Table adapted with permission from Depta *et al.* [224] under CC-BY 4.0 license.

Case	**Model**	**c [wt.%]**	**G cont. (H/L)**	**M_w [kDa]**	**f [-]**	**Annealing**	**t_{sim} [µs]**
1	IM1	0.5	H	200	0.1	no	10
2	IM1	0.5	H	200	0.2	no	10
3	IM1	0.5	H	200	0.35	no	10
4	**IM1**	**0.5**	**H**	**200**	**0.5**	**no**	**10**
5	IM1	0.5	H	200	1.0	no	10
6	IM1	0.5	L	200	0.1	no	10
7	IM1	0.5	L	200	0.2	no	10
8	IM1	0.5	L	200	0.35	no	10
9	IM1	0.5	L	200	0.5	no	10
10	IM1	0.5	L	200	1.0	no	10
11	IM1	0.5	H	100	0.1	no	10
12	IM1	0.5	H	100	0.2	no	10
13	IM1	0.5	H	100	0.35	no	10
14	IM1	0.5	H	100	0.5	no	10
15	IM1	0.5	H	100	1.0	no	10
16	IM2	0.5	H	200	0.1	no	40
17	IM2	0.5	H	200	0.2	no	40
18	IM2	0.5	H	200	0.35	no	40
19	IM2	0.5	H	200	0.5	no	40
20	IM2	0.5	H	200	1.0	no	40
21	IM1	1.0	H	200	0.1	no	9.4
22	IM1	1.0	H	200	0.2	no	10
23	IM1	1.0	H	200	0.35	no	10
24	IM1	1.0	H	200	0.5	no	10
25	IM1	1.0	H	200	1.0	no	10
26	IM1	0.5	H	200	0.1	AN2	15
27	IM1	0.5	H	200	0.2	AN2	15
28	IM1	0.5	H	200	0.5	AN2	20
29	IM1	0.5	H	200	1.0	AN2	20
30	IM1	0.5	L	200	0.5	AN2	20
31	IM1	0.5	H	100	0.5	AN2	20
32	IM2	0.5	H	200	0.5	AN2	40
33	IM1	1.0	H	155	0.5	no	10

6.4.1 Base Case

The fully gelled case 4 marked grey in Tab. 6.4 was chosen as a base case. A visualization of a 200 nm cross-section shown in Fig. 6.5 highlights the difference in structural features between the non-gelled (A and B) and gelled state (C and D). Additionally, the transient development of system properties is shown in Fig. 6.6 including all potential components (A), ion balances (B), average coordinate number (C), distribution of coordination number (D), normalized contact types (E), and volume fractions of dimer concentrations (F).

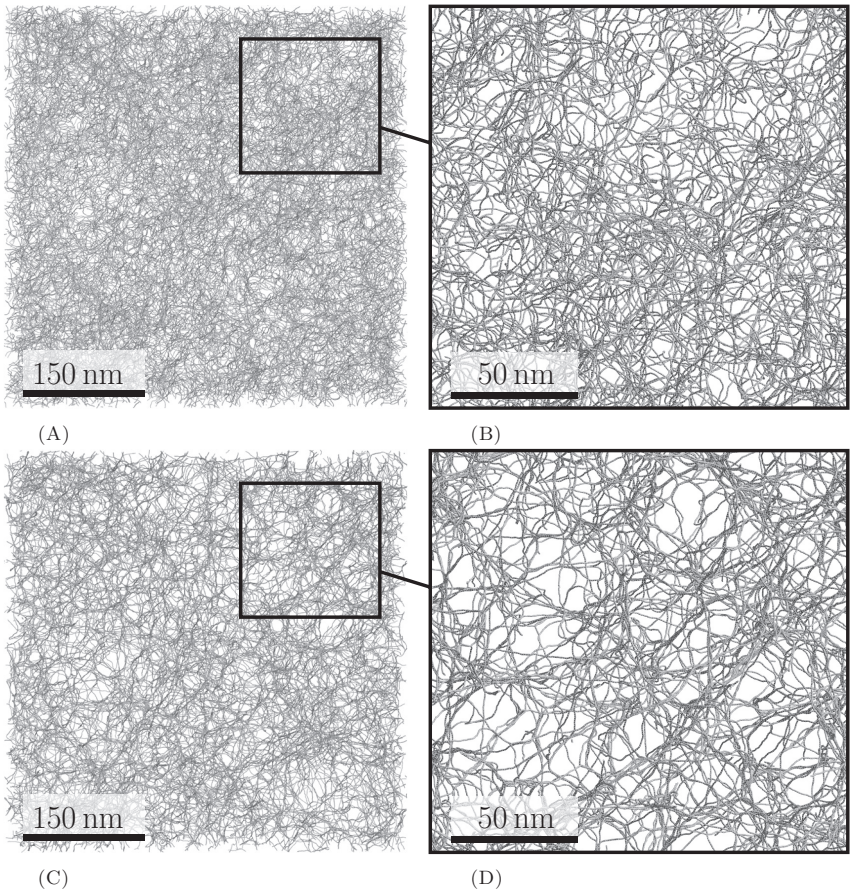


FIGURE 6.5: Visual comparison of a polymer solution (individual fibers) at concentration of 0.5 wt.% with high G content in its **non-gelled state (A and B) and its gelled state using f = 0.5 ions (C and D)**, see base case 4. A 200 nm depth section is shown and bonds visualized with a 0.7 nm diameter. Figure adapted with permission from Depta *et al.* [224] under CC-BY 4.0 license.

As it can be seen in Fig. 6.5 A and B, the non-gelled solution displays a very homogeneous state of individual fibers with fine spaces between them. After gelation of the system through the addition of calcium ions, the system (see Fig. 6.5 C and D) exhibits a network formation of the fibers into bundles, as well as the formation of pores. As shown in Fig. 6.6 D through the distribution of the coordination number, approximately half of all dimers participate in a connection indicating a densely interconnected gel. The consequently formed pores are highly anisotropic and largely on scales of O(10) nm with the largest fraction of pores at a size of approximately 45 nm as defined by the chord length distribution. Due to the high anisotropy and inter-connectivity of pores a precise quantification is challenging. In order to alleviate this in the following, the transient data in Fig. 6.6 will be discussed in more detail including the volume fraction of empty pores, which represents a more universal measure.

As it can be observed in the transient behavior of all system properties shown in Fig. 6.6, the system reaches steady-state in approximately 4 µs. At this point the kinetics of gelation significantly slow down and exhibit a steady-state with regard to the fast modes of equilibration. It can be observed that only $f = 0.34$ ions are used during the gelation process, while $f = 0.16$ ions remain available (unbound in solution). It should be noted that at the same time $f = 0.17$ ions are desired by intermolecular GG-GG and MM-MM interactions. This result indicates that during the zipping mechanism responsible for bundle formation approximately every third GG-GG and MM-MM interaction remains without a calcium ion as none was available until closure and the (modeled) inability to received an ion. It appears plausible that over time through thermodynamic effects such connections might still receive an ion and that such a process occurs both numerically and physically. However, the respective time scales of such a process are likely well beyond current numerical possibilities. Additionally, deviations of the simplified model from the physical system are likely to some extend in this respect.

Moreover, structural aspects of the bundle and network formation can be observed through the coordination number shown in Fig. 6.6 C and D. In the initial non-gelled state virtually no intermolecular contacts exist. Over time, the average coordination number increases to 1.3 with 46 % of dimers having a coordination number of at least one. The majority of coordination numbers with 93 % remains below 5 and 99.98 % below 10. Individual coordination numbers can go as high as 16, which is attributed to individual compressed junction zones of multiple bundles. Furthermore, it can be seen in Fig. 6.6 E that an average of 2.74 attractive 'egg-box' contacts are entered by each GG-dimer, indicating a strong multi-fiber bundle formation, which is in agreement with experimental expectations (more detail on experimental validation in Sec. 6.5).

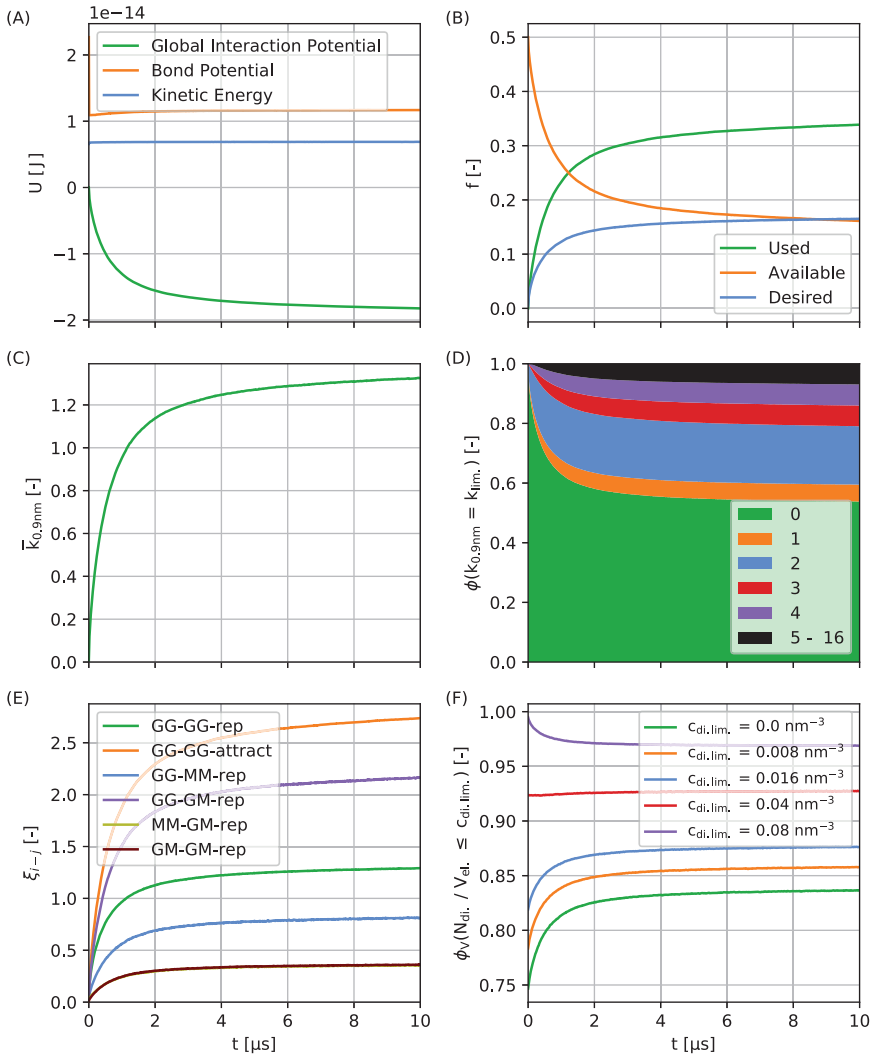


FIGURE 6.6: Transient simulation results of the base case 4. See Sec. 6.3 for more details on the properties. Figure adapted with permission from Depta *et al.* [224] under CC-BY 4.0 license.

In addition, the volume fractions at specific concentration limits in Fig. 6.6 F show the development of pore and bundle volume fractions. It can be observed that the pore volume fraction (empty cells) increases from 0.75 to 0.84, while the volume fraction of concentrations higher than 0.08 nm^{-3} (10 dimers per 5 nm cell) also increases from 0.005 to 0.03. These trends of pore volume generation through bundle formation and high concentration regions is in agreement with expectations during the gelation process.

In summary, it should be noted that the presented simulation results represent the ideally homogeneous case of gelation. Experimental conditions are expected to contain more inhomogeneities resulting from material composition (e.g. molecular weight), processing (e.g shear, dispersion of ions), and impurities, thus leading to larger bundle and pore sizes. More details will be discussed in relation to experimental validation in Sec. 6.5.

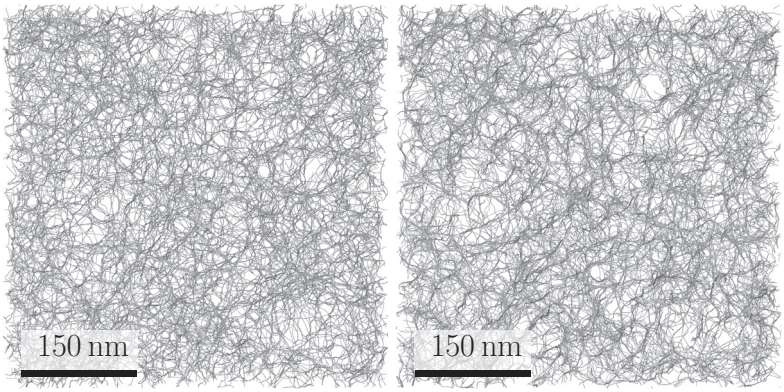
6.4.2 Case Studies

6.4.2.1 Constant Temperature (No Annealing)

During the network formation of polymer gelation, the system's temperature is a critical parameter. In this regard, numerical studies at constant temperature without annealing present the lower limit of network formation during the polymer gelation process due to being more limited by local potential minima and consequently equilibration at the time scales numerically reachable. In this context, these studies are crucial in understanding the mechanisms of gelation and its numerical modeling. The gelled systems at constant temperature are visualized in Fig. 6.7 and will be discussed first qualitatively before a quantitative analysis of gelation properties, which are shown in Fig. 6.8.

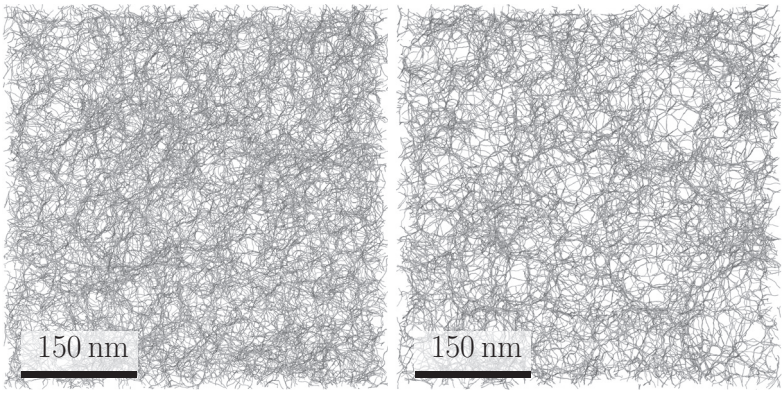
The visualizations in Fig. 6.7 show the various system conditions studied at sufficiently large ion concentration to enable gelation ($f = 0.5$). As it can be observed when comparing the binding model IM1 (A) and unbinding model IM2 (B), both provide visually similar polymer networks with somewhat larger pores and bundle sizes of the unbinding model IM2 (B). Additionally, it can be observed that a composition with less G content leads to smaller bundles and pore sizes (C). In an opposite fashion, a decrease in molecular weight to 100 kDa appears to increase bundle and pore sizes (D). Lastly, an increase in polymer concentration to 1.0 wt.% (E and F) visually leads to a denser polymer network and no clear difference between a molecular weight of 155 kDa and 200 kDa (E and F) is visible. As will be discussed with regards to experimental validation in Sec. 6.5, literature has largely addressed how the degree of cross-linking through ion concentration f influences the bundle sizes, but the role of material properties such as molecular weight and composition (i.e. G content) has not been addressed in literature to the author's best knowledge [224]. Thus, predictions using the proposed model may lay the basis for future experimental works. In the following, the effects of polymer concentration, molecular weight, fiber composition, and ion concentration on gelation will be discussed in more quantitative detail based on the proposed model.

With regard to the **200 kDa and high G composition system** (see Fig. 6.7A and blue circular data in Fig. 6.8), all system properties show an asymptotic behavior beyond



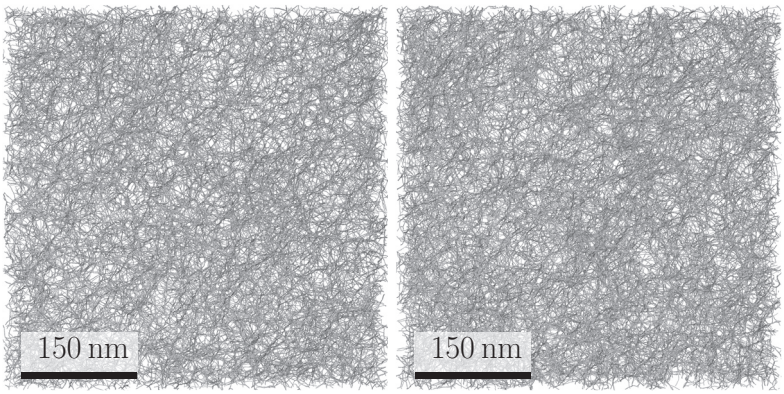
(A) Case 4: IM1, 0.5 wt.%, H, 200 kDa.

(B) Case 19: IM2, 0.5 wt.%, H, 200 kDa.



(C) Case 9: IM1, 0.5 wt.%, L, 200 kDa.

(D) Case 14: IM1, 0.5 wt.%, H, 100 kDa.



(E) Case 24: IM1, 1.0 wt.%, H, 200 kDa.

(F) Case 33: IM1, 1.0 wt.%, H, 155 kDa.

FIGURE 6.7: Visualization of gelled polymer cases **without annealing** with an ion concentration of $f = 0.5$. A 200 nm depth section is shown and polymer fiber bonds visualized with a 0.7 nm diameter. Partial data has been previously published and is reused with permission from Depta *et al.* [224] under CC-BY 4.0 license.

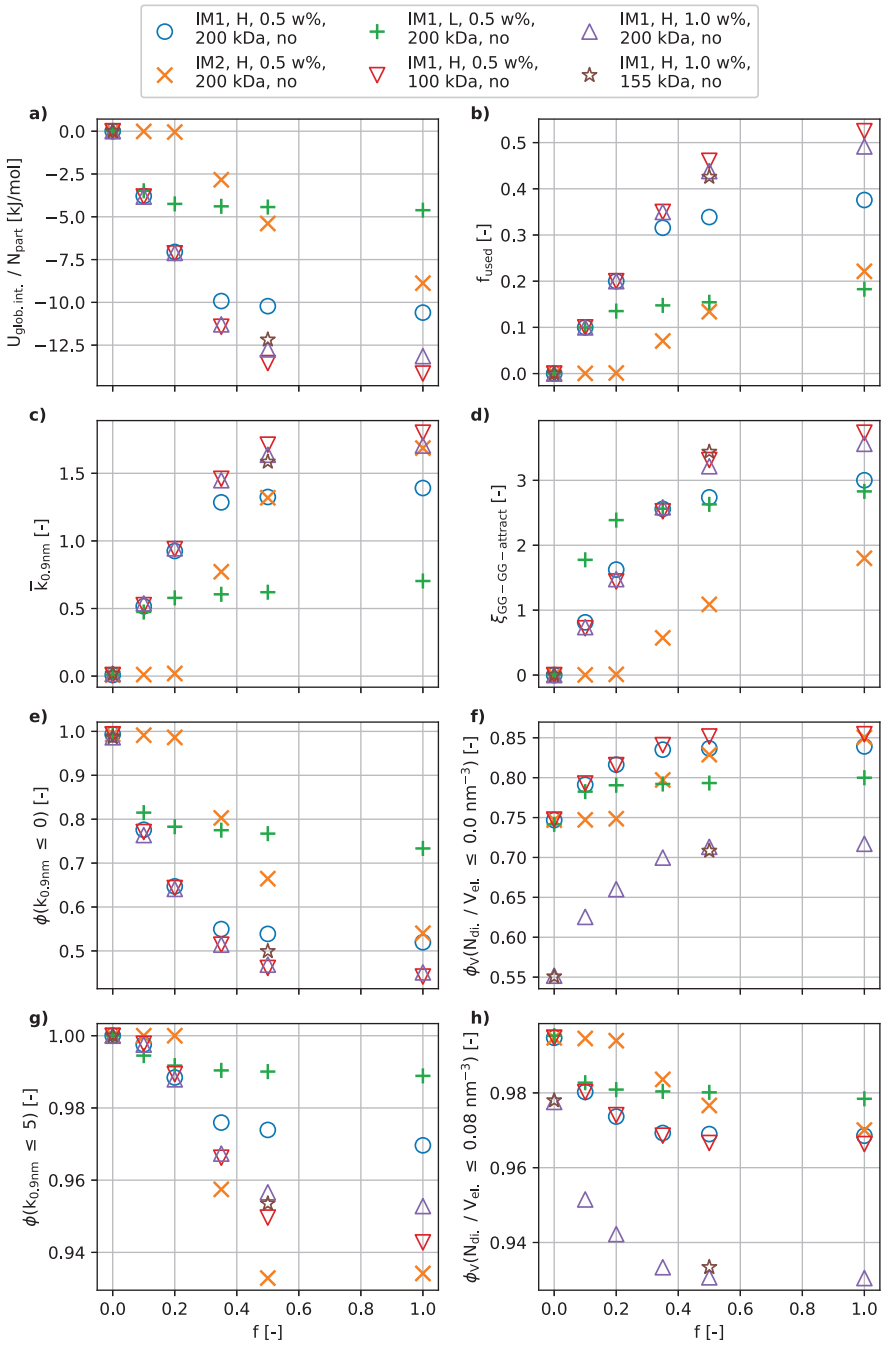


FIGURE 6.8: Simulation results of all case studies **without annealing**. Partial data has been previously published and is reused with permission from Depta *et al.* [224] under CC-BY 4.0 license.

a calcium concentration of $f = 0.35$. This transition point is further characterized and supported by the uptake of calcium ions between $f_{\text{used}} = 0.34 - 0.38$ for $f = 0.5$ and 1.0, respectively. While the gelled system appears to be saturated of calcium ions at this point, closed off interaction zones remain resulting from the limited availability during the zipping mechanism and might bind further calcium ions over time. Nonetheless, no changes in gel structure are expected. Consequently, the discussion of the base case in Sec. 6.4.1 applies to the asymptotic case.

With regard to the **200 kDa and low G composition system** (see Fig. 6.7C and green cross data in Fig. 6.8), it can be observed that the asymptotic contact number $\xi_{\text{GG-GG-attract}}$ is approximately the same as for the high G composition indicating the same relative number of 'egg-box' interactions per GG dimer, thus leading to an overall reduction with decreasing G content. This leads to a less densely connected gel structure as also highlighted by the lower coordination number (C) and reduced pore volume (D), as well as decreased uptake of ions (B) and increased global interaction potential (A). These effects of G composition are expected as the network formation of alginate gelation is known to be strongly dependent on the formation of GG-GG 'egg-box' connections [235] and thus requiring a large G composition for densely connected gels. Furthermore, an earlier asymptotic behavior with regard to ion concentration at approximately $f = 0.1$ - 0.2 can be observed and is in line with the reduced uptake of ions (B) at $f_{\text{used}} = 0.15$.

With regard to the system of reduced molecular weight at **100 kDa and high G composition** (see Fig. 6.7D and red lower triangle data in Fig. 6.8), it can be observed that the reduced molecular weight leads to an increased bundle thickness and more densely connected gel as indicated by the increase average coordination number (C) and increased fraction of coordination numbers above five (G). In a similar direction, the lower molecular weight uses significantly more ions (B), which are largely bound by additional 'egg-box' interactions (D) and further lower the global interaction potential (A). Furthermore, the pore volume increases slightly (F) highlighting the densification of the gel. These effects are in line with expectations resulting from the increased mobility of polymer fibers caused by decreased length. As the polymer fibers are less constrained due to cross-linking, network formation through bundle formation is supported and in turn increases pore size.

With regard to an **increased polymer concentration of 1 wt.% at constant composition** (see Fig. 6.7E and violet upper triangle data in Fig. 6.8), it can be observed that concerning bundle thickness nearly all parameters show similar values to those of a decreased molecular weight of 100 kDa at 0.5 wt.%. This applies to the average coordination number (C), fraction of coordination numbers at zero (E), number of 'egg-box' formations (D), uptake of ions (B), and global interaction potential (A). Only for the

fraction of higher coordination numbers (G) a less pronounced increase relative to the lower molecular weight can be observed, which is in line with the more constrained polymer fibers. However, the differences in pore volume relative to all previous systems are significant. Due to the increased polymer concentration the initial pore volume without ions causing gelation is significantly lower with a fraction of 0.55 (F) and also the pore volume of the geled system is lower with a fraction of 0.71. At the same time, the volume fraction of bundles above a concentration of $0.08\,\mathrm{nm}^{-3}$ (10 dimers per 5 nm cell) is more than doubled. Consequently, the increased polymer concentration leads to a decrease in pore volume and increase in total number of polymer bundles, while the properties of bundles are similar to a lower molecular weight of 100 kDa at 0.5 wt.%.

Concerning the **increased polymer concentration of 1 wt.% at intermediate molecular weight of 155 kDa** (see Fig. 6.7F and brown star data in Fig. 6.8), no significant differences from the change in molecular weight relative to 200 kDa can be noted.

For the **ion unbinding model IM2** (see Fig. 6.7B and orange cross data in Fig. 6.8), some difference concerning polymer bundle and pore formation are notable. The unbinding model IM2 attempts to describe the loss of a bound ion from interacting dimers and thus model the dynamic equilibrium of ion binding and unbinding. Convergence studies with respect to equilibration times found an increase in required simulation time by a factor 2-4, thus leading to computational requirements of approximately 17 days for a data point of 40 µs on a Nvidia A100 GPU. Therefore, only selected simulations were carried out for the base case system (orange cross data in Fig. 6.8). Due to the comparative nature between IM1 and IM2, some experimental literature references are provided below and extended in Sec. 6.5. As it can be seen, between $f = 0.2$ - 0.35 ions are required to initiate gelation. Below this point, ion unbinding is too frequent for permanent network formation to take place. This minimum ion concentration is in agreement with experimental findings by Fang *et al.* [239] (dilute systems) who found a minimum $f = 0.16$ - 0.35 for 'egg-box' formation to take place at the studied polymer composition. Similarly, Fang *et al.* [239] predicts an asymptotic behavior starting around $f = 0.5$. While this agrees well with IM1, the unbinding model IM2 shows less of an asymptotic behavior indicating a possible overestimation of the unbinding probability. Nonetheless, the unbinding model predicts largely similar trends for the average gel structure (bundles and pores): While the ion uptake and 'egg-box' interactions with bound ions are reduced (A, B, D), the average coordination number (C) and pore volume (F) indicate similarity in average gel structure. However, the distribution of bundles shows slight differences. The fraction of dimers with a coordination number of zero (E) increases indicating a decreased cross-linkage of individual fibers, while in turn also leading to an increased fraction of thicker bundles (G). Note that these thicker bundles fill a smaller

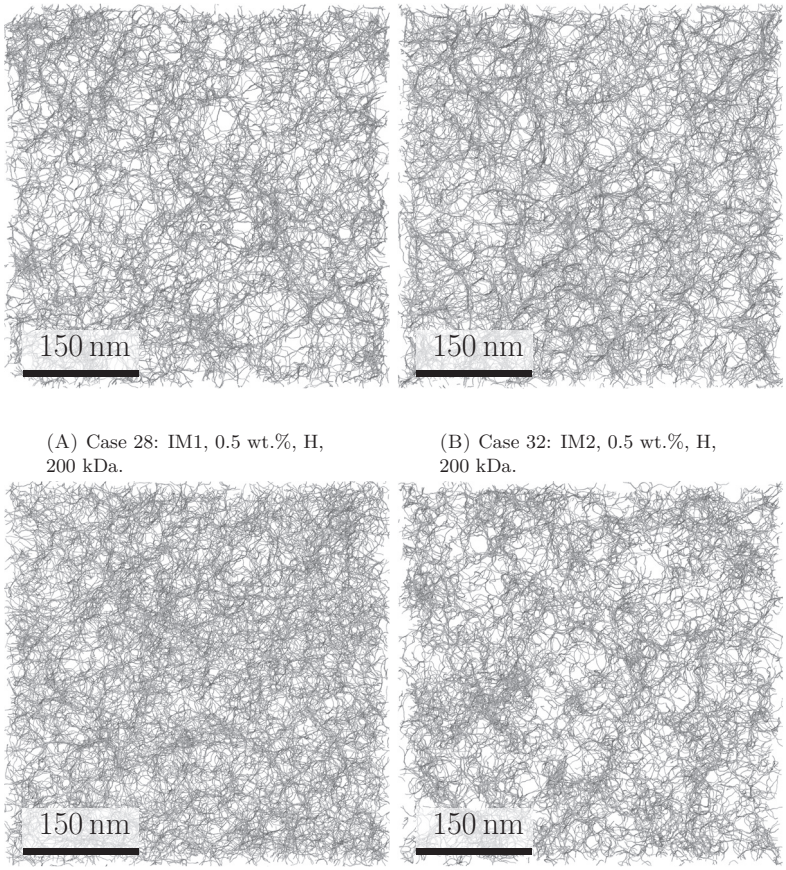
space fraction in contrast to the base case as shown in H. In summary, the developed unbinding model IM2 predicts slightly different gels. Concerning ion uptake, the unbinding model captures the physically more reasonable case of requiring a minimum ion concentration for gelation to initiate. For network formation, average structures are overall similar with respect to pore volume and coordination numbers, but result in a distribution of slightly thicker polymer bundles. In light of limited experimental insight into the detailed mechanisms, no clear assessment as to which model is more appropriate can be made at this point. Indications from more limited asymptotic behavior in relation to experimental literature data [239] suggest that unbinding might be overestimated by IM2 and thus overall be between IM1 and IM2. Detailed investigation of these mechanisms requires further research beyond the scope of this work. In view of computational cost and the prediction of reasonable experimental and empirical trends, the simple binding model IM1 was chosen as the primary ion model.

6.4.2.2 Annealing Procedure (AN2)

As the constant temperature procedure represents largely the lower limit of gelation with regards to network and bundle formation, annealing attempts to overcome these limitations by temporarily increasing the kinetic energy through temperature and thus transitioning local potential minima towards a more equilibrated and natural state (see also Sec. 3.6 and 6.1.2.1). At the same time, annealing comes at the cost of additional computational requirements and possible introduction of artifacts. Accordingly, the results of annealing are presented separately for a range of conditions in this method-developing work. Visualizations of the gelled systems with annealing procedure AN2 are provided in Fig. 6.9 and changes in system properties in comparison to constant temperature provided in Fig. 6.10, which will be discussed in the following.

Before discussing the results quantitatively, a qualitative visual comparison of Fig. 6.9 at $f = 0.5$ with the constant temperature results in Fig. 6.7 indicates that the unbinding model IM2 (B) and low G composition (C) provide largely similar results, while the base case (A) and to a lesser extend the lower molecular weight case (D) produce slightly larger pore and bundle sizes. Additionally, the junction zones at lower molecular weight (D) appear slightly more disorganized, indicating structural changes and possibly artifacts. These differences will be quantified in the following and an overview is provided in Fig. 6.10.

As Fig. 6.10 shows, there are no significant differences of any system properties for low ion concentrations ($f \leq 0.2$) between the constant temperature and annealing procedure. As at these low ion concentration gelation is very limited and consequently low energy



(A) Case 28: IM1, 0.5 wt.%, H, 200 kDa.

(B) Case 32: IM2, 0.5 wt.%, H, 200 kDa.

(C) Case 30: IM1, 0.5 wt.%, L, 200 kDa.

(D) Case 31: IM1, 0.5 wt.%, H, 100 kDa.

FIGURE 6.9: Visualization of gelled polymer cases with **annealing procedure AN2** and an ion concentration of f = 0.5. A 200 nm depth section is shown and polymer fiber bonds visualized with a 0.7 nm diameter. Figure adapted with permission from Depta *et al.* [224] under CC-BY 4.0 license.

barriers are present in the system, these results are reasonable. Furthermore, at the onset of gelation with $f = 0.5$, systems with a low G content and the unbinding model IM2 produce similar results, while high G content and low molecular weight produce slightly different results in comparison to constant temperature. With regard to low G, this reasonable behavior is attributed to the lower energy barriers of low G systems in comparison to high G systems resulting from the reduced number of GG dimers and thus fewer stable 'egg-box' bindings. With regard to the unbinding model IM2, these largely equivalent results also appear reasonable, as the unbinding of ions effectively also supports equilibration and overcoming of energy barriers. The two different numerical

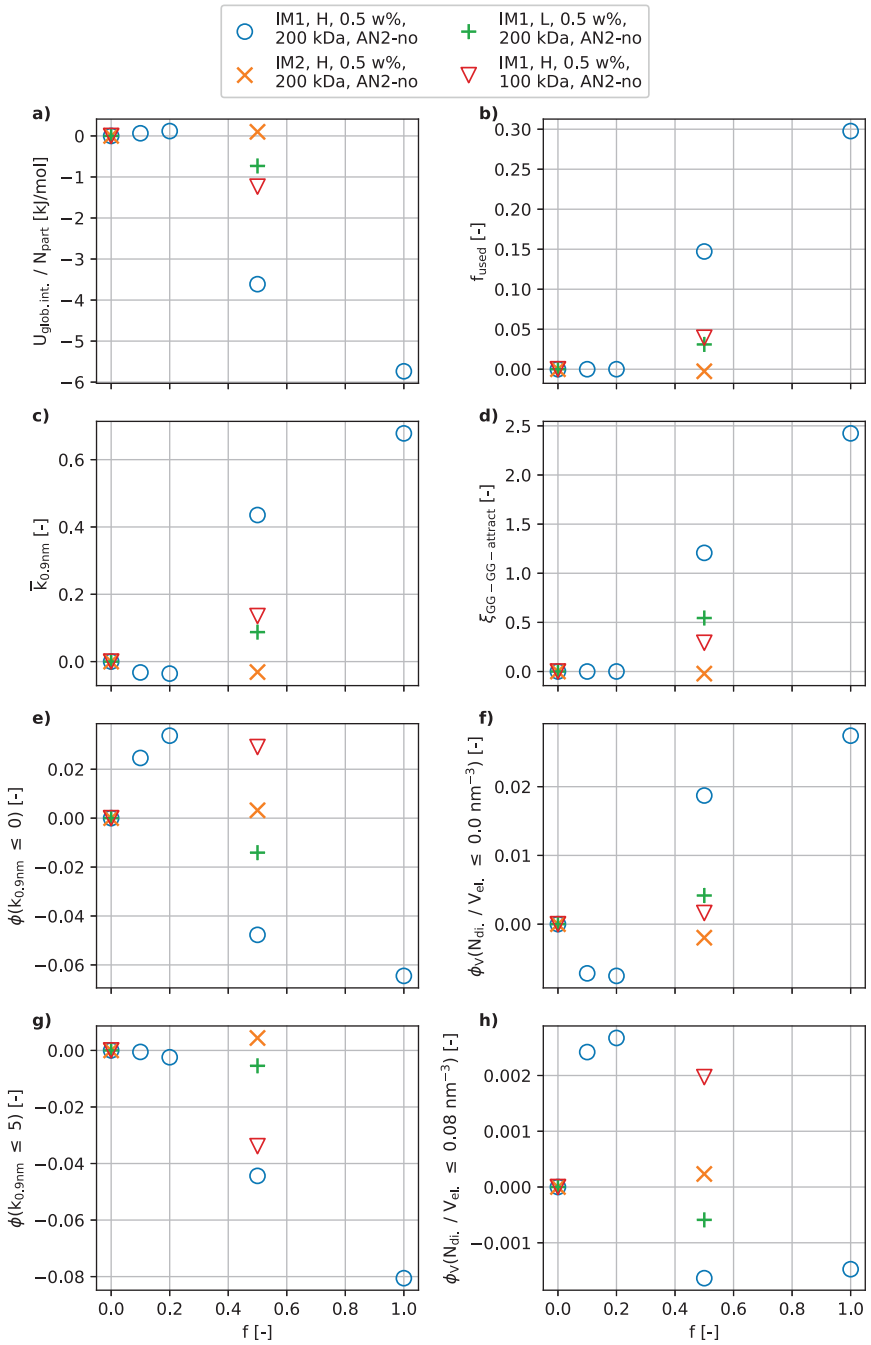


FIGURE 6.10: **Difference** in simulation results between annealing procedure AN2 and no annealing **(AN2 - no)**. Figure adapted with permission from Depta *et al.* [224] under CC-BY 4.0 license.

mechanisms of probabilistic ion unbinding (IM2) and through temporarily increase temperature (annealing AN2) are thus virtually equivalent. These quantitative results are therefore in agreement with the qualitative visual comparison.

With regard to the lower molecular weight of 100 kDa, differences are slightly larger, but remain below that of the base case (IM1, H, 0.5 wt.%, 200 kDa). As discussed in the previous section, reduced cross-linkage of individual/few fibers over bundle formation is the main difference to the base case. Thus, the lower differences of annealing might be attributed to this, as well as the visually higher disorganization resulting from an increased flexibility of polymer fibers. Instead comparing to the base case of high G content and high molecular weight, the annealing procedure causes the largest differences and leads to a higher uptake of ions, increased coordination number / bundle formation, as well as increased pore volume. Thus, the 'egg-box' contacts previously left ion-free during zipping appear to have received an ion (see Sec. 6.4.1). The annealing procedure hence fulfilled its aim of overcoming local energy barriers.

The annealing procedure AN2 overall has an increased effect for alginate systems with higher G content and molecular weight using ion model IM1 at sufficiently high ion concentrations to trigger gelation. As intended, the temporary increase in temperature leads to a momentary breakage of interactions, thus causing a higher acceptance of ions, which then results in a higher coordination number (Fig. 6.10C) and 'egg-box' formation (GG-GG attractive interaction, Fig. 6.10D). Consequently, bundle formation (Fig. 6.10G) and pore volume generation (Fig. 6.10F) is increased.

Overall, annealing was found to support gelation and lead to increased bundle and pore sizes in some cases (high G composition and molecular weight at gelation conditions using IM1). However, as the impact at other conditions and generally on gelation trends is limited while essentially doubling the computational requirements, the annealing procedure can only be recommended to a limited extend.

6.5 Comparison with Literature and Collaborator Data

Experimental validation and comparison with literature data was performed in the context of Depta *et al.* [224] for a polymer concentration of 1.0 wt.%, molecular weight of 155 kDa, high G composition, and saturated calcium concentration of $f = 0.5$. Validation was performed concerning calcium uptake, pore size distributions, and bundle thickness. Credit for the experimental works in this context is given to the respective collaborators and for additional details on experimental procedures the reader is referred to ref. [224].

Hydrogel samples were prepared by Pavel Gurikov, Baldur Schroeter, and Parnpailin Jeansathawong using the jet-cutting process reported in refs. [335, 336] for various solutions of 0.5, 1.0, 5.0 wt.% $CaCl_2$ and remained in solution for at least 12 hours to ensure completed cross-linking. The produced hydrogel droplets were either preserved in 0.02 wt.% sodium azide prior to analysis or converted into aerogels using solvent exchange and supercritical drying as described in refs. [335, 336]. All samples were prepared with a polymer concentration of 1.0 wt.% as preliminary results with a polymer concentration of 0.5 wt.% exhibited easily breakable hydrogels.

The experimentally used sodium alginate is the same as the 'high G' alginate used by Agulhon *et al.* [234] with a reported G and M content of 0.63 and 0.37, respectively (no uncertainty estimation provided). For this study, the G content was independently characterized by Geo Paul and Leonardo Marchese using ^{13}C cross-polarization magic-angle-spinning nuclear magnetic resonance (CPMAS NMR) as described in refs. [224, 337–340] for respective hydrogels in 0.5 and 1.0 wt.% $CaCl_2$ solution. Results found the G content to be 0.66 ± 0.10 and 0.70 ± 0.10 for the two calcium concentrations, respectively. In light of these minor differences, the G content of 0.63 by Agulhon *et al.* [234] was kept for consistency. Molecular weight of alginate samples was not independently measured, but instead used from manufacturer specifications as 155 kDa.

Calcium Uptake/Binding The calcium uptake of hydrogels was measured by Pavel Gurikov and Baldur Schroeter performing first a processing to the respective aerogel, followed by an inductively coupled plasma optical emission spectroscopy (ICP-OES) with dissolution in HNO_3/H_2O_2. Results found the calcium content to be 81.6 ± 1.8, 91.1 ± 2.1, and 85.0 ± 0.90 g Ca/kg-aerogel for the gelation solution concentrations of 0.5, 1.0, and 5.0 wt.% $CaCl_2$, respectively. Consequently, calcium uptake is approximately equal for all gelation bath concentrations with a mean value of 85.9 ± 4.4 g Ca/kg-aerogel. The corresponding cross-linking degree $f_{\text{used,exp}} = 0.38 \pm 0.02$ is consequently in good agreement with simulation predictions of the developed framework as $f_{\text{used,sim}} = 0.43$.

Additionally, calcium binding during gelation of alginate has been studied by other authors in literature. In this regard, Fang *et al.* [239] studied the gelation mechanisms at increasing ion concentrations in dilute systems using isothermal titration calorimetry (ITC) and relative viscosity measurements for almost identical alginate compositions of $\phi_G = 0.64$. They found that a minimum ion concentration of $f = 0.16 - 0.35$ is necessary for 'egg-box' formation to take place and asymptotic behavior starts around $f = 0.5$ (each for both short and long chained alginates). In a similar direction, Lu *et al.* [341] investigated alginate gelation at various polymer concentrations using rheological measurements finding a logarithmic scaling between the critical ion concentration

f_{crit} for gelation and polymer concentration. When extrapolating their data (2 - 6 wt.% for the same $\phi_G = 0.63$ and a lower molecular weight of 62 kDa) to the polymer concentrations investigated in this work, this leads to a f_{crit} of 0.13 and 0.17 for polymer concentrations of 1.0 wt.% and 0.5 wt.%, respectively, thus being at the lower end of the range by Fang *et al.* [239]. These experimental findings are in excellent agreement with the model predictions presented in Sec. 6.4.2.1. In addition to the ion binding model IM1, the unbinding model IM2 also captures nicely the delayed 'egg-box' formation in agreement with the experimental data by Fang *et al.* [239].

Pore Size Distribution Furthermore, the pore size distribution (PSD) of experimental hydrogels was measured by Attila Forgács and József Kalmár using nuclear magnetic resonance (NMR) cryoporometry [224, 342–344] (as previously established for hydrogel-like systems [345]). In addition, the PSD of corresponding aerogels was measured by Pavel Gurikov and Baldur Schroeter using N_2 porosimetry and the Barrett–Joyner–Halendia (BJH) method for desorption analysis. Measurements of corresponding aerogels are expected to produce comparable results to those of the hydrogel, as the solvent exchange is known to minimally influence the pore network and gel structure [346]. Results of all three calcium concentrations produced largely similar pore size distributions [3] and were consequently averaged given the very similar degree of cross-linking. Results of average PSD can be seen in Fig. 6.11 and indicate a good agreement between the different measurement techniques for both hydrogels and aerogels. NMR cryoporometry results found pore sizes of diameter 30 - 36 nm regardless of calcium concentration, while results for corresponding aerogels using N_2 porosimetry (BJH method) were shifted to slightly larger values around approx. 42 nm. Consequently, the most probable pore diameter of the studied hydrogels lies in the range of 33 - 42 nm.

In addition, small angle neutron scattering (SANS) measurements of hydrogel PSD were performed by Attila Forgács and József Kalmár as specified in [224, 339, 347] (Beaucage model for data fitting [348, 349])[4]. Based on the scattering profile (data in Depta *et al.* [224]) and corresponding Beaucage model, the mean pore diameter was estimated as approximately 29.4 ± 3.0 nm, thus being slightly below estimations through NMR cryoporometry and N_2 porosimetry with BJH. However, such reasonable agreement is very positive as it should be noted that precise definition and measurement of pores is inherently challenging due to the anisotropic and interconnected nature of pores within the gel network.

[3] As detailed discussion of these minor differences is out the scope of this work, the reader is referred to Depta *et al.* [224].

[4] Note that for SANS preparation hydrogel beads were placed in D_2O solution with 0.5 wt.% $CaCl_2$ for two weeks, refreshing the solvent every three days.

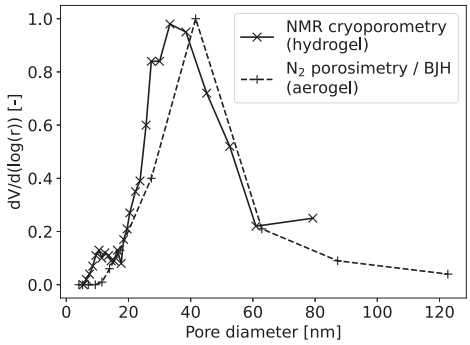
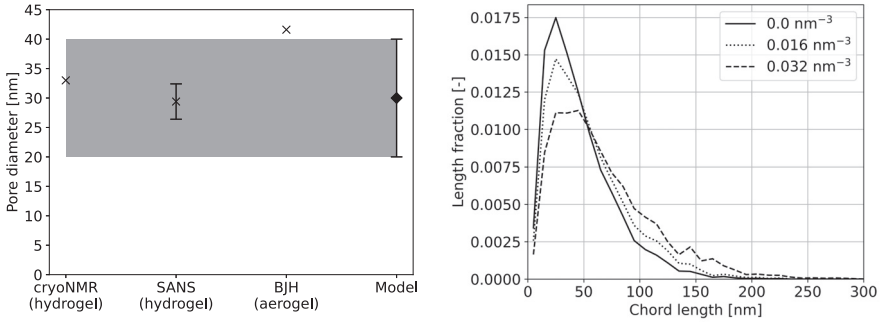


FIGURE 6.11: Experimental pore size distributions based on NMR cryoporometry measured by Attila Forgács and József Kalmár and N_2 porosimetry measured by Pavel Gurikov and Baldur Schroeter. Figure adapted with permission from Depta *et al.* [224] under CC-BY 4.0 license.

Before going into model validation concerning PSD, the measured values should be placed into perspective with existing literature. Generally speaking, the pore size distribution is a key property of aerogels as it defines their specific surface area - most notably influenced by meso pores of range between 2 - 50 nm. Consequently, precise knowledge and prediction is crucial. Available literature data for alginate aerogels across various publications indicates a broad range with a median of 23 nm and an interquartile range of 17 nm ($n = 7$ from refs. [350–353], see also supplementary in ref. [224]). Consequently, the measured values for this study are towards the higher end. Note that to date there is no clear insight as to the factors influencing structural properties of hydrogels and aerogels. Furthermore, most literature data lacks complete characterization of alginates and/or documentation of precise process conditions, specifically the calcium content, making direct comparison of the model to literature data difficult.

In order to compare the numerical predictions of pore sizes to experimental measurements and thus validate the model, the chord length distributions of pores can be used (see Sec. 6.3). The comparison between experimental PSD in Fig. 6.11 and numerical PSD in Fig. 6.12B indicates generally good agreement with slighly smaller pore sizes of the simulation results. As it can be seen in Fig. 6.12B, the definition of a pore (i.e. whether a single dimer in a 5 nm voxel occupies a pore or not) influences the PSD leading to a most probable pore size in the range of 20 - 40 nm. The results of NMR cryoporometry, N_2 porosimetry, and SANS are all in agreement and lie within this range as it can be seen in Fig. 6.12A. This agreement is very reasonable considering the physically different experimental measurement approaches, as well as possible inhomogeneities of the samples.



(A) Most probable pore diameter (peak of pore size distribution; model shaded).

(B) Numerical pore size distribution for varying pore concentration limits.

FIGURE 6.12: A) Comparison of experimental and numerical pore sizes (most probable pore size, PSD peak). B) Pore size distribution of numerical results for the corresponding alginate system of c = 1 wt.%, M_w = 155 kDa, high G composition, f = 0.5, using IM1. Figure adapted with permission from Depta *et al.* [224] under CC-BY 4.0 license.

Bundle/Fibril Diameter Lastly, the bundle diameter, also called fibril diameter, is another fundamental measure extensively reported in literature. Due to this extensive reporting, no independent measurements were performed and validation will be performed with existing data based on small-angle X-ray scattering (SAXS). An extensive overview of more than forty literature data points of SAXS measurements for $CaCl_2$ cross-linked alginate hydrogels and aerogels is provided in Tab. 6.5. As it can be observed, the bundle diameter varies by an order of magnitude as a function of polymer concentration, polymer composition, molecular weight, ion concentration, and other process conditions.

Based on the data shown in Tab. 6.5, some tendencies of bundle formation can be observed: With regard to alginate concentration, there appears to be no strong dependency on bundle diameter. For alginate composition / G fraction ϕ_G, there are limited indications showing a increase in bundle sizes with decreasing ϕ_G. This might be attributed to the decreased cross-linking from a reduced number of 'egg-box' interaction and thus increased mobility during gelation. Such a behavior could not be observed by the proposed model (see Sec. 6.4.2.1) and indications in experimental data are fairly limited. Considering molecular weight M_w, the bundle diameter appears to increase with decreasing M_w. This trend is in agreement with simulation results and attributed to the increase in flexibility resulting from shorter polymer fibers. For ion concentration, strong dependencies can be observed: While over-saturated gels are in the range 10 - 20 nm (data points are also at lower ϕ_G), lower ion concentrations do not exceed 10 nm although being near estimated ion saturation. Furthermore, the majority of data is in the range 2 – 3 nm. In a similar regard, Stokke *et al.* [354] have found a near linear strong correlation between bundle diameter and relative ion concentration $[Ca^{2+}]/[G]$. Furthermore, it should be

noted that lower bundle diameters appear to be present for controlled release of calcium ions, while uncontrollable and large quantities of ions appear to promote larger bundle diameters (arbitrarily assigned $f = 1.0$ in Tab. 6.5).

With regard to the simulation results, the alginate system with a polymer concentration of 1.0 wt.%, a molecular weight of 155 kDa, and an ion concentration of $f = 0.5$ exhibits bundle diameters between 0.6 nm (single polymer fibers) and 5.8 nm. The resulting average bundle diameter is 1.1 ± 0.5 nm including individual polymer fibers and 1.5 ± 0.3 nm when excluding individual fibers. Thus, a large fraction of individual fibers remain in the solution, which is attributed to restrained individual fibers due to cross-linking in the context of a highly homogeneous solution. Additionally, precise definition of contact distances between bundled fibers relatable to SAXS is difficult: By increasing the center of mass distance between two dimers on different fibers composing a bundle to 1.5 nm, the average bundle diameter increases to 1.7 ± 1.1 nm including individual polymer fibers and 2.4 ± 0.8 nm when excluding individual fibers. Nonetheless, simulation results are at the lower end in comparison to SAXS data from literature. With regard to specific comparison, the results marked in grey by Stokke *et al.* [354] provide the most similar experimental conditions (denoted as InG_{155} in ref. [354] with a slightly lower $\phi_G = 0.53$ and $f = 0.44$). Stokke *et al.* found bundle diameters between 1.96 - 4.76 nm ($R_1 - R_2$, average 3.4 nm) for these conditions, which is slightly above the numerically predicted bundle diameters, but still in reasonable agreement especially considering the multiplicity of junction zones in SAXS [354].

When analyzing the scattering profile of SAXS data, the Guinier approximation is frequently used, which causes a multiplicity of junction zones [354] (zones of high density, i.e. polymer thickness). This has been found especially challenging in the high G and high calcium content systems favorable for alginate gelation, as these conditions cause curvature effects of Guinier plots [354]. Thus, direct comparison of SAXS data with the numerically predicted structures is difficult. In this regard, future studies might explore the modeling of scattering profiles based on the numerically predicted structures. Additionally, the simulated polymer solution represents the ideally homogeneous case, thus forming ideal gel networks with individual fibers cross-linked along the network. In contrast, experimental conditions inherently contain more inhomogeneities resulting from material composition (e.g. molecular weight), processing (e.g shear, dispersion of ions, density gradients), and impurities (e.g. gas bubbles), thus leading to larger bundle sizes.

Overall, good agreement between simulation predictions and experimental results was achieved by the proposed model. In this context, ion uptake/binding, pore size distributions, and bundle/fibril diameter were compared indicating the network formation

during gelation and good agreement shown for each property. Future studies might extend the model validation to additional conditions with regard to polymer concentration, ion concentration, and polymer composition.

TABLE 6.5: SAXS measurements for $CaCl_2$ cross-linked hydrogels and aerogels. Table reprinted with permission from Depta *et al.* [224] under CC-BY 4.0 license. *) f of 1.0 is assumed given a large excess of Ca^{2+} over alginate, but not directly measured in the cited paper; **) calculated from alginate concentration in g/L, assuming density of the alginate solution of 1 g/cm^3; ***) estimated from: $R_1 = 8.0$ Å with corresponding weights from the cited work; ****) data for aerogels derived from the corresponding alginate solution; *****) averaged from two values of the characteristic size given in the original publication.

Alg. conc. [wt.%]	**f [−]**	**ϕ_G [−]**	**M_w [kDa]**	**Bundle diameter [nm]**	**Ref.**
2	1.0 *)	0.42	250	11.6	[355]
3	1.0 *)	0.42	250	12.8	[355]
5	1.0 *)	0.42	250	14.2	[355]
2	1.0 *)	0.35	n.d.	10.2	[346]
1	1.0 *)	0.35	n.d.	9.8	[356]
2	0.46	0.63	n.d.	6.7±0.3	[234]
2	0.46	0.63	n.d.	9.1±0.3 ****)	[234]
2	0.52	0.33	n.d.	21.8±0.3	[234]
2	0.52	0.33	n.d.	15.4±0.3 ****)	[234]
1.0 **)	0.175	0.68	160	2.2 ***)	[357]
1.0 **)	0.175	0.39	230	1.4 ***)	[357]
0.5 **)	0.263	0.53	155	2.9 *****)	[354]
0.5 **)	0.525	0.53	155	3.6 *****)	[354]
0.75 **)	0.263	0.53	155	2.6 *****)	[354]
0.55 **)	0.350	0.53	155	3.4 *****)	[354]
1.0 **)	0.175	0.39	230	2.0 *****)	[354]
1.0 **)	0.175	0.39	230	2.2 *****)	[354]
1.0 **)	0.350	0.39	230	2.5 *****)	[354]
1.0 **)	0.175	0.53	155	2.2 *****)	[354]
1.0 **)	0.350	0.53	155	2.9 *****)	[354]
1.0 **)	0.131	0.53	155	1.5 *****)	[354]
1.0 **)	0.175	0.53	155	2.3 *****)	[354]
1.0 **)	0.219	0.53	155	2.5 *****)	[354]
1.0 **)	0.263	0.53	155	2.8 *****)	[354]
1.0 **)	0.438	0.53	155	3.4 *****)	[354]
1.25 **)	0.210	0.53	155	2.4 *****)	[354]
1.25 **)	0.259	0.53	155	2.9 *****)	[354]
1.5 **)	0.175	0.53	155	2.3 *****)	[354]
0.5 **)	0.175	0.50	455	2.2 *****)	[354]
0.5 **)	0.350	0.50	455	2.7 *****)	[354]
1.0 **)	0.175	0.50	455	2.1 *****)	[354]
1.0 **)	0.350	0.50	455	2.9 *****)	[354]
0.5 **)	0.175	0.50	455	1.6 *****)	[354]
1.0 **)	0.175	0.70	51	3.5 *****)	[354]
1.0 **)	0.350	0.70	51	4.4 *****)	[354]
1 **)	0.175	0.68	160	2.5 *****)	[354]
1 **)	0.525	0.68	160	3.2 *****)	[354]
1 **)	0.175	0.68	160	2.9 *****)	[354]
1 **)	0.350	0.68	160	3.6 *****)	[354]
1 **)	0.525	0.68	160	6.0 *****)	[354]
0.5 **)	0.175	0.66	465	2.1 *****)	[354]
0.5 **)	0.350	0.66	465	2.9 *****)	[354]
1 **)	0.175	0.66	465	2.4 *****)	[354]
1 **)	0.35	0.66	465	3.1 *****)	[354]
1 **)	0.175	0.66	465	1.9 *****)	[354]

7

Results: HBcAg System

This chapter is based on the following publications:

P. N. Depta, M. Dosta, W. Wenzel, M. Kozlowska, and S. Heinrich. Hierarchical Coarse-Grained Strategy for Macromolecular Self-Assembly: Application to Hepatitis B Virus-Like Particles. *Int. J. Mol. Sci.*, 23(23):14699, 2022

P. N. Depta, M. Dosta, and S. Heinrich. Data-Driven Multiscale Modeling of Self-Assembly and Hierarchical Structural Formation in Biological Macro-Molecular Systems. In W. E. Nagel, D. H. Kröner, and M. M. Resch (editors), *High Performance Computing in Science and Engineering '21*. Springer International Publishing, Cham, 2023

7.1 Model Parameters

7.1.1 Structural Model

In order to study the self-assembly of virus-like particles (VLP) on the example of the hepatitis B core antigen (HBcAg), a dimer of HBcAg monomers (subsequently termed $HBcAg_2$), is abstracted as the smallest unit object as presented in Sec. 1.3.1 and Ch. 2. The $HBcAg_2$ dimer can be assumed stable on the time scales studied and its reference conformation can be found in Sec. 1.3.1, which is used for the parameterization of all model components. The dimer posses a mass of $33.6\,\mathrm{kDa} = 5.586 \times 10^{-23}\,\mathrm{kg}$ and a radius of gyration in x, y, and z of 1.31 nm, 1.85 nm, and 1.97 nm, respectively. The dimers are freely interacting with each other and the environment, i.e. implicit solvent and ions. They possess a position, orientation, as well as spatial extend and other anisotropic properties through its functional models, which will be specified in the following.

7.1.2 Functional Model

7.1.2.1 Diffusion and Thermodynamics

Anisotropic diffusion and the respective thermodynamics of the desired canonical ensemble were modeled and parameterized for the $HBcAg_2$ dimer using the method presented in Ch. 3. The model captures the interaction of the structurally flexible macromolecule with the solvent, as well as ions. The determined diffusion coefficients for $HBcAg_2$ at the desired process conditions of 293 K and 150 mM NaCl in aqueous solution are given in Tab. 7.1. Hydrodynamic interaction was accounted for through reduction of the effective viscosity by a factor of 0.1 as discussed in Sec. 3.2 and App. A.3.

TABLE 7.1: Diffusion coefficients for $HBcAg_2$ dimer at 293 K and 150 mM NaCl in aqueous solution. Table adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

D_t [µm^2 s^{-1}]			D_r [Mrad^2 s^{-1}]		
x	y	z	α	β	γ
87.69	72.27	71.48	12.05	7.46	7.00

7.1.2.2 Intermolecular Interaction

The intermolecular interaction of $HBcAg_2$ with $HBcAg_2$ was modeled and parameterized using the method presented in Sec. 4.3. Initial MD sampling was performed based on distances classes as specified in Sec. 4.3.4.3 and detailed in Tab. 7.2 for a total of 95'000 samples. Based on these initial samples, iterative refinement was performed as specified in Sec. 4.3.4.3 using the following sequence of refinement criteria and number of samples:

- Iteration 1 - 10: Variance minimization using 5'000 samples for each iteration.
- Iteration 11- 20: Normalized variance minimization using 5'000 samples for each iteration.
- Iteration 21, 24, 27: Potential minima resampling using 15'000 samples for main extrema points and 5'000 samples for first-level neighborhood points.
- Iteration 22, 25, 28: Potential maxima resampling using 15'000 samples for main extrema points and 5'000 samples for first-level neighborhood points.
- Iteration 23, 26, 29: Gradient maxima resampling using 15'000 samples for main extrema points and 5'000 samples for first-level neighborhood points.

Overall, 375'000 MD data points were sampled and analyzed for estimation of the intermolecular interaction potential of $HBcAg_2$ with $HBcAg_2$. The results of iterative refinement will be presented in detail in Sec. 7.4.1.1. During resampling, a 0.63 nm grid was employed and after finalization refined to 0.5 nm for the remaining works.

TABLE 7.2: Random sampling over distances classes for $HBcAg_2$ – $HBcAg_2$ data set before iterative resampling. Note that the number of samples is essentially double (or interaction space half in volume) due to the symmetry resulting from molecule A = B. Reprinted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

d_{A-B} [nm]		# samples
lower	upper	
0.4	0.5	20'000
0.5	0.7	5'000
0.7	0.9	5'000
0.9	1.1	5'000
1.1	1.3	5'000
1.3	1.5	5'000
1.5	1.7	5'000
1.7	1.9	5'000
1.9	2.1	5'000
2.1	2.3	5'000
2.3	2.5	5'000
2.5	3.0	5'000
3.0	3.5	5'000
3.5	4.0	5'000
4.0	4.5	5'000
4.5	5.0	5'000
Sum		95'000

Furthermore, biased MD simulations were performed at the binding locations of the HBcAg capsid for extended simulation times. As non-colliding start configurations are required for MD, the algorithm described in Sec. 4.3.5 was employed to determine the configurations closest to known binding locations. The determined non-overlapping binding configurations can be found in Tab. 7.3. At each binding location, 1'016 simulations were performed for 10 ns each. The results will be discussed in detail in Sec. 7.4.1.2.

7.1.2.3 Bonded Interaction

No bonded interaction was modeled for the HBcAg system. All intermolecular interactions for the self-assembly process were captured by the interaction model in Sec. 7.1.2.2.

TABLE 7.3: $HBcAg_2$ – $HBcAg_2$ binding locations from reference capsid and determined non-colliding locations for dimer reference structure. x/y/z in nano-meter and $\alpha/\beta/\gamma$ in radian with respect to body frame of reference of molecule A. Each binding location is present 60 times in the reference 120 dimer capsid and the mean and maximum δ_r distance to any instance is 0.005 nm and 0.12 nm respectively. Table adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

#	Original			Non-colliding		
	x	y	z	x	y	z
1	-2.74	-0.74	-3.10	-2.74	-0.94	-3.50
2	1.47	-0.91	-4.14	1.67	-0.91	-4.34
3	-3.01	-0.70	-3.08	-3.01	-1.10	-3.48
4	-0.65	-0.77	4.25	-0.65	-0.57	4.45
	α	β	γ	α	β	γ
1	-0.48	0.98	-0.32	-0.13	0.81	-0.06
2	-0.88	-1.05	0.67	-0.53	-0.79	0.50
3	-2.72	-1.05	3.03	-3.07	-0.96	-3.08
4	2.72	0.92	2.76	2.81	0.83	2.76

7.1.2.4 Critical Time Step

In order to estimate the critical time step τ_{crit} necessary for a convergent and numerically stable solution, the previously discussed methods for the individual model components were used, see Sec. 3.4.1 and 4.5. Based on this, the critical time step can be estimated as 2.9×10^{-13} s for the diffusion model (2.9×10^{-12} s when accounting for hydrodynamic interaction through reduced viscosity as discussed in Sec. 3.2 and App. A.3), 5.3×10^{-12} s for the translational component of the intermolecular interaction model, and 2.0×10^{-12} s for the rotational component of the intermolecular interaction model. Consequently, the diffusion model constrains the overall simulation time step. Unless otherwise indicated, a simulation time step of 10^{-13} s was used leading to an error of 1.0 % with regards to RMS displacement (see Sec. 3.4.1).

7.2 Simulation Setup and Procedure

A variety of simulation setups and procedures were performed in order to study the derived macromolecular interaction potentials, VLP stability, and VLP self-assembly. Details on the MDEM implementation can be found in Sec. 2.2. Unless otherwise indicated, a temperature of 293 K, dynamic viscosity of 1.0074×10^{-3} Pa s, time step of 10^{-13} s, and cubic simulation domain with periodic boundary conditions at all boundaries was used. The following three simulation procedures can be differentiated.

VLP Binding Location Agreement and Stability (SP-VLP-1) First, in order to study the derived macromolecular interaction potentials and agreement of binding locations with literature data, the stability of the smallest structural assembly of the $HBcAg_2$, a trimer, was investigated. The used reference trimer (of $HBcAg_2$ dimers) was extracted from the reference capsid and is shown in Fig. 7.1A. Based on the trimer in free solution with an open domain, equilibration simulations were performed for 25 ns at a temperature of 0 K to avoid thermostat effects and enable a pure equilibration of the structure to its (local) equilibrium based on the macromolecular interaction potential. Consequently, agreement of binding positions with the potential minima of the interaction potential can be investigated and visualized.

VLP Capsid Stability (SP-VLP-2) Second, the stability of individual HBcAg capsids was evaluated using the objective function for structural stability defined in App. C.3. For these simulations, an assembled capsid based on its reference assembly was placed in an open domain and simulation performed for 250 ns at normal process conditions. Structural stability was evaluated every 1 ns and calculated as specified in App. C.3.

VLP Self-Assembly (SP-VLP-3) Third, self-assembly of $HBcAg_2$ was investigated based on randomly initialized systems. $HBcAg_2$ dimers were placed and oriented randomly in a cubic box of specified size and concentration with periodic boundary conditions. During placement, overlap of dimers was permitted and then corrected during the simulation through the intermolecular interaction model. Four concentrations of 5 µM, 10 µM, 50 µM, and 100 µM were investigated. The simulations of the two lower concentrations were performed in $1\,\mu m^3$ (1 µm edge) cubic boxes and the two higher concentrations in $0.125\,\mu m^3$ (0.5 µm edge) cubic boxes to achieve similar run times and maintain comparable statistics. In order to account for the non-dilute state during self-assembly, the effective viscosity was reduced to 1.0074×10^{-4} Pa s as discussed in Sec. 3.2 and App. A.3. This enabled an increase of the simulation time step to 10^{-12} s. Based on this, self-assembly simulations were then performed for 5 ms with a savings interval of 500 ns.

7.3 Analysis and Postprocessing

Simulation results were analyzed both online and in postprocessing to study a variety of features with focus on structural properties. The analyzed system properties include the following and are taken from Depta *et al.* [225]:

- objective function of capsid stability f_{stab} during stability analysis of an individual capsid (SP-VLP-2, see Appendix C.3 for definition);
- self-assembled structures (SAS) were identified using a network search algorithm distinguishing between structured contacts ($\delta_r \leq 1\,\text{nm}$ from a known binding location, see Tab. 7.3) and unstructured contacts ($\delta_m \leq 0.3\,\text{nm}$ and $\delta_r > 1\,\text{nm}$ from a known binding location). Based on these, the following properties were investigated and can be readily compared to known capsids:
  - N_{SAS} is the number of dimers / particles in each SAS;
  - $d_{\text{SAS,gyr}}$ is the diameter of gyration of each SAS;
  - ξ_{struc} is the number of structured contacts normalized by the number of dimers in the system, can also be used per SAS or per dimer;
  - ξ_{unstruc} is the number of unstructured contacts normalized by the number of dimers in the system, can also be used per SAS or per dimer;
  - $\Phi_{\text{struc}} = \xi_{\text{struc}}/(\xi_{\text{struc}} + \xi_{\text{unstruc}})$ is the fraction of structured contacts in the system, can also be used per SAS or per dimer;
- assembly kinetics were quantified by exponential fitting of average N_{SAS}, see Sec. 4.3.3 eq. 4.17 with time constant τ_{SAS} (r in eq. 4.17) and asymptotic structure size $N_{\text{SAS,asymp}}$ (s in eq. 4.17);
- transitions of dimers between size classes based on N_{SAS} were summarized and normalized by the number of dimers / particles to capture assembly mechanisms (note saving step of 500 ns). Bi-directional and net transitions between classes were distinguished and visualized using chord diagrams [358];
- lifetimes t_{life} of structures were analyzed based on their duration of existence until either class change or end of simulation.

In addition, global properties such as potential and kinetic energies were analyzed. However, as these provided little additional insight focus was placed on the structural features.

7.4 Results

In the following section, the results of VLP stability and self-assembly will be presented based on the $HBcAg_2$ dimer. In this context, first the intermolecular interaction potential and its impact on stability of VLP sub-structures will be discussed. Second, VLP self-assembly using the fully parameterized model will investigated for varying concentrations of dimers.

7.4.1 Intermolecular Interaction Potential and VLP Stability

With regard to both the stability and self-assembly of VLPs, the intermolecular interaction potential is the most crucial aspect in modeling as it defines binding locations and hence the overall structural formation. In order to investigate the behavior of the interaction potential on stability, the smallest sub-component of the HBcAg VLP – a trimer of $HBcAg_2$ dimers – will be investigated as it is visualized in Fig. 7.1. The intermolecular interaction potential will be explored based on pure MD sampling, including biased MD sampling at the binding locations, and with additional empirical data as outlined in Sec. 4.3 and 7.1.2.2.

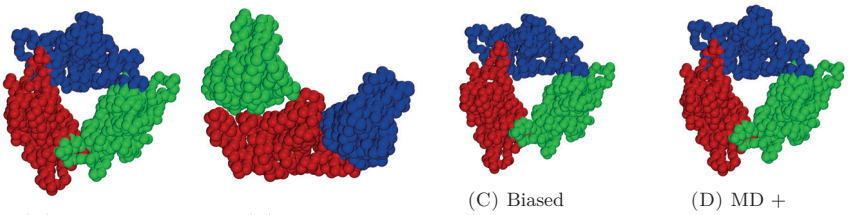


FIGURE 7.1: Visual comparison of the trimer equilibrium positions between the capsid reference structure (A) and MD-based interaction potentials (B-D) after equilibration using simulation protocol SP-VLP-1. Shown cases are for the pure MD-based interaction potential (B), biased MD-based at binding locations (C), and MD-based with empirical data (D). Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

7.4.1.1 Pure MD-Based Interaction Potential

MD Data Trends and Statistics During iterative resampling of the supervised learning algorithm for the intermolecular interaction potential, the MD data was analyzed adaptively by the algorithm for trends and correlation statistics. An overview of the full trend and statistical analysis of the final data set for $HBcAg_2$ – $HBcAg_2$ interaction is provided in Appendix D.3. The intermolecular interaction between two $HBcAg_2$ shows an attractive trend for potential A–B (170 kJ/mol), repulsive for potential A–PW + B–PW (200 kJ/mol), attractive for potential PW–PW (49 kJ/mol), repulsive for potential A–ion + B–ion (11 kJ/mol), attractive for potential PW–ion (5 kJ/mol), and repulsive for bond potential (4 kJ/mol). In contrast, the potential contributions A–A + B–B, ion–ion, G96-angles, improper dihedral angles, and Coulomb reciprocal contain no significant trend. Consequently, the results indicate a largely stable conformation of $HBcAg_2$ during interaction and only a slight repulsion resulting from bonded interactions with each molecule. Furthermore, long-range electrostatic effects captured in the

reciprocal Coulomb term appear negligible, which is especially notable in the context of the large concentration of ions known to mediate intermolecular interaction. Overall, the dominating influences are caused by electrostatic and Lennard-Jones interaction directly between molecules, through solvent effects, and ion mediation. The overall trend sum has an interaction range of $\delta_m \approx 1.2$ nm and contains a local minimum of -32 kJ/mol at $\delta_m \approx 0.45$ nm before increasing to -9 kJ/mol at $\delta_m \approx 0$ nm (contact). These values are in agreement with theoretical models [359, 360] and slightly higher than experimental association energies [254, 361], which explicitly account for allosteric modulation effects.

After accounting for the identified potential trends, the remaining residual potential contained valid correlations only for the potential component A–B with variogram values up to 1'000 – 10'000 kJ^2/mol^2 over interaction ranges between 2 – 4 nm increasing towards sections of larger δ_m. All other potential components contained either significantly more inherent noise and / or very short correlation distances prohibiting further interpretation beyond the trend. All data is provided in App. D.3 including additional spatial descriptors in App. D.1.

Convergence The convergence behavior concerning potential changes and estimation variance resulting from the iterative resampling strategy is shown in Fig. 7.2. The three criteria of resampling for variance, normalized variance, and extrema (potential minima, potential maxima, gradient maxima) can be clearly distinguished in the changes of the potential field and exhibit a decreasing mean change in potential for each criteria. At the beginning of each criteria, the mean change in potential increases indicating the change in emphasis within the interaction space. This is especially notable during extrema resampling, when the overall field changes significantly indicating stronger deviations from the trend through e.g. binding locations. At the same time, the maximum change in potential decreases only slightly from approximately 300 kJ/mol to the range of 100 – 200 kJ/mol, indicating locally higher changes in potential overall. Furthermore, the estimation variance decreases slightly within each criteria of resampling and exhibits a strong increase at the beginning of extrema resampling followed by a decrease and stabilization. This increase is attributed to a change in variogram model resulting from a larger variance of the potential samples during extrema resampling.

Overall, the convergence results underline the challenges of sampling in the six dimensional interaction space and inherent noise caused by lower-scale changes, e.g. thermodynamics. While increased sampling would be beneficial, the algorithm performs reasonably well in the context of limited computational resources, as well as the trade-off between sampling extrema and learning the overall field, as can be observed in the reduction of potential changes.

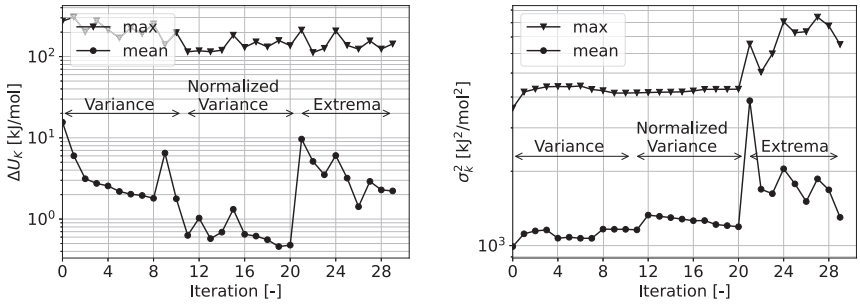
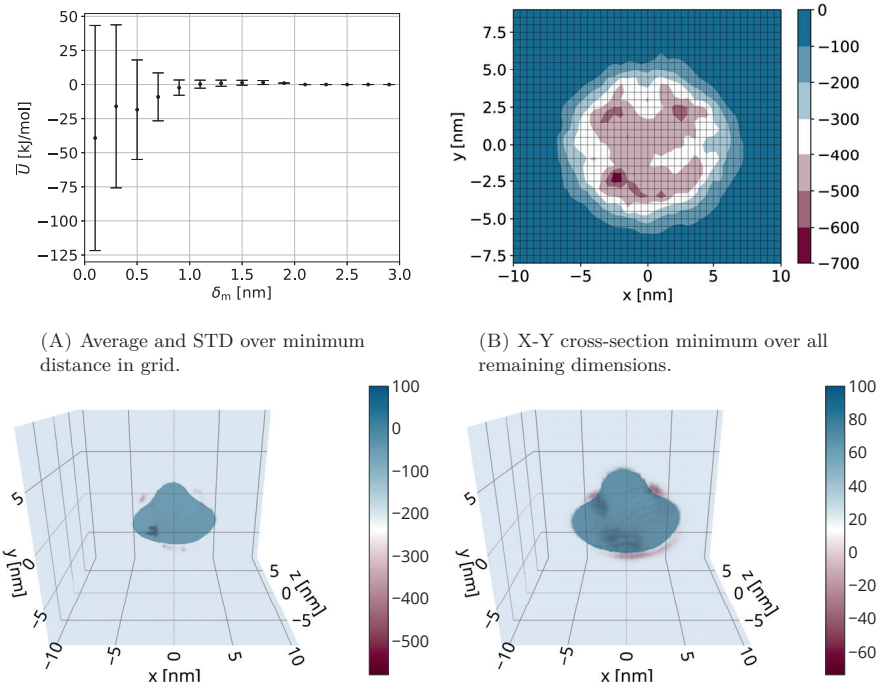


FIGURE 7.2: Iterative resampling convergence concerning potential field changes (left, $U_i - U_{i-1}$) and field variance (right). Adapted with permission from Springer Nature regarding Depta *et al.* [226].

Resulting Field The interaction potential field resulting from the pure MD-based sampling is visualized in Fig. 7.3 using a set of 1D, 2D, and 3D projections out of the 6D space. As it can be seen in the 1D projection into δ_m space, the interaction range between two $HBcAg_2$ molecules is approximately $\delta_m \approx 2$ nm. After entering the interaction region, a slight potential barrier between 0 – 5 kJ/mol exists at $\delta_m \approx 1.5$ nm followed by a local potential minimum at $\delta_m \approx 0.45$ nm, which results from the minimum in potential trends. After a slight increase in potential with lower δ_m, the average potential between $\delta_m = 0 - 0.2$ nm decreases to -39.0 ± 82.5 kJ/mol leading to possible binding. The specific potential values for each relative configuration can vary significantly from the overall trend especially for small δ_m (note increasing variance with lower δ_m), which is attributed to the strong Coulomb and Lennard-Jones interaction between molecules, solvent, and ions leading to both binding and repulsion depending on the specific relative configuration.

As it can be seen in the 2D minimum projection in Fig. 7.3B, 3 – 4 primary binding locations are recognized. These binding locations can be identified more clearly in the 3D projections in Fig. 7.3C - 7.3D and are located beneath the dimer in negative y-direction, as well as in positive and negative x-direction from the dimer spike (upper y-axis part of molecule). As it can be seen in the corresponding trimer equilibration shown in Fig. 7.1B, the underneath of the dimer essentially binds next to the spike on either side. Note that these recognized binding locations are different from the expected binding locations present in the capsid and trimer (see Fig. 7.1A and Tab. 7.3). Consequently, the derived interaction potential does not produce stable capsids.

Investigations of binding details and the underlying MD data show that these differences and limitations in capturing binding can be attributed to conformational changes required for binding in the context of reference conformation, MD simulation time, and



(A) Average and STD over minimum distance in grid.

(B) X-Y cross-section minimum over all remaining dimensions.

(C) 3D minimum over orientations.

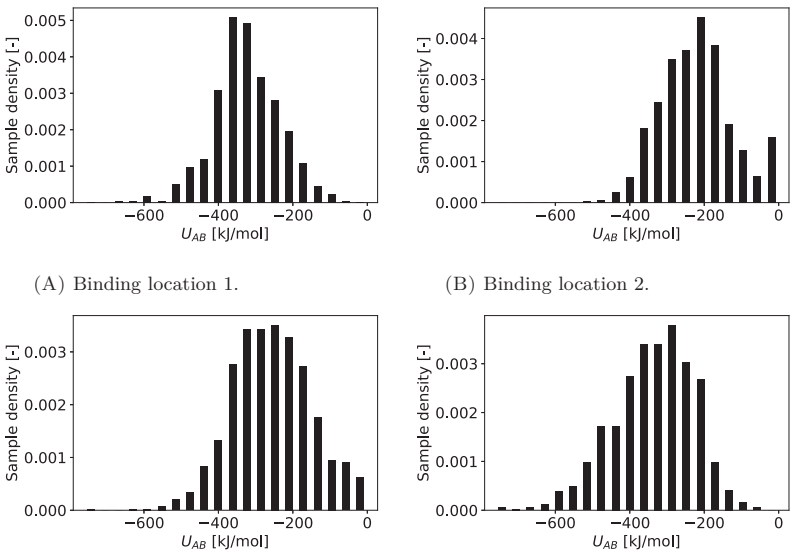
(D) 3D mean over orientations.

FIGURE 7.3: Interaction potential field from **pure MD**-based sampling strategy using a 0.5 nm grid (units of kJ/mol). A and B adapted with permission from Springer Nature regarding Depta *et al.* [226].

possibly the employed *Martini* force-field. As it can be seen in Fig. 7.1A, the reference structure determined by representative clustering during free diffusion runs differs from the binding conformation causing overlapping side-chains. Consequently, during binding the conformation is required to change to enable the strong binding interaction. While the MD model explicitly permits these degrees of freedom during potential sampling, the studied simulation times appear to be either to short for the conformational difference to occur with sufficient probability or the employed *Martini* force-field is not suitable enough for the HBcAg system. In order to study these effects in more detail, biased MD simulations are performed for extended simulation times of 10 ns in the following. Note that a study investigating the suitability of different force-fields goes beyond the scope and intent of this work.

7.4.1.2 Biased MD-Based Interaction Potential

In order to address the challenges in binding recognition, at each of the four locations (see Tab. 7.3 and Fig. 7.1A) 1'016 replicates were performed with 10 ns simulation time each as specified in Sec. 4.3.5 and 7.1.2.2. Post-processing was conducted using the last 0.1 ns as before (see Sec. 4.3.1). Results show an improved binding recognition with A–B potentials as low as -762 kJ/mol and conformations similar to literature [12, 244, 246] as it can be seen in Fig. 7.5. However, probability of strong binding remains low as it can be seen in the wide probability distribution of potential A–B shown in Fig. 7.4 with an average potential at -285 kJ/mol and peaks in probability between -200 and -400 kJ/mol. Furthermore, conformational difference of the C-terminal region remain for the majority of binding simulations.



(A) Binding location 1.

(B) Binding location 2.

(C) Binding location 3.

(D) Binding location 4.

FIGURE 7.4: Probability distribution of U_{AB} for biased MD simulations at binding locations 1-4 of $HBcAg_2$ system, see Tab. 7.3 and Fig. 7.1A.

The resulting interaction potential field can be found in App. D.2 Fig. D.2 and exhibits no visual differences to the pure MD-based interaction potential. A detailed comparison shows a minimal average potential decrease of the interaction region (within cutoff) by -0.05 kJ/mol and local changes include a decrease of up to -299 kJ/mol, as well as an increase of up to 234 kJ/mol. While these changes appear visually minimal, the interaction potential produces stable trimers in equilibration studies as it can be seen in Fig. 7.1C. Consequently, the extended simulations at the binding locations improve the interaction potential notably. However, capsid structures remain unstable as the

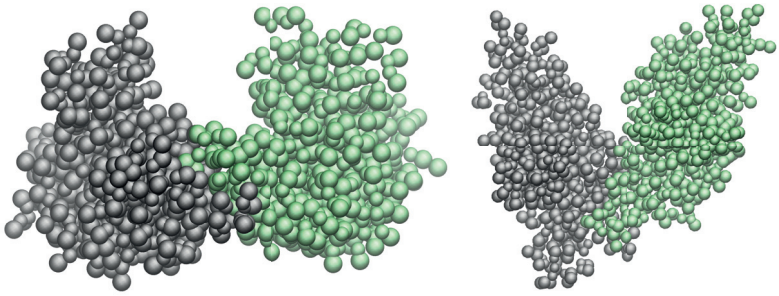


FIGURE 7.5: Visualization from side (left) and top (right) of biased MD simulation with lowest potential A-B after 10 ns. Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

global potential minima (i.e. global binding locations) remain unchanged and are not coinciding with the known binding location. Consequently, while recognition of binding is improved, the extended simulation runs are not sufficient in recognizing the specificity and strength of binding due to the remaining low probability attributed to conformation changes required.

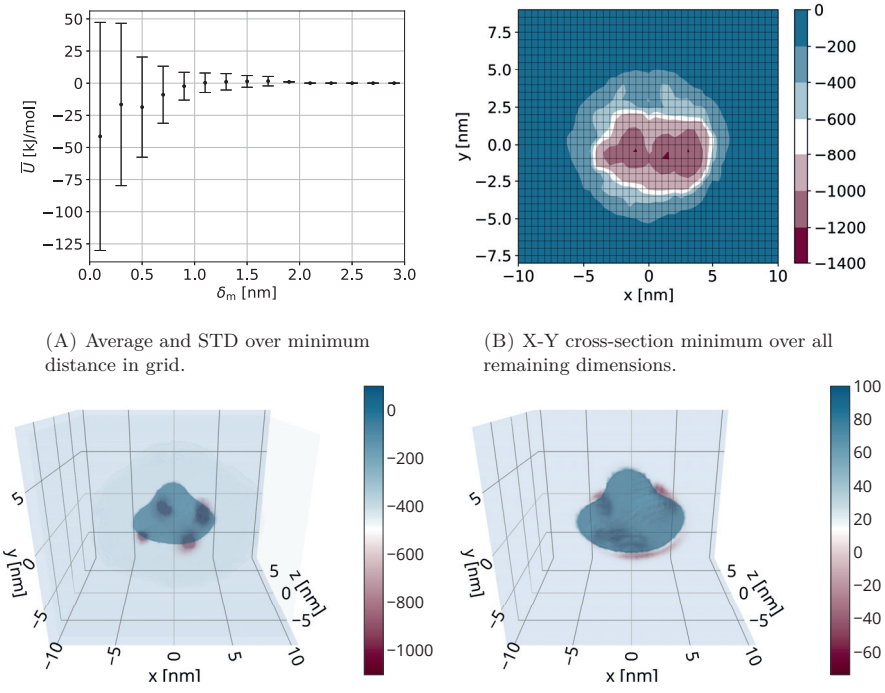
In the context of effective surrogate modeling and force-field development using lower-scale models, such limitations are well known [30]. While lower-scale methods such as MD provide a wealth of information, their detail, scales, and sampling capabilities are limited. Furthermore, the abstraction and formulation of the effective surrogate model is limited. In order to address this in the context of surrogate modeling, additional information is often included to address such limitations and improve the overall model. The following section will explore such an approach through insertion of empirical data.

7.4.1.3 MD-Based Interaction Potential with Empirical Data

In order to address the limitations of the underlying MD model during parameterization of the macromolecular interaction potential, a variety of approaches to insert additional empirical data were explored. The main goal in this context was to perform adaptations only locally, while retaining the remaining information from MD. Good results were achieved through the insertion of empirical data points with specific potential values at defined locations. These empirical data points influence the interaction potential estimation in their proximity through Universal Kriging in the same manner a data point from MD sampling does. However, they are considered virtual as they do not influence the trend and variogram model (i.e. correlations), leaving the remaining molecular information of the macromolecular interaction potential unchanged.

A variety of insertion methods were explored for improving the definition of $HBcAg_2$ – $HBcAg_2$ binding sites in the context of VLP / capsid stability and self-assembly. For this, the shape and potential depth of inserted empirical data points were varied, as well as the repulsion model for molecular overlap, which correlates closely. This was performed through parameter studies, optimizations, as well as manual testing. It was found that the binding locations have to be specified lower than the potential minima of the pure MD-based potential, i.e. at least -800 to -1000 kJ/mol at the binding locations. Improved stability and self-assembly of capsids was achieved by decreasing potentials to approximately -1400 kJ/mol. While these potentials are considerably low, they are not overly different from the lower potentials found during extended MD simulations in Sec. 7.4.1.2, especially in the context of remaining conformational differences of the C-terminal [362]. Additionally, the unbinding barrier through the conformational interconnection of molecules during binding is captured efficiently through this deepened interaction potential. Lastly, it was found that the most suitable spatial extend of binding locations for achieving capsid stability is approximately 1 nm in δ_r space with increasing potentials of a Gaussian shape. Both smaller and larger spatial extends lead to unstable capsids through lacking spatial specificity.

Overall, the best solution concerning capsid stability and self-assembly achieved an objective stability criterion $f_{\text{stab}} = 0.725$ over 250 ns (see simulation procedure SP-VLP-2) indicating a near perfect capsid. Similarly, the reference trimer was kept stable during equilibration as it can be seen in Fig. 7.1D. The empirical data points were inserted as specified in Sec. 4.3.5 at the binding locations in Tab. 7.3 including symmetry points using $U_{\text{emp,bind,center}} = -1400$ kJ/mol, $d_{\text{step,center}} = 0.1$ nm, $U_{\text{emp,bind,outer}} = -1000$ kJ/mol, $r_{\text{emp,bind}} = 1.0$ nm, and $d_{\text{step,outer}} = 0.2$ nm as parameters. Molecular collisions were found to be optimally accounted for by the repulsion model (see Sec. 4.3.6) using $c_{\text{rep,bb}} = 50$ kJ/mol, $N_{\text{min,bb}} = 0.25$, $d_{\text{MD,min}} = 0.4$ nm, $w_{\text{MD,min}} = 0.5$ nm, $w_{\text{coll}} = 0.3$ nm, and $c_{\text{rep,side}} = 0$ kJ/mol as parameters. These parameters for the repulsion model were used consistently for pure MD and biased MD results in Sec. 7.4.1.1 and 7.4.1.2. The resulting interaction potential from the insertion of empirical data points together with MD-based samples is visualized in Fig. 7.6. As it can be seen in Fig. 7.6B and 7.6C in comparison to the pure MD-based potential in Fig. 7.3, four specific potential minima are introduced coinciding with the desired binding locations. At the same time, the overall potential seen most prominently in Fig. 7.6A and 7.6D is visually unchanged indicating a majority of information resulting from the MD-based sampling, e.g. the potential barrier at $\delta_m \approx 1.5$ nm. Consequently, the MD-based information is merged with empirical information at specific locations underlining feasibility of the approach.



(A) Average and STD over minimum distance in grid.

(B) X-Y cross-section minimum over all remaining dimensions.

(C) 3D minimum over orientations.

(D) 3D mean over orientations.

FIGURE 7.6: Interaction potential field from **MD with empirical data** (units of kJ/mol). Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

7.4.2 VLP Self-Assembly

In the following section, the self-assembly of $HBcAg_2$ into virus-like particles (VLPs, also termed virus capsids) will be investigated using the proposed framework and MD-based interaction potential with empirical data shown in Sec. 7.4.1.3. Four $HBcAg_2$ concentrations of 5 µM, 10 µM, 50 µM, and 100 µM were investigated at constant conditions of 293 K and 150 mM sodium chloride employing the procedure SP-VLP-3 presented in Sec. 7.2. The two lower concentrations were performed in a cubic simulation domain of 1 μm^3, while the two larger concentrations were performed in 0.125 μm^3 domains to ensure comparable run-times and statistics. Exemplary for the 1 μm^3 system with 10 µM concentration, the transition from its initially disordered random state at the beginning to self-assembled capsid structures is visualized in Fig. 7.7. In the following, the results of self-assembly will be discussed in detail for all concentrations with regard to assembly properties, kinetics, and pathways.

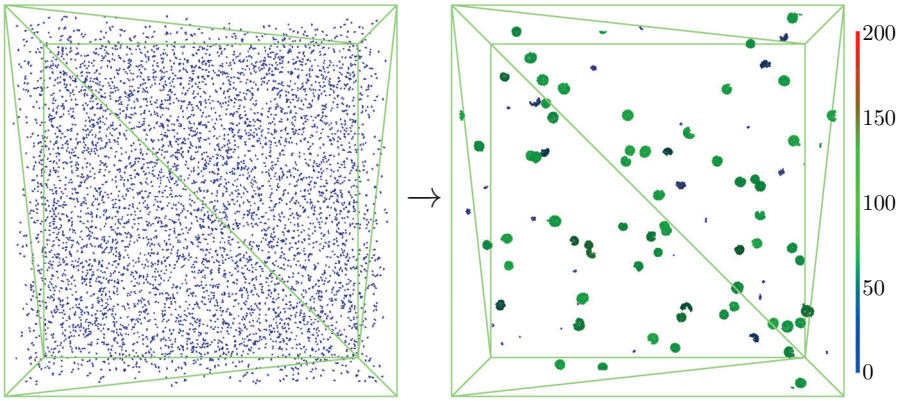
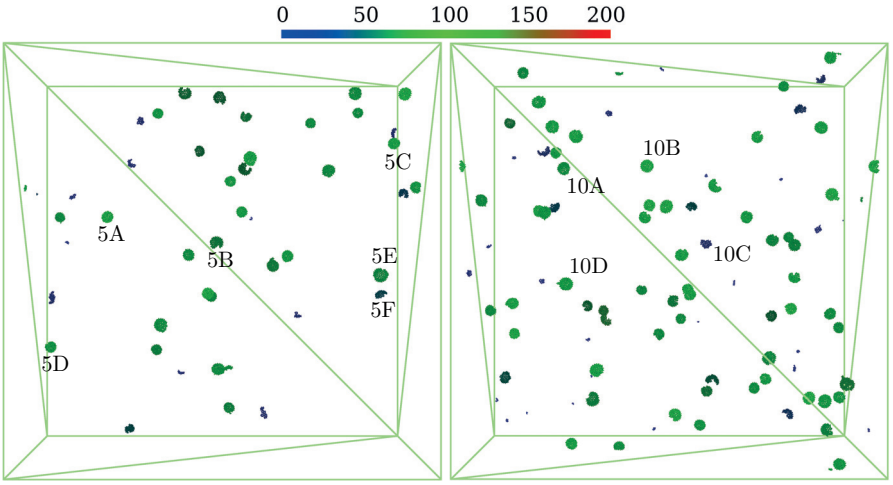


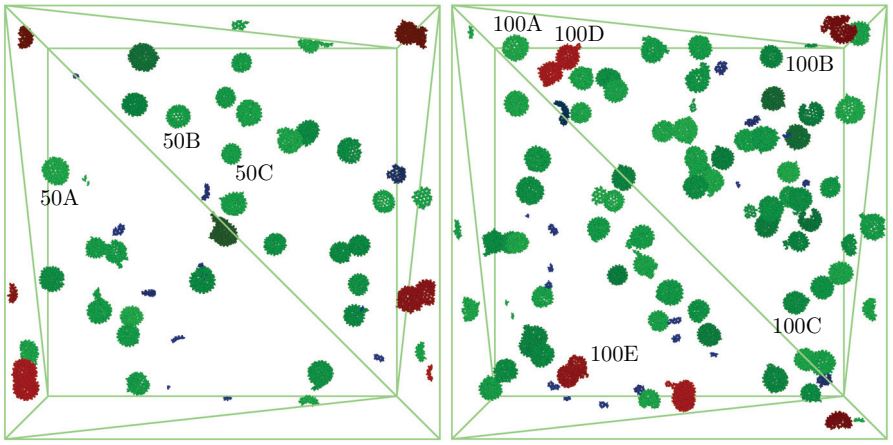
FIGURE 7.7: VLP self-assembly for 1 µm^3 box (1 µm edge) with 10 µM concentration before (left) and after (right) 5 ms with simulation protocol SP-VLP-3. Back-bone carbon structures are visualized and color indicates self-assembled structure sizes by number of dimer particles (N_{SAS}).

Assembly Properties As it can be seen in the overall visualization of systems (Fig. 7.8) and closeups of structures (Fig. 7.9), the individual $HBcAg_2$ dimers self-assemble into regular spherical capsids, which are visually in good agreement with the expected icosahedral VLPs [244] (e.g., 5A, 5C, and 10B in Fig. 7.9). Each dimer forms on average $\xi_{\mathrm{struc}} = 3.5$ structured connections for all concentrations, which is close to the expected value of $\xi_{\mathrm{struc}} = 4.0$ for perfect T = 4 capsids (120 dimers), thus supporting agreement with icosahedral VLPs. The majority of the formed capsid population contains around 100 dimers (green coloring) and possesses a diameter of gyration between 24 - 30 nm as is shown in the population distribution in Fig. 7.10. As indicated by the marked regions, a smaller portion of the population agrees with the T = 3 capsid with regard to these properties (24.0 %, 20.7 %, 19.2 %, and 11.5 % for concentrations of 5 µM - 100 µM, respectively), while the majority of the capsid population can be considered pre-stages of T = 4 capsids missing up to 20 dimers. Such excess of T = 4 capsids over T = 3 capsids is in agreement with literature [246, 363–365], which predicts more than 90 % of T = 4 capsids. This assessment is further supported by closeup visualizations of capsids (e.g. 5B in Fig. 7.9), which tend to show defects in the form of missing dimers or dimer segments. It appears the finalization to perfect T = 4 capsids is incomplete at the end of the simulations, which is attributed to the decreased equilibration kinetics resulting from low availability of individual dimers or small structures at this stage (discussed subsequently in assembly kinetics, visible through comparatively low number of blue colored structures). These observations are in line with experimental insights [254, 366]. In addition to defects of missing individual dimers and dimer segments, rare defects of misalignments in structure can be observed such as 100B in Fig. 7.9, where the ring assembly at the center contains seven instead of six dimers. However, as can



(A) 1 µm^3 box (1 µm edge) with 5 µM.

(B) 1 µm^3 box (1 µm edge) with 10 µM.



(C) 0.125 µm^3 box (0.5 µm edge) with 50 µM.

(D) 0.125 µm^3 box (0.5 µm edge) with 100 µM.

FIGURE 7.8: VLP self-assembly for four concentrations after 5 ms using simulation protocol SP-VLP-3. Back-bone carbon structures are visualized and color indicates self-assembled structure sizes by number of dimer particles (N_{SAS}). Note that red structure at the bottom left of C exceeds the color scale with a structure size of 221, as well as structure 100D with 233 dimers. Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

be seen, such defects appear to be rare. With regard to the influence of concentration, a slightly higher population in the range of 100 – 120 dimers can be observed for the higher concentrations (see Fig. 7.10) supporting experimental findings [367, 368].

In addition to the primary capsid population around 90 - 120 dimers, structures of smaller and larger sizes can be observed in lower quantities. Smaller structures closely

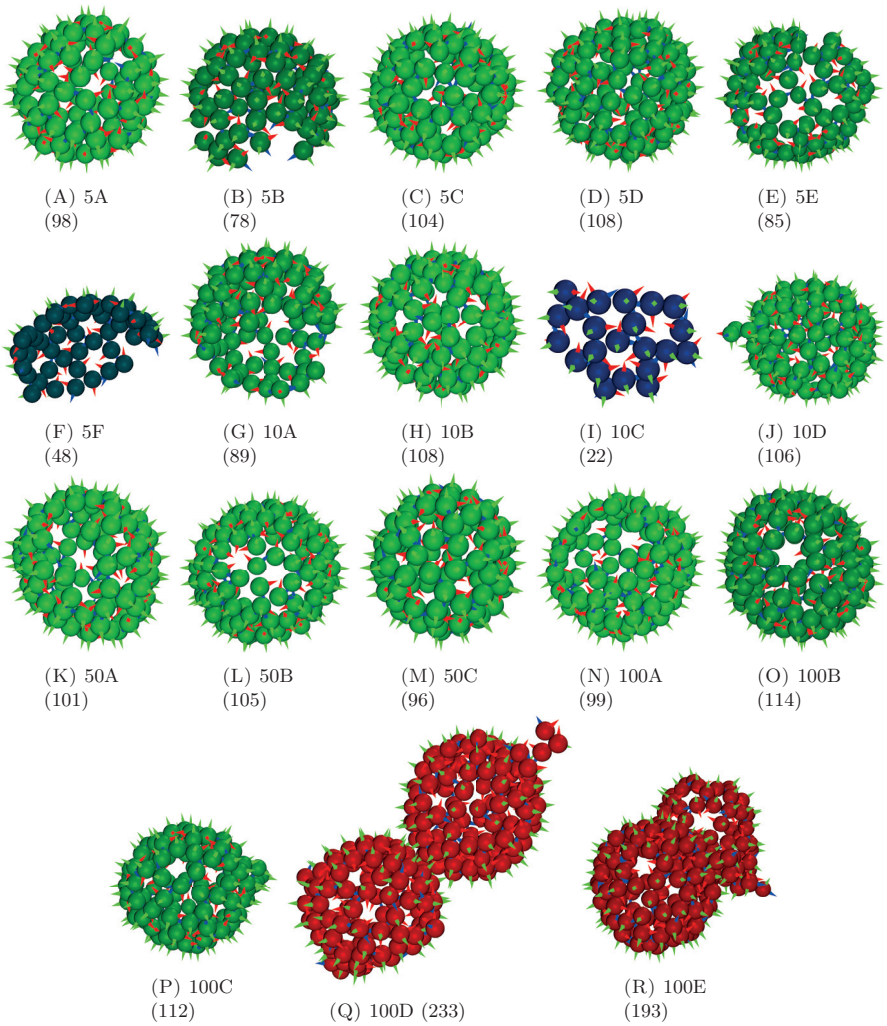
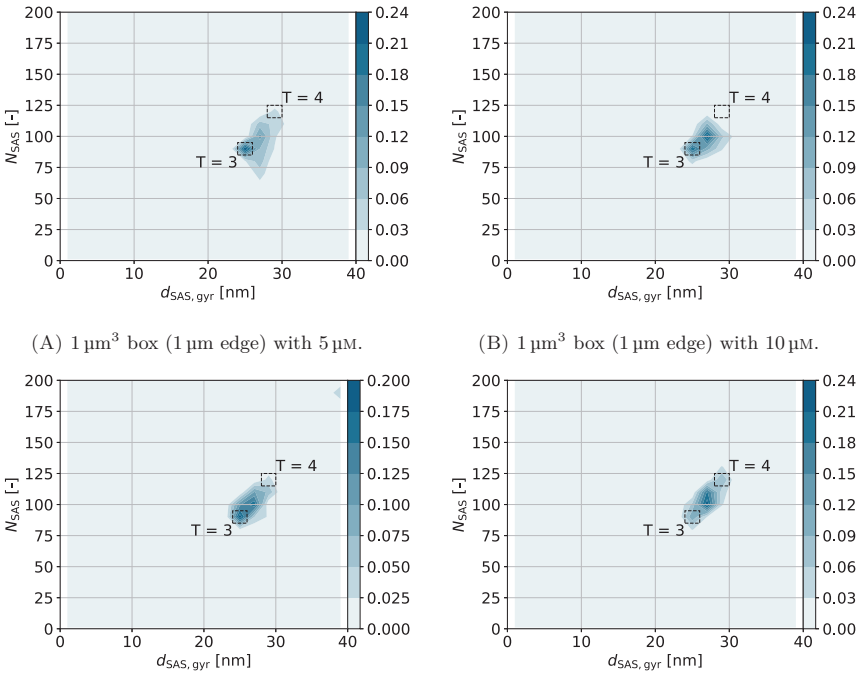


FIGURE 7.9: Close-up visualizations of capsids marked in Fig. 7.8. Spheres with orientation arrows used for visualization of each dimer (x-axis red, y-axis green, z-axis blue). Numbers in brackets behind identifier specify number of $HBcAg_2$ in structure. Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

resemble pre-stages of capsids (e.g. 5F and 10C in Fig. 7.9) and appear to be more frequent for lower concentrations. As will be shown subsequently with regard to assembly kinetics, this can be attributed to a diffusion limitation at these concentrations resulting from decreased diffusivities with increasing structural assembly, as also observed experimentally [254]. In contrast, larger structures beyond 120 dimers can almost exclusively be observed for the higher concentrations starting with 50 µM. While some

over-growing appears to be involved, these structures are also largely caused by contacting capsids (e.g. 100D and 100E in Fig. 7.9), which are generally stable. Nonetheless, this leads to an increase of unstructured contacts per dimer from $\xi_{\text{unstruc}} = 0.34$ for 5 µM to $\xi_{\text{unstruc}} = 0.42$ for 100 µM. These variations of capsid assembly dependent upon the initial concentration, particularly the increase of unstructuredness and kinetic traps for higher concentrations, are well known from experiments [254, 367, 369]. Overall, the self-assembled structures are highly regular and match the expected icosahedral structure of HBcAg capsids [244].



(A) 1 µm^3 box (1 µm edge) with 5 µM.

(B) 1 µm^3 box (1 µm edge) with 10 µM.

(C) 0.125 µm^3 box (0.5 µm edge) with 50 µM.

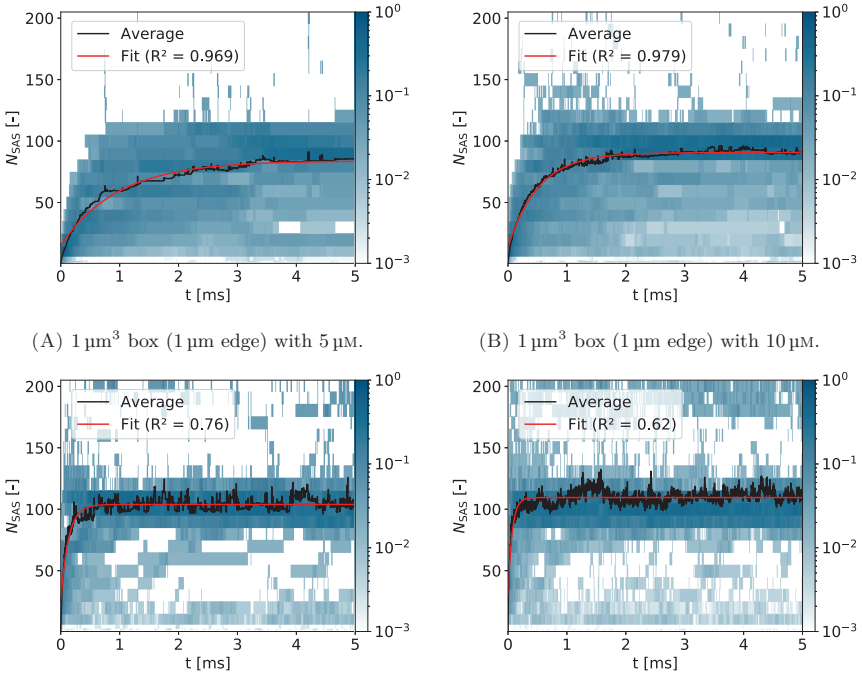
(D) 0.125 µm^3 box (0.5 µm edge) with 100 µM.

FIGURE 7.10: Size fraction of self-assembled structures versus their diameter of gyration at the end of the simulation (average over last ten saving steps). Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

Assembly Kinetics Furthermore, the presented methodology enables detailed insight into the assembly kinetics starting from individual $HBcAg_2$ dimers. In this regard, Fig. 7.11 and 7.12 show the development of population sizes and structuredness of contacts over time, respectively. After an initial growth to the previously discussed primary population of structures around 100 $HBcAg_2$ in size, finalization towards perfect T = 3 and T = 4 capsids significantly slows down exhibiting drastically longer time scales.

These finalization procedures are known from experiments to take between seconds and days [366, 369, 370] and are consequently beyond current computational capabilities.

Regarding the population sizes (Fig. 7.11), it can be observed in agreement with aforementioned assembly properties that for lower concentrations there is a comparatively larger fraction of smaller structures below 90-mers (29.0 % of $HBcAg_2$ for 5 µM and 7.5 % for 100 µM at the end of the simulation), which additionally decreases more slowly than for higher concentrations ($\tau_{SAS} = 2.9$ ms, see Tab. 7.4). Both is attributed to the more dilute solution and thus increased diffusion lengths for self-assembly with decreasing concentration, i.e. a diffusion limitation [254]. Furthermore, for these lower concentrations virtually no assemblies exist above 120-mers. In contrast, with increasing concentration



(A) 1 µm³ box (1 µm edge) with 5 µM.

(B) 1 µm³ box (1 µm edge) with 10 µM.

(C) 0.125 µm³ box (0.5 µm edge) with 50 µM.

(D) 0.125 µm³ box (0.5 µm edge) with 100 µM.

FIGURE 7.11: Histogram of self-assembled structures sizes by number over time. Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

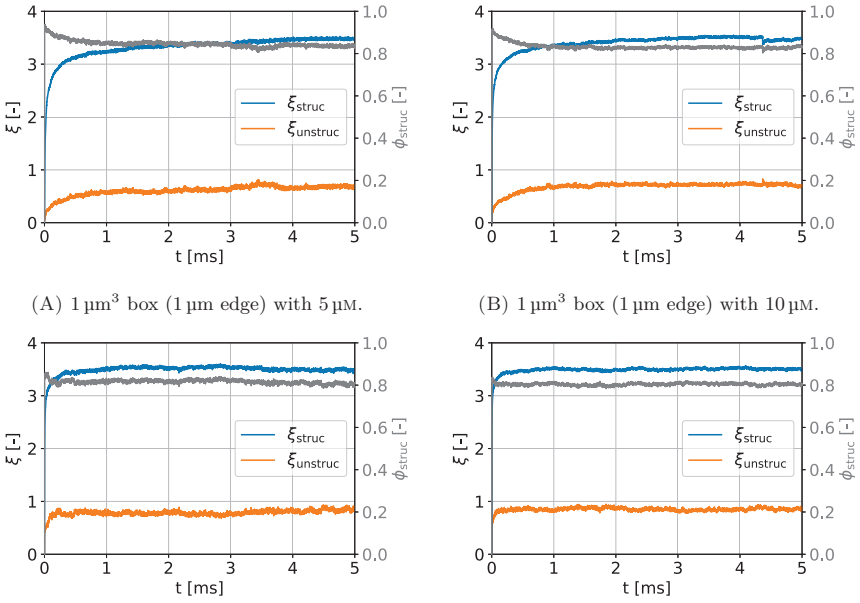
TABLE 7.4: Properties of assembly kinetic fits in Fig. 7.11.

c [µM]	τ_{SAS} [ms]	$N_{SAS,asymp}$ [-]
5	2.91	83.9
10	1.55	90.7
50	0.39	103.8
100	0.21	109.5

the number of larger structures increases and a frequent formation of such is visible in agreement with experimental observations [367, 368]. The lifetime of structures beyond 120-mers is, however, very low on the micro-second scale supporting the aforementioned interpretation of in part colliding capsids, which are otherwise generally stable, additionally to temporarily overgrown capsids. This is further supported by the development of ξ_{struc} and ξ_{unstruc} in Fig. 7.12, which exhibits an increase of ξ_{unstruc} from 0.34 to 0.42 with the concentration change from 5 µM to 100 µM, while ξ_{struc} remains similar for all concentrations. Furthermore, the shift in size distribution with concentration leads to an increase in the average asymptotic structure size $N_{\text{SAS,asymp}}$ from 83.9 to 109.5, as can be seen in Tab. 7.4. Similar observations have been the subject of previous reviews [371].

With regard of the timescales of self-assembly, it can be observed that the critical time constant τ_{SAS} for initial structural assembly increases with decreasing concentration as shown in Tab. 7.4[1]. While it takes $\tau_{\text{SAS}} = 2.9$ ms for initial structural assembly to take

[1]Note that the simulation timescales are only comparable between each other and not real-world timescales as the coarse-grained simulations are inherently accelerated through their abstraction.



(A) 1 µm³ box (1 µm edge) with 5 µM.

(B) 1 µm³ box (1 µm edge) with 10 µM.

(C) 0.125 µm³ box (0.5 µm edge) with 50 µM.

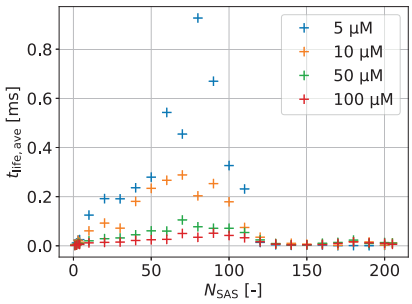
(D) 0.125 µm³ box (0.5 µm edge) with 100 µM.

FIGURE 7.12: Average number of structured and unstructured connections per dimer particle including their relation (right axis). Note a perfect T = 4 (120 dimer) capsid features four structured connections per dimer particle ($\xi_{\text{struc}} = 4$). Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

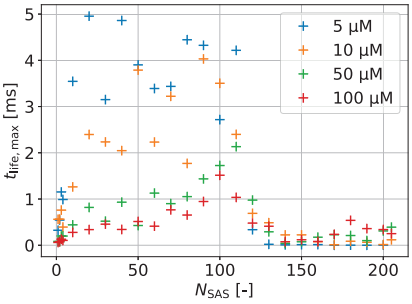
place and ξ_{struc} to converge to 3.5 at a concentration of 5 µM, the same process takes with $\tau_{\text{SAS}} = 0.2$ ms more than one order of magnitude less time for a concentration of 100 µM. Concerning the functional relationship of average structure size over time, good agreement with the fitted asymptotic exponential behavior (R^2 see Fig. 7.11) can be observed for all concentrations, which is in line with literature [371–374], including modeling approaches [375]. Similar to the formation of structures, the fraction of individual dimers falls below 1% in 40 µs for 5 µM and 2 µs for 100 µM. This acceleration is equally attributed to the reduced distances having to be covered via diffusion for self-assembly and comes at the price of a slightly increased unstructuredness of assemblies as seen in Φ_{struc}. While initial self-assembled structures at low concentrations are very structured for low concentrations with $\Phi_{\text{struc}} = 0.84$, this is slightly reduced to $\Phi_{\text{struc}} = 0.80$ for the higher concentrations.

Furthermore, self-assembly can be better understood by observing the distribution of structure lifetimes t_{life} in Fig. 7.13. The average lifetime $t_{\text{life,ave}}$ (Fig. 7.13A) increases from the micro-second scale for small structures with increasing size until reaching a maximum in the range of $t_{\text{life,ave}} = 0.1 - 0.9$ ms between 70-mers and 100-mers ($N_{\text{SAS}} = 70 - 100$) - thus, being in agreement with expectations of capsid assembly [244]. In the range 90-mers to 120-mers, $t_{\text{life,ave}}$ decreases slightly, which is attributed to a re-organization and finalization during T = 4 assembly. Beyond 120-mers, the lifetime decreases drastically to the micro-second scale indicating instability of overgrown structures. Correspondingly, the maximum lifetime $t_{\text{life,max}}$ (Fig. 7.13B) exhibits a sharp decrease above 120-mers, while being in the range of multiple milli-seconds (up to simulation time of 5 ms) for structures between 10-mers and 120-mers, especially at low concentrations. Lastly, with increasing concentration the average lifetime $t_{\text{life,ave}}$ decreases drastically by one order of magnitude. This effect is most prominently shown for the range between 90-mers and 120-mers in Fig. 7.13C and highlights the increasing unstructuredness and necessary re-organization of capsids with increasing concentration. Consequently, the proposed model clearly captures increases in kinetic traps and capsid aggregation for higher protein concentrations well known from experimental VLP yields [254, 367, 369].

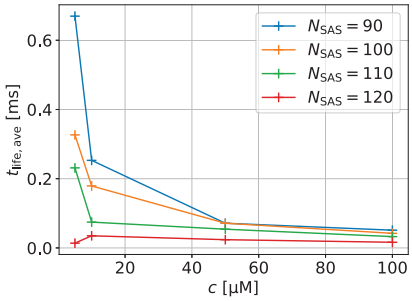
Assembly Pathways Understanding the self-assembly and disassembly pathways of virus capsids is an essential aspect, especially for the development of medical treatments against virus infections, such as antivirals and vaccines. However, while experimental methods provide various insights, e.g. regarding effects of environmental conditions and variations in pathways [361, 369], detailed understanding and development of the knowledge necessary for rational design and prediction of new capsids is limited, e.g. regarding mutations of core proteins [370, 376, 377]. These limitations are even more significant



(A) Average Lifetime vs. Size



(B) Maximum Lifetime vs. Size

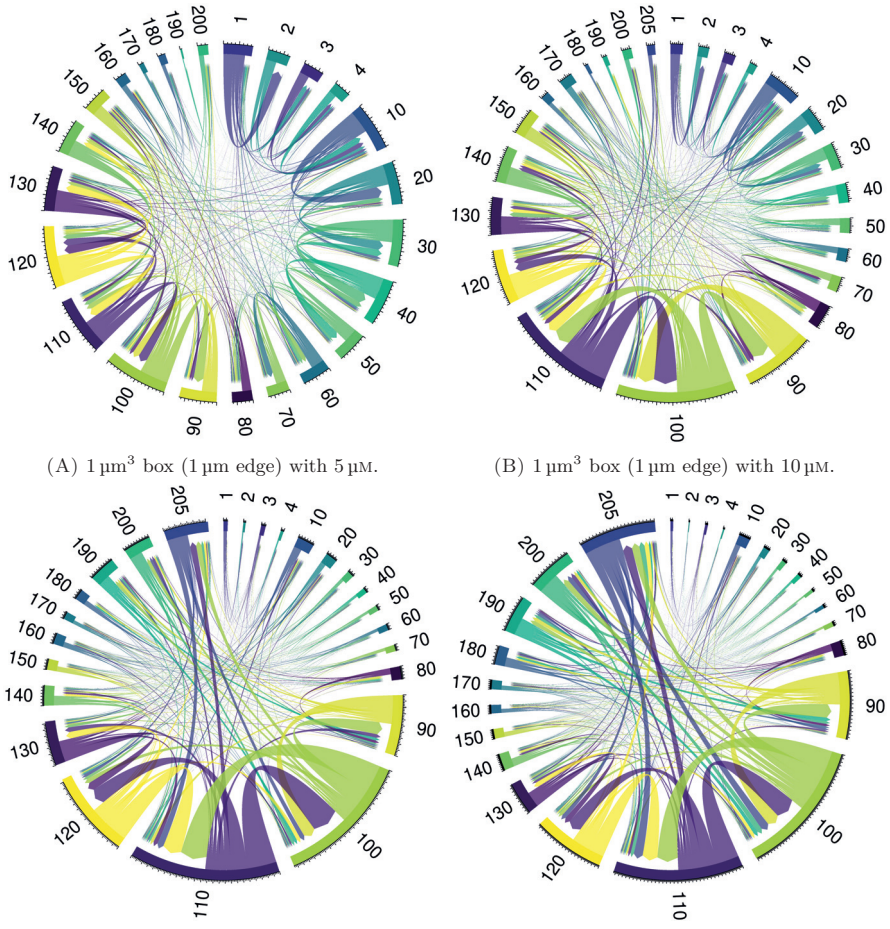


(C) Average Lifetime vs. Concentration

FIGURE 7.13: Lifetimes of structures with regard to their averages (A, C) and maxima (B). Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

than for assembly properties and kinetics as experimental trapping of intermediates is virtually impossible in the nucleation-limited self-assembly [253, 254]. Consequently, the proposed simulation method is highly desired as it not only enables detailed insights into the assembly pathway, but also allows for variations of the core protein (e.g. mutations) explicitly through its parameterization approach. In order to visualize the complex self-assembly and disassembly pathways, chord diagrams [358] are employed showing the transitions between size classes (1, 2, 3, 4, 5 - 14 [10], ..., 195 - 204 [200], ≥205 [205]) normalized by the number of $HBcAg_2$ dimers. In this regard, bi-directional (Fig. 7.14) and net transition (Fig. 7.15) are provided for the four concentrations of the studied $HBcAg_2$ core protein.

As it can be observed in the chord diagrams of Fig. 7.14 and 7.15, the self-assembly process is highly complex and includes various pathways of different probabilities and intermediates. Throughout all concentrations and particularly for lower concentrations, such as 5 µM, a hierarchical step-wise self-assembly takes place from smaller to larger structures until reaching capsid-like sizes around 90-mers to 120-mers. Initially in this



(A) 1 µm^3 box (1 µm edge) with 5 µM.

(B) 1 µm^3 box (1 µm edge) with 10 µM.

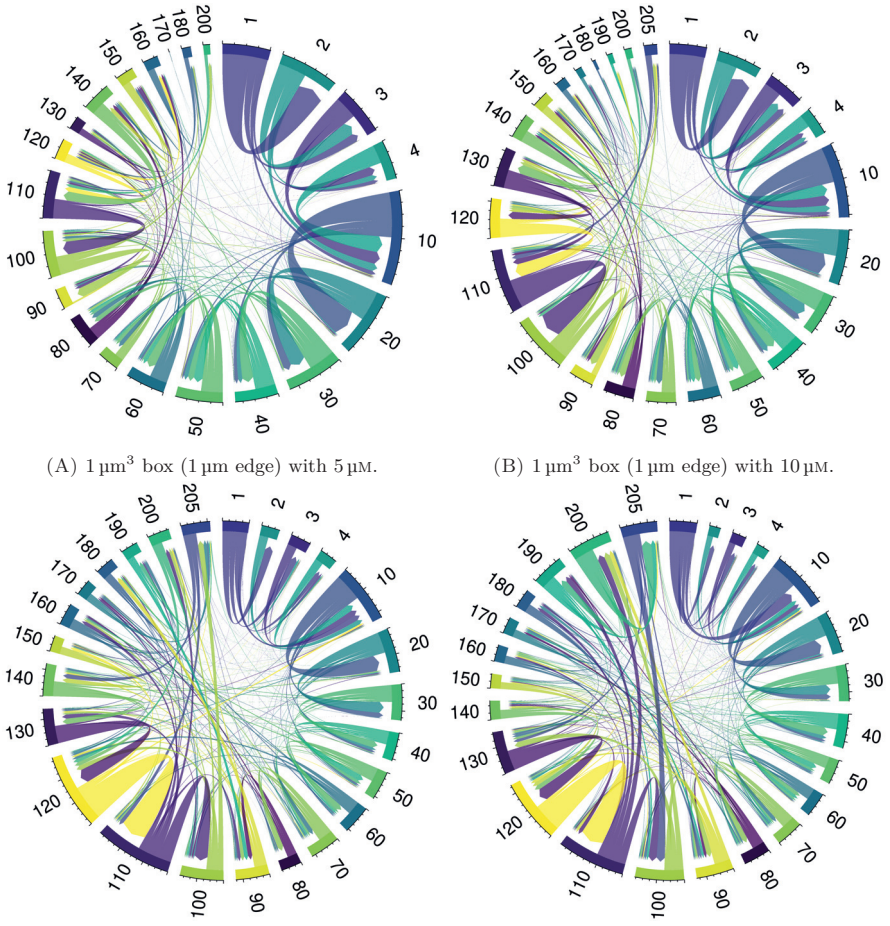
(C) 0.125 µm^3 box (0.5 µm edge) with 50 µM.

(D) 0.125 µm^3 box (0.5 µm edge) with 100 µM.

FIGURE 7.14: Chord diagrams showing transitions between size classes (bi-directional) with multiple transitions possible per dimer particle. Transitions are normalized by the number of dimers (major ticks outer scale are unit one). Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

process, the individual $HBcAg_2$ assemble from their disorganized state into structure of two, three, four, and ten (5 - 14, 10-mer) with decreasing probability. The transition probabilities are concentration dependent with higher concentrations leading to an increased probability of a direct transition from single $HBcAg_2$ (1) to the 10-mer, thus indicating an accelerated assembly with increasing concentration[2]. In the following, larger structures up to the 30-mer (25 - 34) are built primarily from these 10-mer structures (see light purple/blue transitions from 10-mers, particularly in Fig. 7.15), which is

[2]Note the savings interval of 500 ns for class transition analysis.



(A) 1 µm^3 box (1 µm edge) with 5 µM.

(B) 1 µm^3 box (1 µm edge) with 10 µM.

(C) 0.125 µm^3 box (0.5 µm edge) with 50 µM.

(D) 0.125 µm^3 box (0.5 µm edge) with 100 µM.

FIGURE 7.15: Chord diagrams showing net transitions between size classes (i.e. only one direction between all classes) with multiple transitions possible per dimer particle. Transitions are normalized by the number of dimers (major ticks outer scale are unit one). Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

in line with recent experimental observations [361]. The importance of the 10-mer population is further highlighted by being the first population class of significantly higher average and maximum lifetimes (see Fig. 7.13).

Once these initial small intermediates have formed, subsequent assembly continues primarily through addition towards the next size class until approximately the 80-mer population. Consequently, highlighting the hierarchical nature of the self-assembly pathway, which can also be observed visually in the chord diagrams with few transitions directly

to larger structures, i.e. through the center of the diagram (Fig. 7.14 and 7.15). Subsequently larger structures, such as the 90-mer, exhibit a shift towards a pathway of disassembly from overgrown structures (see e.g. light green transition from 100-mer to 90-mer accounting for the largest transition class to 90-mer in Fig. 7.14). Nonetheless, combined pathways through both step-wise growth and overgrowth can also be observed, particularly for the 120-mer population. This multiplicity of pathways to capsid assembly including overgrowth has previously been observed by Lutomski *et al.* [369].

In this context, the large number of intermediates and transitions in the range of 100-mers and 110-mers, i.e. 95 - 114 $HBcAg_2$ (e.g. 5A, 10B, and 100C in Fig. 7.9), are crucial to be noted for the assembly pathway. Similarly, experimental investigation have shown such intermediates in the assembly pathway [378] (e.g. 104/105-mer similarly to 5C and 50B in Fig. 7.9; and 110/111-mer similarly to 100C in Fig. 7.9). As discussed previously, these intermediates are in addition to T = 3 capsids most likely pre-stages of the T = 4 capsid, which can be considered semi-stable. On their self-assembly pathway, a partial disassembly and re-organization takes place including separation and addition of smaller structures, if available in proximity. Consequently, finalization of the capsid, particularly T = 4 capsid, becomes a question of simulation time and availability of small structures, i.e. a thermodynamic process of structured docking and self-assembly versus unstructured and thus unstable contacts. Generally, also T = 3 appear to have formed (see Fig. 7.10 and 7.13), which is experimentally known to be very sensitive with regard to initial assembly conditions [361].

Beyond sizes of the 120-mer, formation of structures is notably concentration dependent with high concentrations exhibiting temporary formation of large aggregates (>205 $HBcAg_2$). Particularly, for the high concentrations of 50 µM and 100 µM the majority of structures follow this pathway during capsid assembly with short term overgrowth (lifetimes on the micro-second scale, see Fig. 7.13). However, note as previously discussed the influence of the contact of otherwise well-structured capsids. In contrast, for the low concentrations of 5 µM and 10 µM this pathway is significantly less pronounced with few structures overgrowing beyond the range of 130-mer to 150-mer.

In summary, the self-assembly pathway is observed to employ a hierarchical buildup of sequentially larger structures combined with a semi-stable equilibrium between 90-mers and 120-mers for finalization of capsids through addition and hole closure. Larger concentrations, specifically above 50 µM, appear to cause temporary overgrowth above 120-mers highlighting another pathway to capsid formation through overgrowth. These findings are in good agreement with experimental insights, e.g. regarding pathway through overgrowth [369] and intermediates [378], and provide many additional insights through the high level of detail of the proposed modeling approach.

8

Results: PDC System

This chapter is based on the following publications:

P. N. Depta, U. Jandt, M. Dosta, A.-P. Zeng, and S. Heinrich. Toward Multiscale Modeling of Proteins and Bioagglomerates: An Orientation-Sensitive Diffusion Model for the Integration of Molecular Dynamics and the Discrete Element Method. *J. Chem. Inf. Model.*, 59(1):386–398, 2019

P. N. Depta, M. Dosta, and S. Heinrich. Data-Driven Multiscale Modeling of Self-Assembly and Hierarchical Structural Formation in Biological Macro-Molecular Systems: Pyruvate Dehydrogenase Complex. In W. E. Nagel, D. H. Kröner, and M. M. Resch (editors), *High Performance Computing in Science and Engineering '22*. Springer International Publishing, Cham, 2024 (in print)

P. N. Depta, M. Dosta, and S. Heinrich. Multiscale Model-Based Investigation of Functional Macromolecular Agglomerates for Biotechnological Applications. In A. Kwade and I. Kampen (editors), *Dispersity, Structure and Phase Changes of Proteins and Bio Agglomerates in Biotechnological Processes*. Springer International Publishing, Cham, 2024 (in print)

8.1 Model Parameters

8.1.1 Structural Model

In order to study the structural self-assembly of the pyruvate dehydrogenase complex (PDC), each of the four components E1, E2, E3, and E3BP is abstracted as an anisotropic unit object as presented in Sec. 1.3.3 and Ch. 2. More details on PDC and the respective reference conformations of the individual components can be found in

Sec. 1.3.3. Masses and radii of gyration are specified in Tab. 8.1. The abstracted four components are freely interacting with each other and the implicit solvent environment. Each abstracted object possesses a position, orientation, as well as spatial extend and other anisotropic properties through its functional models, which will be specified in the following.

TABLE 8.1: Masses and radii of gyration of PDC components based on Depta *et al.* [223]. Table adapted with permission from Depta *et al.* [223]. Copyright 2019 American Chemical Society.

Kind	m [kg]	$r_{g,x}$ [nm]	$r_{g,y}$ [nm]	$r_{g,z}$ [nm]
E1	2.42×10^{-22}	2.51	2.60	2.95
E2	9.81×10^{-23}	1.44	9.96	9.93
E3	1.67×10^{-22}	1.98	2.74	2.76
E3BP	7.95×10^{-23}	1.60	10.68	10.72

8.1.2 Functional Model

8.1.2.1 Diffusion and Thermodynamics

Anisotropic diffusion and the respective thermodynamics of the desired canonical ensemble were modeled and parameterized for each PDC component using the method presented in Ch. 3. The model captures the interaction of the structurally flexible macromolecules with the solvent. The determined diffusion coefficients for the PDC components at conditions of 300 K without additional salts in aqueous solution can be found in Tab. 8.2. Hydrodynamic interaction was accounted for through reduction of the effective viscosity by a factor of 0.1 as discussed in Sec. 3.2 and App. A.3.

TABLE 8.2: Diffusion coefficients of PDC components at 300 K without additional salts. Table adapted with permission from Depta *et al.* [223]. Copyright 2019 American Chemical Society. Note slight changes in the diffusion coefficients of E2 due to an update of the reference structure in comparison to Depta *et al.* [223].

Kind	D_t [µm^2 s^{-1}]			D_r [Mrad^2 s^{-1}]		
	x	y	z	α	β	γ
E1	54.9	52.5	52.0	2.96	2.78	2.35
E2	49.6	42.9	41.0	10.6	0.60	0.55
E3	66.5	53.0	58.1	5.16	3.33	3.08
E3BP	53.8	45.4	45.3	12.3	0.48	0.50

8.1.2.2 Intermolecular Interaction

The intermolecular interaction of PDC components was modeled and parameterized as presented in Sec. 4.3. Overall, ten pairwise interaction permutations exist for the four

PDC components. Due to the computational cost of parameterization, not all of the ten interaction pairs could be parameterized. In the context of PDC, the most important interactions are E2 – E2 for assembly of the 60-mer core and trimers, as well as E2 – E1 and E3BP – E3 for binding of enzymes necessary for catalytic activity. Most of the remaining interaction pairs are less relevant and thus subject to simplification. In the context of this work and the available computational resources (13.25 million CPU core-hours on the Hawk system at HLRS, Acid 44178 [227]), the following simplifications were chosen in line with literature (see Sec. 1.3.3) and are summarized in Tab. 8.3:

- As the interactions of all combinations of E1 and E3 are not known to attractively interact, their interaction is assumed to be repulsive only (i.e. due to atomic overlap). Specifically, this includes E1 – E1, E1 – E3, and E3 – E3.
- As E2 and E3BP are similar in structure / function (E3BP is an additional enzyme in most eukaryotic PDC, e.g. human, for specific binding of E3 [259, 260, 265, 379–381]), their interaction with each other (E2 – E2, E2 – E3BP, E3BP – E3BP) is assumed equivalent and the potential field for E2 – E2 used for all such interactions.
- As E3BP is known to specifically bind E3 [259, 260, 265, 379–381], the interaction E3BP – E1 is assumed to be repulsive only (i.e. due to atomic overlap).
- As E3 is typically thought to not bind to E2 for human PDC [18, 263, 379, 382], but has been found to possess residual affinity in E3BP deficient cases [383], the interaction E2 – E3 is assumed to be a weaker version of E2 – E1. For this, the relative binding strength was determined as a fraction of E2 – E1 to be 0.746 through statistical binding analysis in MD by Jacobi *et al.* [384] (E2 – E3 = 0.746 × E2 – E1). Results will later show that this still leads to a higher binding specificity of E3 to E3BP and to a lesser extent to E2 as expected [18, 263, 379, 382].

Table 8.3: Overview of PDC component intermolecular interaction simplifications. A+R = attractive + repulsive; R = repulsive-only based on molecular collisions model; FP = fully parameterized from MD; D = derived from a similar interaction parameterized from MD.

	E1	E2	E3	E3BP
E1	R	A+R (FP)	R	R
E2	A+R (FP)	A+R (FP)	A+R (D)	A+R (D)
E3	R	A+R (D)	R	A+R (FP)
E3BP	R	A+R (D)	A+R (FP)	A+R (D)

For the remaining interaction pairs, initial MD sampling was performed based on distances classes as specified in Sec. 4.3.4.3 and detailed in Tab. 8.4 for a total of 100'000 to 150'000 samples per interaction pair. Furthermore, proximity resampling was performed

TABLE 8.4: Random sampling over distances classes for PDC data sets before iterative resampling. Note that the number of samples is essentially double (or interaction space half in volume) for interaction partners A = B, i.e. E2 = E2.

d_{A-B} [nm]		# samples		
lower	upper	E2 – E2	E2 – E1	E3BP – E3
0.4	0.5	25'000	25'000	25'000
0.5	0.7	15'000	5'000	5'000
0.7	0.9	15'000	5'000	5'000
0.9	1.1	15'000	5'000	5'000
1.1	1.3	15'000	5'000	5'000
1.3	1.5	15'000	5'000	5'000
1.5	1.7	5'000	5'000	5'000
1.7	1.9	5'000	5'000	5'000
1.9	2.1	5'000	5'000	5'000
2.1	2.3	5'000	5'000	5'000
2.3	2.5	5'000	5'000	5'000
2.5	3.0	5'000	5'000	5'000
3.0	3.5	5'000	5'000	5'000
3.5	4.0	5'000	5'000	5'000
4.0	4.5	5'000	5'000	5'000
4.5	5.0	5'000	5'000	5'000
Sum		150'000	100'000	100'000

to improve variogram statistics at short distances as specified in Sec. 4.3.4.3 and detailed in Tab. 8.5 for a total of 25'000 samples per interaction pair. Proximity resampling becomes increasingly necessary with the large interaction spaces of extended molecules such as E2 and E3BP. Starting from these initial samples, iterative refinement was performed as specified in Sec. 4.3.4.3 using the following sequence of refinement criteria and number of samples:

- Iteration 1 - 5: Variance minimization using 5'000 samples for each iteration.
- Iteration 6 - 7: Extrema resampling using 7'500 samples for each iteration.
  - Potential minima resampling using 2'000 samples for main extrema points and 500 samples for first-level neighborhood points,
  - Potential maxima resampling using 2'000 samples for main extrema points and 500 samples for first-level neighborhood points,
  - Gradient maxima resampling using 2'000 samples for main extrema points and 500 samples for first-level neighborhood points.

Overall, between 165'000 and 215'000 MD data points were sampled and analyzed for estimation of each intermolecular interaction potential. Fields were discretized with a

resolution of 0.8 nm for E2 – E1 / E3BP – E3, 1.3 nm for E2 – E2, and 1.5 nm for all all repulsive potentials – leading to an overall size of 36.2 GB. Similarly to the HBcAg VLP system, in addition to the pure MD-based interaction potentials supplementation by inclusion of empirical data from known binding locations was performed, which are specified in Appendix E.1. The resulting potentials will be presented in detail in Sec. 8.4.1.

TABLE 8.5: Proximity resampling for improved statistical correlation information at short distances. At each location the number of replicates is run. These data points do not influence the trend or are used for Kriging besides Variogram modeling. Note that the number of samples is essentially double (or interaction space half in volume) for interaction partners A = B, i.e. E2 = E2.

d_{A-B} [nm]		# rep.	# locations			total # samples		
lower	upper	/ loc.	E2-E2	E2-E1	E3BP-E3	E2-E2	E2-E1	E3BP-E3
0.5	0.7	50	50	50	50	2'500	2'500	2'500
0.7	0.9	50	50	50	50	2'500	2'500	2'500
0.9	1.1	50	50	50	50	2'500	2'500	2'500
1.1	1.3	50	50	50	50	2'500	2'500	2'500
1.3	1.5	50	50	50	50	2'500	2'500	2'500
1.5	1.7	50	50	50	50	2'500	2'500	2'500
1.7	1.9	50	50	50	50	2'500	2'500	2'500
1.9	2.1	50	50	50	50	2'500	2'500	2'500
2.1	2.3	50	50	50	50	2'500	2'500	2'500
2.3	2.5	50	50	50	50	2'500	2'500	2'500
Sum						25'000	25'000	25'000

8.1.2.3 Bonded Interaction

No bonded interaction was modeled for the PDC system. All intermolecular interactions for the self-assembly process were captured by the interaction model in Sec. 8.1.2.2.

8.1.2.4 Critical Time Step

In order to estimate the critical time step τ_{crit} necessary for a convergent and numerically stable solution, the previously discussed methods for the individual model components were used, see Sec. 3.4.1 and 4.5. Based on this, the critical time step can be estimated as 5.2×10^{-12} s for the diffusion model dominated by E2 when accounting for hydrodynamic interaction through reduced viscosity as discussed in Sec. 3.2 and App. A.3 (5.2×10^{-13} s without hydrodynamic interaction at regular dynamic viscosity), as well as 1.3×10^{-12} s for the intermolecular interaction model dominated by E2 – E2. Consequently, without hydrodynamic interaction the diffusion model constrains the overall simulation time step and when accounting for hydrodynamic interaction the intermolecular interaction model

does (default). Therefore, unless otherwise indicated, a simulation time step of 10^{-12} s was used leading to an error of 0.6 % with regards to RMS displacement (see Sec. 3.4.1).

8.2 Simulation Setup and Procedure

PDC Self-Assembly (SP-PDC-1) Self-assembly of PDC components was investigated at a constant temperature of 300 K without any additional ions etc. based on randomly initialized system according to the composition specified for each case. All components were placed and oriented randomly in a cubic box of 1 µm^3 and concentration of 1 mg/mL with periodic boundary conditions applied throughout. During placement, overlap of units was permitted and then corrected during the simulation through the intermolecular interaction model. In order to account for the non-dilute state during self-assembly, the effective viscosity was reduced by a factor of 0.1 to 8.5416×10^{-5} Pa s as discussed detailed in Sec. 3.2 and App. A.3 enabling a simulation time step of 10^{-12} s. Based on this, self-assembly simulations were performed for 5 ms with a savings interval of 500 ns.

PDC Self-Assembly with Simulated Annealing (SP-PDC-1-AN) As the SP-PDC-1 simulation procedure at constant temperature was found to be prone to trapping in local potential minima during structural formation similarly to alginate gelation (Ch. 6), various simulated annealing procedures were tested for PDC assembly. Based on a self-assembly analysis of pure E2 at 1 mg/mL concentration, it was found that the following parameter set corresponding to the annealing model provided in Sec. 3.6 provides a good balance between temporary breakage due to temperature increase and stability: $\tau_{\text{an,cool}} = 2\,\mu\text{s}$, $\tau_{\text{an,period}} = 5\,\mu\text{s}$, $t_{\text{an,finished}} = 4\,\text{ms}$, linear temperature decay, $T_{\text{an,max}} = 1300\,\text{K}$, total simulation time 5 ms with a savings interval of 500 ns.

8.3 Analysis and Postprocessing

Simulation results were analyzed both online and in postprocessing to study a variety of features with focus on structural properties. The analyzed system properties include the following and are supplemented from the HBcAg model system in Sec. 7.3 and Depta *et al.* [225]:

- self-assembled structures (SAS) were identified using a network search algorithm distinguishing between structured contacts ($\delta_r \leq 3.5\,\text{nm}$ from a known binding

location, see Tab. E.1) and unstructured contacts ($\delta_m \leq 1\,\text{nm}$ and $\delta_r > 3.5\,\text{nm}$ from a known binding location). Based on these, the following properties were investigated and can be readily compared to known structures like the 60-mer:

  - N_{SAS} is the number of molecules in each SAS;
  - $d_{\text{SAS,ave}}$ is the average extent of backbone atoms of each SAS (averaged over a discretization of 162 orientations). Size of reference 60-mer provided by Hezaveh *et al.* [271] based on MD is 51.2 nm and marked ±5 nm in plot;
  - $d_{\text{SAS,gyr}}$ is the diameter of gyration of each SAS (provided in App. E.5). Size of reference 60-mer provided by Hezaveh *et al.* [271] based on MD is 43.0 nm and marked ±5 nm in plot;
  - $d_{\text{SAS,max}}$ is the maximum extent of backbone atoms of each SAS (provided in App. E.5). Size of reference 60-mer provided by Hezaveh *et al.* [271] based on MD is 68.0 nm and marked ±5 nm in plot;
  - ξ_{struc} is the number of structured contacts normalized by the number of molecules in the system, can also be used per SAS or per molecule;
  - ξ_{unstruc} is the number of unstructured contacts normalized by the number of molecules in the system, can also be used per SAS or per molecule;
  - N_{kind} is the stoichiometry by number of molecules relative to the E2 content in a 60-mer (determination restricted to $N_{\text{SAS}} \geq 5$);
  - Φ_{kind} is the stoichiometry by molar fraction for each SAS;

- assembly kinetics were quantified by exponential fitting of average N_{SAS} excluding monomers, see Sec. 4.3.3 eq. 4.17 with time constant τ_{SAS} (r in eq. 4.17) and asymptotic structure size $N_{\text{SAS,asymp}}$ (s in eq. 4.17);
- transitions of molecular assemblies between size classes (1, 2, 3, 4, 5 - 14 [10], 15 - 24 [20], ...) based on N_{SAS} were summarized and normalized by the number of molecules to capture assembly mechanisms (note saving step of 500 ns). Bi-directional and net transitions between classes were distinguished and visualized using chord diagrams [358];
- independent of structural recognition, molecular interaction was analyzed based on:
  - f_{bind} is the (un-)binding rate per molecule type (uniform smoothing applied over 10 µs);
  - $N_{\text{i-j}}/N_{\text{E2}}$ is the number of reactant / active-site (AS) combinations i and j at a specific center-of-mass distance d_{COM} normalized by the number of E2 molecules in the system;

  - N_{cont} is the number of contacts for each molecule type with other molecule types (provided in App. E.5);

- K_D is the equilibrium dissociation constant, also called affinity constant, calculated over the last 0.1 ms as the ratio $(c_A \times c_B)/c_{AB}$, where c_A is the concentration of species A (e.g. unbound E2), c_B is the concentration of species B (e.g. unbound E1), and c_{AB} is the concentration of structured compound AB (e.g. E2+E1).

In addition, global properties such as potential and kinetic energies were analyzed. However, as these provided little additional insight focus was placed on the structural features.

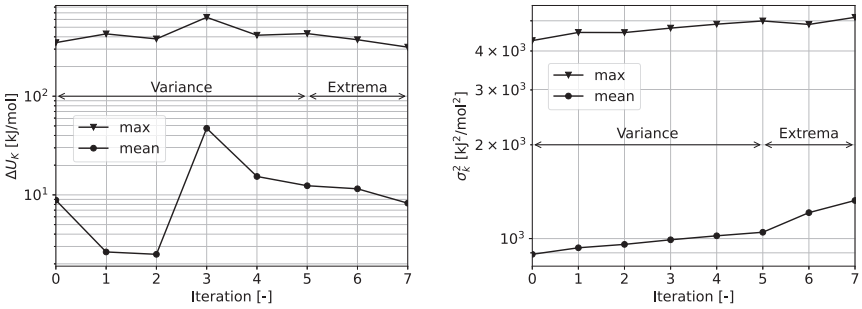
8.4 Results

In the following section, the results of PDC self-assembly will be presented based on the four components E1, E2, E3, and E3BP. In this context, first the intermolecular interaction potentials will be discussed building upon the insights of the $HBcAg_2$ system. Second, structural self-assembly will be investigated using the fully parameterized model.

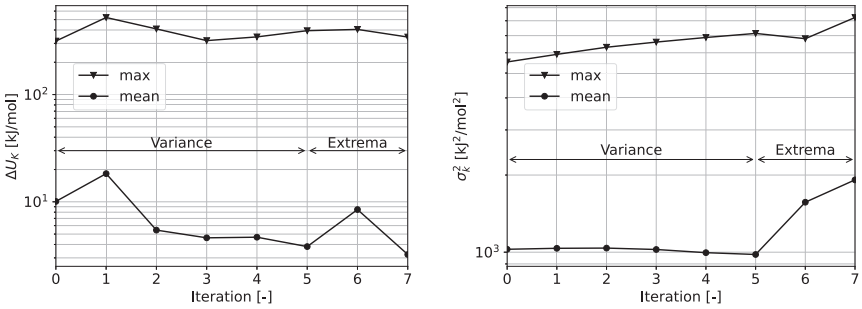
8.4.1 Intermolecular Interaction Potentials

As previously discussed regarding the determination of the $HBcAg_2$ – $HBcAg_2$ interaction potential (Sec. 7.4.1), limited sampling in the context of available computational resources in combination with remaining conformational changes during binding drives the need for additional data beyond pure MD and can be addressed by inclusion of empirical data (e.g. binding locations from crystallography, see Sec. 7.4.1.3). While slightly more computational resources were available for the PDC system (13.25 million CPU core-hours on the Hawk system at HLRS, Acid 44178 [227]), similar limitations applied due to the increased complexity of the PDC system. As it can be seen in the convergence plots for PDC component interaction in Fig. 8.1, such sampling limitations are even more pronounced for this model system, which is attributed to the spatial extent of the involved molecules (specifically E2 and E3BP, see introduction Fig. 1.4 and 1.5) and consequently decreased sampling density of the interaction region between two molecules. Furthermore, the high flexibility of linker arms requires larger numbers of samples to sufficiently cover the various binding states. Consequently, inclusion of empirical data from known binding sites becomes necessary and was directly performed similarly to the HBcAg system, which will be presented subsequently. Full data of the

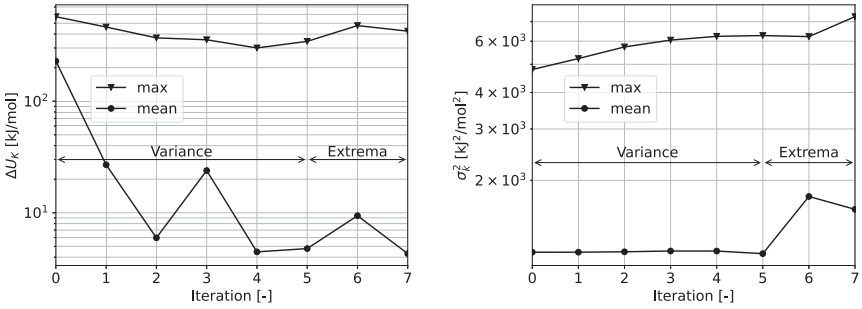
pure MD-based sampling including statistical data from Kriging is provided in App. E.4 and E.2, respectively.



(A) E2 - E2 Refinement.



(B) E2 - E1 Refinement.



(C) E3BP - E3 Refinement.

FIGURE 8.1: Convergence of iterative resampling procedure concerning potential changes (left, $U_i - U_{i-1}$) and variance development (right). Note that iterations 1-5 perform resampling for variance minimization (5'000 samples each) and iterations 6-7 for extrema identification (e.g. binding or repulsive locations, 7'500 samples each). Adapted with permission from Springer Nature regarding Depta *et al.* [227].

In order to incorporate the empirical data on binding sites (App. E.1), the approach determined for $HBcAg_2$ – $HBcAg_2$ (Sec. 4.3.5 and 7.4.1.3) was slightly updated to

the stability of an E2 60-mer, as well as bound states of E1 to E2 and E3 to E3BP. As expected for the highly flexibly linker arms involved in binding and the resulting more flexibly spatial organization, this leads to an increase of the binding radius with $r_{\text{emp,bind}} = 2.0$ nm and $d_{\text{step,outer}} = 0.8$ nm as parameters for E2 – E2 ($HBcAg_2$ – $HBcAg_2$ was $r_{\text{emp,bind}} = 1.0$ nm and $d_{\text{step,outer}} = 0.2$ nm). Furthermore, due to the binding domain being located directly on the flexible linker arm of E2 – E1 and E3BP – E3, for these interaction pairs the binding radius was increased to $r_{\text{emp,bind}} = 8.0$ nm and $d_{\text{step,outer}} = 1.6$ in addition to a potential reduction to $U_{\text{emp,bind,center}} = -3000$ kJ/mol and $U_{\text{emp,bind,outer}} = -2000$ kJ/mol.

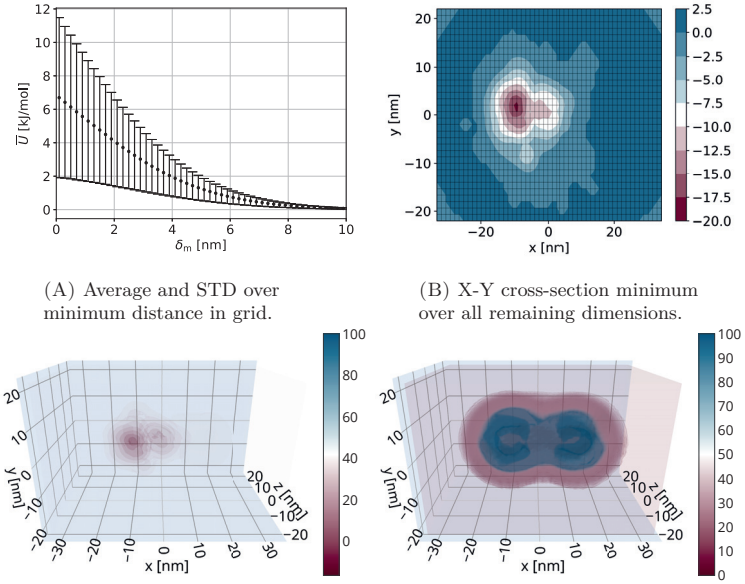
Similarly motivated, the molecular repulsion model (see Sec. 4.3.6) was updated to $c_{\text{rep,bb}} = 100$ kJ/mol, $N_{\text{min,bb}} = 1$, $d_{\text{MD,min}} = 0.3$ nm, $w_{\text{MD,min}} = 0.5$ nm, $w_{\text{coll}} = 0.3$ nm, and $c_{\text{rep,side}} = 0$ kJ/mol as parameters, as well as including a smoothing of the Kriging potential with $w_{\text{krig}} = 3$ nm. These parameters for the repulsion model were used consistently for all pairwise interaction of PDC components. Note that these changes in comparison to HBcAg are primarily motivated by the high flexibility of the E2 and E3BP linker arms, as these conformational features and their flexibility are at the limit of the proposed methodology, which assumes semi-stable macromolecules during structural assembly.

However, while the inclusion of empirical data is necessary overall, it can be noted that the pure MD-based potential provides qualitatively similar results for the E2 – E2 interaction potential (see Fig. 8.2 for pure MD-based vs. Fig. 8.3 with empirical data). As it can be seen in Fig. 8.2 B and C, the global potential minimum of pure MD-based E2 – E2 interaction is similarly located to the binding location of catalytic domains inside the 60-mer core (for reference see MD with empirical data in Fig. 8.3 B and C). However, the empirical potential minimum is significantly deeper and wider resulting from the large variation in binding locations in turn resulting from flexibility of the linker arm. Furthermore, the pure MD-based potential contains a slightly repulsive potential over the minimum distance with approximately 6 kJ/mol and significantly less variations (standard deviations of up to approximately 4 kJ/mol vs. up to 30 kJ/mol), which is attributed to the emphasized binding location by the empirical data. Overall, positive agreement between the purely MD-based interaction potential and with additional enhancements from empirical data can be noted in light of limited computational resources (6.22 million CPU core-hours for E2 – E2).

Regarding E2 – E1 and E3BP – E3, the interaction potentials with additional empirical data are displayed in Fig. 8.4 and 8.5, while pure MD-based potentials are provided in App. E.2. As it can be seen, the interaction potentials with empirical data nicely outline binding at the respective binding domains along the linker arms (see subfigures

B – D in comparison to molecular structure provided in introduction Fig. 1.4 and 1.5). These potential minima lead to an overall slightly attractive behavior over the minimum distance with significant variation depending upon specific location (subfigure A). Regarding these two interaction pairs, the pure MD-based potentials exhibit more challenges attributed to sampling as in contrast to E2 – E2 no symmetry exists (interaction partners are not equal). Particularly, for the case E3BP – E3 an additional strong interaction of E3 with the catalytic domain of E3BP is found (Fig. E.2), as well as for E2 – E1 an interaction of E1 with the catalytic domain of E2 and to a lesser extend the inner lipoyl domain dominate (Fig. E.1). Consequently, for these interaction partners the additional empirical data is similarly needed as for the HBcAg system due to limited sampling and computational resources (2.71 and 2.38 million CPU core-hours for E2 – E1 and E3BP – E3, respectively).

Enhanced E2 – E2 Arm Interaction Furthermore, in this context the precise interaction of two E2 linker arms (catalytic domains on opposite sites) and possibly a subsequent structural assembly beyond the 60-mer is an additional topic of interest. While PDC is almost exclusively known to form the 60-mer core [259, 264, 266, 385], Guo *et al.* [269] have observed the additional presence of a larger size fraction around r_h = 75.2 nm using dynamic light scattering (DLS) on wild-type human E2/E3BP systems, i.e. without any genetic modifications. In order to improve understanding, a variation of the E2 – E2 interaction potential with enhanced arm interaction has been derived and is presented in App. E.5.3 along with its impact on PDC assembly. Summarized briefly, those simulation studies find it unlikely that such large agglomerates can maintain similar catalytic activity as binding of E1 and E3 necessary for the reaction pathway was found to be severely inhibited. Consequently, the following section will present the self-assembly of PDC without such variations and in line with the established 60-mer core.



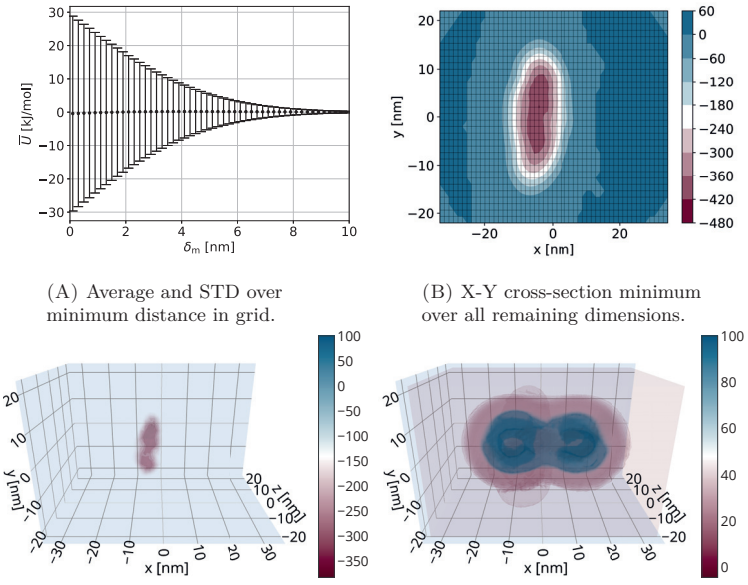
(A) Average and STD over minimum distance in grid.

(B) X-Y cross-section minimum over all remaining dimensions.

(C) 3D min. over orientations.

(D) 3D mean over orientations.

FIGURE 8.2: Interaction potential field of **E2 with E2 from pure MD**-based sampling strategy (units of kJ/mol).



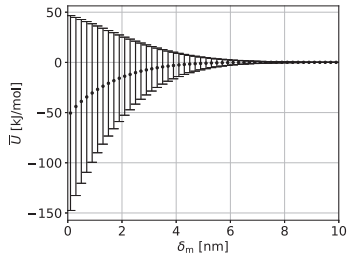
(A) Average and STD over minimum distance in grid.

(B) X-Y cross-section minimum over all remaining dimensions.

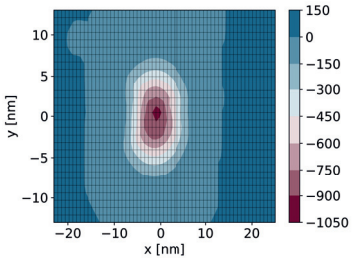
(C) 3D min. over orientations.

(D) 3D mean over orientations.

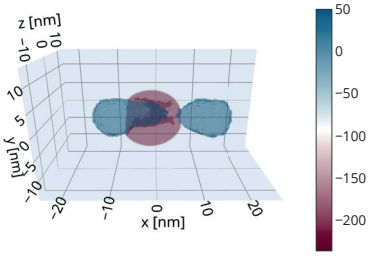
FIGURE 8.3: Interaction potential field of **E2 with E2 from MD with empirical data** (units of kJ/mol).



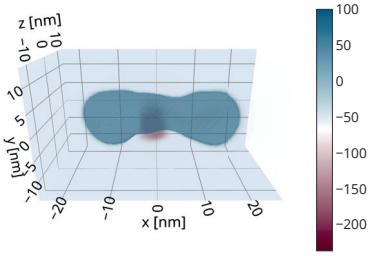
(A) Average and STD over minimum distance in grid.



(B) X-Y cross-section minimum over all remaining dimensions.

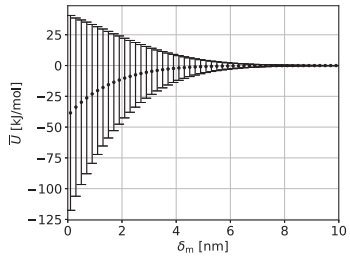


(C) 3D min. over orientations.

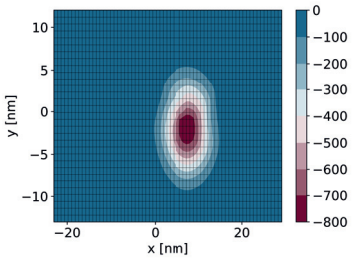


(D) 3D mean over orientations.

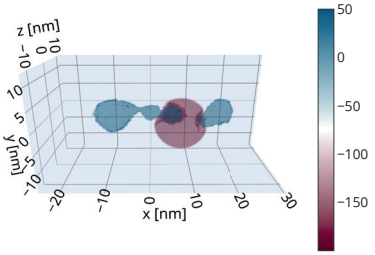
FIGURE 8.4: Interaction potential field of **E2 with E1 from MD with empirical data** (units of kJ/mol). C and D adapted with permission from Springer Nature regarding Depta *et al.* [228].



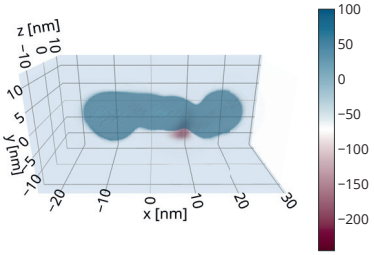
(A) Average and STD over minimum distance in grid.



(B) X-Y cross-section minimum over all remaining dimensions.



(C) 3D min. over orientations.



(D) 3D mean over orientations.

FIGURE 8.5: Interaction potential field of **E3BP with E3 from MD with empirical data** (units of kJ/mol).

8.4.2 PDC Self-Assembly

In the following section, the self-assembly of PDC will be investigated using the proposed model framework and MD-based interaction potential with empirical data shown in Sec. 8.4.1. First, self-assembly of the E2 component into the known 60-mer cores of PDC will be investigated at 1 mg/mL employing simulation procedure SP-PDC-1-AN (Sec. 8.2). Beforehand, it was found this simulated annealing procedure is necessary for overcoming local potential minima in 60-mer formation leading to early core closure. For reference, simulations without annealing (SP-PDC-1) are provided in App. E.5. Second, the full PDC system will be investigated with a stoichiometry $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at 1 mg/mL employing simulation procedure SP-PDC-1-AN. In the following, the results of self-assembly will be discussed in detail.

8.4.2.1 Pure E2 System

With regard to the self-assembly of the well-known 60-mer core of PDC, the E2 enzyme is the most crucial component. While human PDC is known to also include E3BP in its 60-mer core for specific binding of E3 [265], many organisms (e.g. bacteria) lack this specific enzyme while still forming the 60-mer core [386] – subsequently providing the structural features for enabling catalytic function. Hence, the pure E2 system provides a reasonably simplified model system for investigating the self-assembly process. The pentagonal dodecahedral structure of the 60-mer core is well known from small-angle X-ray scattering (SAXS), small-angle neutron scattering (SANS), and cryoelectron microscopy (cryo-EM) [265, 381, 387], similarly supported by size measurements from analytical ultracentrifugation (AUC) [260] and dynamic light scattering (DLS) [269], as well as supported by molecular dynamics (MD) studies [271]. Consequently, this available knowledge can be used for validation of the proposed model, while the model additionally provides molecular details on the self-assembly process difficult to establish experimentally – thus improving understanding through enhanced modeling capabilities.

The self-assembly of E2 is visualized qualitatively and quantitatively in Fig. 8.6. As it can be seen in Fig. 8.6 A and B, the randomly distributed E2 enzymes assemble into regular spherical structures with the catalytical domains of E2 at the center of the core. While individual monomers and other small structures are still present in the solution, the majority of units is structured into cores matching the expected pentagonal dodecahedral [265] visually well. This is further supported quantitatively by Fig. 8.6 H indicating $\xi_{\text{struc}} = 2.5$ structured contacts per enzyme close to the theoretical value of $\xi_{\text{struc}} = 3$ for the pentagonal dodecahedral core [388].

Concerning the self-assembly pathway, Fig. 8.6 C and D show the chord diagrams of transitions between size classes (1, 2, 3, 4, 5 - 14 [10], 15 - 24 [20], ...) indicating an initial formation of dimers and then trimers out of the originally present monomers. The formation of trimers has been identified experimentally [389] to be a key intermediate in the formation of 60-mers and hence matches current results well. Subsequently, the formation of larger oligomers takes place in a sequentially hierarchical buildup through addition of trimers and other small units from solution, i.e. growing from one size class to the neighboring larger one (see particularly Fig. 8.6 D). This self-assembly through addition continues until reaching the size of 50-mers and 60-mers, which exhibit a continuous back-and-forth transition between the two classes (50-/60-mer) on their pathway to equilibrium (see particularly Fig. 8.6 C). Beyond the 60-mer, only few larger structures form and if so only temporarily - highlighting the 60-mer structure as expected [265].

Regarding the self-assembly kinetics, Fig. 8.6 E shows the process matches the functional fitting of an asymptotic exponential behavior well ($R^2 = 0.972$) similarly to the case of VLPs (Sec. 7) for which it is also in line with literature [371–375]. A rapid structural growth takes place with a critical time constant $\tau_{SAS} = 0.8$ ms[1] towards an asymptotic value of $N_{SAS} = 55.4$ closely matching the 60-mer. Additionally, it can be noted that after the end of the annealing procedure (4 ms), small structures detach from some of the 60-mers, which subsequently re-assemble to larger structures and also cause a transition of 60-mers to 50-mers. It can hence be estimated that the 60-mers still contain individual defects and longer timescales are necessary for perfect assembly. Similar conclusions can be made from the binding rates shown in Fig. 8.6 G.

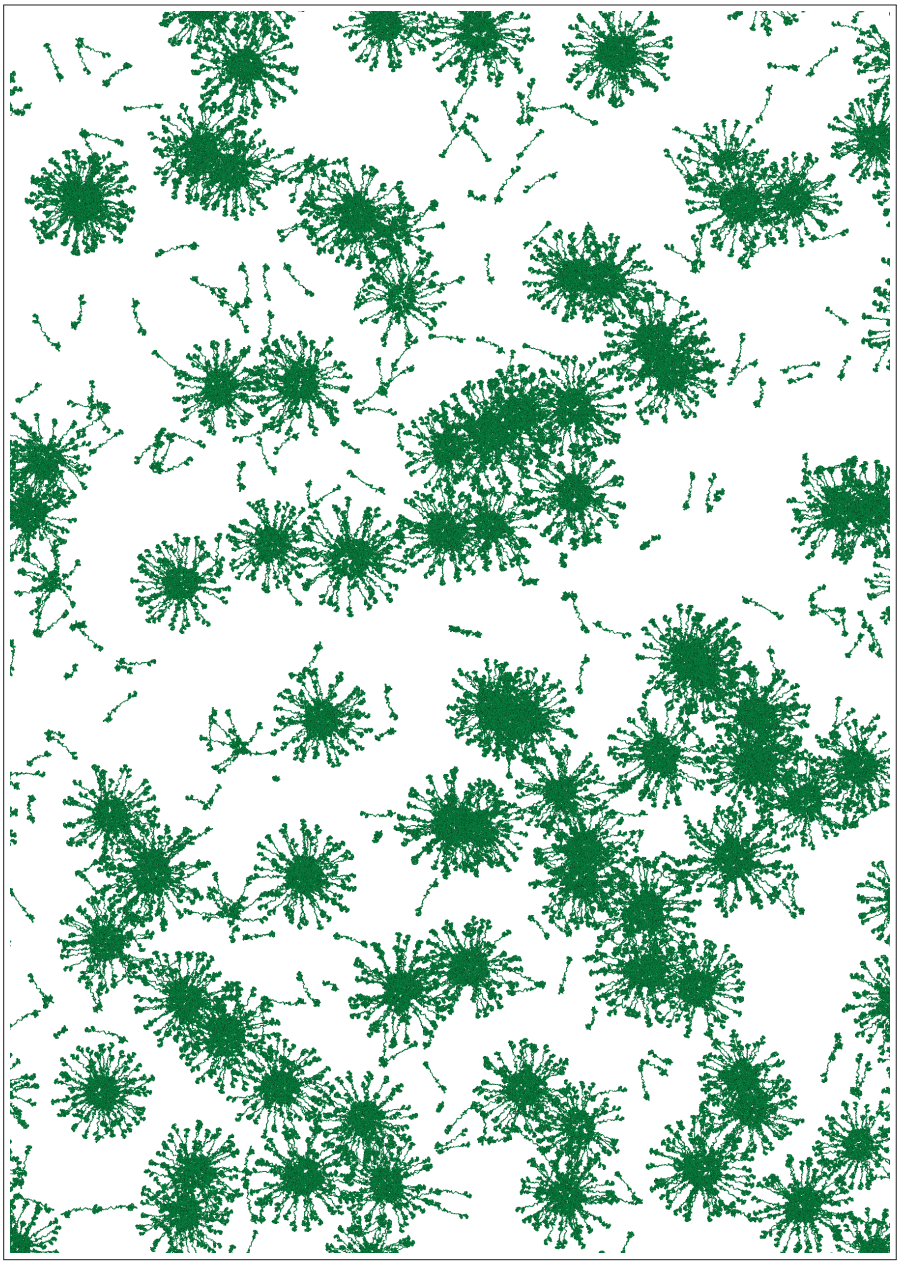
Concerning the reactant distance distribution, Fig. 8.6 I and J show that the self-assembly leads to a much closer distance distribution between E2 units resulting from the 60-mer formation. While this is primarily of interest for the full PDC system in the following section, it can already be noted regarding the pure E2 system that the distance distribution increases almost linearly between $d_{COM} = 4$ nm and $d_{COM} = 20$ nm with two slight discontinuities at $d_{COM} \approx 6$ nm and $d_{COM} \approx 12$ nm, which might be related to trimer substructures and curvature of the 60-mer, respectively.

Overall, self-assembly of the pure E2 system of PDC leads to a composition of 50-mers and 60-mers at the end of the simulation. In addition to matching the expected 60-mer well by number of enzyme copies and pentagonal dodecahedral structural organization [265, 388], good agreement can also be found with regard to the diameter of structures ($d_{SAS,ave} \approx 50$ nm, see Fig. 8.6 F) in comparison to various experimental measurements of comparable systems (52.2 nm from DLS [269], 44.8 nm to 57.6 nm from AUC [387, 390],

[1] Note that the simulation timescales are only comparable between each other and not real-world timescales as the coarse-grained simulations are inherently accelerated through their abstraction.

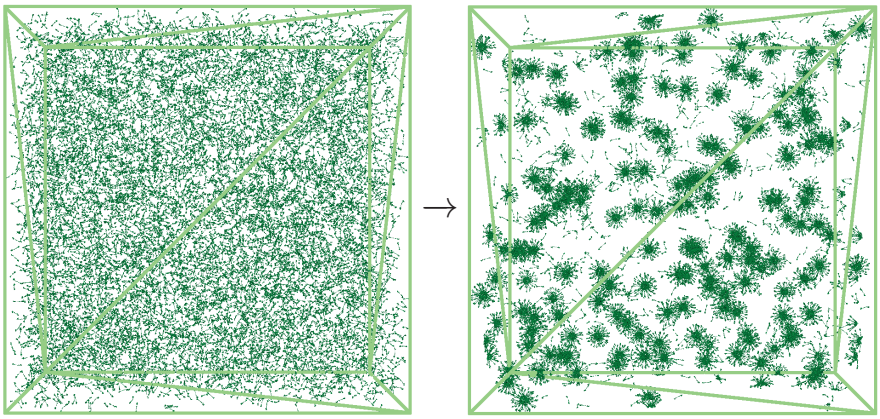
as well as 51.2 nm from MD [271]). Hence, the proposed modeling approach matches experimental data from literature well in addition to providing further insights regarding assembly pathway and dynamics through the high level of detail.

Detached from this self-assembly simulation of the pure E2 system, a stability analysis of the preassembled 60-mer based purely on E2 was performed to assess fluctuations in the core over a time of 1 µs difficult to investigate during assembly. Results showed that the core remained stable throughout the simulation and the average diameter fluctuated over time as $d_{\mathrm{SAS,ave}} = 52.7 \pm 1.2$ nm (49.0 – 56.4 nm). Similar variations of the 60-mer core size have been previously reported experimentally [391] (approximately 1.0 nm standard deviation of flucations with 4 nm range between minimum and maximum diameter observed on 60-mer core without linker arms) and in MD simulations [268, 271]. These fluctuations were attributed by Zhou *et al.* [391] to a 'breathing' process of the core via contractions and retractions between the composing trimers connected via the C-terminal of E2 and E3BP (ball-and-socket joint). Consequently, the proposed model also matches experimental data of such stability fluctuations well.

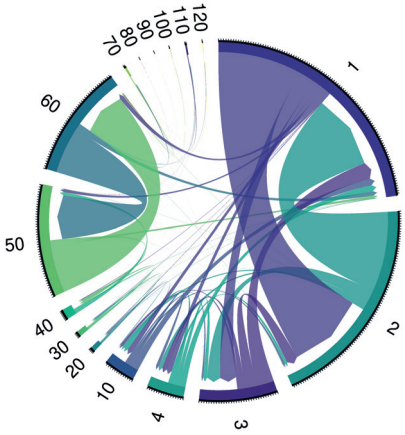


(A) Close-up visualization of self-assembled structures by enzyme type: E1, E3, E2, E3BP.

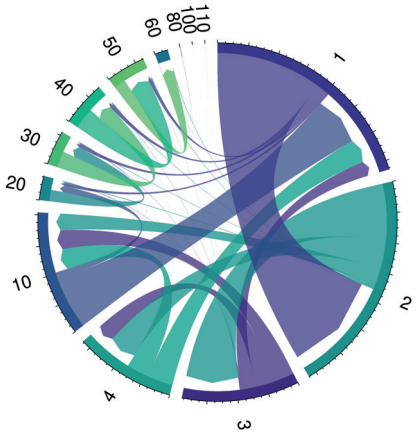
FIGURE 8.6: PDC self-assembly for a pure E2 system at concentration of $1\,\mathrm{mg\,mL^{-1}}$ in $1\,\mu\mathrm{m}^3$ cubic box ($1\,\mu\mathrm{m}$ edge) using simulation protocol SP-PDC-1-AN.



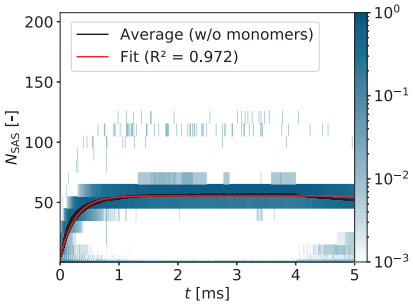
(B) Visualization of before (left) and after (right) self-assembly using back-bone carbon structures. Color indicates enzyme type: E1, E3, E2, E3BP.



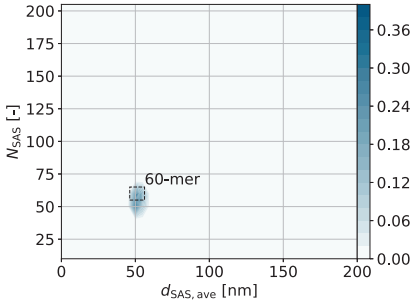
(C) Assembly pathway by bi-directional transitions between size classes.



(D) Assembly pathway by net transitions between size classes.

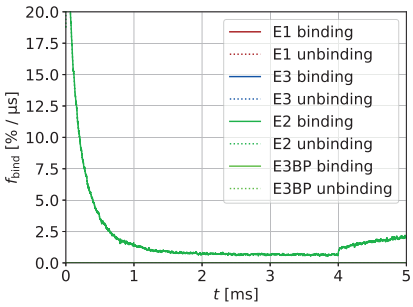


(E) Histogram of self-assembled structures by numbered size over time.

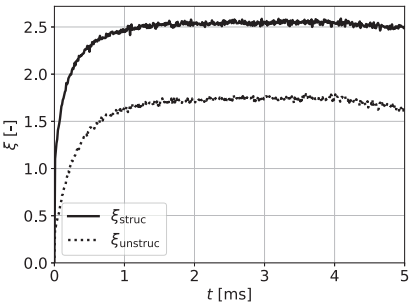


(F) Numbered size versus average extent without monomers (final time).

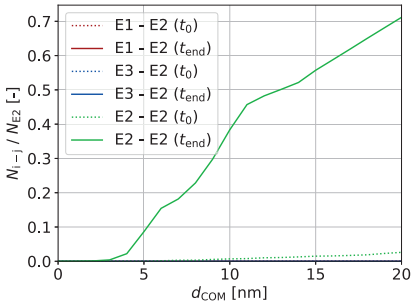
FIGURE 8.6: PDC self-assembly for a pure E2 system at concentration of 1 mg mL^{-1} in 1 µm^3 cubic box (1 µm edge) using simulation protocol SP-PDC-1-AN.



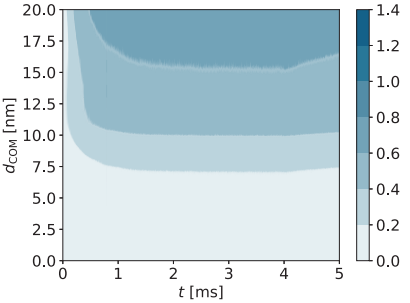
(G) (Un-)binding rates per enzyme type with uniform smoothing over 10 µs.



(H) Average (un-)structured contacts per enzyme (10 µs saving).



(I) Reactant distance distribution at beginning and end of self-assembly.



(J) Reactant distance distribution for E2 - E2 over time.

FIGURE 8.6: PDC self-assembly for a pure E2 system at concentration of 1 mg mL^{-1} in 1 µm^3 cubic box (1 µm edge) using simulation protocol SP-PDC-1-AN.

8.4.2.2 Full Component PDC System

Secondly, self-assembly of the full component PDC system was investigated. As outlined in the introduction, recent studies suggest $40 \times \text{E2} + 20 \times \text{E3BP}$ being the most likely stoichiometry of 60-mer core [265], which was consequently used for the ratio of E2 and E3BP. Regarding E1 and E3 binding, ranges of $20 - 30 \times \text{E1}$ and $6 - 12 \times \text{E3}$ have been widely reported [18, 263, 266, 267, 271] aside from the maximum occupancy of $40 \times \text{E1} + 20 \times \text{E3}$ [259, 265]. Based on this, $30 \times \text{E1}$ and $10 \times \text{E3}$ have been chosen, thus leading to an overall stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ in the solution. For this stoichiometry, simulations were performed at a concentration of $1\,\mathrm{mg\,mL^{-1}}$ (similarly to DLS studies [269]) in a $1\,\mu\mathrm{m}^3$ cubic box ($1\,\mu\mathrm{m}$ edge). Note that as before results with simulated annealing protocol SP-PDC-1-AN (see Sec. 8.2) will be presented directly and results without annealing (SP-PDC-1) are provided in App. E.5.2.

The self-assembly of this full component PDC system is visualized qualitatively and quantitatively in Fig. 8.7 with additional data provided in App. E.5.2 Fig. E.22. As it can be seen qualitatively in Fig. 8.7 A and B, the randomly distributed E1, E2, E3, and E3BP enzymes assemble into regular sphere-like structures with E2 and E3BP (green) at the core (similar to earlier 60-mer), while E1 (red) and E3 (blue) attach to the outside. The core structure of E2 and E3BP matches the 60-mer core visually well with some structures having formed only a partial core. This structural organization matches the experimental expectations of 60-mer core of E2 and E3BP with E1 and E3 attached [18, 263, 266, 267, 271]. At the end of the simulation, a larger fraction of E1 and E3 appear to remain in solution with few E2 and E3BP bound to them, while the majority of E2 and E3BP appears to have formed 60-mers (or structurally similar oligomers). In the following, these qualitative visual observations will be discussed quantitatively.

Concerning the self-assembly pathway, Fig. 8.7 C and D show the chord diagrams of transitions between size classes (1, 2, 3, 4, 5 - 14 [10], 15 - 24 [20], ...) indicating an initial formation of dimers and trimers out of the originally present monomers similarly to the pure E2 system (Sec. 8.4.2.1) and as experimentally observed as key intermediates in systems without E1 and E3 [389]. Subsequently, the formation of larger oligomers takes place in a similar sequential hierarchical buildup through addition as for the pure E2 system. In this regard, it can be noted that in contrast to pure E2, the model predicts that for the full PDC system dimers play a key role as the additive unit in structural growth (see particularly net transitions away from '2' in Fig. 8.7 D, possibly being E2+E1 or E3BP+E3). The self-assembly through addition subsequently continues until reaching sizes between 50-mers and 70-mers, which exhibit the same back-and-forth transition between classes on the pathway to equilibrium as pure E2 – only also including

the larger 70-mer fraction. Hence, the presence of E1 and E3 seems to modify the self-assembly pathway only slightly besides the increased role of dimers, which might be attributed to the binding of dimers of E2+E1 or E3BP+E3 to larger structures. Above 70-mers, i.e. more than 75 enzymes, only few structure form indicating the full 60-mer plus $20 - 30 \times$ E1 and $6 - 12 \times$ E3 is not finalized, which will be further discussed subsequently.

In addition to the self-assembly pathway, binding and unbinding of monomers to assembled structures like the 50-mer and 60-mer is especially notable in the chord diagram of total transitions (Fig. 8.7 C). These bindings and unbindings are particularly attributed to E1 and approximately half as many E3 (absolute numbers, note relative flux in Fig. 8.7 G combined with E1/E3 ratio of 3). Thus, indicating a high (un-)binding flux of E1 and E3 supporting the reaction pathway enabled by PDC [261]. The higher absolute flux of E1 is also in line with E1 known to catalyze the rate-limiting step in PDC activity [18, 264]. Regarding binding kinetics, it can be noted that experimental knowledge on macroscopic protein binding is available through methods such as isothermal titration calorimetry (ITC) [259, 260, 380, 381], surface plasmon resonance (SPR) [379, 382], and bio-layer interferometry (BLI) [392]. However, the determined equilibrium binding constants are difficult to compare (between each other and with simulations) as there is a measurement dependency in addition to various truncations of enzymes [392]. Experimental works have found depending on measurement method and truncation of enzymes dissociation constants K_D for E1 to E2 between 9.5 – 218 nM [260, 382] and for E3 to E3BP between 0.8 – 102 nM [259, 260, 379–381] with upper bounds being the most similar system, where E1 and E3 are separately bound to a 60-mer core of E2+E3BP [260]. Simulation results in this work find intermediate values of 23.2 nM for E1 to E2 and 81.1 nM for E3 to E3BP (average over last 0.1 ms). The higher value for E3 to E3BP indicates a decreased binding affinity of E3 relative to E1, which might be attributed by the simultaneous presence of E2 in the simulation setup. Overall, the dissociation constants are within the range of experimental values and thus in reasonable agreement considering the measurement dependency of related methods and setup differences [392].

Regarding the self-assembly kinetics, Fig. 8.7 E shows the same asymptotic exponential behavior ($R^2 = 0.995$) as was found for the pure E2 assembly, but with considerably longer timescales of $\tau_{SAS} = 2.1$ ms (pure E2 to 60-mer was $\tau_{SAS} = 0.8$ ms) and a slightly lower asymptotic structure size $N_{SAS} = 46.0$ (pure E2 to 60-mer was $N_{SAS} = 55.4$). A detailed investigation shows that the majority of enzymes (dark blue region in Fig. 8.7 E) self-assembles into structure of 50-mers to 70-mers within 1 – 2 ms with considerably larger size variations than for the pure E2 system (Fig. 8.6) in addition to more small structures. While this indicates partially incomplete structural assembly, which could be addresses through longer simulation times and possibly higher annealing

temperatures, these larger variations are also expected with E1 and E3 dynamically binding and unbinding to and from the 60-mer core of E2 and E3BP – hence exchanging reactants with the solvent. Regarding the diameter of structures (Fig. 8.7 F), it can be observed that it is slightly larger than that of the pure E2 60-mer and in line with experimental sedimentation studies finding Stoke's diameters of up to 59.5 nm [390].

Concerning the stoichiometry, Fig. 8.7 I and J provide compositions of structures over time and size. It can be observed that E1 and E3 bind correctly to the core provided by E2 and E3BP with initially particularly E2+E1 forming complexes, which supports the earlier argument of E2+E1 providing the dimer units in the assembly pathway. It can further be observed that E3 binds generally to both E3BP and E2, but relative to the stoichiometry (E2/E3BP = 2/1) primarily to E3BP and only half as much to E2 (App. E.5.2 Fig. E.22 C), which is even more significant without annealing (App. E.5.2 Fig. E.21 C). Consequently, although E3 is permitted to bind to E2 via a weaker interaction derived through statistical binding analysis in MD (see Sec. 8.1.2.2), the binding specificity of E3 to E3BP expected for human PDC in literature is largely reproduced [18, 263, 379, 382]. With regard to the overall stoichiometry in Fig. 8.7 I, it can be seen that a complex is on average composed of $40 \times \text{E2} + 23 \times \text{E3BP} + 10 \times \text{E1} + 5 \times \text{E3}$ (E2 fixed according to original stoichiometry). Consequently, the E3BP fraction is slightly overestimated, while the E1 (and partially E3) fraction are slightly underestimated compared to expectations of $20 \times \text{E3BP}$, $20-30 \times \text{E1}$, and $6-12 \times \text{E3}$ [18, 263, 266, 267, 271]. With increasing structure size (see Fig. 8.7 J), less E1 and E3 are found to bind[2], which is likely attributed to the large flexibility of E2 and E3BP linker arms not being fully captured by the proposed modeling approach and hence limiting the number of E1 and E3BP bound by atom collisions with the (at least primarily) rigid model of the linker arm. While the model still permits capturing such flexible systems to a large extent, this aspect might present an opportunity for further developments, e.g. by decreasing repulsion of the linker arm in the molecular repulsion model or sampling different conformations. Nonetheless, even with fewer E1 and E3 binding the essential aspects of PDC self-assembly are still captured.

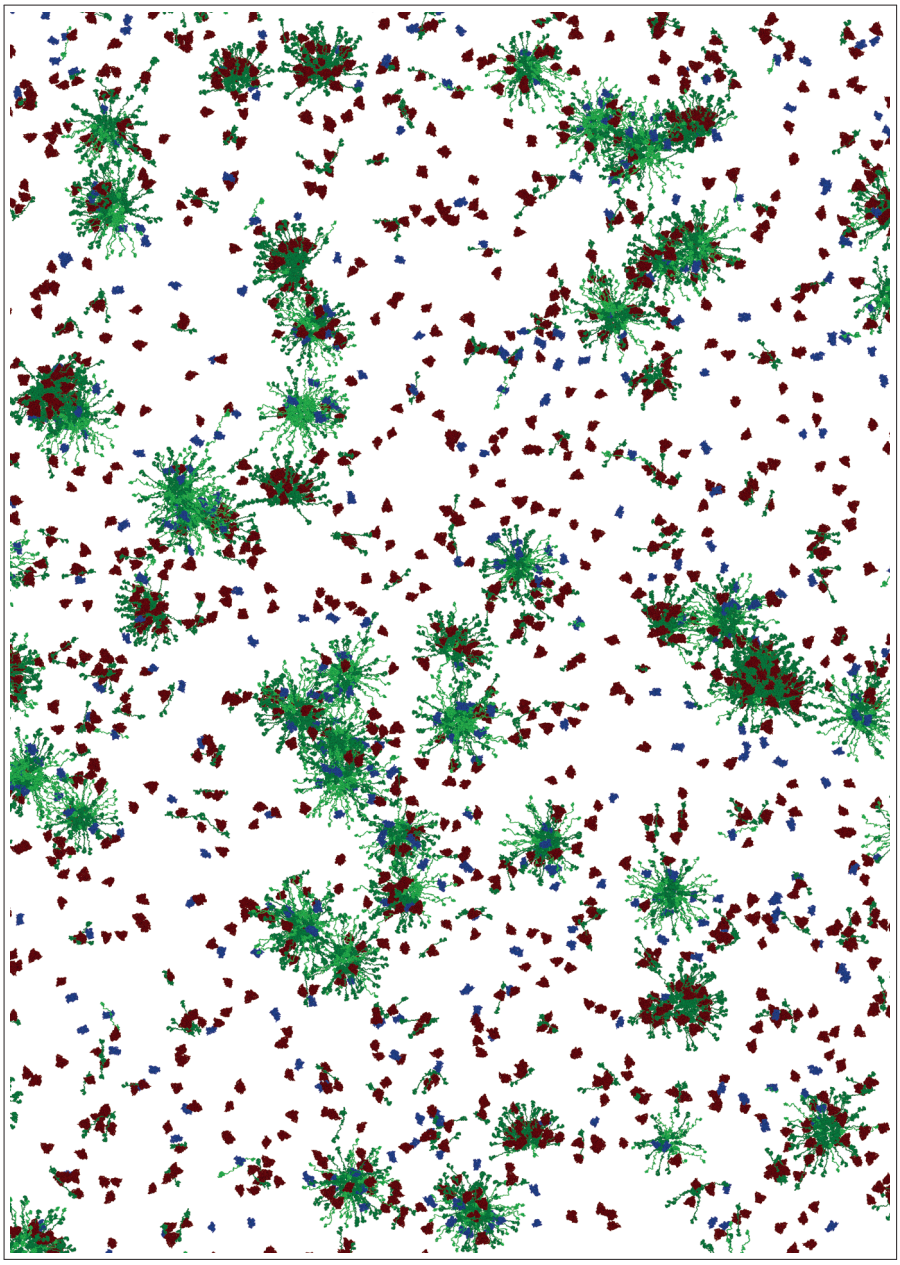
In this regard, the reactant distance distribution shown in Fig. 8.7 K is particularly important for enabling PDC's catalytic activity through active-site coupling and metabolic channeling [261]. In brief, E1 binds the reactant pyruvate, which is subsequently transferred to E2 where the product acetyl coenzyme A (acetyl-CoA) is released, followed by a regeneration of the active site of E2 at E3 [18]. Consequently, for catalytic activity the active sites of E1 and E2, as well as E2 and E3 have to be in proximity. As it can be seen in Fig. 8.7 K, the structural self-assembly leads to a peak in E1 – E2 (red) and E3 –

[2]Excess E1 and E3 are mainly unbound in solution as shown in Fig. 8.7 J ($N_{\text{SAS}} = 0$) and visually in Fig. 8.7 A and B, which further leads to a decreased number of structured contacts in Fig. 8.7 H.

E2 (blue) around $d_{\mathrm{COM}} \approx 5$ nm. Consequently, the formed structures enable active-site coupling and hence ensure catalytic activity of the complex. Note that high activity is further supported by a continuous binding and unbinding of E1 and E3 (Fig. 8.7 G), which further enhances exchange of reactants with the solvent environment.

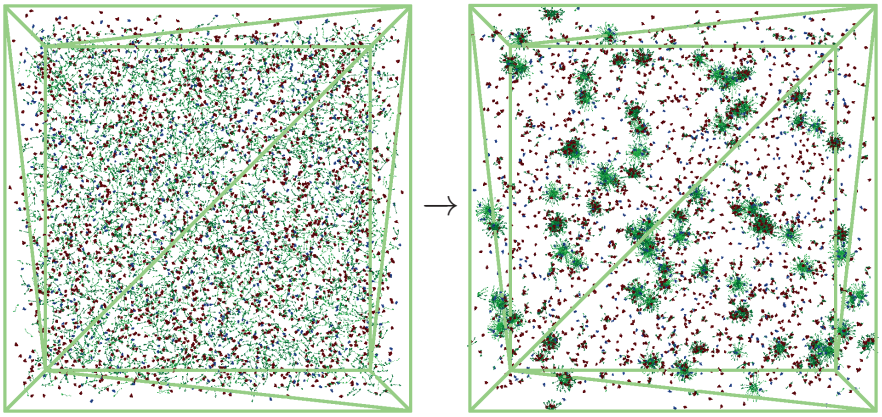
In a similar context, Prajapati *et al.* [265] have found local clusters of E1 and E3 around the 60-mer core. These findings are supported by the proposed model through the distance distribution of E1 – E1, E3 – E3, E1 – E3 (number of contacts at peak located around $d_{\mathrm{COM}} \approx 5$ nm, normalized by number of enzymes N_{E1}, N_{E3}, and $\sqrt{N_{\mathrm{E1}} N_{\mathrm{E3}}}$, respectively). This shows clustering of E1 – E1 with 8.5 %, while E1 – E3 and E3 – E3 only have approximately 2.3 % and 1.7 % at peak. Hence, particularly E1 – E1 clusters are present (also beyond pure stoichiometry considerations) similarly to the observations of Prajapati *et al.* [265].

In summary, the complex self-assembly of the multi-enzymatic human PDC system into pentagonal dodecahedral cores of E2 and E3BP with attached E1 and E3 enzymes necessary for catalytic activity is well reproduced by the proposed modeling approach. While slightly fewer amounts of E1 and E3 were found to bind ($10 \times$ E1 and $5 \times$ E3 compared to $20-30 \times$ E1 and $6-12 \times$ E3 expected, likely associated with the limited flexibility of E2 and E3BP linker arms in the model), the fundamental properties of complex size, composition, and buildup were reproduced correctly in comparison to available literature data. Hence, for the first time it has been possible to model the full self-assembly of such complex multi-enzymatic machines underpinning the structural features enabling catalytic activity and thus going towards predictive modeling of enzymatic reaction cascades.

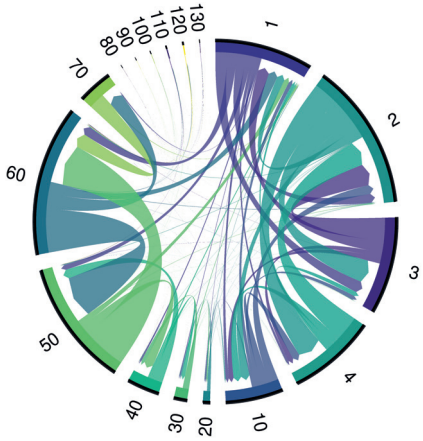


(A) Close-up visualization of self-assembled structures by enzyme type: E1, E3, E2, E3BP.

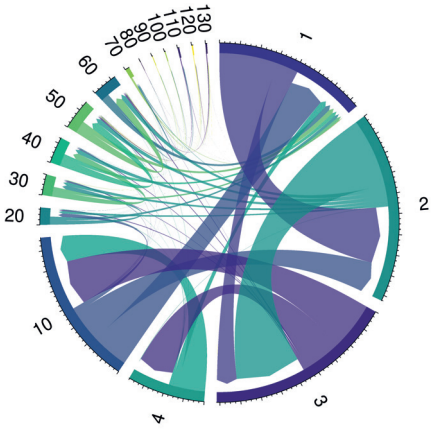
FIGURE 8.7: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1-AN.



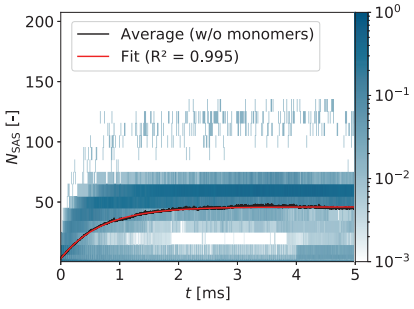
(B) Visualization of before (left) and after (right) self-assembly using back-bone carbon structures. Color indicates enzyme type: E1, E3, E2, E3BP.



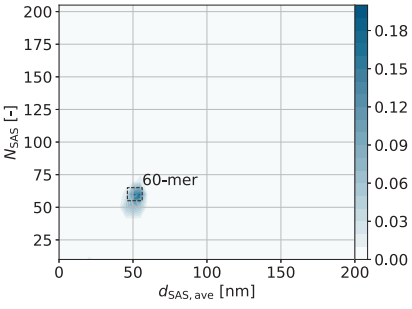
(C) Assembly pathway by bi-directional transitions between size classes.



(D) Assembly pathway by net transitions between size classes.

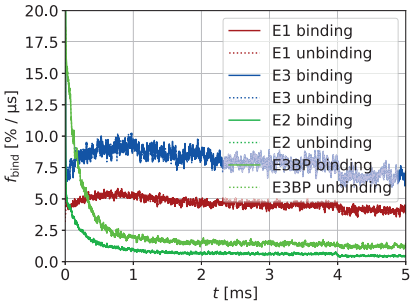


(E) Histogram of self-assembled structures by numbered size over time.

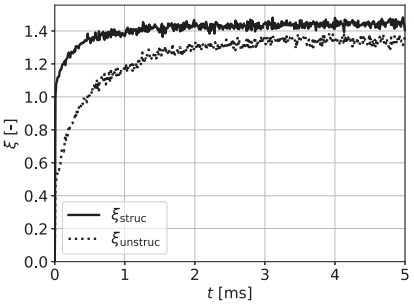


(F) Numbered size versus average extent without monomers (final time).

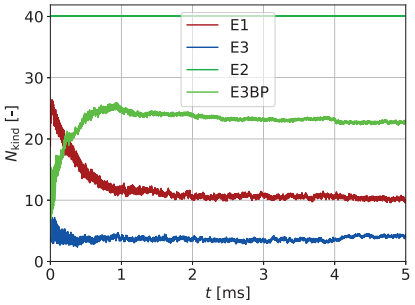
FIGURE 8.7: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box ($1\,\mu\text{m}$ edge) using simulation protocol SP-PDC-1-AN.



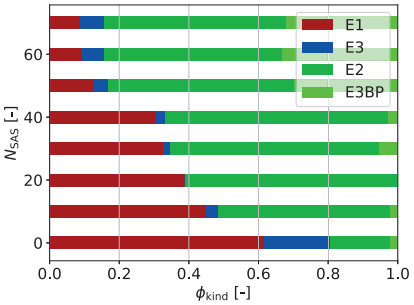
(G) (Un-)binding rates per enzyme type with uniform smoothing over 10 µs.



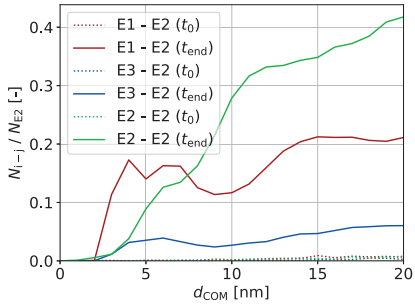
(H) Average (un-)structured contacts per enzyme (10 µs saving).



(I) Stoichiometry of struc. ($N_{SAS} \geq 5$) relative to E2 content in 60-mer.



(J) Stoichiometry of structures by molar fraction for each N_{SAS} (final time).



(K) Reactant distance distribution at beginning and end of self-assembly.

FIGURE 8.7: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1-AN.

9

Conclusions

In conclusion, a framework for hierarchical multiscale modeling of structural formation in macromolecular systems using a physics-based and data-driven parameterization from atomic scales has been presented. The proposed framework resembles an ultra coarse-grained molecular dynamics approach based on Langevin dynamics and abstracts each macromolecule as an object with anisotropic properties concerning interaction with the environment, i.e. solvent, as well as other molecules. For this, the framework builds upon three main model components: (I) a diffusion and thermodynamic model describes the interaction with the implicit solvent and enforces the correct thermodynamics (Ch. 3); (II) intermolecular interaction models describe the effective anisotropic interaction between molecules (Ch. 4); (III) optional bonded interaction models allow for specific/decoupled treatment of stable molecular structures (e.g. through chemical bonds, see Ch. 5). Using the proposed framework, large length and time scales in the order of micro-meters and milli-seconds can be reached well beyond traditional simulation methods such as (coarse-grained) molecular dynamics (CG-MD).

Focus of this work was placed on method development with regards to a generic model formulation through data-driven approaches (e.g. interaction potential fields), but also specialized functional models. In addition, methods for parameterization from molecular dynamics were developed to enable a bottom-up modeling approach, while performing top-down validation and enabling optional incorporation of empirical information. The framework with its various model components has been applied to three model systems from material science and biology to study their structural self-assembly and test the framework. First, the polymer network formation during gelation was investigated for alginate in $CaCl_2$ solution. Second, self-assembly of the hepatitis B core antigen (HBcAg) into virus-like particles (VLPs) was studied, which is essential for its biological function of virus transport and infection. Third, self-assembly and agglomeration of the pyruvate

dehydrogenase complex (PDC) was investigated, which features hierarchical assembly critical to enable catalytic activity through phenomena such as metabolic channeling. Thus, the developed framework was tested on a variety of applications and systems, which are highly interesting from a scientific and commercial point of view.

With regard to the modeling of alginate gelation, a specialized version of the proposed framework has been applied to study the mechanisms of structural network formation using calcium cations in solution. In order to describe the calcium mediated interaction of alginate polymer chains, a probabilistic ion model has been developed and integrated into the framework based on literature parameters and theoretical considerations. Results were able to capture the mechanisms and dependencies with regard to polymer concentration, polymer composition, and ion concentration on gelation. Validation was performed in comparison to literature data and collaborator experimental data, which yielded good agreement with respect to calcium uptake, gel pore sizes, and bundle formation of polymer chains. Thus, the model aids understanding of the underlying gelation mechanisms and supports predictions in the design of gels. Details can be found in Ch. 6 and Depta *et al.* [224].

With regard to modeling the self-assembly of HBcAg dimers ($HBcAg_2$) into VLPs, the generic model framework building upon MD parameterization was employed. Mechanics of hierarchical capsid formation and intermediates were studied at protein concentrations of 5 µM, 10 µM, 50 µM, and 100 µM with ion concentrations of 150 mM sodium chloride. Results provided detailed insight into stage-wise assembly and highlighted importance of small agglomerates (10 - 35 $HBcAg_2$) in the assembly process. Furthermore, concentration dependent effects could be observed including diffusion limitations at low concentrations and overgrowing, i.e. increased kinetic traps, at high concentrations. Validation was performed in comparison to literature data showing good agreement with respect to structural properties and limited experimental insights on formation mechanics. Thus, the model aided understanding of the self-assembly processes and enabled predictions for future works, e.g. at different process conditions. At the same time, limitations of the model were investigated and addressed including allostery-induced conformational changes during pairwise $HBcAg_2$ binding through biased MD parameterization or incorporation of empirical data. Details can be found in Ch. 7 and Depta *et al.* [225, 226].

With regard to modeling PDC self-assembly and agglomeration, the same generic model framework was employed as used for the HBcAg system. Mechanics of hierarchical structural formation and intermediates were studied at different component stoichiometries for the more complex system comprised of up to four enzyme types E1, E2, E3, and E3BP. Besides the investigation of assembly kinetics similar to HBcAg, analysis

highlighted the structural composition and detailed organization of formed agglomerates. Furthermore, binding and unbinding of E1 and E3 to agglomerate structures was investigated in detail, which is essential in enabling the catalytic activity of PDC. Consequently, the model aided detailed understanding of PDC structure in combination with functional requirements, which were discussed in combination with available literature data showing good agreement. At the same time, inherent challenges and possible future extension of the model framework resulting from the complex structural features of PDC components, specifically the flexible linker arms of E2 and E3BP, were discussed. Details can be found in Ch. 8 and Depta *et al.* [223, 227, 228].

In summary, a framework for physics-based and data-driven multiscale modeling of macromolecular structural formation has been proposed and tested on three highly interesting model systems to show wide applicability. The developed framework enables novel scales to be investigated using numerical simulations and includes a supervised-learning bottom-up parameterization, thus paving the way towards physically-mechanistic modeling of such structural assembly processes. As a result, the framework can be readily applied to better understand and test modifications of a variety of natural and synthetic molecular systems including for example other viruses (e.g. adenovirus, coronaviruses), enzymatic complexes (e.g. glutamine synthase), self-assembled monolayers (e.g. thiolates on metals), or colloids. Future studies might also improve and extend the framework in various ways, including for example: improved interaction potential parameterization through e.g. novel MD force-fields, higher-dimensional interaction potentials including conformational changes, further abstractions to larger multimers (e.g. trimer), incorporation of fluid flow, coupling to population balance modeling, structure-dependent reaction kinetics through e.g. finite volume modeling approaches, and more.

Appendix A

General Appendix

A.1 Euler Angle Definition

The order of Euler angles in this work used is first α, then β, and γ. The value range is from $-\pi$ to π for α and γ, as well as from $-\pi/2$ to $\pi/2$ for β. Conversion from a unit quaternion with components q_0, q_1, q_2, q_3 can be calculated as

$$\alpha = \text{atan2}(2(q_2q_3 + q_0q_1), q_0q_0 - q_1q_1 - q_2q_2 + q_3q_3), \tag{A.1}$$

$$\beta = \text{asin}(2(q_0q_2 - q_1q_3)), \tag{A.2}$$

$$\gamma = \text{atan2}(2(q_1q_2 + q_0q_3), q_0q_0 + q_1q_1 - q_2q_2 - q_3q_3). \tag{A.3}$$

A.2 Detailed Framework Overview

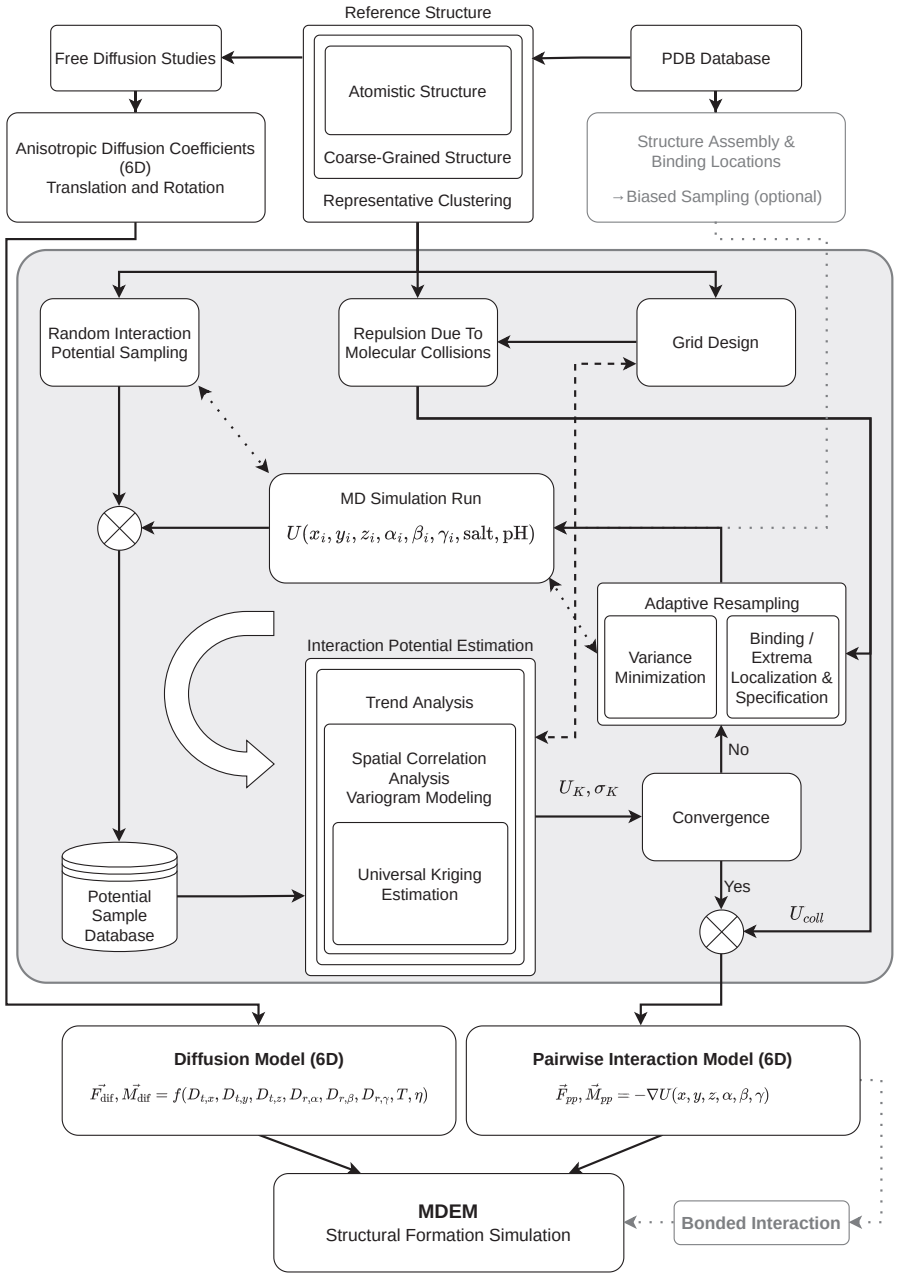


FIGURE A.1: Detailed visualization of the physics-based and data-driven framework for macromolecular structural formation.

A.3 Hydrodynamic Interaction

Introduction and Literature For the simulation of non-dilute solutions *hydrodynamic interaction*, i.e. forces resulting from the relative velocities of coarse-grained beads coupled through the solvent, can become important in addition to systematic forces between beads, i.e. forces resulting from relative position and atomic structure. This is for example more widely employed for polymer systems [393] using Brownian Dynamics (see Sec. 1.2.2) with the Ermak-McCammon algorithm [394, 395]. The friction coefficient (see eq. 1.7) becomes a tensor coupling the DOF of all coarse-grained beads. Common forms are the Oseen tensor [396] and the Rotne-Prager-Yamakawa tensor [393, 397] assuming stationary flow fields[1]. While these are derived for the translation of uniformly sized spheres in solution, extensions for rotation-translation coupling [394], different sphere radii [399, 400], and ellipsoids [401] have been researched as well. However, while mathematical incorporation of anisotropic particles leading to a $6N \times 6N$ friction tensor is straightforward [172], derivation of the individual friction tensors between arbitrarily anisotropic objects is not and requires detailed investigation [277]. Subsequent solution of the hydrodynamic interaction requires complexity $\mathrm{O}(N^3)$ for direct solution through factorization of the tensor, but can be approximated by the truncated expansion ansatz (TEA) of Geyer and Winter [402] with complexity $\mathrm{O}(N^2)$. For reviews see e.g. refs. [277, 403]. Alternatively on larger scales of macroscopic objects outside the focus of this work, e.g. the discrete element method (DEM), hydrodynamic interaction and generally fluid flow is often described by coupling to continuum methods such as computational fluid dynamics (CFD) to solve the Navier-Stokes equation numerically.

Relation to this Work With regard to (anisotropic) macromolecular assemblies in the context of Langevin Dynamics (LD) and implicit solvation, effects are nontrivial. While in dilute systems the macromolecules are fully surrounded by solvent molecules (typically water), these are replaced by the respective binding partner during assembly. As a result, the effective friction and random forces in LD are reduced[2] in the respective direction (similarly to an effective reduction in viscosity) and replaced by intermolecular interaction. During binding, i.e. direct structural contact between two macromolecules, it is reasonable to assume that this is primarily dominated by their relative position and orientation and not their relative velocities. For larger distances as well as with respect to mutual movement through the solvent, however, hydrodynamic interaction is likely relevant, as it essentially shields macromolecules from the solvent and couples the

[1] Thus they are strictly speaking not applicable to LD, which requires an explicit time-dependent modeling, see e.g. refs. [277, 398].

[2] Note that these also incorporate internal DOF of the coarse-grained bead and cannot directly be separated (see Sec. 1.2.2).

respective DOF of the macromolecules through the friction tensor in LD. As discussed in the previous paragraph, methods have been developed in the context of BD primarily with regard to spherical objects. **Due to the complexity of extending this to general anisotropic and rotation dependent objects, hydrodynamic interaction has been neglected throughout this work and is left for future works as refinement - an approximation which is in line with literature [277]. Furthermore, as the primary argument for inclusion of hydrodynamic interaction is that shielding through neighboring elements limits the effective friction/resistance of the molecule with respect to the fluid flow [396], decreasing the effective viscosity leads to a partial account of the non-dilute state of the solution. Consequently, this strategy of reducing the effective viscosity is employed with respect to assembly studies in the results.** Note that such a reduction of effective viscosity is widely employed in literature as it additionally leads to accelerated dynamics, while expected to largely maintaining equilibrium [30]. Reduction is typically performed using a factor between 10 - 1000× [110, 127, 128].

In order to quantify the reduction in effective viscosity beyond literature values, an example of hydrodynamic interaction and influence on drag is provided in the following. This example is employing the Rotne-Prager-Yamakawa (RPY) tensor [393, 397], i.e. following Stoke's law at low Reynolds numbers. Let five spherical particles be positioned with one at the center and each of the remaining four at 2.2 times the radius in $\pm x$ or $\pm y$ (formation of a plus-like structure '+')[3]. Let all of them have a uniform velocity in x direction and no fluid velocity. Their respective drag in comparison to a diluted particle is 5.8% for the center particle, 41.8% for the particles in $\pm x$, and 57.9% for the particles in $\pm y$. Alternatively, for a flow in z direction it is 12.9% for the center particle and 56.2% for the outer particles. Both cases highlight the shielding effect of hydrodynamic interaction and support a reduction of the effective viscosity by a factor between 8 - 20×. **Consequently, an effective viscosity reduction by a factor of 10× was chosen, which is at the lower end of literature scaling (10 - 1000×) [110, 127, 128].** Note additionally that for opposite movement between particles strong restoring forces due to hydrodynamic interaction and consequently coupled DOF occur: In case of a zero velocity central particle e.g. 208.9% its normal drag in x direction and 142.4% its normal drag in z direction. As discussed before, these aspects are not further modeled, but would additionally stabilize the assembled structures.

[3]Note that the chosen simplified positioning resembles part of the structure on the HBcAg VLP capsid. While being simplified, it provides insight into the qualitative effects. The actual HBcAg VLP structure contains direct contacts between anisotropic dimers (thus making direct application of the RPY tensor less accurate), as well as repeating structural elements over the VLP - thus leading to further shielding of the solvent and drag reduction.

Appendix B

Diffusion Model Comparison with Molecular Dynamics Data

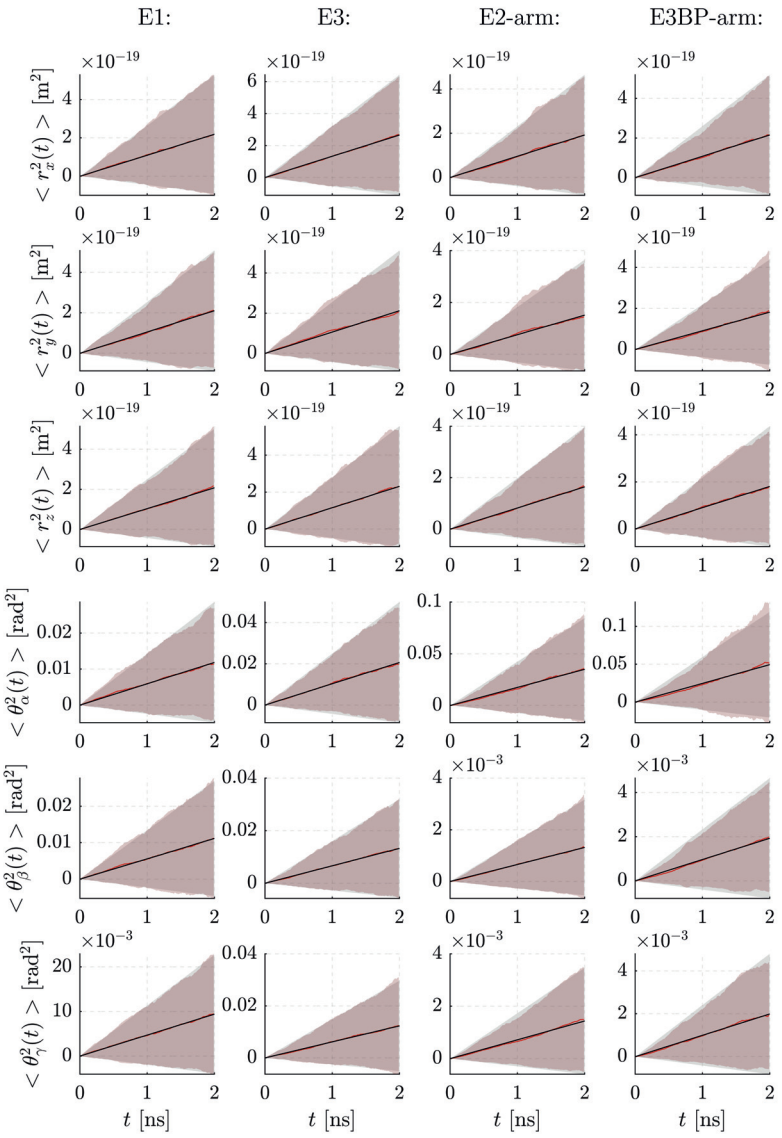


FIGURE B.1: Comparison of MSD plots from molecular dynamics data (red) and diffusion model results (black). Solid lines indicate mean values (MSD) and shaded regions the standard deviation. The MD statistic consists of 600 independent replicates, while the DEM statistic consists of 500'000 replicates to provide a more accurate trend for comparability. For a close-up of the x and α DOF of E2-arm as an example please see Fig. 3.3. Reprinted with permission from supplementary of Depta *et al.* [223]. Copyright 2019 American Chemical Society.

Appendix C

Kriging Algorithm Components

This chapter is based on the following publications:

P. N. Depta, M. Dosta, W. Wenzel, M. Kozlowska, and S. Heinrich. Hierarchical Coarse-Grained Strategy for Macromolecular Self-Assembly: Application to Hepatitis B Virus-Like Particles. *Int. J. Mol. Sci.*, 23(23):14699, 2022

P. N. Depta, M. Dosta, and S. Heinrich. Data-Driven Multiscale Modeling of Self-Assembly and Hierarchical Structural Formation in Biological Macro-Molecular Systems. In W. E. Nagel, D. H. Kröner, and M. M. Resch (editors), *High Performance Computing in Science and Engineering '21*. Springer International Publishing, Cham, 2023

C.1 Variogram Binning Algorithm

During analysis of spatial correlation and subsequent variogram fitting, a large number of samples exist due to the scaling of the problem with $O(N^2)$ (resulting from the correlation between all points), where N is the number of samples. The number of correlations is typically in the order of 10^{10} for a fully sampled field and becomes infeasible for direct fitting. In order to overcome this, an adaptive binning algorithm was implemented. The fundamental idea is to bin all correlation samples according to their distance, calculate averages and standard-deviation in each bin, and perform weighted-least-squares fitting based on these bins as described in Sec. 4.3.3.

For this, the continuous variogram definition in equation 4.25 is discretized as

$$\gamma_Y^*(\delta_r) = \frac{1}{2|N(\delta_r)|} \sum_{N(\delta_r)} \left(Y(\vec{x}_i, \mathbf{q}_i) - Y(\vec{x}_j, \mathbf{q}_j)\right)^2 , \tag{C.1}$$

where $N(\delta_r)$ denotes the set of pairwise distances in distance class δ_r between data points i and j, which is also known as the method-of-moment estimator or classical estimator of the experimental variogram [295, 297, 404]. Note that this formulation is sensitive to outliers due to the square of the difference. To address this, more robust estimators against outliers have been proposed in literature [405]. This is typically achieved by transformations to determine the expected value of the squared difference under the assumption of a Gaussian distribution, see e.g. refs. [294, 405]. This assumption is fulfilled by many processes, especially in the context of the central limit theorem. However, as in this case for large δ_m the distribution of the residuum Y tends towards a Laplace distribution and the number of samples at small lag distances is small, applicability of such an estimator is questionable. Initial tests on the HBcAg system using the estimator proposed by Cressie *et al.* [405] showed no significant differences and consequently, the classical estimator defined in eq. C.1 was used.

In order to automatically perform binning and avoid statistical issues at large sample distances (e.g. a decrease of the experimental variogram), the following algorithm was developed: First, all sample correlations are calculated within a limit of 4 nm to reduce the number of overall samples. Second, initial binning is performed between the minimum distance and maximum distance with 100 bins including the calculation of the mean and standard-deviation. At least 200 samples are required for a bin to be considered valid. Third, a smoothed version of the bin averages is calculated (0.25 weight for bin before and after, 0.5 for center bin) and based on this, the sum of all increments calculated. If this sum is positive, an increasing function will be deduced and decreasing otherwise (this is kept generic although a decreasing function is not a valid variogram). Furthermore, the extrema of the bin average will be determined along with its (bin) position. Motivation for this is that the experimental variogram often tends to decrease again after reaching a maximum. Fourth, the minimum (bin) distance at which either the last binning element is reached or the value has dropped to 95 % of the extrema value will be determined. Binning will be re-run (second step) based on this distance as the upper limit until convergence is reached (upper limit change less than 0.1 %) or ten iterations are exceeded. Once the last binning was run, binned average and standard-deviations are used for weighted-least-squares fitting as described in Sec. 4.3.3.

C.2 Kriging Neighborhood Search and Convergence

In order to perform the Kriging estimation, the data set has to be searched for the subset of nearby points, also called neighborhood points (number N_K). As calculating the distance between points from the RMSD δ_r between two structure of B is computationally

demanding, a more efficient search algorithm was implemented to avoid a brute force search through the data set to determine the closest data points. This algorithm will be explained in the following and is based on incrementally increasing the cartesian search volume:

1. Pre-calculate the cartesian distance from the estimation point to all data points in the current section.
2. Search the data set for points within an interaction volume of radius $r_{cur} = 1$ nm in cartesian space around the estimation point.
3. Calculate δ_r from the estimation point to each found point and save the result. Calculate the number of thus far found points with $\delta_r \leq r_{cur}$, termed N_{cur}.
4. Increase r_{cur} by a factor of 1.26 (doubling the cartesian volume) and return to step 2 until $N_{cur} \geq N_K$.
5. Sort all points with $\delta_r \leq r_{cur}$ by δ_r and use the first N_K for estimation.

As calculating the cartesian distance is much faster than calculating the δ_r distance, the used algorithm is significantly more performant.

Note that the estimate is dependent on the number of Kriging points used (N_K). In order to estimate the impact, a **convergence study** was performed varying N_K between 100 and 1000 and comparing the potential field and variance of the field to the reference case of $N_K = 1000$. For this, the random $HBcAg_2$ – $HBcAg_2$ data set was used together with the grid for iterative refinement. Results can be found in table C.1. As it can be seen, the differences to the reference case of $N_K = 1000$ decrease with increasing N_K indicating a convergent behavior. The difference in variance indicates an over-estimation of the variance for lower N_K, which is consistent with the expectation of a reduced estimation accuracy for lower numbers of used samples. Nonetheless, results indicate that the overall variance is captured even for $N_K = 100$. Consequently, in order to reduce computational demand during variance minimization resampling, $N_K = 100$ was chosen. The difference in potential estimate are more significant for lower N_K, but overall balanced around a zero mean. Consequently, $N_K = 500$ was chosen for all remaining Kriging to ensure a balance between computational demand and accuracy.

TABLE C.1: Convergence study of the number of Kriging points N_K in reference to $N_K = 1000$ on the example of the $HBcAg_2$ – $HBcAg_2$ random data set and grid for iterative refinement. Relative comparisons are for portions of the grid within cutoff and outside of collisions. Reprinted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

Parameter	100	250	500	1000
Mean potential dif. [kJ/mol]	-0.064	0.042	0.065	0
Min potential dif. [kJ/mol]	-80.5	-33.0	-19.0	0
Max potential dif. [kJ/mol]	62.3	46.3	21.4	0
Mean variance dif. [kJ^2/mol^2]	5.9	2.5	0.8	0
Min variance dif. [kJ^2/mol^2]	-0.32	-0.19	-0.002	0
Max variance dif. [kJ^2/mol^2]	309.5	103.0	31.6	0

C.3 Objective Function For Quantitative Structural Stability

In order to quantitatively evaluate the stability of the HBcAg VLP and PDC 60-mer core, an objective function was derived. The function is based on the combination of various properties of the structures in comparison to the starting structures, i.e. ground truth (gt). The following properties of the structures were used:

- r_{gyr}: radius of gyration
- $d_{\mathrm{com,min}}$: minimum distance from COM of entire structure to the COM of any molecule / particle
- $d_{\mathrm{com,max}}$: maximum distance from COM of entire structure to the COM of any molecule / particle
- $d_{\mathrm{i,minDistAnyJ,rmsd}}$: root-mean-square of the minimum distance from all i to any $j \neq i$
- $d_{\mathrm{ij,min}}$: minimum distance between the COM of all molecule / particle permutations i and j with $i \neq j$
- $d_{\mathrm{ij,rmsd}}$: root-mean-square distance between the COM of all molecule / particle permutations i and j

For all time points saved, the deviation from the ground truth was calculated and denoted as e.g. Δr_{gyr}. Based on this, the overall objective function f_{stab} is defined as

$$f_{\mathrm{stab}} = \langle 2|\Delta r_{\mathrm{gyr}}| + |\Delta d_{\mathrm{com,min}}|/3 + |\Delta d_{\mathrm{com,max}}|/3 + 2|\Delta d_{\mathrm{i,minDistAnyJ,rmsd}}| + |\Delta d_{\mathrm{ij,min}}| + |\Delta d_{\mathrm{ij,rmsd}}|\rangle_{10} \quad \text{(C.2)}$$

where $\langle\rangle_{10}$ indicates the time-averaging over the last 10 saving points. The objective function consequently captures both the extend of the structure, as well as the inner conformation with respect to the reference structure. Furthermore, by referencing to the original structure (ground truth), a value of zero indicates perfect agreement, while increasing values of f_{stab} indicate increasing structural changes. Negative values are not possible.

C.4 MD Quality Criteria

Criteria to identify errors and quality issues during MD runs were as follows:

- Minimum distance between A and B less than 0.35 nm
- Temperature less than 290 K or more than 305 K
- Minimum distance to a periodic image less than 7 nm
- Minimum distance between A-B less than the distance to a periodic image
- Difference between minimum distance between A-B and distance to the PBC less than 3 nm

MD runs which failed any of these criteria at the end of a run were discarded.

Appendix D

HBcAg Results Supplementary

This chapter is based on the following publications:

P. N. Depta, M. Dosta, W. Wenzel, M. Kozlowska, and S. Heinrich. Hierarchical Coarse-Grained Strategy for Macromolecular Self-Assembly: Application to Hepatitis B Virus-Like Particles. *Int. J. Mol. Sci.*, 23(23):14699, 2022

P. N. Depta, M. Dosta, and S. Heinrich. Data-Driven Multiscale Modeling of Self-Assembly and Hierarchical Structural Formation in Biological Macro-Molecular Systems. In W. E. Nagel, D. H. Kröner, and M. M. Resch (editors), *High Performance Computing in Science and Engineering '21*. Springer International Publishing, Cham, 2023

D.1 Spatial Descriptors

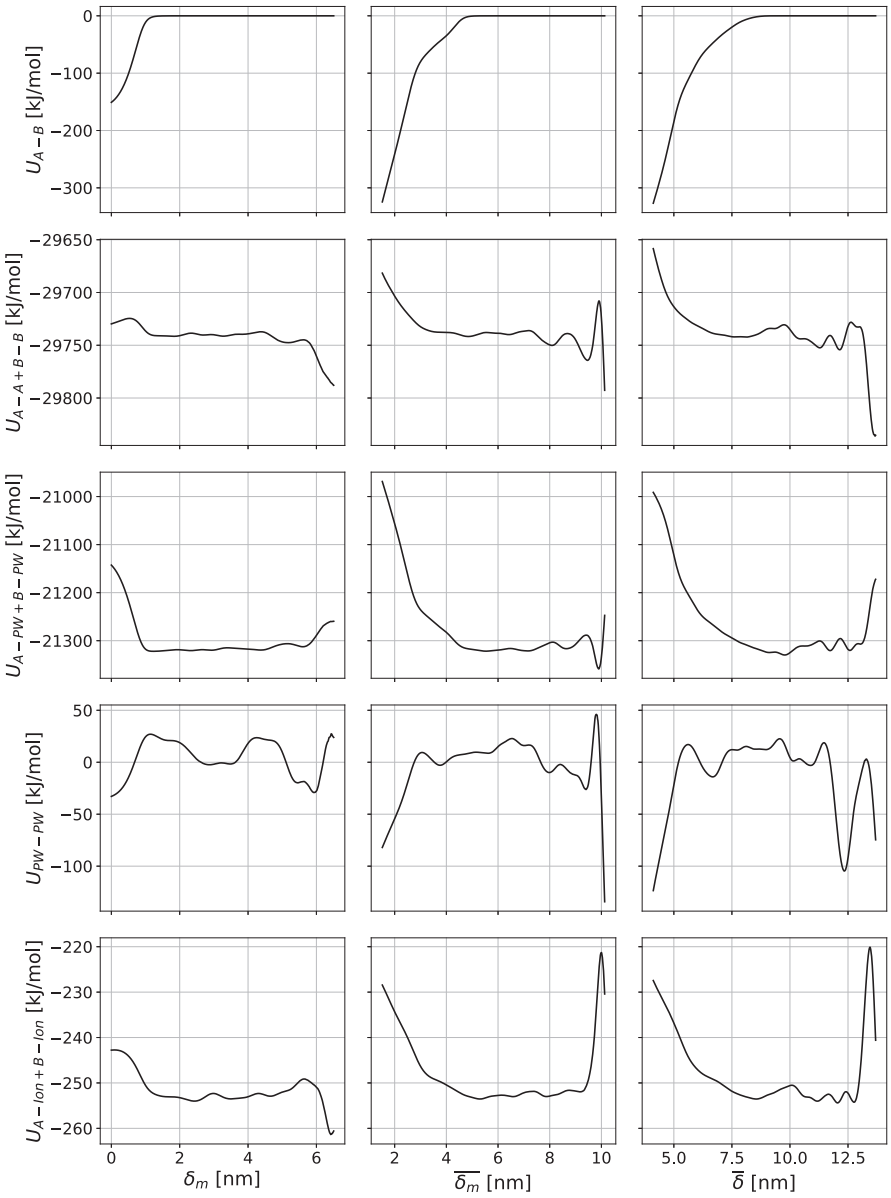
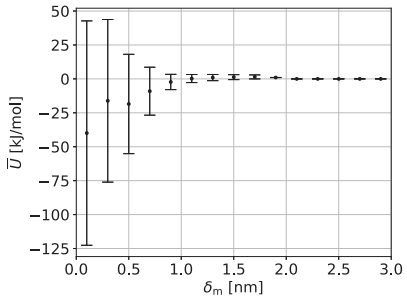
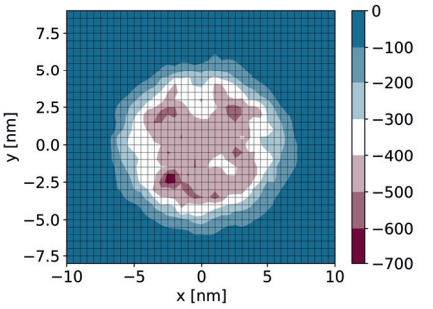


FIGURE D.1: Overview of spatial descriptor trends using Gaussian smoothing for $HBcAg_2$ – $HBcAg_2$ interaction with final MD data set. Note that excitations at large distances are due to low numbers of samples.

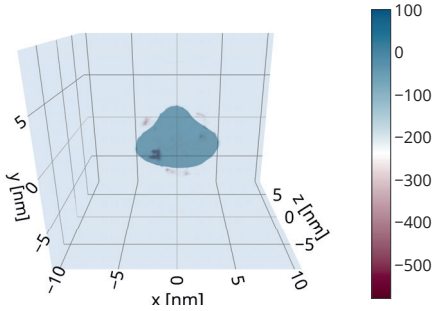
D.2 Biased MD-Based Interaction Potential



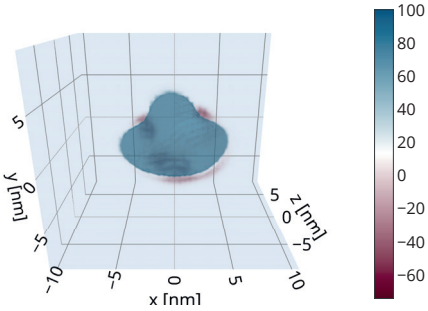
(A) Average and STD over minimum distance in grid.



(B) X-Y cross-section minimum over all remaining dimensions.



(C) 3D minimum over orientations.

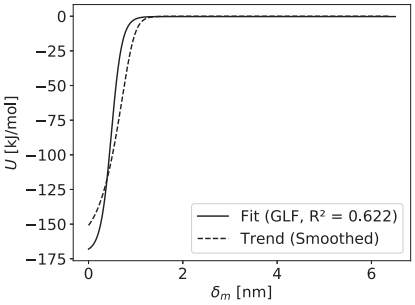


(D) 3D mean over orientations.

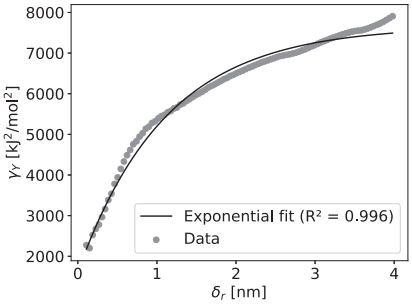
FIGURE D.2: Interaction potential field from **biased MD** sampling at binding locations (units of kJ/mol). Note that trend and variogram models were generated without biased MD data. Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

D.3 Kriging Statistical Data

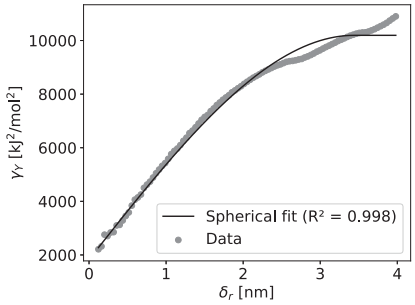
The statistical data of the Kriging algorithm for $HBcAg_2$ is reprinted with permission from Depta *et al.* [225] under CC-BY 4.0 license: "Note that not all shown trend and variogram fits are considered valid for Kriging purposes. See Sec. 4.3.3 for requirements. Valid trends are A–B, A–PW + B–PW, PW–PW, A–Ion + B–Ion, PW–Ion, Bond. Not valid trends are A–A + B–B, Ion–Ion, G96-Angles, improper dihedral angles, reciprocal coulomb potential. Valid variogram models were only those for potential A–B besides that above range."



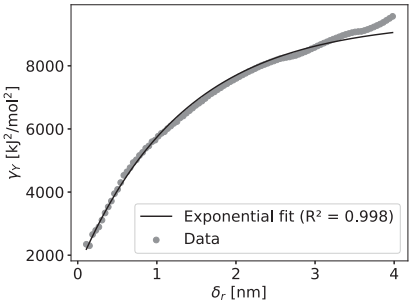
(A) Potential trend.



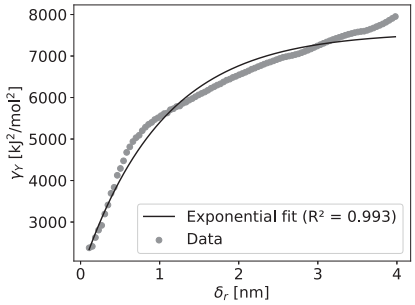
(B) Overall variogram.



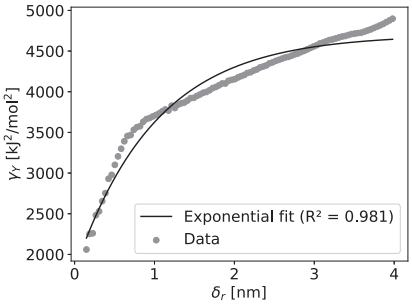
(C) Variogram section 0.



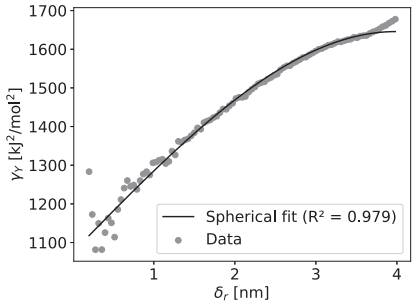
(D) Variogram section 1.



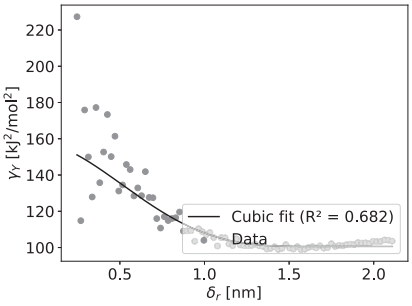
(E) Variogram section 2.



(F) Variogram section 3.

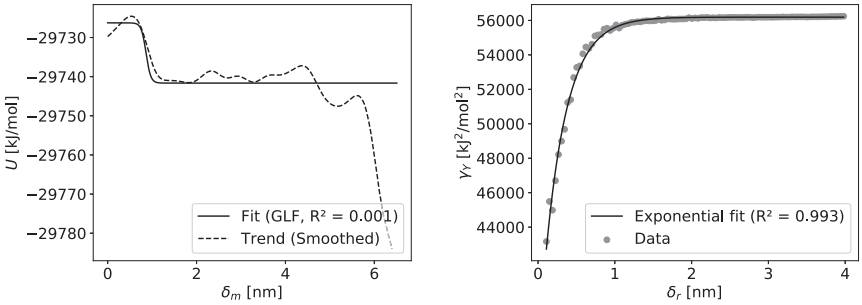


(G) Variogram section 4.

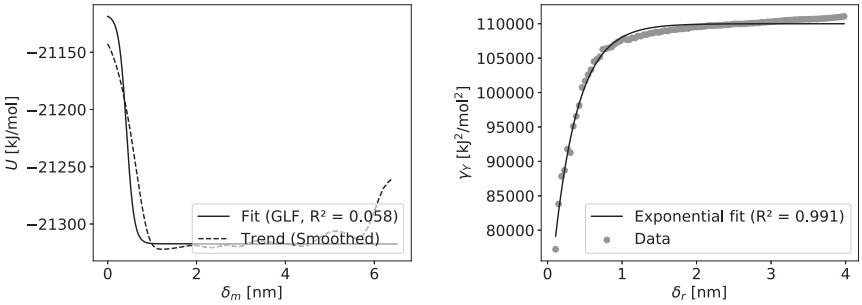


(H) Variogram above range.

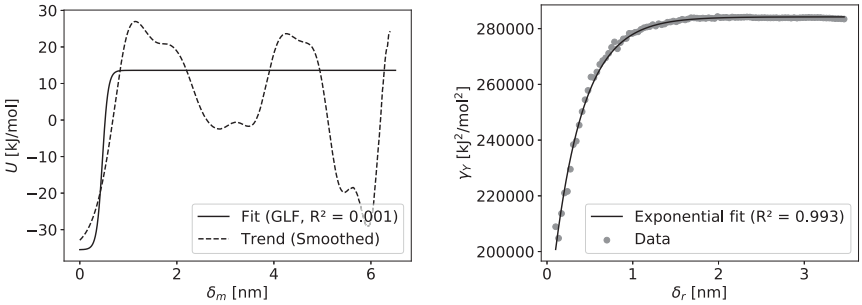
FIGURE D.3: $HBcAg_2$ – $HBcAg_2$ potential A–B. Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.



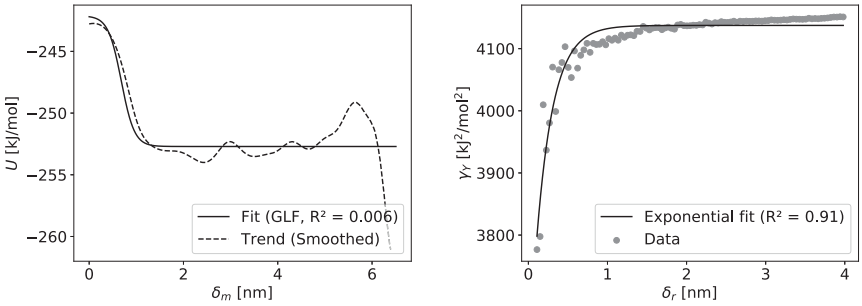
(A) A–A + B–B.



(B) A–PW + B–PW.

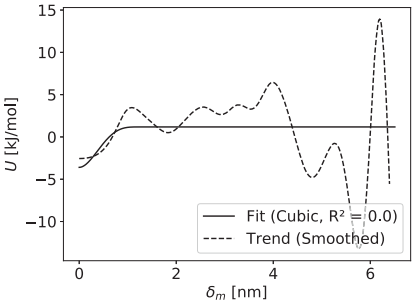


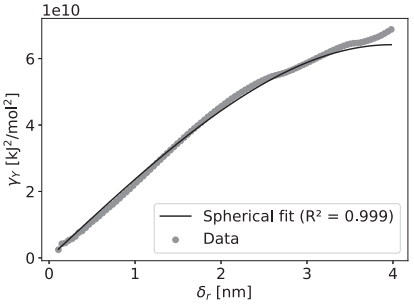
(C) PW–PW.



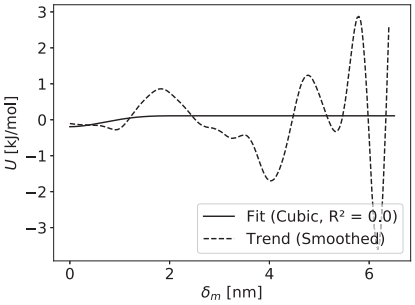
(D) A–Ion + B–Ion.

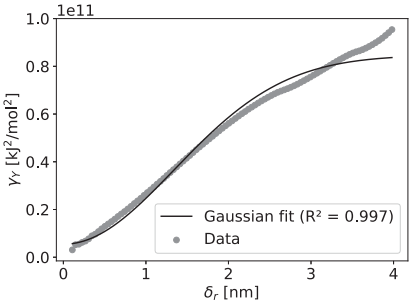
FIGURE D.4: $HBcAg_2$ – $HBcAg_2$ potential trends (left) and overall variogram (right). Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.



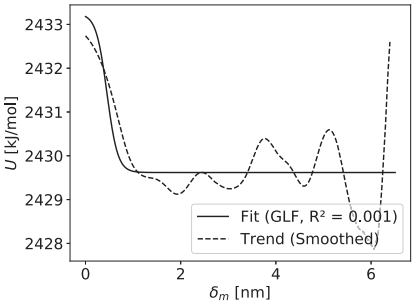


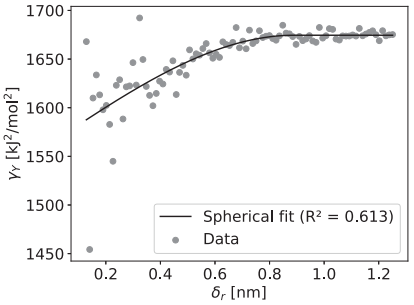
(A) PW–Ion.



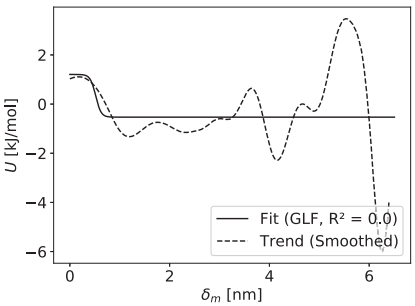


(B) Ion–Ion.





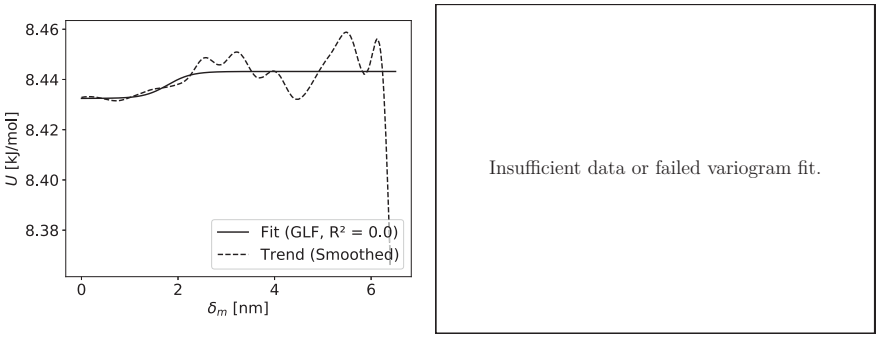
(C) Bonds.



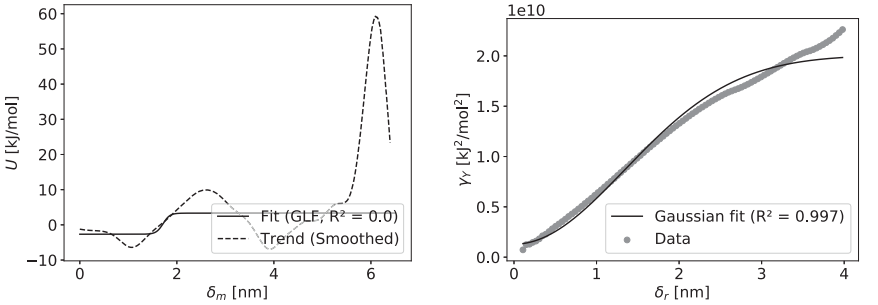
Insufficient data or failed variogram fit.

(D) G96-Angle.

FIGURE D.5: $HBcAg_2$ – $HBcAg_2$ potential trends (left) and overall variogram (right). Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.



(A) Improper dihedral angles.



(B) Coulomb reciprocal.

FIGURE D.6: $HBcAg_2$ – $HBcAg_2$ potential trends (left) and overall variogram (right). Adapted with permission from Depta *et al.* [225] under CC-BY 4.0 license.

Appendix E

PDC Results Supplementary

This chapter is based on the following publications:

P. N. Depta, U. Jandt, M. Dosta, A.-P. Zeng, and S. Heinrich. Toward Multiscale Modeling of Proteins and Bioagglomerates: An Orientation-Sensitive Diffusion Model for the Integration of Molecular Dynamics and the Discrete Element Method. *J. Chem. Inf. Model.*, 59(1):386–398, 2019

P. N. Depta, M. Dosta, and S. Heinrich. Data-Driven Multiscale Modeling of Self-Assembly and Hierarchical Structural Formation in Biological Macro-Molecular Systems: Pyruvate Dehydrogenase Complex. In W. E. Nagel, D. H. Kröner, and M. M. Resch (editors), *High Performance Computing in Science and Engineering '22*. Springer International Publishing, Cham, 2024 (in print)

P. N. Depta, M. Dosta, and S. Heinrich. Multiscale Model-Based Investigation of Functional Macromolecular Agglomerates for Biotechnological Applications. In A. Kwade and I. Kampen (editors), *Dispersity, Structure and Phase Changes of Proteins and Bio Agglomerates in Biotechnological Processes*. Springer International Publishing, Cham, 2024 (in print)

E.1 Binding Locations

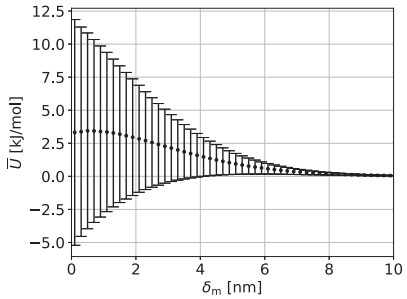
Table E.1: Binding locations of PDC components extracted from structural assemblies in ref. [271] provided by Uwe Jandt (E2 – E2) and Cornelius Jacobi (E2 – E1, E3BP – E3). Note the large flexiblity of E2 – E2.

#	x [nm]	y [nm]	z [nm]	α [rad]	β [rad]	γ [rad]
E2 – E1						
1	-0.87	-0.27	-1.87	-2.46	1.06	2.22
2	-0.87	-0.27	-1.87	-0.85	-0.97	-0.75
E3BP – E3						
1	7.37	-2.62	-1.94	2.48	0.54	1.47
2	7.37	-2.62	-1.94	-2.59	-0.48	1.63
E2 – E2						
1	-4.02	3.27	9.78	1.43	-0.73	0.13
2	-7.89	-1.77	-12.45	-2.48	1.2	0.30
3	-4.48	-5.84	-9.77	3.02	0.82	-0.34
4	-2.94	-4.72	-10.02	2.81	0.72	-0.18
5	-7.28	-9.64	-6.54	2.53	0.6	-0.87
6	-4.63	-4.47	-10.41	3.07	0.91	-0.14
7	-4.82	-7.55	-9.11	2.85	0.78	-0.55
8	-3.52	-0.55	-9.01	-2.82	0.75	0.41
9	-4.97	-5.37	-9.79	2.98	0.87	-0.31
10	-2.67	-7.18	-6.68	2.91	0.55	-0.39
11	-4.06	-6.62	-7.31	2.92	0.55	-0.40
12	-8.05	-2.96	-11.39	-2.9	1.1	-0.05
13	-7.49	-3.11	-10.15	-2.38	0.97	-0.10
14	-1.33	-6.81	-5.44	2.53	0.41	-0.31
15	-5.9	-1.74	-10.18	-2.77	0.96	0.29
16	-3.18	-3.95	-9.5	-2.44	0.75	-0.12
17	-7.77	-5.39	-10.91	-3.11	1.05	-0.59
18	-5.8	6.5	-7.97	-3.09	0.91	0.61
19	-8.95	1.81	-11.98	-2.19	1.03	0.81
20	-4.39	-9.41	-4.15	2.44	0.42	-0.68
21	-5.5	4.74	-8.8	2.4	0.93	0.35
22	-5.84	-9.1	-6.15	2.82	0.62	-0.81
23	-4.39	-6.15	-9.32	2.7	0.78	-0.42
24	-4.09	-4.69	-8.67	2.8	0.64	-0.23
25	-3.04	2.16	-7.9	1.51	0.69	-0.03

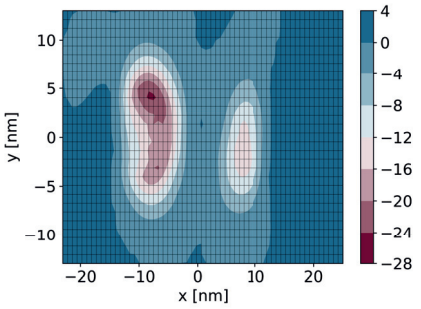
26	-3.19	-7.03	-6.96	2.16	0.5	-0.41
27	-4.17	3.62	-8.88	2.5	0.85	0.15
28	-7.17	5.53	-7.72	2.95	0.96	0.53
29	-3.32	-0.59	-7.4	1.41	0.59	-0.37
30	-3.56	8.63	0.11	-1.79	0.15	0.67
31	-9.38	-2.1	-10.58	-3.08	1.19	0.00
32	-3.93	-0.88	-8.37	1.48	0.74	-0.35
33	-3.83	5.18	-6.44	2.7	0.6	0.35
34	-4.53	-0.82	-8.42	1.17	0.74	-0.39
35	-1.43	6.01	0.18	-1.67	0.09	0.33
36	-5.61	8.91	-4.39	-2.96	0.45	0.82
37	-10.02	1.96	-11.79	-1.44	1.2	0.88
38	-4.09	-1.15	-7.31	0.56	0.66	-0.47
39	-4.46	9.38	1.24	-2.66	0.02	0.81
40	-7.8	8.1	-5.18	-2.27	0.17	1.09
41	-3.92	0.76	-7.98	1.1	0.71	-0.21
42	-1.78	6.46	-1.84	-1.69	0.32	0.37
43	-3.32	7.41	-2.97	-2.99	0.33	0.55
44	-8.93	5.77	-10.51	-1.73	0.84	1.03
45	-2.93	-0.62	-6.23	1.36	0.51	-0.27
46	-2.21	7.11	-0.85	-1.45	0.23	0.55
47	-3.8	5.54	-6.64	-2.98	0.73	0.37
48	-3.65	0.43	-8.68	1.29	0.67	-0.21
49	-1.51	6.66	-0.35	-2.01	0.14	0.39
50	-4.51	8.57	-4.35	2.88	0.44	0.64
51	-8.19	2.47	-11.37	-1.92	1.01	0.69
52	-4.2	2.79	-8.55	2.19	0.76	0.10
53	-4.3	9.81	-1.22	-1.93	0.24	0.83
54	-4.65	0.61	-9	1.72	0.8	-0.16
55	-4.32	1.64	-8.81	1.49	0.86	-0.02
56	-3.07	8.84	-0.93	-1.33	0.24	0.70
57	-5.27	6.54	-6.71	-3.02	0.65	0.60
58	0.7	0.33	2.21	1.68	0.3	0.09
59	-2.67	1.57	-6.74	1.48	0.5	-0.09
60	-4.23	10.29	1.07	-1.65	0	0.87
61	-10.14	-1.51	-10.98	-2.6	1.26	0.02
62	-4.28	11.26	-1.13	-2.05	0.24	0.96
63	-8.57	-3.46	-11.89	-2.85	1.18	-0.50
64	-4.01	-1	-8.61	1.59	0.71	-0.36

65	-1.22	5.73	-0.94	-1.85	0.19	0.31
66	-4.82	5.85	-7.38	2.69	0.67	0.46
67	-3.5	2.91	-7.84	1.8	0.71	0.11
68	-2.92	8.88	-0.35	-2	0.16	0.67
69	-3.94	5.02	-7.17	2.37	0.65	0.39
70	-2.26	2.29	-6.64	1.97	0.51	0.04
71	-2.74	7.87	-0.53	-1.54	0.21	0.59
72	-3.21	2.65	-7.17	2.58	0.59	0.02
73	-3.51	7.72	-2.78	2.8	0.35	0.63
74	-3.83	7.96	6.86	-0.83	-0.46	0.85
75	-3.41	2.64	-8.63	2.7	0.83	0.07
76	-4.15	2.27	-7.01	1.64	0.62	0.10
77	-4.37	9.5	-2.12	-2.09	0.32	0.82
78	-2.47	3.75	-7.49	2.27	0.69	0.15
79	-6.33	5.59	-8.3	2.37	0.83	0.49
80	-4.41	8.9	1.03	-1.01	0.1	0.77
81	-4.66	4.61	-7.65	2.89	0.71	0.33
82	-5.18	-0.04	-9.86	0.9	0.95	-0.35
83	-3.23	7.94	-3.29	2.83	0.3	0.63
84	-4.52	0.85	-8.1	1.84	0.82	-0.08
85	-2.46	8.3	-1.06	-1.55	0.28	0.73
86	-4.13	-2.03	-8.46	1.23	0.66	-0.58
87	-5.91	4.72	-8.47	2.58	0.81	0.45

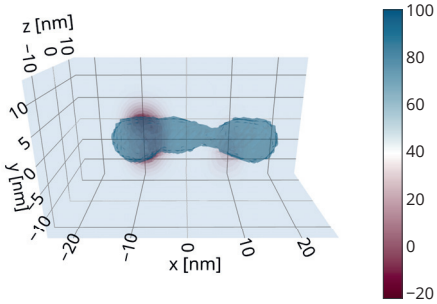
E.2 Pure MD-Based Interaction Potentials



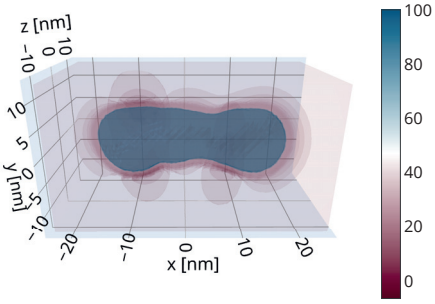
(A) Average and STD over minimum distance in grid.



(B) X-Y cross-section minimum over all remaining dimensions.

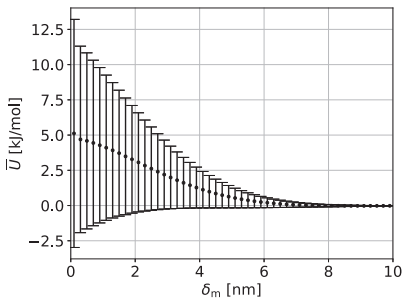


(C) 3D minimum over orientations.

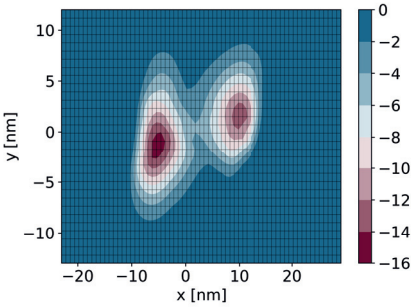


(D) 3D mean over orientations.

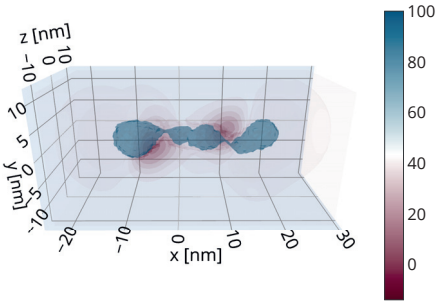
FIGURE E.1: Interaction potential field of **E2 with E1 from pure MD**-based sampling strategy (units of kJ/mol).



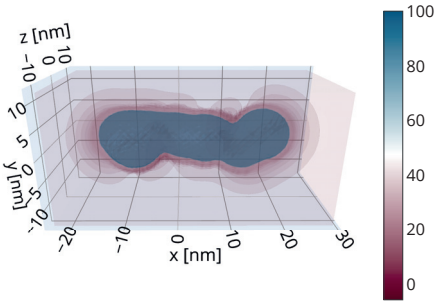
(A) Average and STD over minimum distance in grid.



(B) X-Y cross-section minimum over all remaining dimensions.



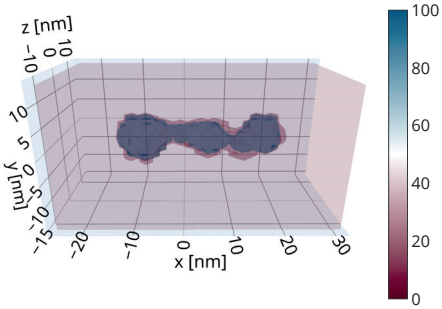
(C) 3D minimum over orientations.



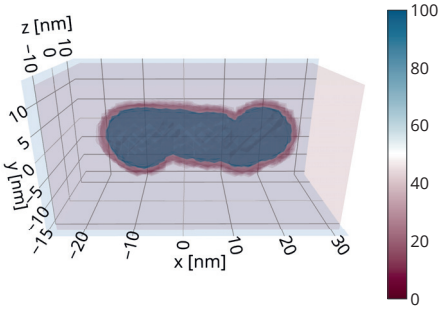
(D) 3D mean over orientations.

FIGURE E.2: Interaction potential field of **E3BP with E3 from pure MD**-based sampling strategy (units of kJ/mol).

E.3 Repulsive-Only Interaction Potentials

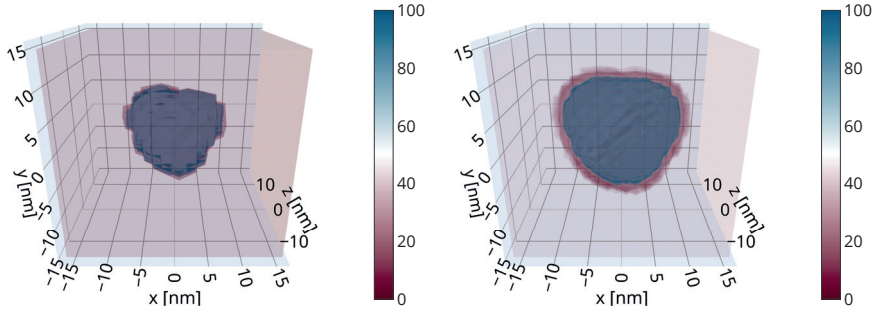


(A) 3D minimum over orientations.



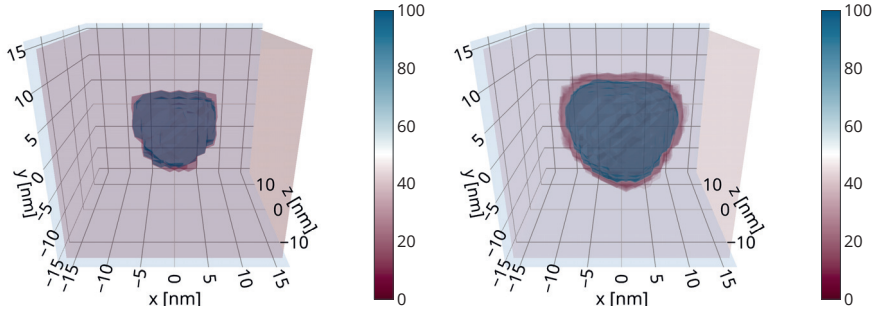
(B) 3D mean over orientations.

FIGURE E.3: Interaction potential field of **E3BP with E1 from repulsion model** (units of kJ/mol).



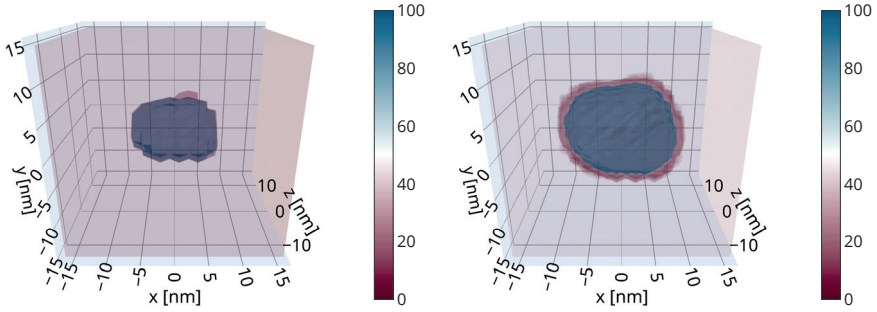
(A) 3D minimum over orientations. (B) 3D mean over orientations.

FIGURE E.4: Interaction potential field of **E1 with E1 from repulsion model** (units of kJ/mol).



(A) 3D minimum over orientations. (B) 3D mean over orientations.

FIGURE E.5: Interaction potential field of **E1 with E3 from repulsion model** (units of kJ/mol).

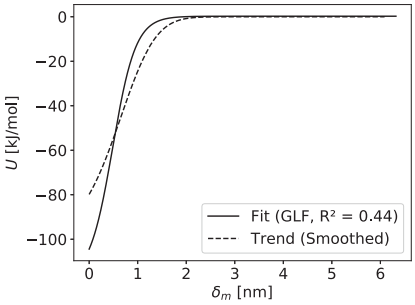


(A) 3D minimum over orientations. (B) 3D mean over orientations.

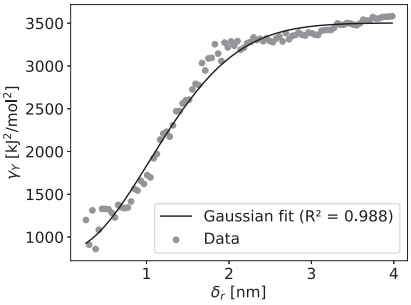
FIGURE E.6: Interaction potential field of **E3 with E3 from repulsion model** (units of kJ/mol).

E.4 Kriging Statistical Data

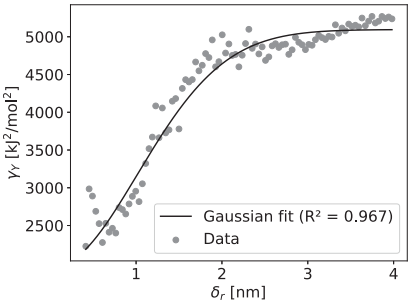
Valid potential trends are for E2 – E2 potentials A – B, A – PW + B – PW, and reciprocal Coulomb; for E2 – E1 potentials A – B and A – PW + B – PW; for E3BP - E3 potentials A – B and A – PW + B – PW. Valid variogram models were those for potential A – B over all distance classes for all molecular interactions. See Sec. 4.3.3 for requirements. Note that fitting jumps at large δ_m (typically beyond 6-7 nm) are caused by significant undersampling. None of such fits were considered valid and used.



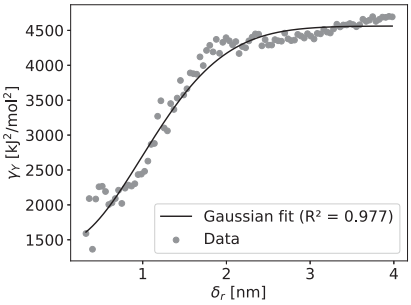
(A) Potential trend.



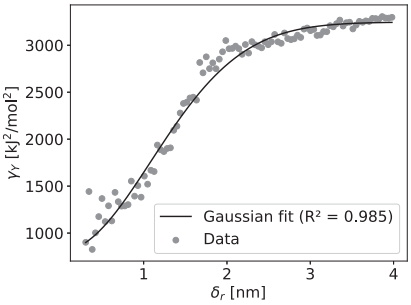
(B) Overall variogram.



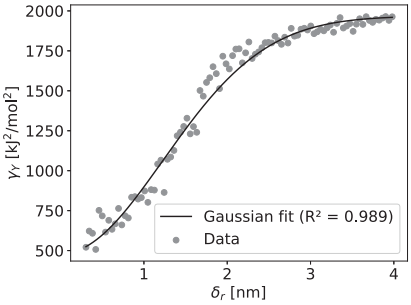
(C) Variogram section 0.



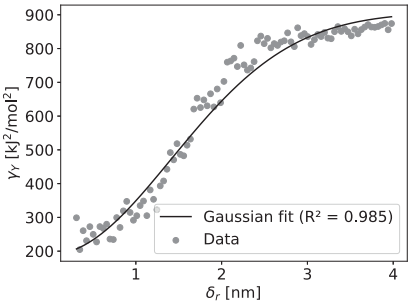
(D) Variogram section 1.



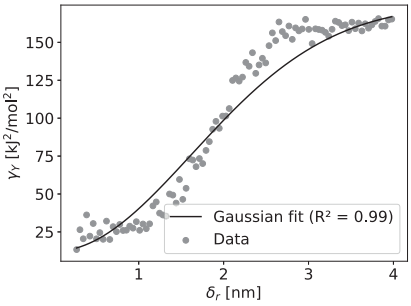
(E) Variogram section 2.



(F) Variogram section 3.

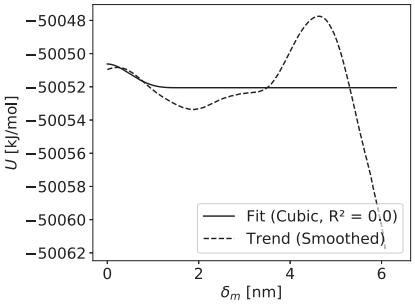


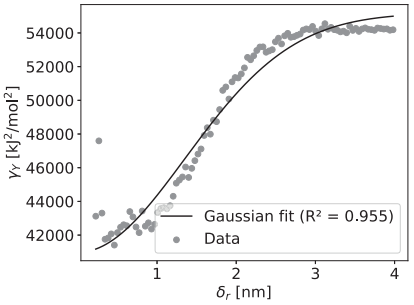
(G) Variogram section 4.



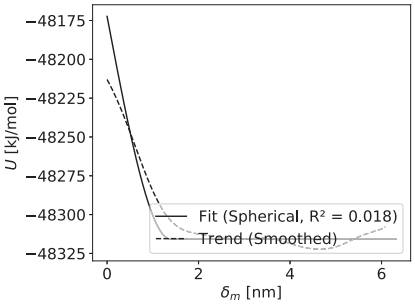
(H) Variogram above range.

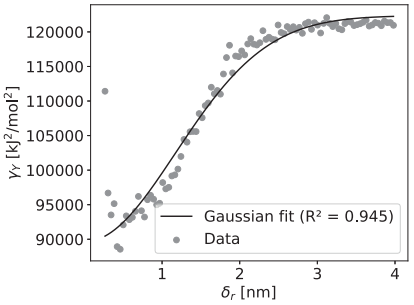
FIGURE E.7: E2 – E2 potential A–B.



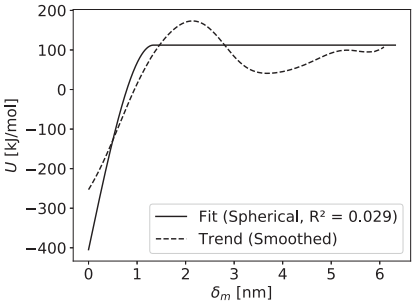


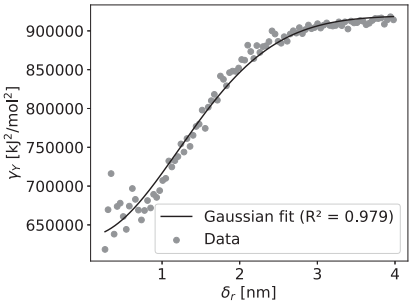
(A) A–A + B–B.



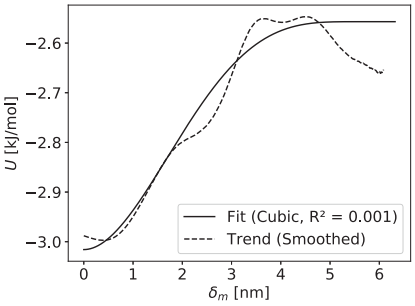


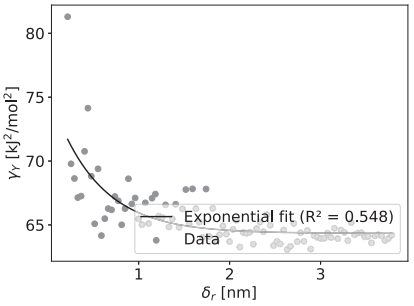
(B) A–PW + B–PW.





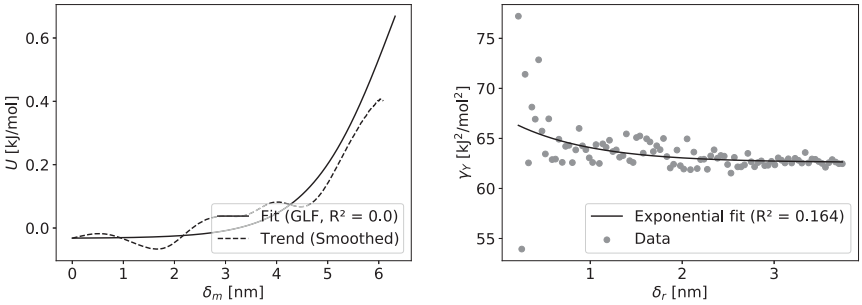
(C) PW–PW.



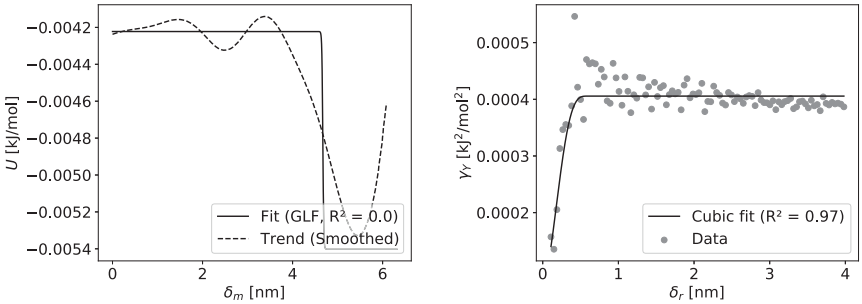


(D) A–Ion + B–Ion.

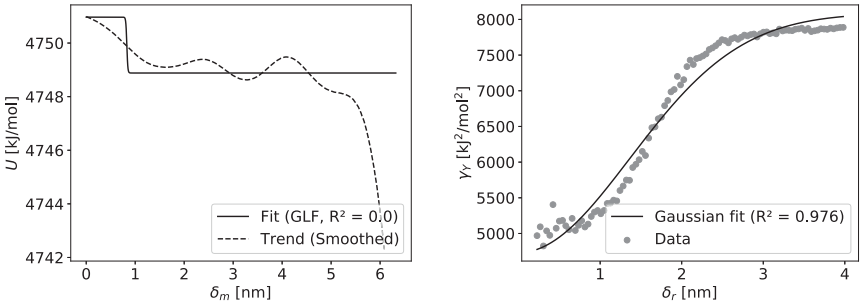
FIGURE E.8: E2 – E2 potential trends (left) and overall variogram (right).



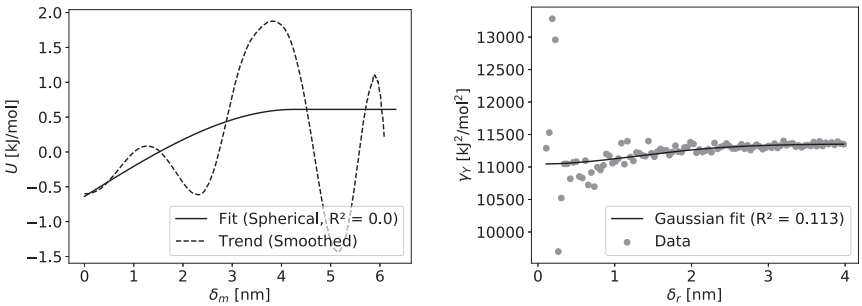
(A) PW–Ion.



(B) Ion–Ion.

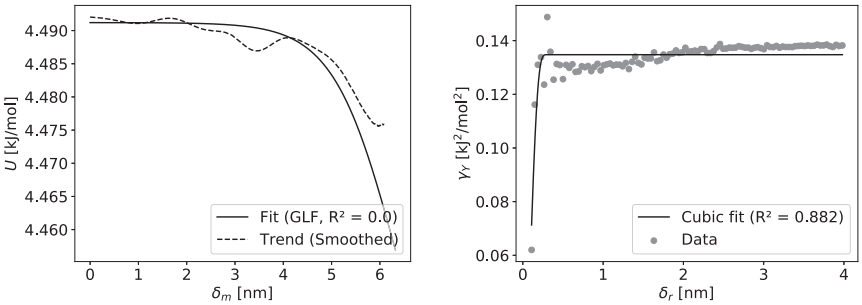


(C) Bonds.

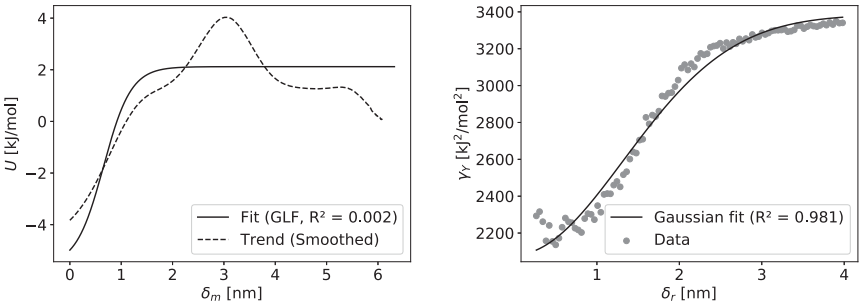


(D) G96-Angle.

FIGURE E.9: E2 – E2 potential trends (left) and overall variogram (right).

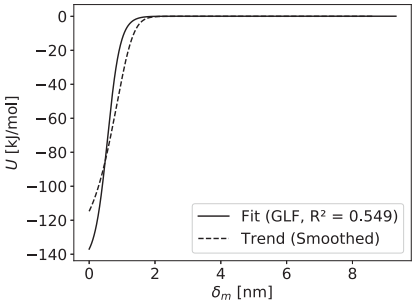


(A) Improper dihedral angles.

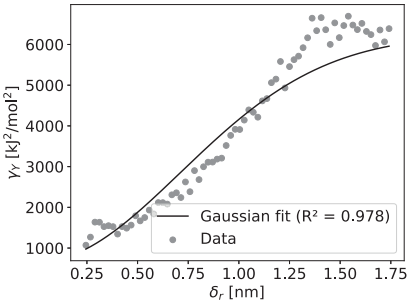


(B) Coulomb reciprocal.

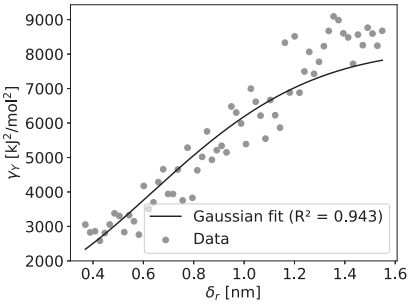
FIGURE E.10: E2 – E2 potential trends (left) and overall variogram (right).



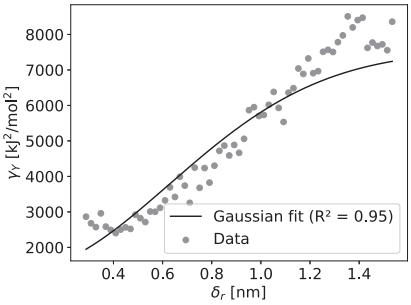
(A) Potential trend.



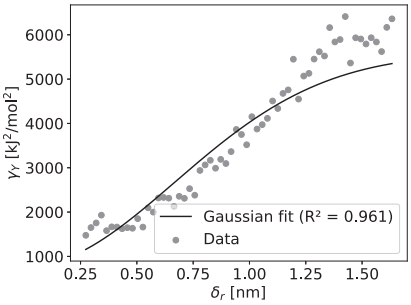
(B) Overall variogram.



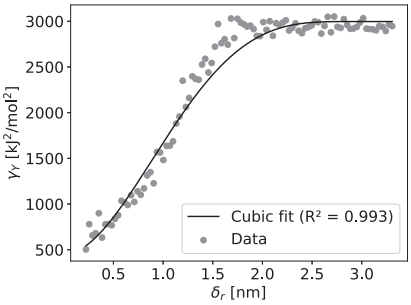
(C) Variogram section 0.



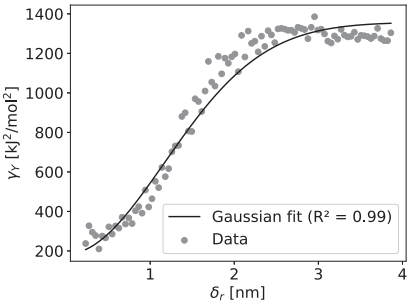
(D) Variogram section 1.



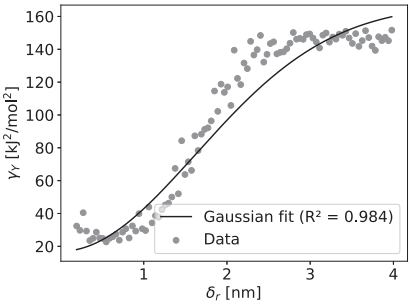
(E) Variogram section 2.



(F) Variogram section 3.

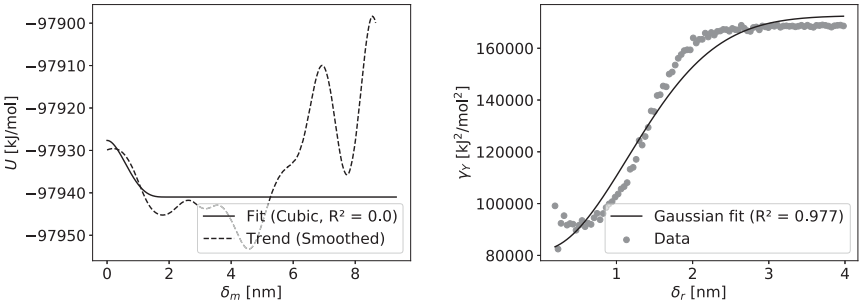


(G) Variogram section 4.

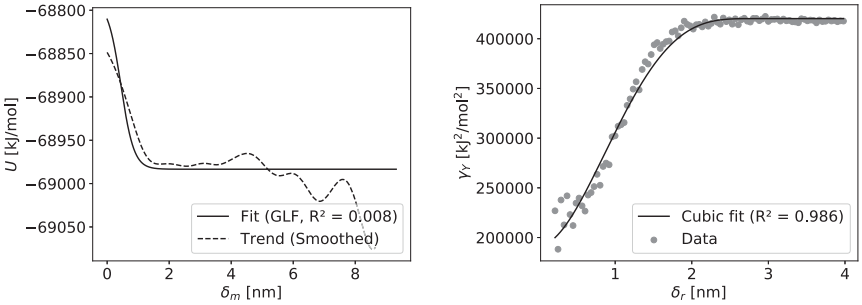


(H) Variogram above range.

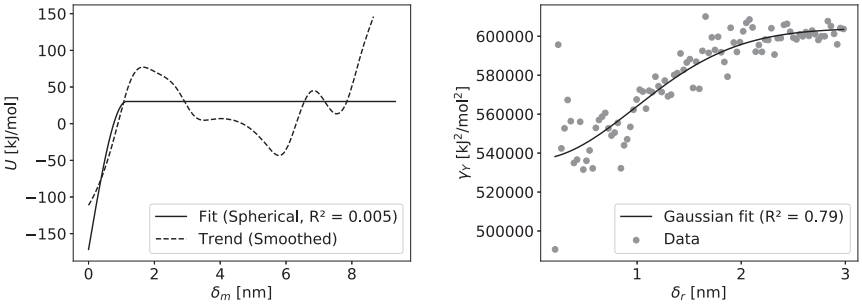
FIGURE E.11: E2 – E1 potential A–B.



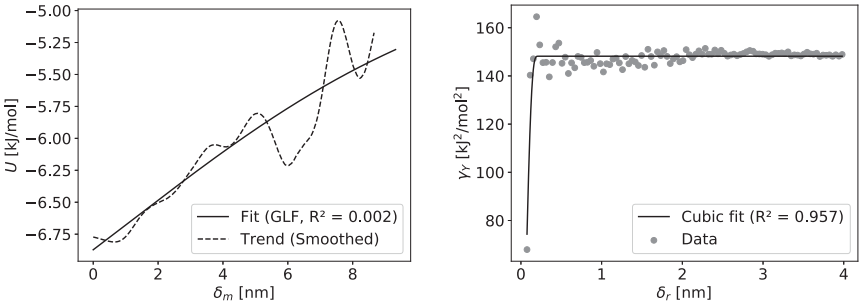
(A) A–A + B–B.



(B) A–PW + B–PW.

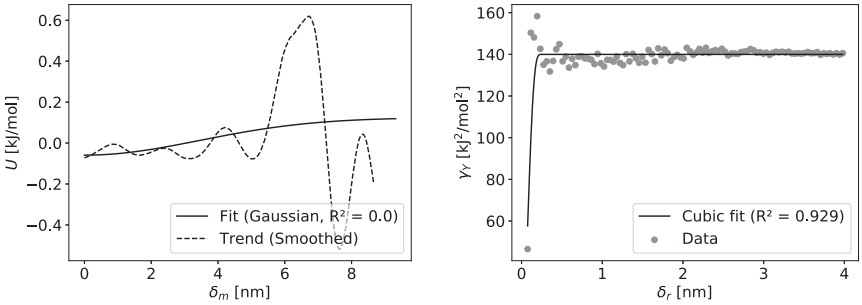


(C) PW–PW.

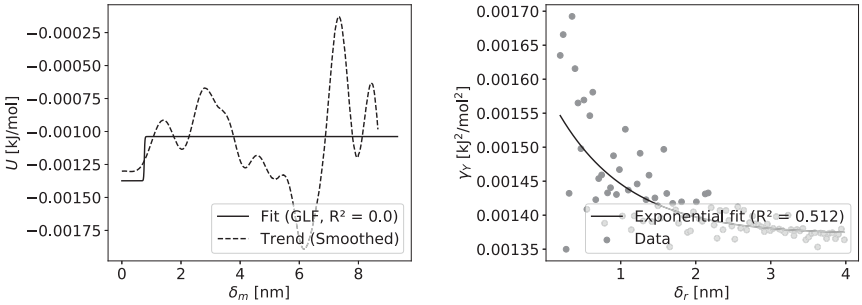


(D) A–Ion + B–Ion.

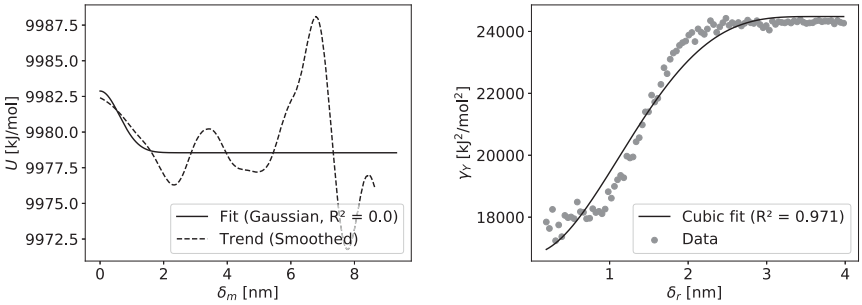
FIGURE E.12: E2 – E1 potential trends (left) and overall variogram (right).



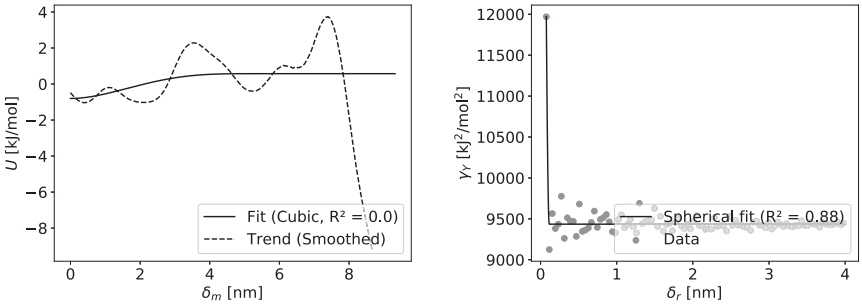
(A) PW–Ion.



(B) Ion–Ion.

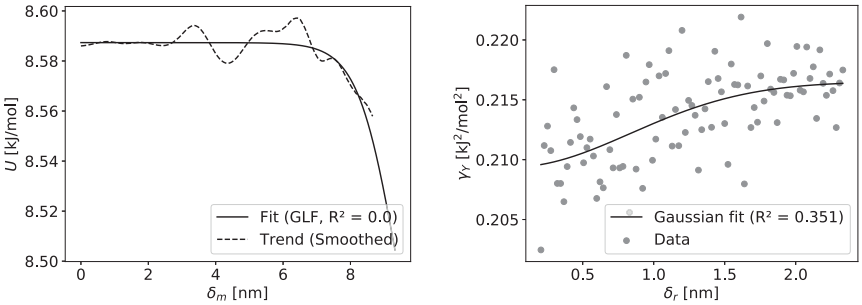


(C) Bonds.

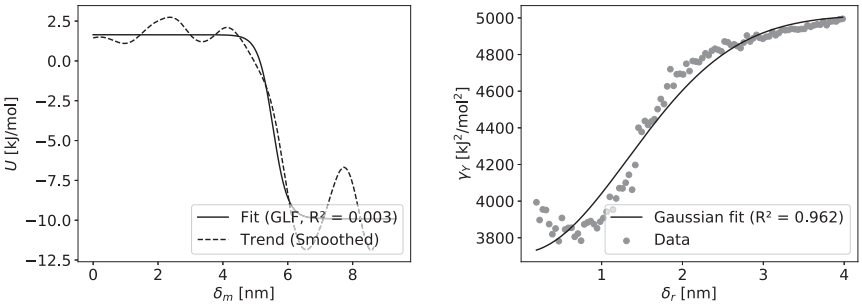


(D) G96-Angle.

FIGURE E.13: E2 – E1 potential trends (left) and overall variogram (right).

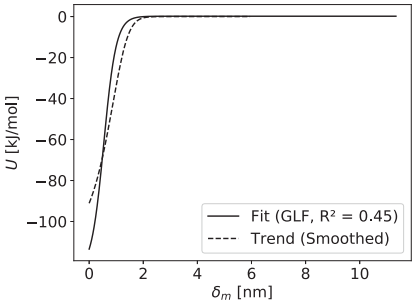


(A) Improper dihedral angles.

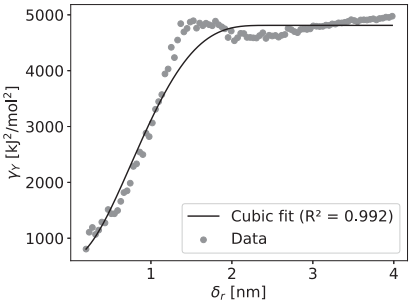


(B) Coulomb reciprocal.

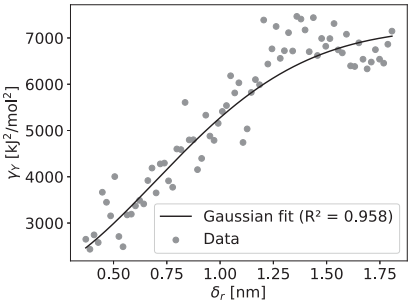
FIGURE E.14: E2 – E1 potential trends (left) and overall variogram (right).



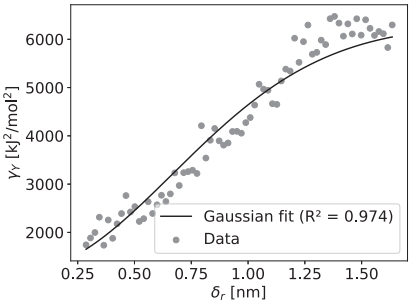
(A) Potential trend.



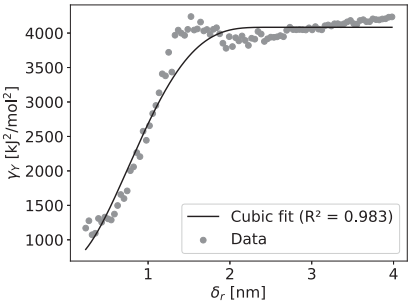
(B) Overall variogram.



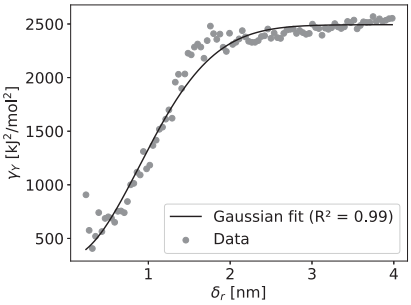
(C) Variogram section 0.



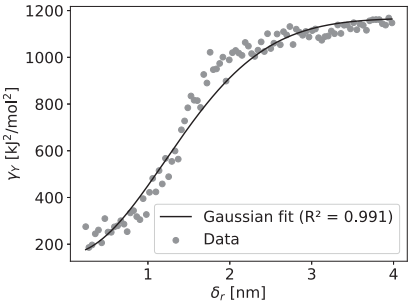
(D) Variogram section 1.



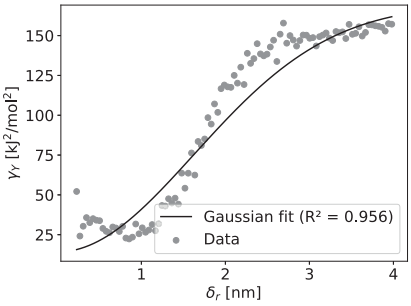
(E) Variogram section 2.



(F) Variogram section 3.

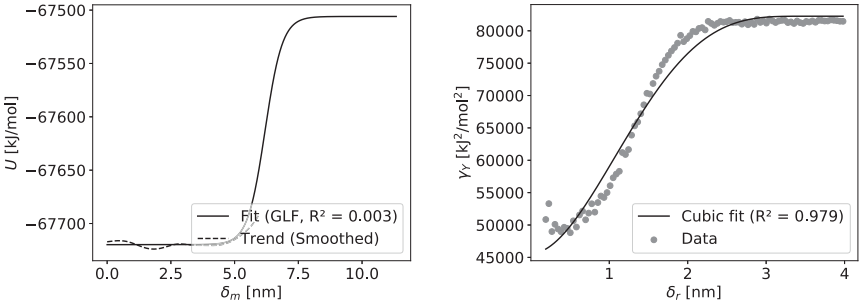


(G) Variogram section 4.

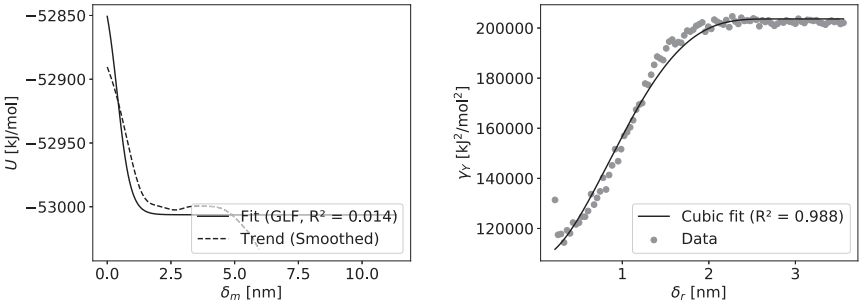


(H) Variogram above range.

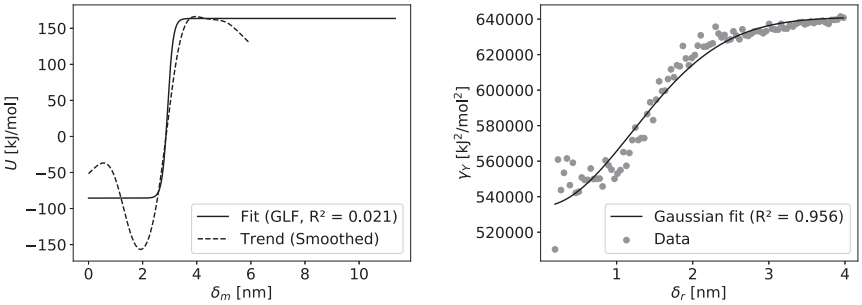
FIGURE E.15: E3BP – E3 potential A–B.



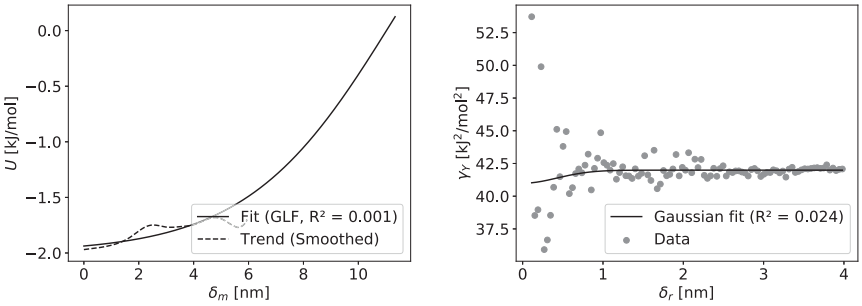
(A) A–A + B–B.



(B) A–PW + B–PW.

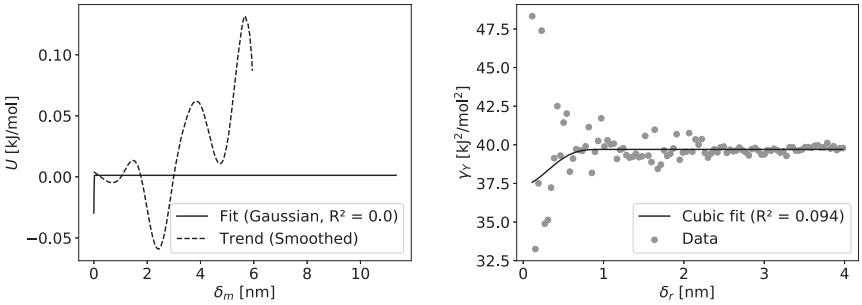


(C) PW–PW.

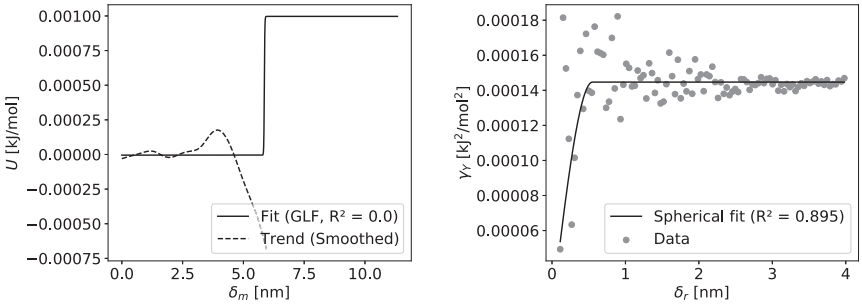


(D) A–Ion + B–Ion.

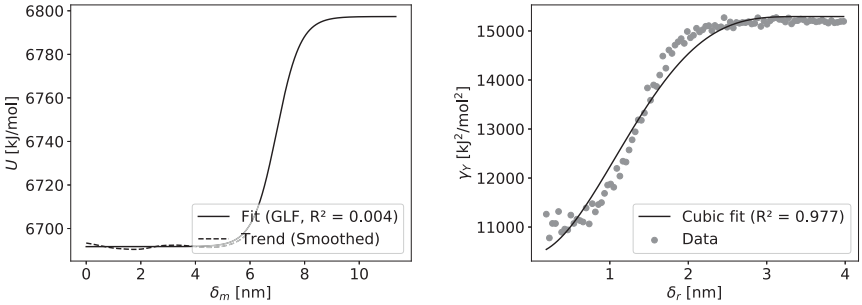
FIGURE E.16: E3BP – E3 potential trends (left) and overall variogram (right).



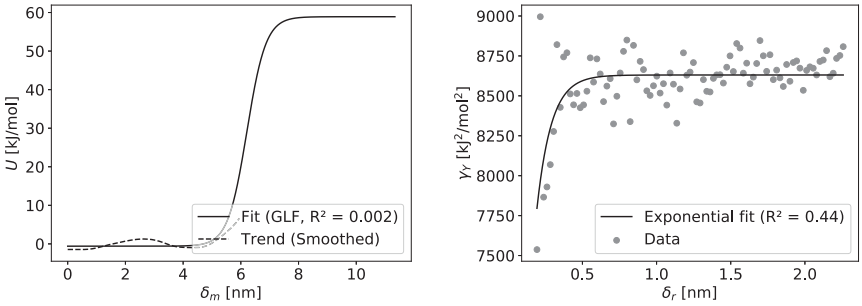
(A) PW–Ion.



(B) Ion–Ion.

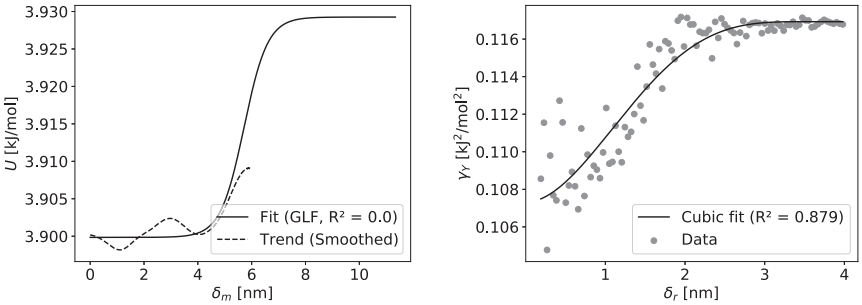


(C) Bonds.

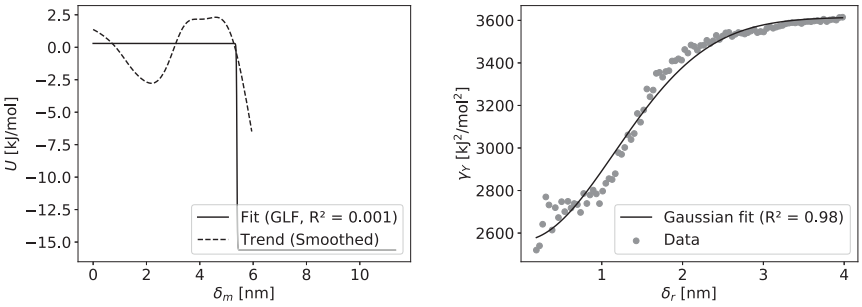


(D) G96-Angle.

FIGURE E.17: E3BP – E3 potential trends (left) and overall variogram (right).



(A) Improper dihedral angles.



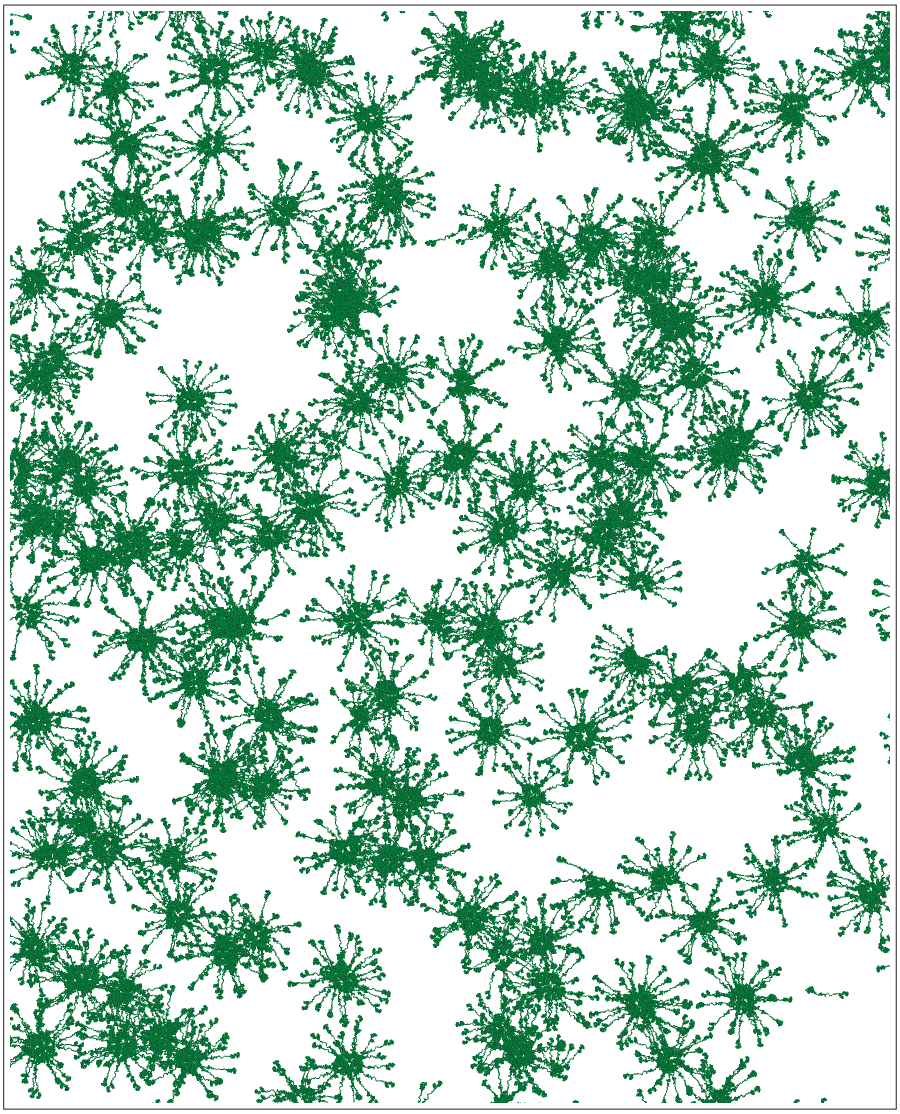
(B) Coulomb reciprocal.

FIGURE E.18: E3BP – E3 potential trends (left) and overall variogram (right).

E.5 PDC Self-Assembly

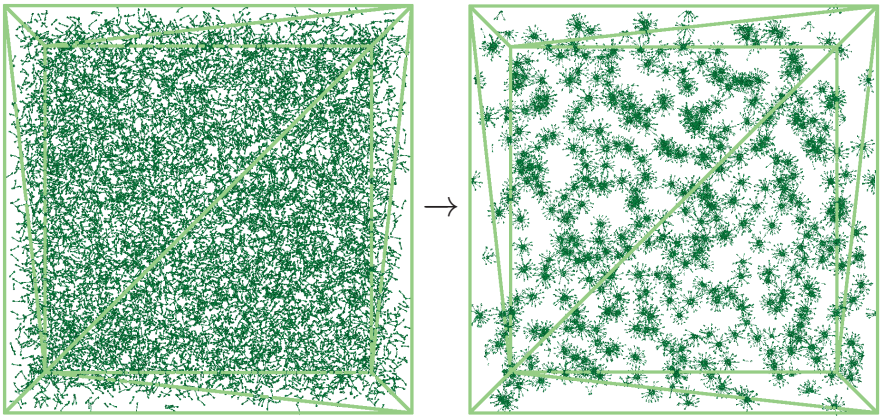
E.5.1 Pure E2 System

SP-PDC-1 (Without Annealing)

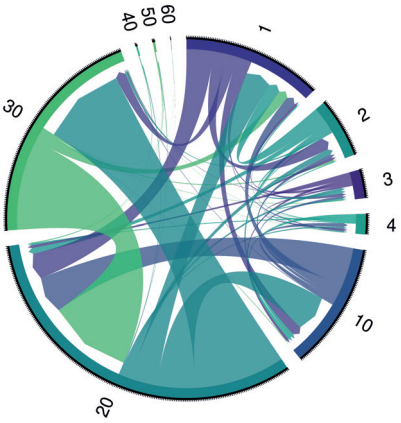


(A) Close-up visualization of self-assembled structures by enzyme type: E1, E3, E2, E3BP.

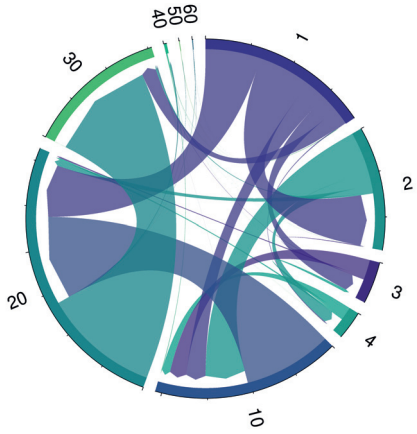
FIGURE E.19: PDC self-assembly for a pure E2 system at concentration of $1\,\mathrm{mg\,mL^{-1}}$ in $1\,\mu\mathrm{m}^3$ cubic box ($1\,\mu\mathrm{m}$ edge) using simulation protocol SP-PDC-1.



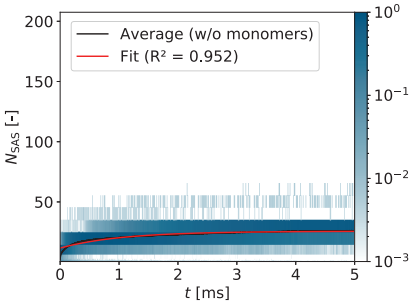
(B) Visualization of before (left) and after (right) self-assembly using back-bone carbon structures. Color indicates enzyme type: E1, E3, E2, E3BP.



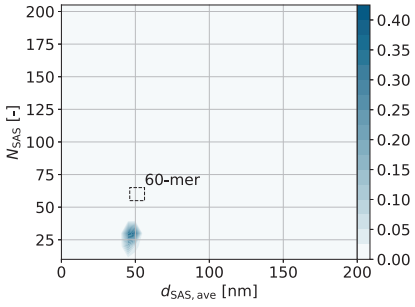
(C) Assembly pathway by bi-directional transitions between size classes.



(D) Assembly pathway by net transitions between size classes.

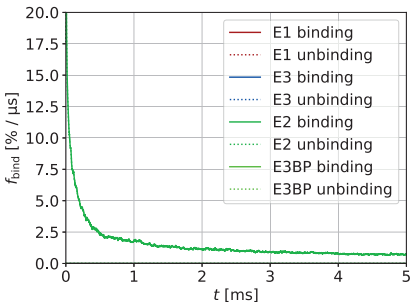


(E) Histogram of self-assembled structures by numbered size over time.

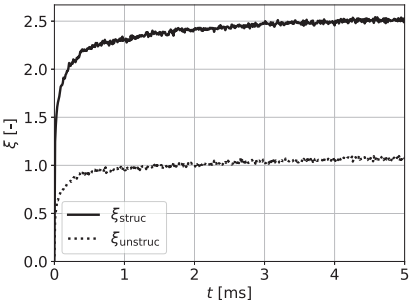


(F) Numbered size versus average extent without monomers (final time).

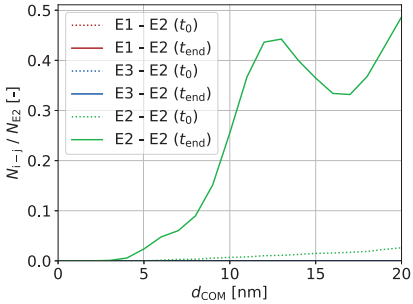
FIGURE E.19: PDC self-assembly for a pure E2 system at concentration of $1\,\mathrm{mg\,mL^{-1}}$ in $1\,\mu\mathrm{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1.



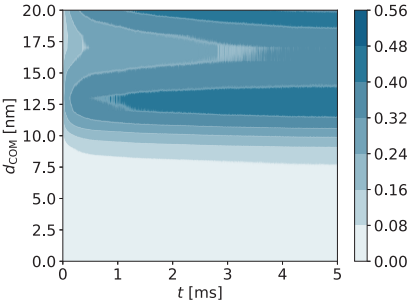
(G) (Un-)binding rates per enzyme type with uniform smoothing over 10 µs.



(H) Average (un-)structured contacts per enzyme (10 µs saving).



(I) Reactant distance distribution at beginning and end of self-assembly.

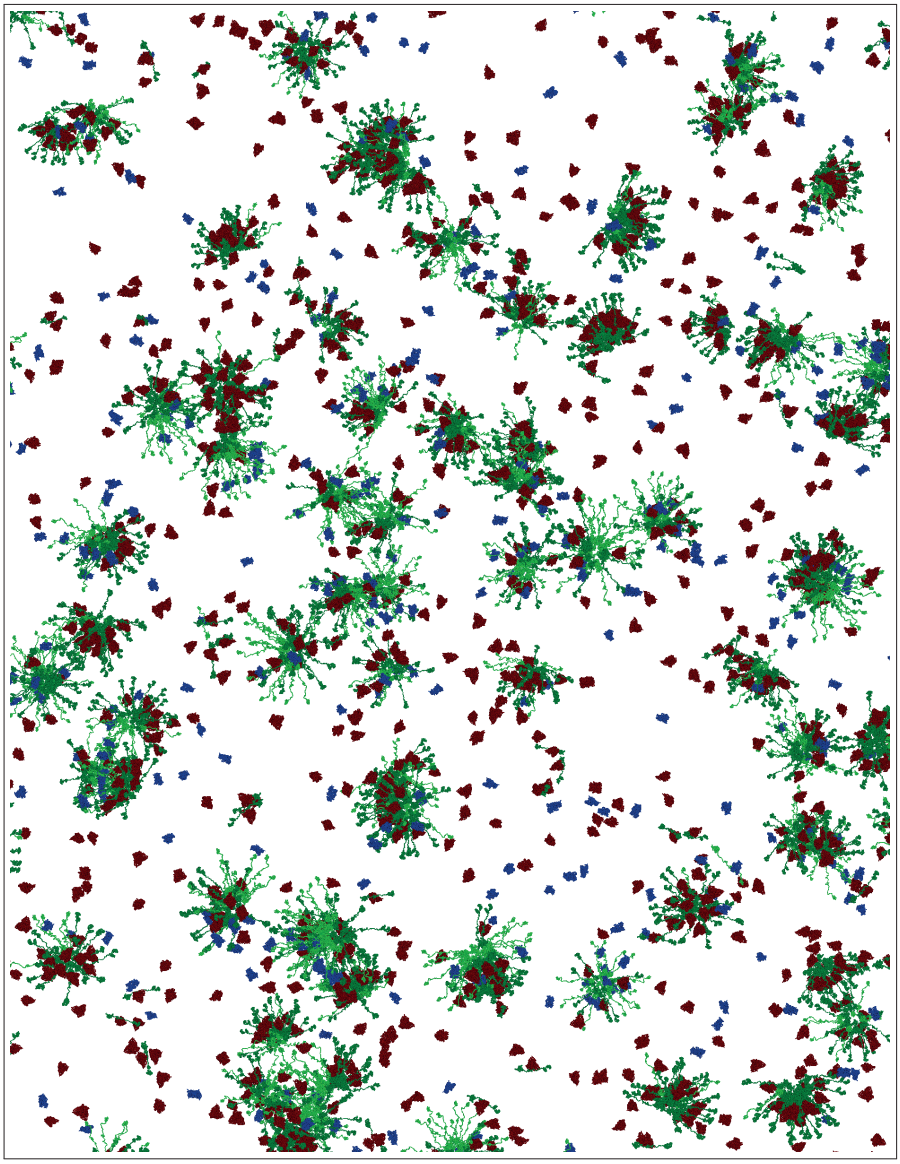


(J) Reactant distance distribution for E2 - E2 over time.

FIGURE E.19: PDC self-assembly for a pure E2 system at concentration of $1\,\mathrm{mg\,mL^{-1}}$ in $1\,\mu\mathrm{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1.

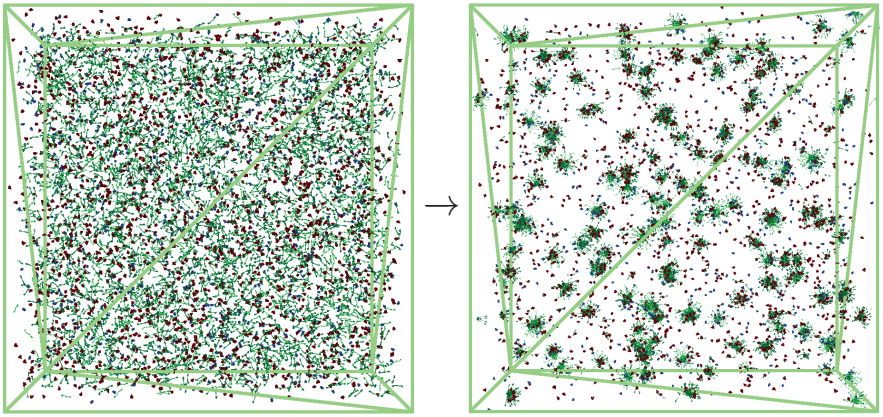
E.5.2 Full Component PDC System

SP-PDC-1 (Without Annealing)

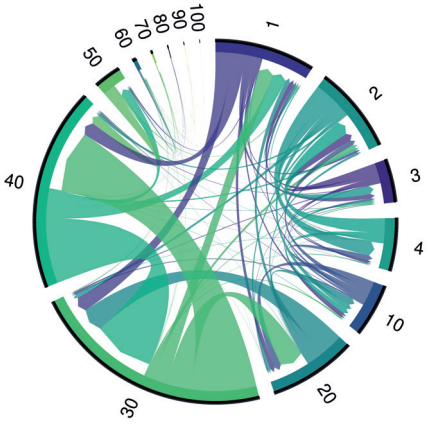


(A) Close-up visualization of self-assembled structures by enzyme type: E1, E3, E2, E3BP.

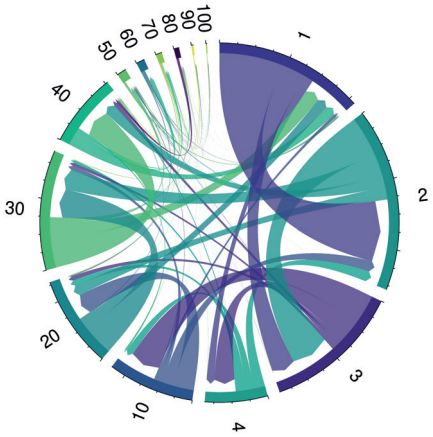
FIGURE E.20: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1.



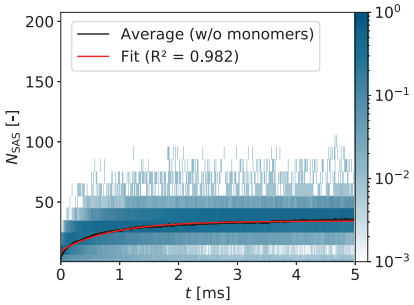
(B) Visualization of before (left) and after (right) self-assembly using back-bone carbon structures. Color indicates enzyme type: E1, E3, E2, E3BP.



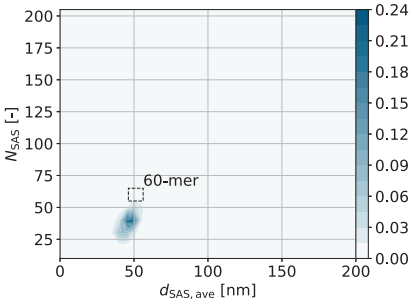
(C) Assembly pathway by bi-directional transitions between size classes.



(D) Assembly pathway by net transitions between size classes.

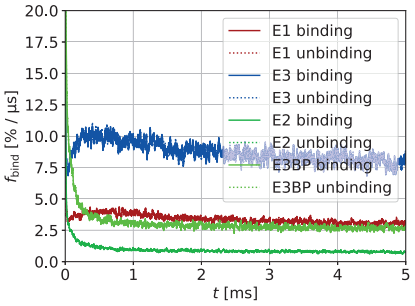


(E) Histogram of self-assembled structures by numbered size over time.

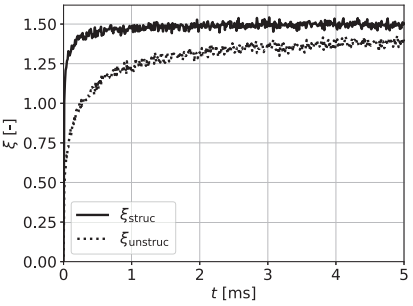


(F) Numbered size versus average extent without monomers (final time).

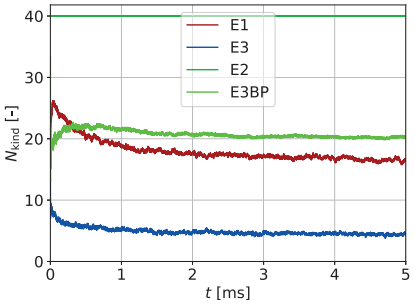
FIGURE E.20: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1.



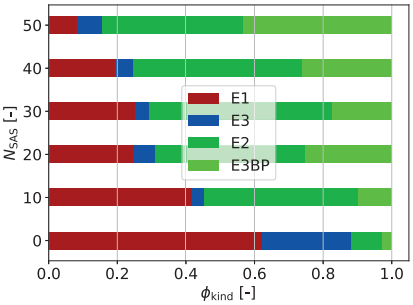
(G) (Un-)binding rates per enzyme type with uniform smoothing over 10 µs.



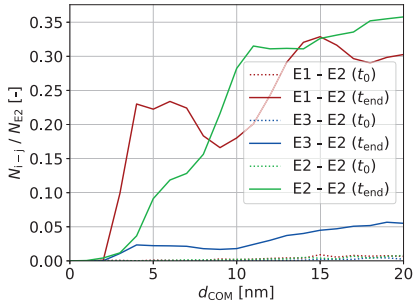
(H) Average (un-)structured contacts per enzyme (10 µs saving).



(I) Stoichiometry of struc. ($N_{SAS} \geq 5$) relative to E2 content in 60-mer.

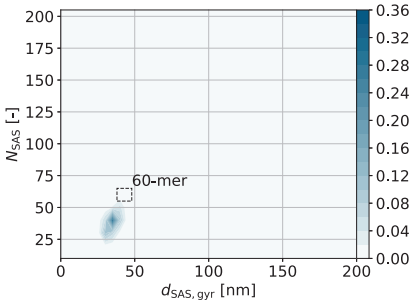


(J) Stoichiometry of structures by molar fraction for each N_{SAS} (final time).

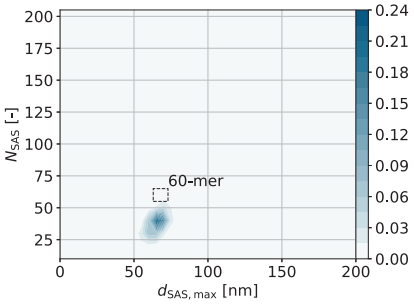


(K) Reactant distance distribution at beginning and end of self-assembly.

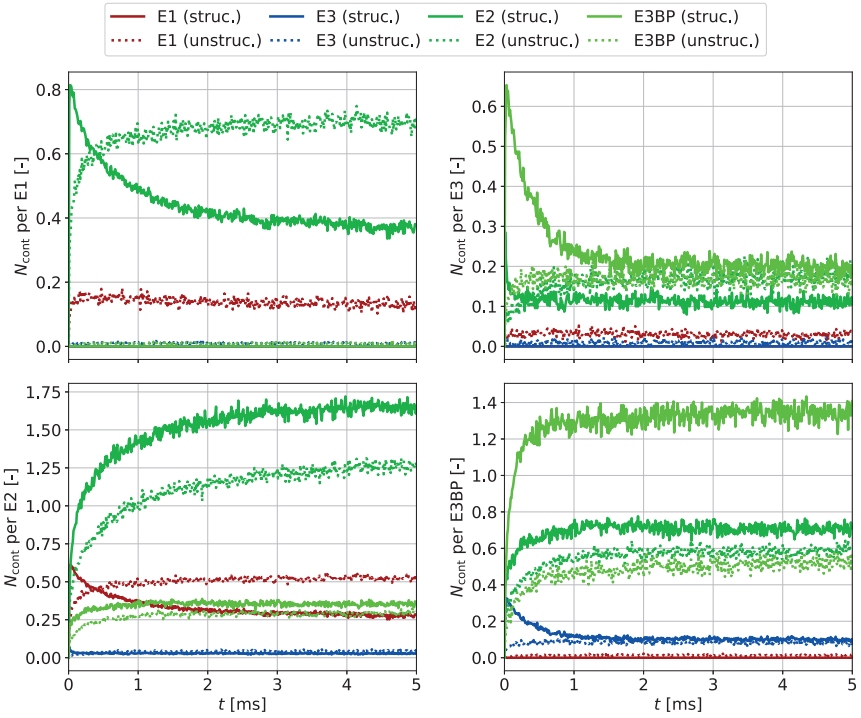
FIGURE E.20: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1.



(A) Numbered size versus diameter of gyration without monomers (final time).

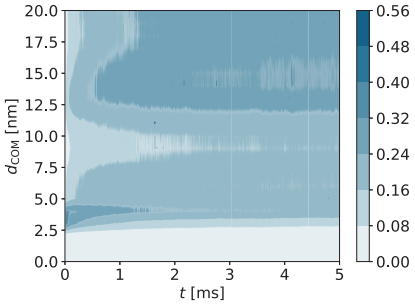


(B) Numbered size versus maximum extent without monomers (final time).

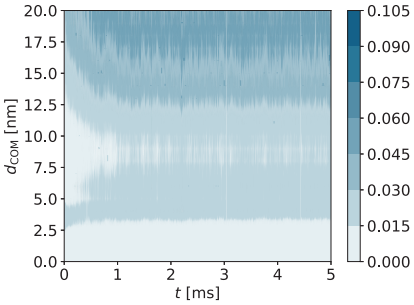


(C) Contacts with each enzyme type over time (10 µs saving).

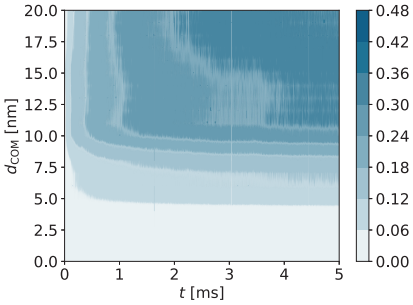
FIGURE E.21: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1.



(D) Reactant distance distribution for E1 - E2.



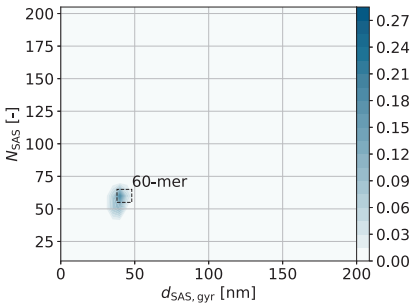
(E) Reactant distance distribution for E3 - E2.



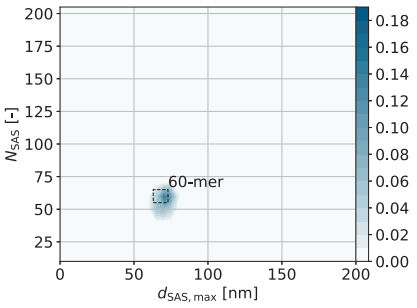
(F) Reactant distance distribution for E2 - E2.

FIGURE E.21: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1.

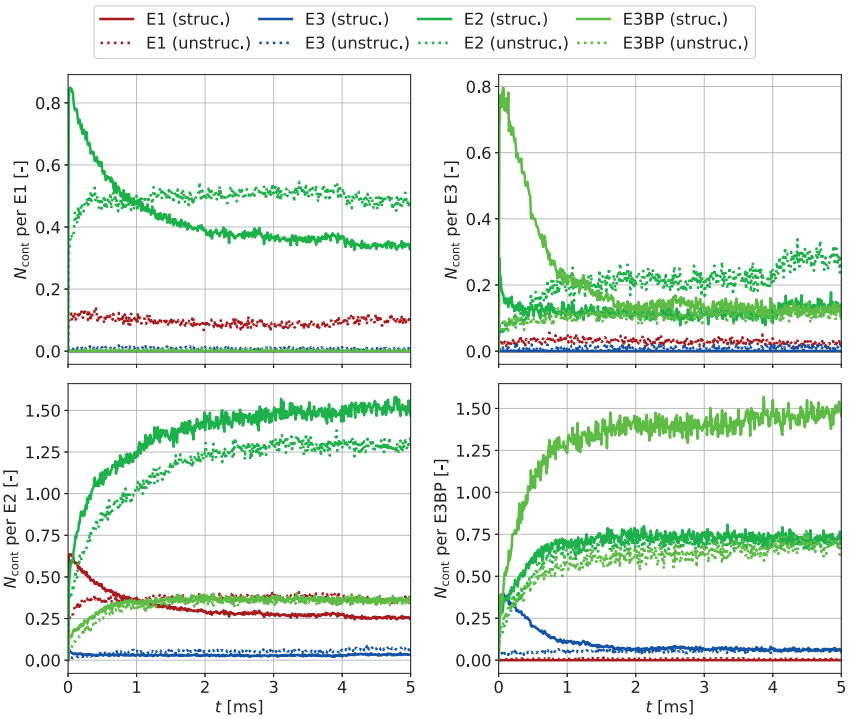
SP-PDC-1-AN (With Annealing)



(A) Numbered size versus diameter of gyration without monomers (final time).

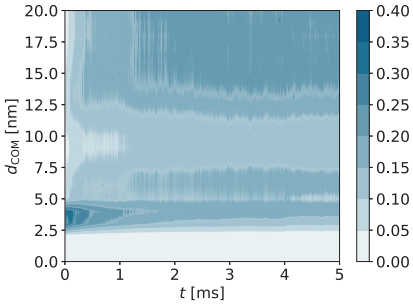


(B) Numbered size versus maximum extent without monomers (final time).

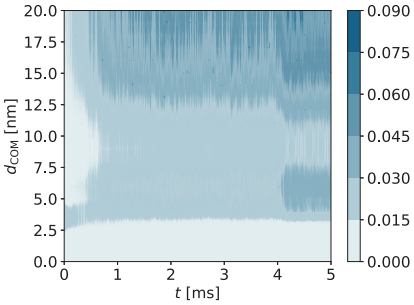


(C) Contacts with each enzyme type over time (10 µs saving).

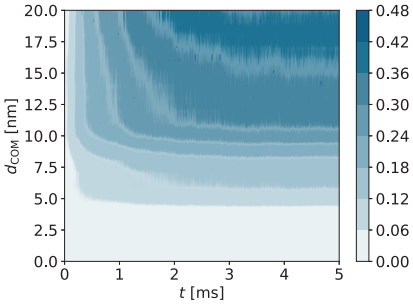
FIGURE E.22: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1-AN.



(D) Reactant distance distribution for E1 - E2.



(E) Reactant distance distribution for E3 - E2.



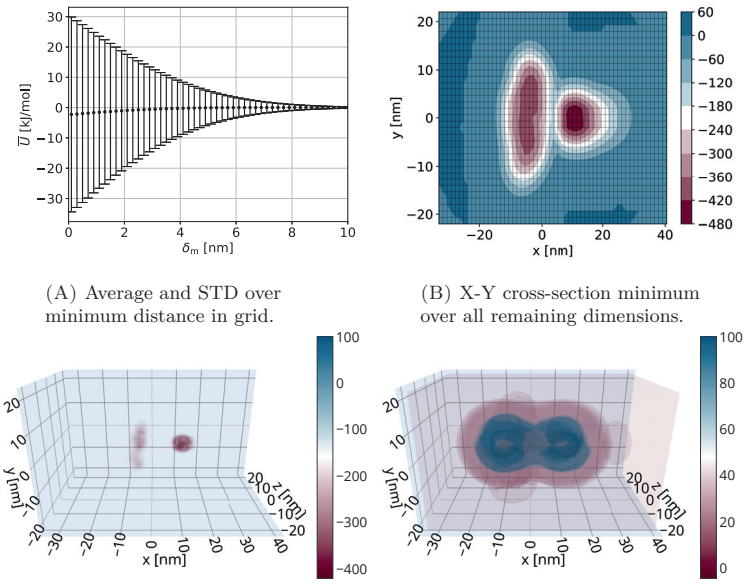
(F) Reactant distance distribution for E2 - E2.

FIGURE E.22: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1-AN.

E.5.3 Enhanced E2 – E2 Arm Interaction

In the context of PDC the precise interaction of two E2 linker arms (catalytic domains on opposite sites) and possibly a subsequent structural assembly beyond the 60-mer is an additional topic of interest. Guo *et al.* [269] have observed in addition to a formation of the 60-mer structure (radius of hydration $r_h = 26.1$ nm) the presence of a fraction of larger size around $r_h = 75.2$ nm (using DLS on wild-type human E2/E3BP systems). However, while activity increases have been observed for a larger population fraction from C-terminal truncations of E2 and E3BP (i.e. reduced catalytic domain binding for a 60-mer leading to more unstructured agglomerates; activity 120.1 % relative to wild-type) [269], it is unknown whether wild-type human PDC (as is investigated in this work) also exhibits an increase in activity for larger structures, which might possibly form beyond the predominantly observed 60-mer. Note that large agglomerates might also be attributed to (partial) unfolding of proteins [406].

In this direction, umbrella sampling of the E2 – E2 arm interaction using the CG-MD model [268] has been performed by Uwe Jandt showing an attractive behavior towards an equilibrium distance between E2 – E2 centers of mass at approximately 10 – 20 nm [407]. Inspired by these results and in an attempt to better understand structural formation



(A) Average and STD over minimum distance in grid.

(B) X-Y cross-section minimum over all remaining dimensions.

(C) 3D min. over orientations.

(D) 3D mean over orientations.

FIGURE E.23: Interaction potential field of **E2 with E2 from MD with additional arm interaction** (units of kJ/mol). C and D adapted with permission from Springer Nature regarding Depta *et al.* [227].

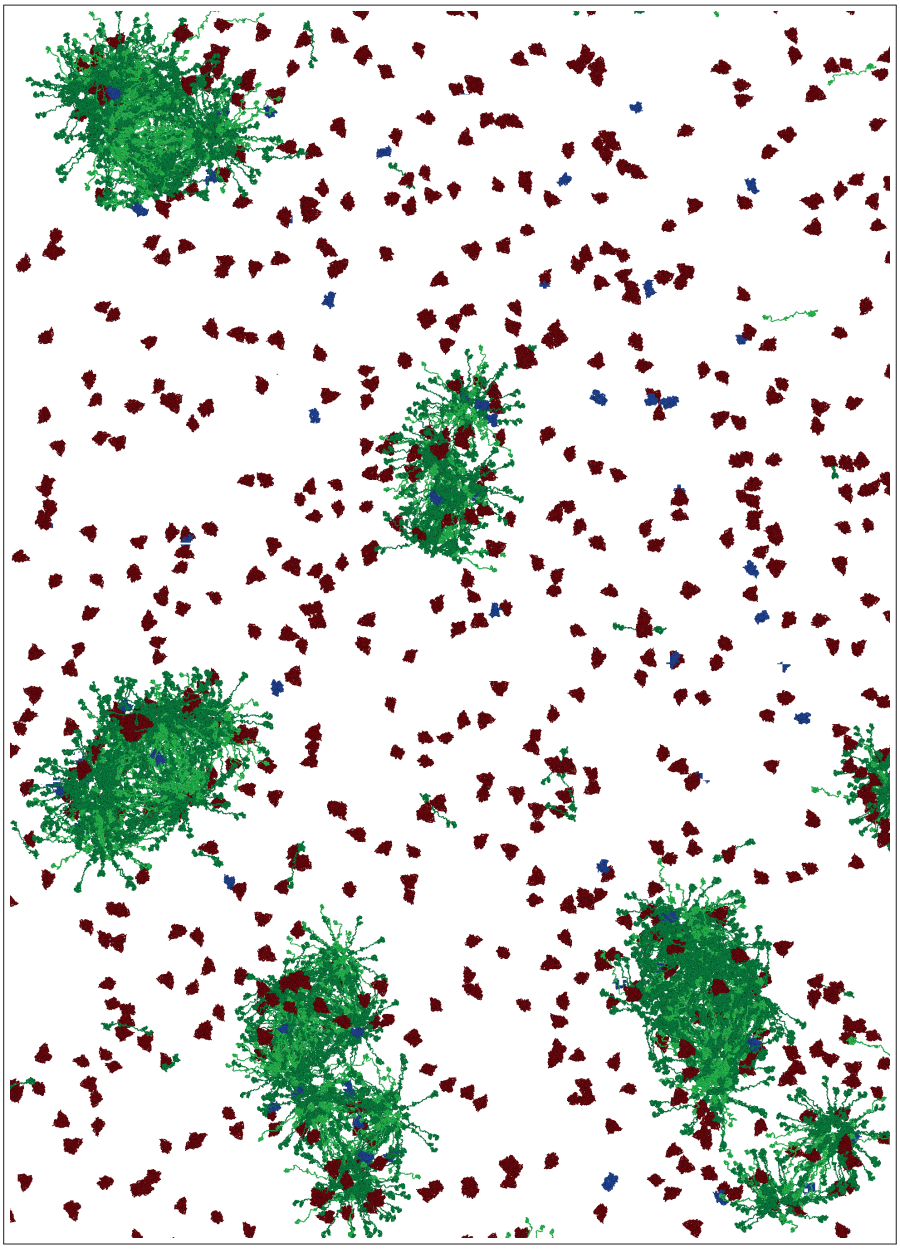
through increased E2 linker arm interaction, a variant of the E2 – E2 interaction potential was derived as shown in Fig. E.23. For this, empirical data points were inserted along the x-axis with arms oriented in opposing directions ($y = z = 0$ nm, $\alpha = -\pi$ to π in $\pi/36$ increments, $\beta = 0$, $\gamma = \pi$) and following a Morse potential as

$$U(x) = -1400 \text{ kJ/mol} + 1344 \text{ kJ/mol} \times (1 - e^{\frac{10 \text{ nm} - x}{30 \text{ nm}/3.676}})^2 \tag{E.1}$$

between $x = 5$ nm and $x = 40$ nm (grid extended to $x = 40.5$ nm accordingly). These data points were then replicated in y and z by ± 1.5 nm and ± 3.0 nm to account for the flexibility of the linker arm in an approximate nature. Fig. E.23 shows the resulting interaction potential with enhanced E2 – E2 linker arm interaction located at $x \approx 10 - 20$ nm. Variations in potential width and depth have been investigated additionally indicating the used values are necessary for meaningful structural assembly beyond the 60-mer as will be discussed subsequently.

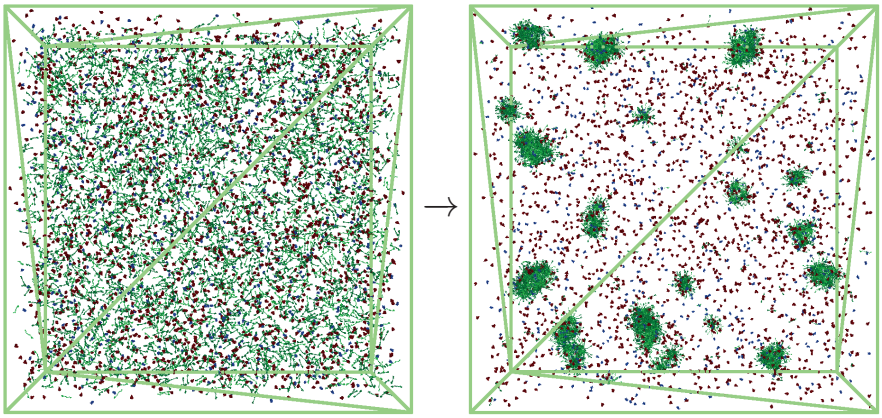
For this variation of the E2 – E2 interaction potential with enhanced arm interaction, the full component PDC system was simulated equivalent to the normal E2 – E2 interaction shown in Sec. 8.4.2.2. The self-assembly of this adaptation is visualized qualitatively and quantitatively in Fig. E.24 and E.25. As it can be seen, large-scale agglomerates form with substructures being similar to the 60-mer core of PDC [265, 388]. These agglomerates are composed of approximately 100 - 400 enzymes with average diameters around 100 nm (Fig. E.24 F) and maximum diameters between 100 - 150 nm (Fig. E.25 B). Consequently, sizes of formed agglomerates are comparable to the population of 150.4 nm diameter observed by Guo *et al.* [269]. However, when investigating the stoichiometry of structures (Fig. E.24 I and J) it has to be noted that much fewer E1 and E3 bind to these structures - particularly the large structures. Visually this can also be observed in Fig. E.24 A with few E1 and E3 located at the center of agglomerates. As a result, the reactant / active-site distribution (Fig. E.24 K) decreases drastically, e.g. for E1 – E2 around $d_{\text{COM}} = 5$ nm by a factor of 2.7 from 0.16 for the normal linker arm interaction to 0.06 for the enhanced linker arm interaction.

Consequently, based on this model of an enhanced E2 – E2 linker arm interaction it is doubtful whether such large agglomerates can maintain the same catalytic activity as the 60-mer core. Similarly, it is questionable whether reactants can be transported into and out of the agglomerate at sufficient rates for catalytic participation of the agglomerate center. Hence, it is considered more plausible that no further agglomeration beyond the predominantly reported 60-mer takes place – at least near physiological conditions enabling catalytic activity.

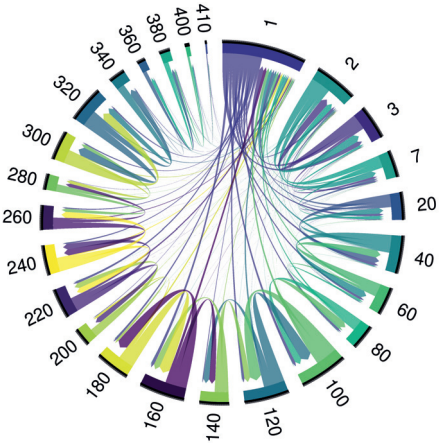


(A) Close-up visualization of self-assembled structures by enzyme type: E1, E3, E2, E3BP.

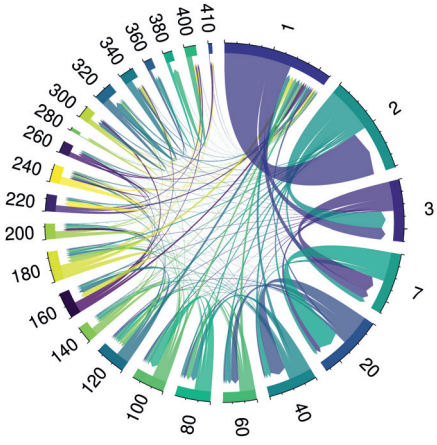
FIGURE E.24: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1 with enhanced E2 – E2 arm interaction.



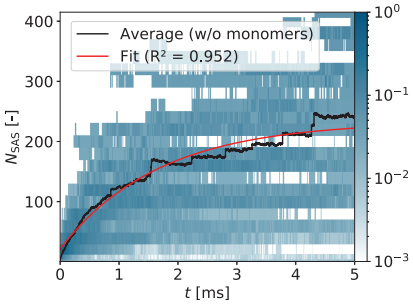
(B) Visualization of before (left) and after (right) self-assembly using back-bone carbon structures. Color indicates enzyme type: E1, E3, E2, E3BP.



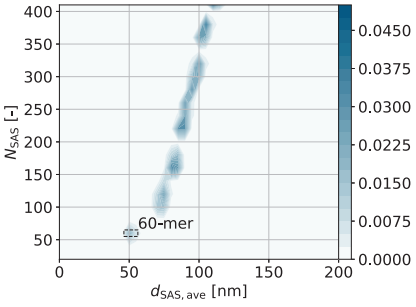
(C) Assembly pathway by bi-directional transitions between size classes.



(D) Assembly pathway by net transitions between size classes.

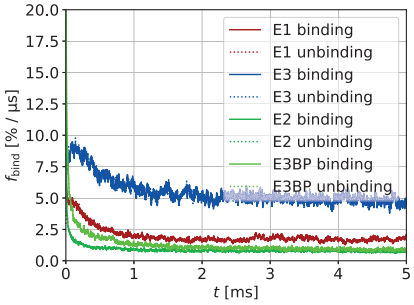


(E) Histogram of self-assembled structures by numbered size over time.

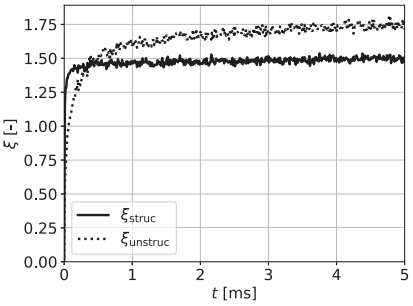


(F) Numbered size versus average extent without monomers (final time).

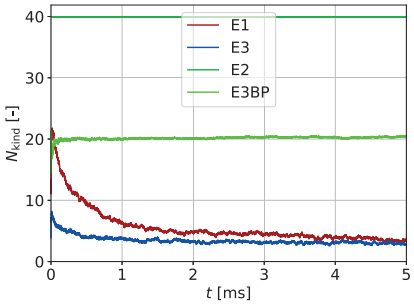
FIGURE E.24: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1 with enhanced E2 – E2 arm interaction.



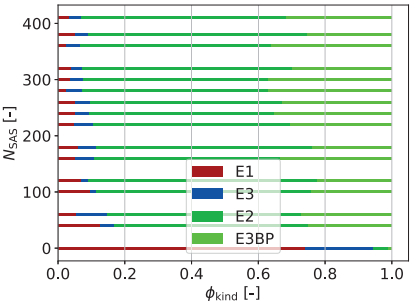
(G) (Un-)binding rates per enzyme type with uniform smoothing over 10 µs.



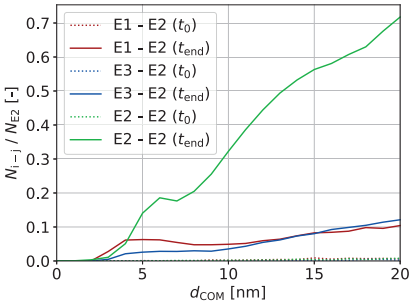
(H) Average (un-)structured contacts per enzyme (10 µs saving).



(I) Stoichiometry of struc. ($N_{SAS} \geq 5$) relative to E2 content in 60-mer.

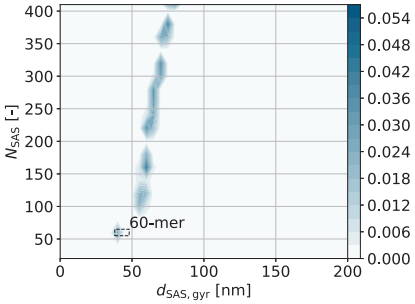


(J) Stoichiometry of structures by molar fraction for each N_{SAS} (final time).

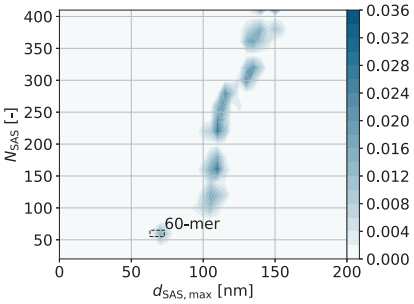


(K) Reactant distance distribution at beginning and end of self-assembly.

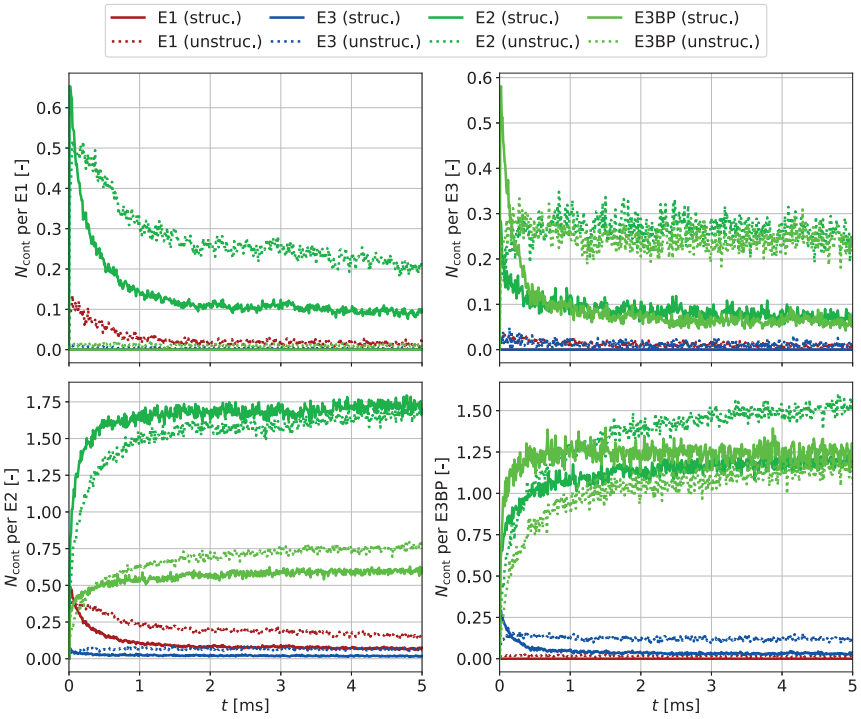
FIGURE E.24: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1 with enhanced E2 – E2 arm interaction.



(A) Numbered size versus diameter of gyration without monomers (final time).

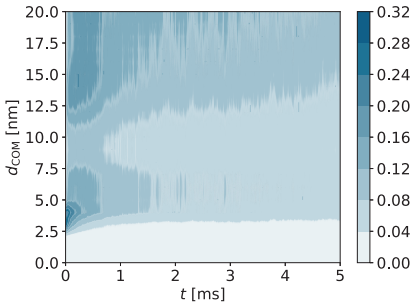


(B) Numbered size versus maximum extent without monomers (final time).

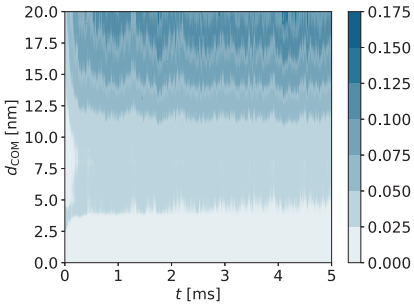


(C) Contacts with each enzyme type over time (10 µs saving).

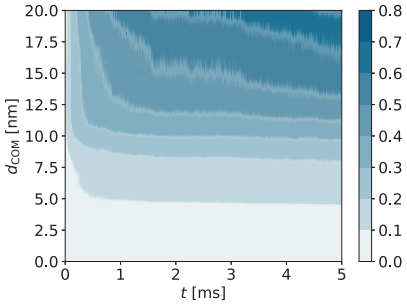
FIGURE E.25: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box (1 µm edge) using simulation protocol SP-PDC-1 with enhanced E2 – E2 arm interaction.



(D) Reactant distance distribution for E1 - E2.



(E) Reactant distance distribution for E3 - E2.



(F) Reactant distance distribution for E2 - E2.

FIGURE E.25: PDC self-assembly for a stoichiometry of $40 \times \text{E2} + 20 \times \text{E3BP} + 30 \times \text{E1} + 10 \times \text{E3}$ at concentration of $1\,\text{mg}\,\text{mL}^{-1}$ in $1\,\mu\text{m}^3$ cubic box ($1\,\mu\text{m}$ edge) using simulation protocol SP-PDC-1 with enhanced E2 – E2 arm interaction.

Bibliography

[1] D. S. Bethune, C. H. Kiang, M. S. de Vries, G. Gorman, R. Savoy, J. Vazquez, and R. Beyers. Cobalt-Catalysed Growth of Carbon Nanotubes with Single-Atomic-Layer Walls. *Nature*, 363(6430):605–607, 1993.

[2] S. Iijima and T. Ichihashi. Single-Shell Carbon Nanotubes of 1 Nanometer Diameter. *Nature*, 363(6430):603–605, 1993.

[3] I. V. Kozhevnikov. Catalysis by Heteropoly Acids and Multicomponent Polyoxometalates in Liquid-Phase Reactions. *Chem. Rev.*, 98(1):171–198, 1998.

[4] M. Misono. Catalytic Chemistry of Solid Polyoxometalates and Their Industrial Applications. *Mol. Eng.*, 3(1-3):193–203, 1993.

[5] B. Alberts. *Essential Cell Biology*. Garland Science, New York, NY, fourth edition, 2013.

[6] J. Moult, J. T. Pedersen, R. Judson, and K. Fidelis. A Large-Scale Experiment to Assess Protein Structure Prediction Methods. *Proteins: Struct., Funct., Gen.*, 23(3):ii–iv, 1995.

[7] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[8] V. Gold (editor). *The IUPAC Compendium of Chemical Terminology: The Gold Book*. International Union of Pure and Applied Chemistry (IUPAC), Research Triangle Park, NC, fourth edition, 2019.

[9] V. I. Minkin. Glossary of Terms Used in Theoretical Organic Chemistry. *Pure Appl. Chem.*, 71(10):1919–1981, 1999.

[10] M. Bustamante-Torres, D. Romero-Fierro, B. Arcentales-Vera, K. Palomino, H. Magaña, and E. Bucio. Hydrogels Classification According to the Physical or Chemical Interactions and as Stimuli-Sensitive Materials. *Gels*, 7(4):182, 2021.

[11] E. V. Grgacic and D. A. Anderson. Virus-Like Particles: Passport to Immune Recognition. *Methods*, 40(1):60–65, 2006.

[12] B. Böttcher, S. A. Wynne, and R. A. Crowther. Determination of the Fold of the Core Protein of Hepatitis B Virus by Electron Cryomicroscopy. *Nature*, 386(6620):88–91, 1997.

[13] C. San Martín. Latest Insights on Adenovirus Structure and Assembly. *Viruses*, 4(5):847–877, 2012.

[14] B. W. Neuman, G. Kiss, A. H. Kunding, D. Bhella, M. F. Baksh, S. Connelly, B. Droese, J. P. Klaus, S. Makino, S. G. Sawicki, S. G. Siddell, D. G. Stamou, I. A. Wilson, P. Kuhn, and M. J. Buchmeier. A Structural Analysis of M Protein in Coronavirus Assembly and Morphology. *J. Struct. Biol.*, 174(1):11–22, 2011.

[15] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science New York, 2014.

[16] G. Rickey Welch and J. S. Easterby. Metabolic Channeling Versus Free Diffusion: Transition-Time Analysis. *Trends Biochem. Sci.*, 19(5):193–197, 1994.

[17] M. Castellana, M. Z. Wilson, Y. Xu, P. Joshi, I. M. Cristea, J. D. Rabinowitz, Z. Gitai, and N. S. Wingreen. Enzyme Clustering Accelerates Processing of Intermediates Through Metabolic Channeling. *Nat. Biotechnol.*, 32(10):1011–1018, 2014.

[18] M. S. Patel, N. S. Nemeria, W. Furey, and F. Jordan. The Pyruvate Dehydrogenase Complexes: Structure-based Function and Regulation. *J. Biol. Chem.*, 289(24):16615–16623, 2014.

[19] A. W. Alberts, A. W. Strauss, S. Hennessy, and P. R. Vagelos. Regulation of Synthesis of Hepatic Fatty Acid Synthetase: Binding of Fatty Acid Synthetase Antibodies to Polysomes. *Proc. Natl. Acad. Sci. U.S.A.*, 72(10):3956–3960, 1975.

[20] D. Eisenberg, H. S. Gill, G. M. Pfluegl, and S. H. Rotstein. Structure–Function Relationships of Glutamine Synthetases. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.*, 1477(1-2):122–145, 2000.

[21] U. Jandt, C. You, Y. H.-P. Zhang, and A.-P. Zeng. Compartmentalization and Metabolic Channeling for Multienzymatic Biosynthesis: Practical Strategies and

Modeling Approaches. In A.-P. Zeng (editor), *Fundamentals and Application of New Bioproduction Systems*, volume 137, pages 41–65. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[22] E. Ricca, B. Brucher, and J. H. Schrittwieser. Multi-Enzymatic Cascade Reactions: Overview and Perspectives. *Adv. Synth. Catal.*, 353(13):2239–2262, 2011.

[23] M. B. Quin, K. K. Wallin, G. Zhang, and C. Schmidt-Dannert. Spatial Organization of Multi-Enzyme Biocatalytic Cascades. *Org. Biomol. Chem.*, 15(20):4260–4271, 2017.

[24] S. Zhao, W. J. Malfait, N. Guerrero-Alburquerque, M. M. Koebel, and G. Nyström. Biopolymer Aerogels and Foams: Chemistry, Properties, and Applications. *Angew. Chem. Int. Ed.*, 57(26):7580–7608, 2018.

[25] G. Skjåk-Bræk, H. Grasdalen, and O. Smidsrød. Inhomogeneous Polysaccharide Ionic Gels. *Carbohydr. Polym.*, 10(1):31–54, 1989.

[26] J. Zhu. Bioactive Modification of Poly(Ethylene Glycol) Hydrogels for Tissue Engineering. *Biomaterials*, 31(17):4639–4656, 2010.

[27] J. C. Love, L. A. Estroff, J. K. Kriebel, R. G. Nuzzo, and G. M. Whitesides. Self-Assembled Monolayers of Thiolates on Metals as a Form of Nanotechnology. *Chem. Rev.*, 105(4):1103–1170, 2005.

[28] S. Anderson, H. L. Anderson, A. Bashall, M. McPartlin, and J. K. M. Sanders. Assembly and Crystal Structure of a Photoactive Array of Five Porphyrins. *Angew. Chem. Int. Ed. Engl.*, 34(10):1096–1099, 1995.

[29] W. A. Freeman. Structures of the *p* -Xylylenediammonium Chloride and Calcium Hydrogensulfate Adducts of the Cavitand 'Cucurbituril', $C_{36}H_{36}N_{24}O_{12}$. *Acta Crystallogr., Sect. B: Struct. Sci.*, 40(4):382–387, 1984.

[30] H. J. Berendsen. *Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*. Cambridge University Press, 2007.

[31] B. Engquist, P. Lötstedt, O. Runborg, T. J. Barth, M. Griebel, D. E. Keyes, R. M. Nieminen, D. Roose, and T. Schlick (editors). *Multiscale Modeling and Simulation in Science*, volume 66 of *Lecture Notes in Computational Science and Engineering*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[32] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Number 1 in Computational Science Series. Academic Press, San Diego, second edition, 2002.

[33] T. Pang. *An Introduction to Computational Physics*. Cambridge Univ. Press, Cambridge, second edition, 2008.

[34] D. C. Rapaport. *The Art of Molecular Dynamics Simulation*. Cambridge University Press, Cambridge, UK ; New York, NY, second edition, 2004.

[35] T. Schlick. Molecular Modeling and Simulation: An Interdisciplinary Guide, 2010.

[36] L. Verlet. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.*, 159(1):98–103, 1967.

[37] R. Hockney and J. Eastwood. *Computer Simulation Using Particles*. CRC Press, 2021.

[38] J. C. Jo and B. C. Kim. Determination of Proper Time Step for Molecular Dynamics Simulation. *Bull. Korean Chem. Soc.*, 21(4):419–424, 2000.

[39] G. King and A. Warshel. A Surface Constrained All-Atom Solvent Model for Effective Simulations of Polar Solutions. *J. Chem. Phys.*, 91(6):3647–3661, 1989.

[40] T. A. Wassenaar and A. E. Mark. The Effect of Box Shape on the Dynamic Properties of Proteins Simulated Under Periodic Boundary Conditions. *J. Comput. Chem.*, 27(3):316–325, 2006.

[41] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. GROMACS: High Performance Molecular Simulations Through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX*, 1–2:19–25, 2015.

[42] E. Lindahl, M. J. Abraham, B. Hess, and D. Van Der Spoel. GROMACS 2020.1 Manual. *Zendo*, 2020.

[43] J. Huang and A. D. MacKerell. Force Field Development and Simulations of Intrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.*, 48:40–48, 2018.

[44] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.*, 106(3):765–784, 1984.

[45] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.*, 4(2):187–217, 1983.

[46] W. F. van Gunsteren and H. J. C. Berendsen. Groningen Molecular Simulation (GROMOS) Library Manual. *Biomos, Groningen, Netherlands*, pages 1–221, 1987.

[47] W. L. Jorgensen and J. Tirado-Rives. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.*, 110(6):1657–1666, 1988.

[48] L. Yang, C.-H. Tan, M.-J. Hsieh, J. Wang, Y. Duan, P. Cieplak, J. Caldwell, P. A. Kollman, and R. Luo. New-Generation Amber United-Atom Force Field. *J. Phys. Chem. B*, 110(26):13166–13176, 2006.

[49] S. Patel and C. L. Brooks. CHARMM Fluctuating Charge Force Field for Proteins: I Parameterization and Application to Bulk Organic Liquid Simulations. *J. Comput. Chem.*, 25(1):1–16, 2004.

[50] S. Patel, A. D. Mackerell, and C. L. Brooks. CHARMM Fluctuating Charge Force Field for Proteins: II Protein/Solvent Properties from Molecular Dynamics Simulations Using a Nonadditive Electrostatic Model. *J. Comput. Chem.*, 25(12):1504–1514, 2004.

[51] A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard. ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A*, 105(41):9396–9409, 2001.

[52] J. W. Ponder and D. A. Case. Force Fields for Protein Simulations. In *Advances in Protein Chemistry*, volume 66, pages 27–85. Elsevier, 2003.

[53] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. De Vries. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B*, 111(27):7812–7824, 2007.

[54] T. Darden, D. York, and L. Pedersen. Particle Mesh Ewald: An N Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[55] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.*, 103(19):8577–8593, 1995.

[56] J. Behler and M. Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98(14):146401, 2007.

[57] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci. Adv.*, 3(5):e1603015, 2017.

[58] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller. Machine Learning Force Fields. *Chem. Rev.*, 121(16):10142–10186, 2021.

[59] R. Car and M. Parrinello. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Phys. Rev. Lett.*, 55(22):2471–2474, 1985.

[60] W. Yang. Direct Calculation of Electron Density in Density-Functional Theory. *Phys. Rev. Lett.*, 66(11):1438–1441, 1991.

[61] K. Laasonen and R. M. Nieminen. Molecular Dynamics Using the Tight-Binding Approximation. *J. Phys.: Condens. Matter*, 2(6):1509–1520, 1990.

[62] J. Gelpi, A. Hospital, R. Goñi, and M. Orozco. Molecular Dynamics Simulations: Advances and Applications. *Adv. Appl. Bioinf. Chem.*, page 37, 2015.

[63] D. Marx and J. Hutter. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*. Cambridge University Press, Cambridge ; New York, 2009.

[64] A. Warshel and M. Levitt. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.*, 103(2):227–249, 1976.

[65] J. Gao and M. A. Thompson (editors). *Combined Quantum Mechanical and Molecular Mechanical Methods*. Number 712 in ACS Symposium Series. American Chemical Society, Washington, DC, 1998.

[66] H. M. Senn and W. Thiel. QM/MM Methods for Biomolecular Systems. *Angew. Chem. Int. Ed.*, 48(7):1198–1229, 2009.

[67] H. C. Andersen. Molecular Dynamics Simulations at Constant Pressure and/or Temperature. *J. Chem. Phys.*, 72(4):2384–2393, 1980.

[68] D. J. Evans. Computer Experiment for Nonlinear Thermodynamics of Couette Flow. *J. Chem. Phys.*, 78(6):3297–3302, 1983.

[69] W. G. Hoover. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A*, 31(3):1695–1697, 1985.

[70] D. J. Evans and G. Morriss. The Isothermal/Isobaric Molecular Dynamics Ensemble. *Phys. Lett. A*, 98(8-9):433–436, 1983.

[71] D. J. Evans and G. Morriss. Isothermal-Isobaric Molecular Dynamics. *Chem. Phys.*, 77(1):63–66, 1983.

[72] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[73] S. Nosé. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Molec. Phys.*, 52(2):255–268, 1984.

[74] S. Nosé. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *J. Chem. Phys.*, 81(1):511–519, 1984.

[75] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[76] Y. Sugita and Y. Okamoto. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.*, 314(1-2):141–151, 1999.

[77] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. New Approach to Monte Carlo Calculation of the Free Energy: Method of Expanded Ensembles. *J. Chem. Phys.*, 96(3):1776–1783, 1992.

[78] H. Nada and J. P. J. M. van der Eerden. An Intermolecular Potential Model for the Simulation of Ice and Water Near the Melting Point: A Six-Site Model of H2O. *J. Chem. Phys.*, 118(16):7401, 2003.

[79] H. Nada. Anisotropy in Geometrically Rough Structure of Ice Prismatic Plane Interface During Growth: Development of a Modified Six-Site Model of H2O and a Molecular Dynamics Simulation. *J. Chem. Phys.*, 145(24):244706, 2016.

[80] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. Interaction Models for Water in Relation to Protein Hydration. In B. Pullman (editor), *Intermolecular Forces*, volume 14, pages 331–342. Springer Netherlands, Dordrecht, 1981.

[81] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.*, 91(24):6269–6271, 1987.

[82] W. L. Jorgensen. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water. *J. Am. Chem. Soc.*, 103(2):335–340, 1981.

[83] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[84] H. Saint-Martin, J. Hernández-Cobos, M. I. Bernal-Uruchurtu, I. Ortega-Blake, and H. J. Berendsen. A Mobile Charge Densities in Harmonic Oscillators (MCDHO) Molecular Model for Numerical Simulations: The Water–Water Interaction. *J. Chem. Phys.*, 113(24):10899–10912, 2000.

[85] B. Guillot. A Reappraisal of What We Have Learnt During Three Decades of Computer Simulations on Water. *J. Mol. Liq.*, 101(1-3):219–260, 2002.

[86] E. C. Allen and G. C. Rutledge. A Novel Algorithm for Creating Coarse-Grained, Density Dependent Implicit Solvent Models. *J. Chem. Phys.*, 128(15):154115, 2008.

[87] G. Brannigan and F. L. Brown. Solvent-Free Simulations of Fluid Membrane Bilayers. *J. Chem. Phys.*, 120(2):1059–1071, 2004.

[88] C.-E. A. Chang, J. Trylska, V. Tozzini, and J. Andrew McCammon. Binding Pathways of Ligands to HIV-1 Protease: Coarse-Grained and Atomistic Simulations. *Chem. Biol. Drug Des.*, 69(1):5–13, 2007.

[89] I. R. Cooke and M. Deserno. Solvent-Free Model for Self-Assembling Fluid Bilayer Membranes: Stabilization of the Fluid Phase Based on Broad Attractive Tail Potentials. *J. Chem. Phys.*, 123(22):224710, 2005.

[90] J. L. Knight and C. L. Brooks. Surveying Implicit Solvent Models for Estimating Small Molecule Absolute Hydration Free Energies. *J. Comput. Chem.*, 32(13):2909–2923, 2011.

[91] T. A. Knotts IV, N. Rathore, D. C. Schwartz, and J. J. de Pablo. A Coarse Grain Model for DNA. *J. Chem. Phys.*, 126(8):084901, 2007.

[92] J. Maupetit, P. Derreumaux, and P. Tuffery. PEP-FOLD: An Online Resource for De Novo Peptide Structure Prediction. *Nucleic Acids Res.*, 37(suppl_2):498–503, 2009.

[93] M. S. Shell, R. Ritterson, and K. A. Dill. A Test on Peptide Stability of AMBER Force Fields with Implicit Solvation. *J. Phys. Chem. B*, 112(22):6878–6886, 2008.

[94] A. Villa, C. Peter, and N. F. A. van der Vegt. Self-Assembling Dipeptides: Conformational Sampling in Solvent-Free Coarse-Grained Simulation. *Phys. Chem. Chem. Phys.*, 11(12):2077–2086, 2009.

[95] L. Wesson and D. Eisenberg. Atomic Solvation Parameters Applied to Molecular Dynamics of Proteins in Solution. *Protein Sci.*, 1(2):227–235, 1992.

[96] D. Bashford and D. A. Case. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.*, 51(1):129–152, 2000.

[97] K. A. Sharp and B. Honig. Calculating Total Electrostatic Energies with the Nonlinear Poisson-Boltzmann Equation. *J. Phys. Chem.*, 94(19):7684–7692, 1990.

[98] J. Chen, C. L. Brooks, and J. Khandogin. Recent Advances in Implicit Solvent-Based Methods for Biomolecular Simulations. *Curr. Opin. Struct. Biol.*, 18(2):140–148, 2008.

[99] S. Decherchi, M. Masetti, I. Vyalov, and W. Rocchia. Implicit Solvent Methods for Free Energy Estimation. *Eur. J. Med. Chem.*, 91:27–42, 2015.

[100] J. Kleinjung and F. Fraternali. Design and Application of Implicit Solvent Models in Biomolecular Simulations. *Curr. Opin. Struct. Biol.*, 25:126–134, 2014.

[101] A. Onufriev. Implicit Solvent Models in Molecular Dynamics Simulations: A Brief Overview. In *Annual Reports in Computational Chemistry*, volume 4, pages 125–137. Elsevier, 2008.

[102] A. V. Onufriev and S. Izadi. Water Models for Biomolecular Simulations. *WIREs Comput. Mol. Sci.*, 8(2), 2018.

[103] F. F. Abraham, R. Walkup, H. Gao, M. Duchaineau, T. Diaz De La Rubia, and M. Seager. Simulating Materials Failure by Using up to One Billion Atoms and the World's Fastest Computer: Work-Hardening. *Proc. Natl. Acad. Sci. U.S.A.*, 99(9):5783–5787, 2002.

[104] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How Fast-Folding Proteins Fold. *Science*, 334(6055):517–520, 2011.

[105] I. Buch, T. Giorgino, and G. De Fabritiis. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U.S.A.*, 108(25):10184–10189, 2011.

[106] S. Hezaveh, A.-P. Zeng, and U. Jandt. Human Pyruvate Dehydrogenase Complex E2 and E3BP Core Subunits: New Models and Insights from Molecular Dynamics Simulations. *J. Phys. Chem. B*, 120(19):4399–4409, 2016.

[107] E. Brini, E. A. Algaer, P. Ganguly, C. Li, F. Rodríguez-Ropero, and N. F. A. van der Vegt. Systematic Coarse-Graining Methods for Soft Matter Simulations – A Review. *Soft Matter*, 9(7):2108–2119, 2013.

[108] S. Y. Joshi and S. A. Deshmukh. A Review of Advancements in Coarse-Grained Molecular Dynamics Simulations. *Mol. Simul.*, 47(10-11):786–803, 2021.

[109] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.*, 116(14):7898–7936, 2016.

[110] A. Liwo, C. Czaplewski, A. K. Sieradzan, A. G. Lipska, S. A. Samsonov, and R. K. Murarka. Theory and Practice of Coarse-Grained Molecular Dynamics of Biologically Important Systems. *Biomolecules*, 11(9):1347, 2021.

[111] N. Singh and W. Li. Recent Advances in Coarse-Grained Models for Biomolecules and Their Applications. *Int. J. Mol. Sci.*, 20(15):3774, 2019.

[112] H. Mori. Transport, Collective Motion, and Brownian Motion. *Prog. Theor. Phys.*, 33(3):423–455, 1965.

[113] R. Zwanzig. Memory Effects in Irreversible Thermodynamics. *Phys. Rev.*, 124(4):983–992, 1961.

[114] R. Zwanzig. *Nonequilibrium Statistical Mechanics*. Oxford University Press, 2001.

[115] P. Langevin. Sur La Théorie Du Mouvement Brownien. *C. R. Acad. Sci.*, 146(530-533):530, 1908.

[116] R. Zwanzig. Nonlinear Generalized Langevin Equations. *J. Stat. Phys.*, 9(3):215–220, 1973.

[117] G. Jung, M. Hanke, and F. Schmid. Generalized Langevin Dynamics: Construction and Numerical Integration of Non-Markovian Particle-Based Models. *Soft Matter*, 14(46):9368–9382, 2018.

[118] V. Balakrishnan. Fluctuation-Dissipation Theorems from the Generalised Langevin Equation. *Pramana*, 12(4):301–315, 1979.

[119] A. Einstein. Zur Theorie Der Brownschen Bewegung. *Ann. Phys.*, 324(2):371–381, 1906.

[120] P. Debye. *Polar Molecules*. Chem. Cat. Comp., Inc., 1929.

[121] F. Perrin. Mouvement Brownien D'un Ellipsoide (I). Dispersion Diélectrique Pour Des Molécules Ellipsoidales. *J. Phys. Radium*, 5(10):497–511, 1934.

[122] F. Perrin. Mouvement Brownien D'un Ellipsoide (II). Rotation Libre Et Dépolarisation Des Fluorescences. Translation Et Diffusion De Molécules Ellipsoidales. *J. Phys. Radium*, 7(1):1–11, 1936.

[123] C. M. Pieper and J. Enderlein. Fluorescence Correlation Spectroscopy as a Tool for Measuring the Rotational Diffusion of Macromolecules. *Chem. Phys. Lett.*, 516(1):1–11, 2011.

[124] P. E. Smith and W. F. van Gunsteren. Translational and Rotational Diffusion of Proteins. *J. Mol. Biol.*, 236(2):629–636, 1994.

[125] D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schäfer, X. Periole, D. P. Tieleman, and S. J. Marrink. Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J. Chem. Theory Comput.*, 9(1):687–697, 2013.

[126] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.*, 4(5):819–834, 2008.

[127] M. Khalili, A. Liwo, A. Jagielska, and H. A. Scheraga. Molecular Dynamics with the United-Residue Model of Polypeptide Chains. II. Langevin and Berendsen-Bath Dynamics and Tests on Model α-Helical Systems. *J. Phys. Chem. B*, 109(28):13798–13810, 2005.

[128] T. Veitshans, D. Klimov, and D. Thirumalai. Protein Folding Kinetics: Timescales, Pathways and Energy Landscapes in Terms of Sequence-Dependent Properties. *Folding Des.*, 2(1):1–22, 1997.

[129] S. T. John and G. Csányi. Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials. *J. Phys. Chem. B*, 121(48):10934–10949, 2017.

[130] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. de Fabritiis, F. Noé, and C. Clementi. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent. Sci.*, 5(5):755–767, 2019.

[131] L. Zhang, J. Han, H. Wang, R. Car, and W. E. DeePCG: Constructing Coarse-Grained Models Via Deep Neural Networks. *J. Chem. Phys.*, 149(3):034101, 2018.

[132] F. Müller-Plathe. Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back. *Chem. Phys. Chem.*, 3(9):754–769, 2002.

[133] S. Takada. Coarse-Grained Molecular Simulations of Large Biomolecules. *Curr. Opin. Struct. Biol.*, 22(2):130–137, 2012.

[134] G. Torrie and J. Valleau. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.*, 23(2):187–199, 1977.

[135] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. the Method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[136] C. Bartels. Analyzing Biased Monte Carlo and Molecular Dynamics Simulations. *Chem. Phys. Lett.*, 331(5-6):446–454, 2000.

[137] R. W. Zwanzig. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.*, 22(8):1420–1426, 1954.

[138] H. Lu, B. Isralewitz, A. Krammer, V. Vogel, and K. Schulten. Unfolding of Titin Immunoglobulin Domains by Steered Molecular Dynamics Simulation. *Biophys. J.*, 75(2):662–671, 1998.

[139] C. Jarzynski. Equilibrium Free-Energy Differences from Nonequilibrium Measurements: A Master-Equation Approach. *Phys. Rev. E*, 56(5):5018–5035, 1997.

[140] C. Jarzynski. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.*, 78(14):2690–2693, 1997.

[141] Y. Wang, W. G. Noid, P. Liu, and G. A. Voth. Effective Force Coarse-Graining. *Phys. Chem. Chem. Phys.*, 11(12):2002, 2009.

[142] E. Brini, V. Marcon, and N. F. A. van der Vegt. Conditional Reversible Work Method for Molecular Coarse Graining Applications. *Phys. Chem. Chem. Phys.*, 13(22):10468, 2011.

[143] D. Reith, M. Pütz, and F. Müller-Plathe. Deriving Effective Mesoscale Potentials from Atomistic Simulations: Mesoscale Potentials from Atomistic Simulations. *J. Comput. Chem.*, 24(13):1624–1636, 2003.

[144] A. P. Lyubartsev and A. Laaksonen. Calculation of Effective Interaction Potentials from Radial Distribution Functions: A Reverse Monte Carlo Approach. *Phys. Rev. E*, 52(4):3730–3737, 1995.

[145] M. S. Shell. The Relative Entropy Is Fundamental to Multiscale and Inverse Thermodynamic Problems. *J. Chem. Phys.*, 129(14):144108, 2008.

[146] J. W. Mullinax and W. G. Noid. A Generalized-Yvon-Born-Green Theory for Determining Coarse-Grained Interaction Potentials. *J. Phys. Chem. C*, 114(12):5661–5674, 2010.

[147] F. Ercolessi and J. B. Adams. Interatomic Potentials from First-Principles Calculations: The Force-Matching Method. *Europhys. Lett.*, 26(8):583–588, 1994.

[148] S. Izvekov and G. A. Voth. A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B*, 109(7):2469–2473, 2005.

[149] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.*, 71(1):361–390, 2020.

[150] R. Alessandri, F. Grünewald, and S. J. Marrink. The Martini Model in Materials Science. *Adv. Mater.*, 33(24):2008635, 2021.

[151] L. Darré, M. R. Machado, A. F. Brandner, H. C. González, S. Ferreira, and S. Pantano. SIRAH: A Structurally Unbiased Coarse-Grained Force Field for Proteins with Aqueous Solvation and Long-Range Electrostatics. *J. Chem. Theory Comput.*, 11(2):723–739, 2015.

[152] M. R. Machado and S. Pantano. SIRAH Tools: Mapping, Backmapping and Visualization of Coarse-Grained Models. *Bioinformatics*, 32(10):1568–1570, 2016.

[153] A. Liwo, M. Baranowski, C. Czaplewski, E. Gołaś, Y. He, D. Jagieła, P. Krupa, M. Maciejczyk, M. Makowski, M. A. Mozolewska, A. Niadzvedtski, S. Ołdziej, H. A. Scheraga, A. K. Sieradzan, R. Ślusarz, T. Wirecki, Y. Yin, and B. Zaborowski. A Unified Coarse-Grained Model of Biological Macromolecules Based on Mean-Field Multipole–Multipole Interactions. *J. Mol. Model.*, 20(8):2306, 2014.

[154] A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J. Phys. Chem. B*, 116(29):8494–8503, 2012.

[155] J. Maupetit, P. Tuffery, and P. Derreumaux. A Coarse-Grained Protein Force Field for Folding and Structure Prediction. *Proteins: Struct., Funct., Bioinf.*, 69(2):394–408, 2007.

[156] T. Cragnolini, Y. Laurin, P. Derreumaux, and S. Pasquali. Coarse-Grained HiRE-RNA Model for Ab Initio RNA Folding beyond Simple Molecules, Including Noncanonical and Multiple Base Pairings. *J. Chem. Theory Comput.*, 11(7):3510–3522, 2015.

[157] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye. DNA Nanotweezers Studied with a Coarse-Grained Model of DNA. *Phys. Rev. Lett.*, 104(17):178101, 2010.

[158] P. Šulc, F. Romano, T. E. Ouldridge, J. P. K. Doye, and A. A. Louis. A Nucleotide-Level Coarse-Grained Model of RNA. *J. Chem. Phys.*, 140(23):235102, 2014.

[159] M. D. Demetriou, W. L. Johnson, and K. Samwer. Coarse-Grained Description of Localized Inelastic Deformation in Amorphous Metals. *Appl. Phys. Lett.*, 94(19):191905, 2009.

[160] J. L. Suter, R. L. Anderson, H. Christopher Greenwell, and P. V. Coveney. Recent Advances in Large-Scale Atomistic and Coarse-Grained Molecular Dynamics Simulation of Clay Minerals. *J. Mater. Chem.*, 19(17):2482, 2009.

[161] M. G. Saunders and G. A. Voth. Coarse-Graining of Multiprotein Assemblies. *Curr. Opin. Struct. Biol.*, 22(2):144–150, 2012.

[162] J. F. Dama, A. V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A. R. Dinner, and G. A. Voth. The Theory of Ultra-Coarse-Graining. 1. General Principles. *J. Chem. Theory Comput.*, 9(5):2466–2480, 2013.

[163] B. Chen and R. Tycko. Simulated Self-Assembly of the HIV-1 Capsid: Protein Shape and Native Contacts Are Sufficient for Two-Dimensional Lattice Formation. *Biophys. J.*, 100(12):3035–3044, 2011.

[164] I. G. Johnston, A. A. Louis, and J. P. K. Doye. Modelling the Self-Assembly of Virus Capsids. *J. Phys.: Condens. Matter*, 22(10):104101, 2010.

[165] A. Levandovsky and R. Zandi. Nonequilibirum Assembly, Retroviruses, and Conical Structures. *Phys. Rev. Lett.*, 102(19):198102, 2009.

[166] D. C. Rapaport. Molecular Dynamics Study of T = 3 Capsid Assembly. *J. Biol. Phys.*, 44(2):147–162, 2018.

[167] S. N. Fejer, T. R. James, J. Hernández-Rojas, and D. J. Wales. Energy Landscapes for Shells Assembled from Pentagonal and Hexagonal Pyramids. *Phys. Chem. Chem. Phys.*, 11(12):2098, 2009.

[168] A. Einstein. Über Die Von Der Molekularkinetischen Theorie Der Wärme Geforderte Bewegung Von in Ruhenden Flüssigkeiten Suspendierten Teilchen. *Ann. Phys.*, 322(8):549–560, 1905.

[169] M. Giani, W. K. den Otter, and W. J. Briels. Early Stages of Clathrin Aggregation at a Membrane in Coarse-Grained Simulations. *J. Chem. Phys.*, 146(15):155102, 2017.

[170] J. G. Hernández Cifre and J. G. de la Torre. Ionic Strength Effect in Polyelectrolyte Dilute Solutions Within the Debye–Hückel Approximation: Monte Carlo and Brownian Dynamics Simulations. *Polym. Bull.*, 71(9):2269–2285, 2014.

[171] I. M. Ilie, W. K. den Otter, and W. J. Briels. Rotational Brownian Dynamics Simulations of Clathrin Cage Formation. *J. Chem. Phys.*, 141(6):065101, 2014.

[172] I. M. Ilie, W. J. Briels, and W. K. den Otter. An Elementary Singularity-Free Rotational Brownian Dynamics Algorithm for Anisotropic Particles. *J. Chem. Phys.*, 142(11):114103, 2015.

[173] D. Palanisamy and W. K. den Otter. Efficient Brownian Dynamics of Rigid Colloids in Linear Flow Fields Based on the Grand Mobility Matrix. *J. Chem. Phys.*, 148(19):194112, 2018.

[174] W. F. van Gunsteren and H. J. C. Berendsen. Algorithms for Brownian Dynamics. *Mol. Phys.*, 45(3):637–647, 1982.

[175] M. Whittle and E. Dickinso. Brownian Dynamics Simulation of Gelation in Soft Sphere Systems with Irreversible Bond Formation. *Molec. Phys.*, 90(5):739–758, 1997.

[176] G. A. Huber and J. A. McCammon. Brownian Dynamics Simulations of Biological Molecules. *Trends Chem.*, 1(8):727–738, 2019.

[177] T. A. Witten and L. M. Sander. Diffusion-Limited Aggregation. *Phys. Rev. B*, 27(9):5686–5697, 1983.

[178] D. A. Weitz and M. Oliveria. Fractal Structures Formed by Kinetic Aggregation of Aqueous Gold Colloids. *Phys. Rev. Lett.*, 52(16):1433–1436, 1984.

[179] D. A. Weitz, J. S. Huang, M. Y. Lin, and J. Sung. Limits of the Fractal Dimension for Irreversible Kinetic Aggregation of Gold Colloids. *Phys. Rev. Lett.*, 54(13):1416–1419, 1985.

[180] R. Jullien, R. Botet, and P. M. Mors. Computer Simulations of Cluster–Cluster Aggregation. *Faraday Discuss. Chem. Soc.*, 83(0):125–137, 1987.

[181] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, Oxford, United Kingdom, second edition, 2017.

[182] K. Binder and D. W. Heermann. *Monte Carlo Simulation in Statistical Physics: An Introduction*. Number 80 in Springer Series in Solid-State Sciences. Springer, Berlin ; New York, fourth edition, 2002.

[183] D. P. Landau and K. Binder. *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press, Cambridge, United Kingdom ; New York, NY, fifth edition, 2020.

[184] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh. Machine Learning for Molecular and Materials Science. *Nature*, 559(7715):547–555, 2018.

[185] M. Haghighatlari and J. Hachmann. Advances of Machine Learning in Molecular Modeling and Simulation. *Curr. Opin. Chem. Eng.*, 23:51–57, 2019.

[186] K. A. Fichthorn and W. H. Weinberg. Theoretical Foundations of Dynamical Monte Carlo Simulations. *J. Chem. Phys.*, 95(2):1090–1096, 1991.

[187] E. Paquet and H. L. Viktor. Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review. *BioMed Res. Int.*, 2015:1–18, 2015.

[188] S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195(2):216–222, 1987.

[189] F. Ding, D. Tsao, H. Nie, and N. V. Dokholyan. Ab Initio Folding of Proteins with All-Atom Discrete Molecular Dynamics. *Structure*, 16(7):1010–1018, 2008.

[190] D. Shirvanyants, F. Ding, D. Tsao, S. Ramachandran, and N. V. Dokholyan. Discrete Molecular Dynamics: An Efficient And Versatile Simulation Method For Fine Protein Characterization. *J. Phys. Chem. B*, 116(29):8375–8382, 2012.

[191] E. A. Proctor and N. V. Dokholyan. Applications of Discrete Molecular Dynamics in Biology and Medicine. *Curr. Opin. Struct. Biol.*, 37:9–13, 2016.

[192] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, and D. Hassabis. Protein Complex Prediction with AlphaFold-Multimer. Preprint, Bioinformatics, 2021.

[193] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.*, 4(2):268–276, 2018.

[194] R. Winter, F. Montanari, A. Steffen, H. Briem, F. Noé, and D.-A. Clevert. Efficient Multi-Objective Molecular Optimization in a Continuous Latent Space. *Chem. Sci.*, 10(34):8016–8024, 2019.

[195] M. Popova, O. Isayev, and A. Tropsha. Deep Reinforcement Learning for De Novo Drug Design. *Sci. Adv.*, 4(7):eaap7885, 2018.

[196] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.*, 9(2):513–530, 2018.

[197] B. Mishra and R. K. Rajamani. The Discrete Element Method for the Simulation of Ball Mills. *Appl. Math. Model.*, 16(11):598–604, 1992.

[198] T. Pöschel and T. Schwager. *Computational Granular Dynamics: Models and Algorithms*. Springer Science & Business Media, 2005.

[199] F. A. Tavarez and M. E. Plesha. Discrete Element Method for Modelling Solid and Particulate Materials. *Int. J. Numer. Meth. Eng.*, 70(4):379–404, 2007.

[200] M. Dosta, K. Jarolin, and P. Gurikov. Modelling of Mechanical Behavior of Biopolymer Alginate Aerogels Using the Bonded-Particle Model. *Molecules*, 24(14):2543, 2019.

[201] D. Potyondy and P. Cundall. A Bonded-Particle Model for Rock. *Int. J. Rock Mech. Min. Sci.*, 41(8):1329–1364, 2004.

[202] D. O. Potyondy. The Bonded-Particle Model as a Tool for Rock Mechanics Research and Application: Current Trends and Future Directions. *Geosyst. Eng.*, 18(1):1–28, 2015.

[203] S. Rybczyński, M. Dosta, G. Schaan, M. Ritter, and F. Schmidt-Döhl. Numerical Study on the Mechanical Behavior of Ultrahigh Performance Concrete Using a Three-Phase Discrete Element Model. *Struct. Concr.*, pages 1–16, 2020.

[204] S. Zellmer, G. Garnweitner, T. Breinlinger, T. Kraft, and C. Schilde. Hierarchical Structure Formation of Nanoparticulate Spray-Dried Composite Aggregates. *ACS Nano*, 9(11):10749–10757, 2015.

[205] C. Schilde, C. F. Burmeister, and A. Kwade. Measurement and Simulation of Micromechanical Properties of Nanostructured Aggregates Via Nanoindentation and DEM-Simulation. *Powder Technol.*, 259:1–13, 2014.

[206] C. Sangrós Giménez, B. Finke, C. Nowak, C. Schilde, and A. Kwade. Structural and Mechanical Characterization of Lithium-Ion Battery Electrodes Via DEM Simulations. *Adv. Powder Tech.*, 29(10):2312–2321, 2018.

[207] C. Sangrós Giménez, B. Finke, C. Schilde, L. Froböse, and A. Kwade. Numerical Simulation of the Behavior of Lithium-Ion Battery Electrodes During the Calendaring Process Via the Discrete Element Method. *Powder Technol.*, 349:1–11, 2019.

[208] J. F. Wendt (editor). *Computational Fluid Dynamics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[209] J. R. Whiteman. *The Mathematics of Finite Elements and Applications*. Academic Press, London New York, 1978.

[210] P. Kieckhefen, S. Pietsch, M. Dosta, and S. Heinrich. Possibilities and Limits of Computational Fluid Dynamics–Discrete Element Method Simulations in Process Engineering: A Review of Recent Advancements and Future Trends. *Annu. Rev. Chem. Biomol. Eng.*, 11(1):397–422, 2020.

[211] M. Schrader, K. Pommerehne, S. Wolf, B. Finke, C. Schilde, I. Kampen, T. Lichtenegger, R. Krull, and A. Kwade. Design of a CFD-DEM-Based Method for Mechanical Stress Calculation and Its Application to Glass Bead-Enhanced Cultivations of Filamentous Lentzea Aerocolonigenes. *Biochem. Eng. J.*, 148:116–130, 2019.

[212] C. Kloss, C. Goniva, A. Hager, S. Amberger, and S. Pirker. Models, Algorithms and Validation for Opensource DEM and CFD-DEM. *Prog. Comput. Fluid Dyn.*, 12(2/3):140, 2012.

[213] A. Munjiza. *The Combined Finite-Discrete Element Method*. Wiley, Hoboken, NJ, 2004.

[214] A. Malevanets and R. Kapral. Mesoscopic Model for Solvent Dynamics. *J. Chem. Phys.*, 110(17):8605–8613, 1999.

[215] R. Kapral. Multiparticle Collision Dynamics: Simulation of Complex Systems on Mesoscales. *Adv. Chem. Phys.*, 140:89, 2008.

[216] T. Ihle and D. M. Kroll. Stochastic Rotation Dynamics: A Galilean-Invariant Mesoscopic Model for Fluid Flow. *Phys. Rev. E*, 63(2):020201, 2001.

[217] G. Gompper, T. Ihle, D. M. Kroll, and R. G. Winkler. Multi-Particle Collision Dynamics: A Particle-Based Mesoscale Simulation Approach to the Hydrodynamics of Complex Fluids. In *Advanced Computer Simulation Approaches for Soft Matter Sciences III*, pages 1–87. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[218] P. J. Hoogerbrugge and J. M. V. A. Koelman. Simulating Microscopic Hydrodynamic Phenomena with Dissipative Particle Dynamics. *Europhys. Lett.*, 19(3):155–160, 1992.

[219] P. Español and P. Warren. Statistical Mechanics of Dissipative Particle Dynamics. *Europhys. Lett.*, 30(4):191–196, 1995.

[220] P. B. Warren. Dissipative Particle Dynamics. *Curr. Opin. Colloid Interface Sci.*, 3(6):620–624, 1998.

[221] H.-J. Qian, C. C. Liew, and F. Muller-Plathe. Effective Control of the Transport Coefficients of a Coarse-Grained Liquid and Polymer Models Using the Dissipative Particle Dynamics and Lowe-Andersen Equations of Motion. *Phys. Chem. Chem. Phys.*, 11(12):1962–1969, 2009.

[222] P. M. Pieczywek, W. Płaziński, and A. Zdunek. Dissipative Particle Dynamics Model of Homogalacturonan Based on Molecular Dynamics Simulations. *Sci. Rep.*, 10(1):14691, 2020.

[223] P. N. Depta, U. Jandt, M. Dosta, A.-P. Zeng, and S. Heinrich. Toward Multiscale Modeling of Proteins and Bioagglomerates: An Orientation-Sensitive Diffusion Model for the Integration of Molecular Dynamics and the Discrete Element Method. *J. Chem. Inf. Model.*, 59(1):386–398, 2019.

[224] P. N. Depta, P. Gurikov, B. Schroeter, A. Forgács, J. Kalmár, G. Paul, L. Marchese, S. Heinrich, and M. Dosta. DEM-Based Approach for the Modeling of Gelation and Its Application to Alginate. *J. Chem. Inf. Model.*, 62(1):49–70, 2022.

[225] P. N. Depta, M. Dosta, W. Wenzel, M. Kozlowska, and S. Heinrich. Hierarchical Coarse-Grained Strategy for Macromolecular Self-Assembly: Application to Hepatitis B Virus-Like Particles. *Int. J. Mol. Sci.*, 23(23):14699, 2022.

[226] P. N. Depta, M. Dosta, and S. Heinrich. Data-Driven Multiscale Modeling of Self-Assembly and Hierarchical Structural Formation in Biological Macro-Molecular Systems. In W. E. Nagel, D. H. Kröner, and M. M. Resch (editors), *High Performance Computing in Science and Engineering '21*. Springer International Publishing, Cham, 2023.

[227] P. N. Depta, M. Dosta, and S. Heinrich. Data-Driven Multiscale Modeling of Self-Assembly and Hierarchical Structural Formation in Biological Macro-Molecular Systems: Pyruvate Dehydrogenase Complex. In W. E. Nagel, D. H. Kröner, and M. M. Resch (editors), *High Performance Computing in Science and Engineering '22*. Springer International Publishing, Cham, 2024 (in print).

[228] P. N. Depta, M. Dosta, and S. Heinrich. Multiscale Model-Based Investigation of Functional Macromolecular Agglomerates for Biotechnological Applications. In A. Kwade and I. Kampen (editors), *Dispersity, Structure and Phase Changes of Proteins and Bio Agglomerates in Biotechnological Processes*. Springer International Publishing, Cham, 2024 (in print).

[229] I. Smirnova and P. Gurikov. Aerogels in Chemical Engineering: Strategies Toward Tailor-Made Aerogels. *Annu. Rev. Chem. Biomol. Eng.*, 8(1):307–334, 2017.

[230] T. Budtova, D. A. Aguilera, S. Beluns, L. Berglund, C. Chartier, E. Espinosa, S. Gaidukovs, A. Klimek-Kopyra, A. Kmita, D. Lachowicz, F. Liebner, O. Platnieks, A. Rodríguez, L. K. Tinoco Navarro, F. Zou, and S. J. Buwalda. Biorefinery Approach for Aerogels. *Polymers*, 12(12):2779, 2020.

[231] I. Donati and S. Paoletti. Material Properties of Alginates. In B. H. A. Rehm (editor), *Alginates: Biology and Applications*, volume 13, pages 1–53. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[232] H. Hecht and S. Srebnik. Structural Characterization of Sodium Alginate and Calcium Alginate. *Biomacromolecules*, 17(6):2160–2167, 2016.

[233] L. Ratke and P. Gurikov. *The Chemistry and Physics of Aerogels: Synthesis, Processing, and Properties*. Cambridge University Press, Cambridge ; New York, NY, 2021.

[234] P. Agulhon, M. Robitzer, L. David, and F. Quignard. Structural Regime Identification in Ionotropic Alginate Gels: Influence of the Cation Nature and Alginate Structure. *Biomacromolecules*, 13(1):215–220, 2012.

[235] G. T. Grant, E. R. Morris, D. A. Rees, P. J. Smith, and D. Thom. Biological Interactions Between Polysaccharides and Divalent Cations: The Egg-Box Model. *FEBS Lett.*, 32(1):195–198, 1973.

[236] I. Braccini and S. Pérez. Molecular Basis of Ca2+ -Induced Gelation in Alginates and Pectins: The Egg-Box Model Revisited. *Biomacromolecules*, 2(4):1089–1096, 2001.

[237] W. Plazinski. Molecular Basis of Calcium Binding by Polyguluronate Chains. Revising the Egg-Box Model. *J. Comput. Chem.*, 32(14):2988–2995, 2011.

[238] L. Cao, W. Lu, A. Mata, K. Nishinari, and Y. Fang. Egg-Box Model-Based Gelation of Alginate and Pectin: A Review. *Carbohydr. Polym.*, 242:116389, 2020.

[239] Y. Fang, S. Al-Assaf, G. O. Phillips, K. Nishinari, T. Funami, P. A. Williams, and L. Li. Multiple Steps and Critical Behaviors of the Binding of Calcium to Alginate. *J. Phys. Chem. B*, 111(10):2456–2462, 2007.

[240] W. Plazinski and M. Drach. Calcium-α-L-Guluronate Complexes: Ca2+ Binding Modes from DFT-MD Simulations. *J. Phys. Chem. B*, 117(40):12105–12112, 2013.

[241] A. Wolnik, L. Albertin, L. Charlier, and K. Mazeau. Probing the Helical Forms of Ca2+-Guluronan Junction Zones in Alginate Gels by Molecular Dynamics 1: Duplexes. *Biopolymers*, 99(8):562–571, 2013.

[242] M. B. Stewart, S. R. Gray, T. Vasiljevic, and J. D. Orbell. Exploring the Molecular Basis for the Metal-Mediated Assembly of Alginate Gels. *Carbohydr. Polym.*, 102:246–253, 2014.

[243] H. Hecht and S. Srebnik. Sequence-Dependent Association of Alginate with Sodium and Calcium Counterions. *Carbohydr. Polym.*, 157:1144–1152, 2017.

[244] S. Wynne, R. Crowther, and A. Leslie. The Crystal Structure of the Human Hepatitis B Virus Capsid. *Molec. Cell*, 3(6):771–780, 1999.

[245] A. Zlotnick, N. Cheng, J. F. Conway, F. P. Booy, A. C. Steven, S. J. Stahl, and P. T. Wingfield. Dimorphism of Hepatitis B Virus Capsids Is Strongly Influenced by the C-Terminus of the Capsid Protein. *Biochemistry*, 35(23):7412–7421, 1996.

[246] B. Böttcher and M. Nassal. Structure of Mutant Hepatitis B Core Protein Capsids with Premature Secretion Phenotype. *J. Mol. Biol.*, 430(24):4941–4954, 2018.

[247] F. Birnbaum and M. Nassal. Hepatitis B Virus Nucleocapsid Assembly: Primary Structure Requirements in the Core Protein. *J. Virol.*, 64(7):3319–3330, 1990.

[248] L. Selzer, S. P. Katen, and A. Zlotnick. The Hepatitis B Virus Core Protein Intradimer Interface Modulates Capsid Assembly and Stability. *Biochemistry*, 53(34):5496–5504, 2014.

[249] C. Ludwig and R. Wagner. Virus-Like Particles—Universal Molecular Toolboxes. *Curr. Opin. Biotechnol.*, 18(6):537–545, 2007.

[250] M. O. Mohsen, L. Zha, G. Cabral-Miranda, and M. F. Bachmann. Major Findings and Recent Advances in Virus–Like Particle (VLP)-Based Vaccines. *Semin. Immunol.*, 34:123–132, 2017.

[251] E. J. Hartzell, R. M. Lieser, M. O. Sullivan, and W. Chen. Modular Hepatitis B Virus-like Particle Platform for Biosensing and Drug Delivery. *ACS Nano*, 14(10):12642–12651, 2020.

[252] M. Somiya and S. Kuroda. Development of a Virus-Mimicking Nanocarrier for Drug Delivery Systems: The Bio-Nanocapsule. *Adv. Drug Delivery Rev.*, 95:77–89, 2015.

[253] L. Selzer and A. Zlotnick. Assembly and Release of Hepatitis B Virus. *Cold Spring Harbor Perspect. Med.*, page a021394, 2015.

[254] A. Zlotnick and S. Mukhopadhyay. Virus Assembly, Allostery and Antivirals. *Trends Microbiol.*, 19(1):14–23, 2011.

[255] A. Arkhipov, P. L. Freddolino, and K. Schulten. Stability and Dynamics of Virus Capsids Described by Coarse-Grained Modeling. *Structure*, 14(12):1767–1777, 2006.

[256] M. R. Machado, H. C. González, and S. Pantano. MD Simulations of Viruslike Particles with Supra CG Solvation Affordable to Desktop Computers. *J. Chem. Theory Comput.*, 13(10):5106–5116, 2017.

[257] M. Cieplak and M. O. Robbins. Nanoindentation of 35 Virus Capsids in a Molecular Model: Relating Mechanical Properties to Structure. *PLoS ONE*, 8(6):e63640, 2013.

[258] J. K. Marzinek, R. G. Huber, and P. J. Bond. Multiscale Modelling and Simulation of Viruses. *Curr. Opin. Struct. Biol.*, 61:146–152, 2020.

[259] C. A. Brautigam, R. M. Wynn, J. L. Chuang, M. Machius, D. R. Tomchick, and D. T. Chuang. Structural Insight into Interactions Between Dihydrolipoamide Dehydrogenase (E3) and E3 Binding Protein of Human Pyruvate Dehydrogenase Complex. *Structure*, 14(3):611–621, 2006.

[260] C. A. Brautigam, R. M. Wynn, J. L. Chuang, and D. T. Chuang. Subunit and Catalytic Component Stoichiometries of an in Vitro Reconstituted Human Pyruvate Dehydrogenase Complex. *J. Biol. Chem.*, 284(19):13086–13098, 2009.

[261] R. N. Perham. Swinging Arms and Swinging Domains in Multifunctional Enzymes: Catalytic Machines for Multistep Reactions. *Annu. Rev. Biochem.*, 69(1):961–1004, 2000.

[262] R. A. Harris, M. M. Bowker-Kinley, P. Wu, J. Jeng, and K. M. Popov. Dihydrolipoamide Dehydrogenase-binding Protein of the Human Pyruvate Dehydrogenase Complex. *J. Biol. Chem.*, 272(32):19746–19751, 1997.

[263] M. S. Patel and L. G. Korotchkina. The Biochemistry of the Pyruvate Dehydrogenase Complex. *Biochem. Mol. Biol. Educ.*, 31(1):5–15, 2003.

[264] S. J. Sanderson, C. Miller, and J. G. Lindsay. Stoichiometry, Organisation and Catalytic Function of Protein X of the Pyruvate Dehydrogenase Complex from Bovine Heart. *Eur. J. Biochem.*, 236(1):68–77, 1996.

[265] S. Prajapati, D. Haselbach, S. Wittig, M. S. Patel, A. Chari, C. Schmidt, H. Stark, and K. Tittmann. Structural and Functional Analyses of the Human PDH Complex Suggest a "Division-of-Labor" Mechanism by Local E1 and E3 Clusters. *Structure*, 27(7):1124–1136, 2019.

[266] S. Vijayakrishnan, P. Callow, M. A. Nutley, D. P. McGow, D. Gilbert, P. Kropholler, A. Cooper, O. Byron, and J. G. Lindsay. Variation in the Organization and Subunit Composition of the Mammalian Pyruvate Dehydrogenase Complex E2/E3BP Core Assembly. *Biochem. J.*, 437(3):565–574, 2011.

[267] Z. H. Zhou, D. B. McCarthy, C. M. O'Connor, L. J. Reed, and J. K. Stoops. The Remarkable Structural and Functional Organization of the Eukaryotic Pyruvate Dehydrogenase Complexes. *Proc. Natl. Acad. Sci. U.S.A.*, 98(26):14802–14807, 2001.

[268] S. Hezaveh, A.-P. Zeng, and U. Jandt. Investigation of Core Structure and Stability of Human Pyruvate Dehydrogenase Complex: A Coarse-Grained Approach. *ACS Omega*, 2(3):1134–1145, 2017.

[269] J. Guo, S. Hezaveh, J. Tatur, A.-P. Zeng, and U. Jandt. Reengineering of the Human Pyruvate Dehydrogenase Complex: From Disintegration to Highly Active Agglomerates. *Biochem. J.*, 474(5):865–875, 2017.

[270] H. Lee and S. Lim. Disassembly and Trimer Formation of E2 Protein Cage: The Effects of C-Terminus, Salt, and Protonation State. *J. Phys. D: Appl. Phys.*, 51(36):365402, 2018.

[271] S. Hezaveh, A.-P. Zeng, and U. Jandt. Full Enzyme Complex Simulation: Interactions in Human Pyruvate Dehydrogenase Complex. *J. Chem. Inf. Model.*, 58(2):362–369, 2018.

[272] H. Pavlu-Pereira, D. Lousa, C. S. Tomé, C. Florindo, M. J. Silva, I. T. de Almeida, P. Leandro, I. Rivera, and J. B. Vicente. Structural and Functional Impact of Clinically Relevant E1α Variants Causing Pyruvate Dehydrogenase Complex Deficiency. *Biochimie*, 183:78–88, 2021.

[273] J. Sgrignani, J. Chen, A. Alimonti, and A. Cavalli. How Phosphorylation Influences E1 Subunit Pyruvate Dehydrogenase: A Computational Study. *Sci. Rep.*, 8(1):14683, 2018.

[274] X. Zhou, S. Yu, J. Su, and L. Sun. Computational Study on New Natural Compound Inhibitors of Pyruvate Dehydrogenase Kinases. *Int. J. Mol. Sci.*, 17(3):340, 2016.

[275] Y. Zhou, M. Cai, H. Zhou, L. Hou, H. Peng, and H. He. Discovery of Efficient Inhibitors Against Pyruvate Dehydrogenase Complex Component E1 with Bactericidal Activity Using Computer Aided Design. *Pestic. Biochem. Physiol.*, 177:104894, 2021.

[276] Y. Li, B. Hu, Z. Wang, J. He, Y. Zhang, J. Wang, and L. Guan. Identification of Pyruvate Dehydrogenase E1 as a Potential Target Against Magnaporthe Oryzae Through Experimental and Theoretical Investigation. *Int. J. Mol. Sci.*, 22(10):5163, 2021.

[277] T. Geyer. Many-Particle Brownian and Langevin Dynamics Simulations with the Brownmove Package. *BMC Biophys.*, 4(1):7, 2011.

[278] H. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation. *Comput. Phys. Commun.*, 91(1):43–56, 1995.

[279] M. Dosta and V. Skorych. MUSEN: An Open-Source Framework for GPU-Accelerated DEM Simulations. *SoftwareX*, 12:100618, 2020.

[280] K. Shoemake. Uniform Random Rotations. In *Graphics Gems III (IBM Version)*, pages 124–132. Elsevier, 1992.

[281] N. Goga, A. J. Rzepiela, A. H. de Vries, S. J. Marrink, and H. J. C. Berendsen. Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.*, 8(10):3637–3649, 2012.

[282] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.*, 26(16):1701–1718, 2005.

[283] E. Lindahl, B. Hess, and D. Van Der Spoel. GROMACS 3.0: A Package for Molecular Simulation and Trajectory Analysis. *J. Mol. Model.*, 7(8):306–317, 2001.

[284] V. Skorych, M. Buchholz, M. Dosta, H. K. Baust, M. Gleiß, J. Haus, D. Weis, S. Hammerich, G. Kiedorf, N. Asprion, H. Nirschl, F. Kleine Jäger, and S. Heinrich. Use of Multiscale Data-Driven Surrogate Models for Flowsheet Simulation of an Industrial Zeolite Production Process. *Processes*, 10(10):2140, 2022.

[285] B. B. Das, S. H. Park, and S. J. Opella. Membrane Protein Structure from Rotational Diffusion. *Biochim. Biophys. Acta, Biomembr.*, 1848(1, Part B):229–245, 2015.

[286] A. C. Bohórquez, C. Yang, D. Bejleri, and C. Rinaldi. Rotational Diffusion of Magnetic Nanoparticles in Protein Solutions. *J. Colloid Interface Sci.*, 506:393–402, 2017.

[287] S. O. Yesylevskyy, L. V. Schäfer, D. Sengupta, and S. J. Marrink. Polarizable Water Model for the Coarse-Grained MARTINI Force Field. *PLoS Comput. Biol.*, 6(6):e1000810, 2010.

[288] G. Bussi, D. Donadio, and M. Parrinello. Canonical Sampling Through Velocity Rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[289] A. Cauchy. Méthode Générale Pour La Résolution Des Systemes D'équations Simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

[290] M. Parrinello and A. Rahman. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[291] S. Nosé and M. Klein. Constant Pressure Molecular Dynamics for Molecular Systems. *Mol. Phys.*, 50(5):1055–1076, 1983.

[292] E. Marinari and G. Parisi. Simulated Tempering: A New Monte Carlo Scheme. *Europhys. Lett.*, 19(6):451–458, 1992.

[293] A. J. Smola and B. Schölkopf. A Tutorial on Support Vector Regression. *Stat. Comp.*, 14(3):199–222, 2004.

[294] N. A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, Inc, Hoboken, NJ, 2015.

[295] R. Webster and M. A. Oliver. *Geostatistics for Environmental Scientists*. Wiley, 2007.

[296] H. Wackernagel. *Multivariate Geostatistics: An Introduction with Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[297] A. Lichtenstern. *Kriging Methods in Spatial Statistics*. Bachelorarbeit, Technische Universität München, 2013.

[298] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA, 2012.

[299] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass, 2006.

[300] D. S. Broomhead and D. Lowe. Multivariable Functional Interpolation and Adaptive Networks. *Complex Syst.*, 2:321–355, 1988.

[301] H. Fang and M. F. Horstemeyer. Global Response Approximation with Radial Basis Functions. *Engineering Optimization*, 38(4):407–424, 2006.

[302] E. Alpaydin. *Introduction to Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass, second edition, 2010.

[303] R. Collobert and J. Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, pages 160–167. ACM Press, Helsinki, Finland, 2008.

[304] Y. Goldberg. *Neural Network Methods for Natural Language Processing*, volume 10 of *Synthesis Lectures on Human Language Technologies*. Springer, first edition, 2017.

[305] R. Alizadeh, J. K. Allen, and F. Mistree. Managing Computational Complexity Using Surrogate Models: A Critical Review. *Res. Eng. Design*, 31(3):275–298, 2020.

[306] M. J. Asher, B. F. W. Croke, A. J. Jakeman, and L. J. M. Peeters. A Review of Surrogate Models and Their Application to Groundwater Modeling: Surrogates of Groundwater Models. *Water Resour. Res.*, 51(8):5957–5973, 2015.

[307] A. Bhosekar and M. Ierapetritou. Advances in Surrogate Based Modeling, Feasibility Analysis, and Optimization: A Review. *Comput. Chem. Eng.*, 108:250–267, 2018.

[308] G. Sun and S. Wang. A Review of the Artificial Neural Network Surrogate Modeling in Aerodynamic Design. *J. Aerosp. Eng.*, 233(16):5863–5872, 2019.

[309] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, Without the Electrons. *Phys. Rev. Lett.*, 104(13):136403, 2010.

[310] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi. Gaussian Process Regression for Materials and Molecules. *Chem. Rev.*, 121(16):10073–10141, 2021.

[311] T. Stecher, N. Bernstein, and G. Csányi. Free Energy Surface Reconstruction from Umbrella Samples Using Gaussian Process Regression. *J. Chem. Theory Comput.*, 10(9):4079–4097, 2014.

[312] L. Mones, N. Bernstein, and G. Csányi. Exploration, Sampling, And Reconstruction of Free Energy Surfaces with Gaussian Process Regression. *J. Chem. Theory Comput.*, 12(10):5100–5110, 2016.

[313] J. Hénin. Fast and Accurate Multidimensional Free Energy Integration. *J. Chem. Theory Comput.*, 17(11):6789–6798, 2021.

[314] E. Schneider, L. Dai, R. Q. Topper, C. Drechsel-Grau, and M. E. Tuckerman. Stochastic Neural Network Approach for Learning High-Dimensional Free Energy Surfaces. *Phys. Rev. Lett.*, 119(15):150601, 2017.

[315] J. S. Smith, O. Isayev, and A. E. Roitberg. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.*, 8(4):3192–3203, 2017.

[316] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. Van Der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris,

A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. Van Mulbregt, SciPy 1.0 Contributors, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. De Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, and Y. Vázquez-Baeza. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods*, 17(3):261–272, 2020.

[317] J. D. Hunter. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.*, 9(3):90–95, 2007.

[318] A. E. Long and D. E. Myers. A New Form of the Cokriging Equations. *Math. Geol.*, 29(5):685–703, 1997.

[319] M. Kumral and P. Dowd. Singular Value Decomposition as an Equation Solver in Co-Kriging Matrices. *J. South. Afr. Inst. Min. Metall.*, 112:853–858, 2012.

[320] S. S. Garud, I. Karimi, and M. Kraft. Smart Sampling Algorithm for Surrogate Model Development. *Comput. Chem. Eng.*, 96:103–114, 2017.

[321] J. J. Gómez-Hernández and E. F. Cassiraga. Theory and Practice of Sequential Simulation. In M. Armstrong and P. A. Dowd (editors), *Geostatistical Simulations*, volume 7, pages 111–124. Springer Netherlands, Dordrecht, 1994.

[322] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.*, 23(3):327–341, 1977.

[323] H. C. Andersen. RATTLE: A "Velocity" Version of the Shake Algorithm for Molecular Dynamics Calculations. *J. Comput. Phys.*, 52(1):24–34, 1983.

[324] S. Miyamoto and P. A. Kollman. SETTLE: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models. *J. Comput. Chem.*, 13(8):952–962, 1992.

[325] B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[326] B. Hess. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.*, 4(1):116–122, 2008.

[327] K. Jarolin and M. Dosta. Linearization-Based Methods for the Calibration of Bonded-Particle Models. *Comp. Part. Mech.*, 8(3):511–523, 2021.

[328] F. L. Markley, Y. Cheng, J. L. Crassidis, and Y. Oshman. Averaging Quaternions. *J. Guid. Contr. Dyn.*, 30(4):1193–1197, 2007.

[329] R. Weast. *CRC Handbook of Chemistry and Physics - 64th Edition*. CRC Press, 1983.

[330] R. D. Shannon. Revised Effective Ionic Radii and Systematic Studies of Interatomic Distances in Halides and Chalcogenides. *Acta Cryst. A*, 32(5):751–767, 1976.

[331] A. C. Ribeiro, M. C. Barros, A. S. Teles, A. J. Valente, V. M. Lobo, A. J. Sobral, and M. Esteso. Diffusion Coefficients and Electrical Conductivities for Calcium Chloride Aqueous Solutions at 298.15K and 310.15K. *Electrochim. Acta*, 54(2):192–196, 2008.

[332] J. J. Cerdà, T. Sintes, and A. Chakrabarti. Excluded Volume Effects on Polymer Chains Confined to Spherical Surfaces. *Macromolecules*, 38(4):1469–1477, 2005.

[333] B. Smit and D. Frenkel. Vapor–Liquid Equilibria of the Two-Dimensional Lennard-Jones Fluid(s). *J. Chem. Phys.*, 94(8):5663–5668, 1991.

[334] S. E. Rodriguez-Cruz, R. A. Jockusch, and E. R. Williams. Hydration Energies of Divalent Metal Ions, Ca2+ (H2O)n (n = 5-7) and Ni2+ (H2O)n (n = 6-8), Obtained by Blackbody Infrared Radiative Dissociation. *J. Am. Chem. Soc.*, 120(23):5842–5843, 1998.

[335] I. Preibisch, P. Niemeyer, Y. Yusufoglu, P. Gurikov, B. Milow, and I. Smirnova. Polysaccharide-Based Aerogel Bead Production via Jet Cutting Method. *Materials*, 11(8):1287, 2018.

[336] C. López-Iglesias, J. Barros, I. Ardao, P. Gurikov, F. J. Monteiro, I. Smirnova, C. Alvarez-Lorenzo, and C. A. García-González. Jet Cutting Technique for the Production of Chitosan Aerogel Microparticles Loaded with Vancomycin. *Polymers*, 12(2):273, 2020.

[337] G. Paul, E. Boccaleri, C. Cassino, D. Gastaldi, L. Buzzi, F. Canonico, and L. Marchese. Fingerprinting the Hydration Products of Hydraulic Binders Using Snapshots from Time-Resolved in Situ Multinuclear MAS NMR Spectroscopy. *J. Phys. Chem. C*, 125(17):9261–9272, 2021.

[338] D. Massiot, F. Fayon, M. Capron, I. King, S. Le Calvé, B. Alonso, J.-O. Durand, B. Bujoli, Z. Gan, and G. Hoatson. Modelling One- and Two-Dimensional Solid-State NMR Spectra: Modelling 1D and 2D Solid-State NMR Spectra. *Magn. Reson. Chem.*, 40(1):70–76, 2002.

[339] A. Forgács, V. Papp, G. Paul, L. Marchese, A. Len, Z. Dudás, I. Fábián, P. Gurikov, and J. Kalmár. Mechanism of Hydration and Hydration Induced Structural Changes of Calcium Alginate Aerogel. *ACS Appl. Mater. Interfaces*, 13(2):2997–3010, 2021.

[340] T. Salomonsen, H. M. Jensen, F. H. Larsen, S. Steuernagel, and S. B. Engelsen. Alginate Monomer Composition Studied by Solution- and Solid-State NMR – A Comparative Chemometric Study. *Food Hydrocolloids*, 23(6):1579–1586, 2009.

[341] Lu, X. Liu, L. Dai, and Z. Tong. Difference in Concentration Dependence of Relaxation Critical Exponent n for Alginate Solutions at Sol-Gel Transition Induced by Calcium Cations. *Biomacromolecules*, 6(4):2150–2156, 2005.

[342] S. G. Allen, P. C. L. Stephenson, and J. H. Strange. Internal Surfaces of Porous Media Studied by Nuclear Magnetic Resonance Cryoporometry. *J. Chem. Phys.*, 108(19):8195–8198, 1998.

[343] J. Dore, J. Webber, and J. Strange. Characterisation of Porous Solids Using Small-Angle Scattering and NMR Cryoporometry. *Colloids Surf. A*, 241(1-3):191–200, 2004.

[344] O. V. Petrov and I. Furó. NMR Cryoporometry: Principles, Applications and Potential. *Prog. Nucl. Magn. Reson. Spectrosc.*, 54(2):97–122, 2009.

[345] M. Kéri, A. Forgács, V. Papp, I. Bányai, P. Veres, A. Len, Z. Dudás, I. Fábián, and J. Kalmár. Gelatin Content Governs Hydration Induced Structural Changes in Silica-Gelatin Hybrid Aerogels – Implications in Drug Delivery. *Acta Biomater.*, 105:131–145, 2020.

[346] M. Robitzer, L. David, C. Rochas, F. Di Renzo, and F. Quignard. Nanostructure of Calcium Alginate Aerogels Obtained from Multistep Solvent Exchange Route. *Langmuir*, 24(21):12547–12552, 2008.

[347] I. Lázár, A. Forgács, A. Horváth, G. Király, G. Nagy, A. Len, Z. Dudás, V. Papp, Z. Balogh, K. Moldován, L. Juhász, C. Cserháti, Z. Szántó, I. Fábián, and J. Kalmár. Mechanism of Hydration of Biocompatible Silica-Casein Aerogels Probed by NMR and SANS Reveal Backbone Rigidity. *Appl. Surf. Sci.*, 531:147232, 2020.

[348] G. Beaucage. Approximations Leading to a Unified Exponential/Power-Law Approach to Small-Angle Scattering. *J. Appl. Crystallogr.*, 28(6):717–728, 1995.

[349] B. Hammouda. Analysis of the Beaucage Model. *J. Appl. Crystallogr.*, 43(6):1474–1478, 2010.

[350] İ. Şahin, E. Uzunlar, and C. Erkey. Investigation of the Effect of Gel Properties on Supercritical Drying Kinetics of Ionotropic Alginate Gel Particles. *J. Supercrit. Fluids*, 152:104571, 2019.

[351] D. Lovskaya, A. Lebedev, and N. Menshutina. Aerogels as Drug Delivery Systems: In Vitro and in Vivo Evaluations. *J. Supercrit. Fluids*, 106:115–121, 2015.

[352] M. Robitzer, A. Tourrette, R. Horga, R. Valentin, M. Boissière, J. Devoisselle, F. Di Renzo, and F. Quignard. Nitrogen Sorption as a Tool for the Characterisation of Polysaccharide Aerogels. *Carbohydr. Polym.*, 85(1):44–53, 2011.

[353] T. Mehling, I. Smirnova, U. Guenther, and R. Neubert. Polysaccharide-Based Aerogels as Drug Carriers. *J. Non-Cryst. Solids*, 355(50-51):2472–2479, 2009.

[354] B. T. Stokke, K. I. Draget, O. Smidsrød, Y. Yuguchi, H. Urakawa, and K. Kajiwara. Small-Angle X-Ray Scattering and Rheological Characterization of Alginate Gels. 1. Ca-Alginate Gels. *Macromolecules*, 33(5):1853–1863, 2000.

[355] Y. Maki, K. Ito, N. Hosoya, C. Yoneyama, K. Furusawa, T. Yamamoto, T. Dobashi, Y. Sugimoto, and K. Wakabayashi. Anisotropic Structure of Calcium-Induced Alginate Gels by Optical and Small-Angle X-Ray Scattering Measurements. *Biomacromolecules*, 12(6):2145–2152, 2011.

[356] M. Robitzer, L. David, C. Rochas, F. Di Renzo, and F. Quignard. Supercritically-Dried Alginate Aerogels Retain the Fibrillar Structure of the Hydrogels. *Macromol. Symp.*, 273(1):80–84, 2008.

[357] Y. Yuguchi, H. Urakawa, K. Kajiwara, K. Draget, and B. Stokke. Small-Angle X-Ray Scattering and Rheological Characterization of Alginate Gels. 2. Time-Resolved Studies on Ionotropic Gels. *J. Mol. Struct.*, 554(1):21–34, 2000.

[358] Z. Gu, L. Gu, R. Eils, M. Schlesner, and B. Brors. Circlize Implements and Enhances Circular Visualization in R. *Bioinformatics*, 30(19):2811–2812, 2014.

[359] M. F. Hagan and D. Chandler. Dynamic Pathways for Viral Capsid Assembly. *Biophys. J.*, 91(1):42–54, 2006.

[360] B. Venkatakrishnan and A. Zlotnick. The Structural Biology of Hepatitis B Virus: Form and Function. *Annu. Rev. Virol.*, 3(1):429–451, 2016.

[361] R. Asor, C. J. Schlicksup, Z. Zhao, A. Zlotnick, and U. Raviv. Rapidly Forming Early Intermediate Structures Dictate the Pathway of Capsid Assembly. *J. Am. Chem. Soc.*, 142(17):7868–7882, 2020.

[362] J. K. Hilmer, A. Zlotnick, and B. Bothner. Conformational Equilibria and Rates of Localized Motion within Hepatitis B Virus Capsids. *J. Mol. Biol.*, 375(2):581–594, 2008.

[363] K. A. Dryden, S. F. Wieland, C. Whitten-Bauer, J. L. Gerin, F. V. Chisari, and M. Yeager. Native Hepatitis B Virions and Capsids Visualized by Electron Cryomicroscopy. *Mol. Cell*, 22(6):843–850, 2006.

[364] A. M. Roseman, J. A. Berriman, S. A. Wynne, P. J. G. Butler, and R. A. Crowther. A Structural Model for Maturation of the Hepatitis B Virus Core. *Proc. Natl. Acad. Sci. U.S.A.*, 102(44):15821, 2005.

[365] S. Seitz, S. Urban, C. Antoni, and B. Böttcher. Cryo-Electron Microscopy of Hepatitis B Virions Reveals Variability in Envelope Capsid Interactions. *EMBO J.*, 26(18):4160–4167, 2007.

[366] S. Katen and A. Zlotnick. Chapter 14 - The Thermodynamics of Virus Capsid Assembly. In *Biothermodynamics, Part A*, volume 455 of *Methods in Enzymology*, pages 395–417. Academic Press, 2009.

[367] Z. D. Harms, L. Selzer, A. Zlotnick, and S. C. Jacobson. Monitoring Assembly of Virus Capsids with Nanofluidic Devices. *ACS Nano*, 9(9):9087–9096, 2015.

[368] C. A. Lutomski, N. A. Lyktey, Z. Zhao, E. E. Pierson, A. Zlotnick, and M. F. Jarrold. Hepatitis B Virus Capsid Completion Occurs through Error Correction. *J. Am. Chem. Soc.*, 139(46):16932–16938, 2017.

[369] C. A. Lutomski, N. A. Lyktey, E. E. Pierson, Z. Zhao, A. Zlotnick, and M. F. Jarrold. Multiple Pathways in Capsid Assembly. *J. Am. Chem. Soc.*, 140(17):5784–5790, 2018.

[370] N. Hillebrandt, P. Vormittag, A. Dietrich, C. H. Wegner, and J. Hubbuch. Process Development for Cross-Flow Diafiltration-Based Vlp Disassembly: A Novel High-Throughput Screening Approach. *Biotechnol. Bioeng.*, 118(10):3926–3940, 2021.

[371] R. F. Bruinsma, G. J. L. Wuite, and W. H. Roos. Physics of Viral Dynamics. *Nat. Rev. Phys.*, 3(2):76–91, 2021.

[372] D. Endres and A. Zlotnick. Model-Based Analysis of Assembly Kinetics for Virus Capsids or Other Spherical Polymers. *Biophys. J.*, 83(2):1217–1230, 2002.

[373] M. F. Hagan and O. M. Elrad. Understanding the Concentration Dependence of Viral Capsid Assembly Kinetics—the Origin of the Lag Time and Identifying the Critical Nucleus Size. *Biophys. J.*, 98(6):1065–1074, 2010.

[374] A. Zlotnick. Distinguishing Reversible from Irreversible Virus Capsid Assembly. *J. Mol. Biol.*, 366(1):14–18, 2007.

[375] A. Zlotnick, J. M. Johnson, P. W. Wingfield, S. J. Stahl, and D. Endres. A Theoretical Model Successfully Identifies Features of Hepatitis B Virus Capsid Assembly. *Biochemistry*, 38(44):14644–14652, 1999.

[376] J. Schumacher, T. Bacic, R. Staritzbichler, M. Daneschdar, T. Klamp, P. Arnold, S. Jägle, Ö. Türeci, J. Markl, and U. Sahin. Enhanced Stability of a Chimeric Hepatitis B Core Antigen Virus-Like-Particle (HBcAg-VLP) by a C-Terminal Linker-Hexahistidine-Peptide. *J. Nanobiotechnol.*, 16(1):39, 2018.

[377] T. Klamp, J. Schumacher, G. Huber, C. Kühne, U. Meissner, A. Selmi, T. Hiller, S. Kreiter, J. Markl, Ö. Türeci, and U. Sahin. Highly Specific Auto-Antibodies against Claudin-18 Isoform 2 Induced by a Chimeric HBcAg Virus-Like Particle Vaccine Kill Tumor Cells and Inhibit the Growth of Lung Metastases. *Cancer Res.*, 71(2):516–527, 2011.

[378] E. E. Pierson, D. Z. Keifer, L. Selzer, L. S. Lee, N. C. Contino, J. C.-Y. Wang, A. Zlotnick, and M. F. Jarrold. Detection of Late Intermediates in Virus Capsid Assembly by Charge Detection Mass Spectrometry. *J. Am. Chem. Soc.*, 136(9):3536–3541, 2014.

[379] M. S. Patel, L. G. Korotchkina, and S. Sidhu. Interaction of E1 and E3 Components with the Core Proteins of the Human Pyruvate Dehydrogenase Complex. *J. Mol. Catal. B: Enzym.*, 61(1-2):2–6, 2009.

[380] Y.-H. Park and M. S. Patel. Characterization of Interactions of Dihydrolipoamide Dehydrogenase with Its Binding Protein in the Human Pyruvate Dehydrogenase Complex. *Biochem. Biophys. Res. Commun.*, 395(3):416–419, 2010.

[381] M. Smolle, A. E. Prior, A. E. Brown, A. Cooper, O. Byron, and J. G. Lindsay. A New Level of Architectural Complexity in the Human Pyruvate Dehydrogenase Complex. *J. Biol. Chem.*, 281(28):19772–19780, 2006.

[382] L. G. Korotchkina and M. S. Patel. Binding of Pyruvate Dehydrogenase to the Core of the Human Pyruvate Dehydrogenase Complex. *FEBS Lett.*, 582(3):468–472, 2008.

[383] C. Marsac, D. Stansbie, G. Bonne, J. Cousin, P. Jehenson, C. Benelli, J.-P. Leroux, and G. Lindsay. Defect in the Lipoyl-Bearing Protein X Subunit of the Pyruvate Dehydrogenase Complex in Two Patients with Encephalomyelopathy. *J. Ped.*, 123(6):915–920, 1993.

[384] C. Jacobi and A.-P. Zeng. EnzymAgglo - Multiskalige Modellgestützte Untersuchungen Funktionaler Enzym- Und Proteinagglomerate Für Biotechnologische Anwendung. Teil 2: Von Der Struktur Zur Funktion. In *SPP DiSPBiotech*. Deutsche Forschungsgemeinschaft, 2022.

[385] N. G. Wallis, M. D. Allen, R. Broadhurst, I. A. Lessard, and R. N. Perham. Recognition of a Surface Loop of the Lipoyl Domain Underlies Substrate Channelling in the Pyruvate Dehydrogenase Multienzyme Complex. *J. Mol. Biol.*, 263(3):463–474, 1996.

[386] T. Izard, A. Aevarsson, M. D. Allen, A. H. Westphal, R. N. Perham, A. de Kok, and W. G. J. Hol. Principles of Quasi-Equivalence and Euclidean Geometry Govern the Assembly of Cubic and Dodecahedral Cores of Pyruvate Dehydrogenase Complexes. *Proc. Natl. Acad. Sci. U.S.A.*, 96(4):1240–1245, 1999.

[387] S. Vijayakrishnan, S. Kelly, R. Gilbert, P. Callow, D. Bhella, T. Forsyth, J. Lindsay, and O. Byron. Solution Structure and Characterisation of the Human Pyruvate Dehydrogenase Complex Core Assembly. *J. Mol. Biol.*, 399(1):71–93, 2010.

[388] J. Jiang, F. L. Baiesc, Y. Hiromasa, X. Yu, W. H. Hui, X. Dai, T. E. Roche, and Z. H. Zhou. Atomic Structure of the E2 Inner Core of Human Pyruvate Dehydrogenase Complex. *Biochemistry*, 57(16):2325–2334, 2018.

[389] R. Behal, M. DeBuysere, B. Demeler, J. Hansen, and M. Olson. Pyruvate Dehydrogenase Multienzyme Complex. Characterization of Assembly Intermediates by Sedimentation Velocity Analysis. *J. Biol. Chem.*, 269(50):31372–31377, 1994.

[390] Y. Hiromasa, T. Fujisawa, Y. Aso, and T. E. Roche. Organization of the Cores of the Mammalian Pyruvate Dehydrogenase Complex Formed by E2 and E2 Plus the E3-binding Protein and Their Capacities to Bind the E1 and E3 Components. *J. Biol. Chem.*, 279(8):6921–6933, 2004.

[391] Z. Zhou, W. Liao, R. Cheng, J. Lawson, D. McCarthy, L. J. Reed, and J. K. Stoops. Direct Evidence for the Size and Conformational Variability of the Pyruvate Dehydrogenase Complex Revealed by Three-dimensional Electron Microscopy. *Journal of Biological Chemistry*, 276(24):21704–21713, 2001.

[392] S. Ilhan. *Novel Strategies for Automated Engineering of Enzymatic Systems: Structural and Functional Insights to Human Pyruvate Dehydrogenase Complex*. Doctoral Thesis, Technische Universität Hamburg, 2021.

[393] H. Yamakawa. Transport Properties of Polymer Chains in Dilute Solution: Hydrodynamic Interaction. *J. Chem. Phys.*, 53(1):436–443, 1970.

[394] E. Dickinson, S. A. Allison, and J. A. McCammon. Brownian Dynamics with Rotation–Translation Coupling. *J. Chem. Soc., Faraday Trans. 2*, 81(4):591–601, 1985.

[395] D. L. Ermak and J. A. McCammon. Brownian Dynamics with Hydrodynamic Interactions. *J. Chem. Phys.*, 69(4):1352–1360, 1978.

[396] J. G. Kirkwood and J. Riseman. The Intrinsic Viscosities and Diffusion Constants of Flexible Macromolecules in Solution. *J. Chem. Phys.*, 16(6):565–573, 1948.

[397] J. Rotne and S. Prager. Variational Treatment of Hydrodynamic Interaction in Polymers. *J. Chem. Phys.*, 50(11):4831–4837, 1969.

[398] P. Ahlrichs, R. Everaers, and B. Dünweg. Screening of Hydrodynamic Interactions in Semidilute Polymer Solutions: A Computer Simulation Study. *Phys. Rev. E*, 64(4):040501, 2001.

[399] J. G. De La Torre and V. A. Bloomfield. Hydrodynamic Properties of Macromolecular Complexes. I. Translation. *Biopolymers*, 16(8):1747–1763, 1977.

[400] J. G. De La Torre and V. A. Bloomfield. Hydrodynamics of Macromolecular Complexes. II. Rotation. *Biopolymers*, 16(8):1765–1778, 1977.

[401] S. Kim. Singularity Solutions for Ellipsoids in Low-Reynolds-Number Flows: With Applications to the Calculation of Hydrodynamic Interactions in Suspensions of Ellipsoids. *Int. J. Multiphase Flow*, 12(3):469–491, 1986.

[402] T. Geyer and U. Winter. An O(N2) Approximation for Hydrodynamic Interactions in Brownian Dynamics Simulations. *J. Chem. Phys.*, 130(11):114905, 2009.

[403] R. R. Schmidt, J. G. H. Cifre, and J. G. de la Torre. Comparison of Brownian Dynamics Algorithms with Hydrodynamic Interaction. *J. Chem. Phys.*, 135(8):084116, 2011.

[404] G. Matheron. *The Theory of Regionalized Variables and Its Applications*. Cahiers. École national supérieure des mines, 1971.

[405] N. Cressie and D. M. Hawkins. Robust Estimation of the Variogram: I. *J. Int. Assoc. Math. Geol.*, 12(2):115–125, 1980.

[406] Y. Wang, Z. Gong, H. Fang, D. Zhi, and H. Tao. The N-Terminal 1–55 Residues Domain of Pyruvate Dehydrogenase from Escherichia Coli Assembles as a Dimer in Solution. *Protein Eng., Des. Sel.*, 32(6):271–276, 2019.

[407] U. Jandt, P. N. Depta, S. Ilhan, C. Müller, M. Dosta, S. Heinrich, and A.-P. Zeng. Multiscale Modeling of Mixtures of Catalytically Active and Inactive Enzymatic Aggregates. In *Himmelfahrtstagung 2018*. Dechema, 2018.