Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering

Doctoral Thesis

Development of Paralell Algorithms for Molecular Dynamics Simulation of Heterogeneous Atomistic Systems

Prague 2023

David Celný

This thesis is submitted to the Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Ph.D.) in Mathematical Engineering.

Bibliografický záznam

Autor	Ing. David Celný, České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská, Katedra matematiky
Název práce	Vývoj paralelních algoritmů pro molekulovou dynamiku heterogenních atomistických systémů
Studijní program	Aplikace přírodních věd
Studijní obor	Matematické inženýrství
Školitel	prof. RNDr. Jiří Kolafa CSc., Vysoká škola chemicko–technologická v Praze, Fakulta chemicko–inženýrská, Ústav fyzikální chemie
Školitel specialista	Ing. Václav Vinš, Ph.D., Akademie věd České republiky, Ústav termomechaniky, Oddělení termodynamiky
Akademický rok	2023
Počet stran	196
Klíčová slova	Nukleace, Fázové rozhranní, gradientní teorie, Molekulární Dynamika, Supersonická expanze, Kapka, Detekce klusterů a voidů, metastabilní stav, Stavová rovnice

Bibliographic Entry

Author	Ing. David Celný, Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Department of Mathematics
Title of dissertation	Development of parallel algorithms for molecular dynamics simulation of heterogeneous atomistic systems
Degree programme	Application of Natural Sciences
Field of study	Mathematical Engineering
Supervisor	prof. RNDr. Jiří Kolafa CSc., University of Chemistry and Technology Prague, Faculty of Chemical Engineering, Department of Physical Chemistry
Supervisor specialist	Ing. Václav Vinš, Ph.D., Czech Academy of Sciences, Institute of Thermomechanics, Department of Thermodynamics
Academic year	2023
Number of pages	196
Keywords	Nucleation, Phase interface, Gradient theory, Molecular dynamics, Supersonic expansion, Droplet, Cluster and void identification, Metastable state, Equation of State

Abstrakt

Výzkum nukleace, jakožto procesu fázové přeměny, má v termodynamice dlouhou tradici, sahající až k prvním pozorování kondenzace vodních par. Zájem o samotný proces byl ještě zesílen s nástupem molekulární dynamiky (MD), díky níž lze proces nukleace sledovat napřímo ve větším přiblížení, nežli je možné v přírodě. Avšak tyto ambice naráží na výpočetní omezení spojená s velikostí simulovatelného systému. V důsledku toho jsou výsledky simulací, experimentů a teoretických modelů porovnatelné pouze kvalitativně, přičemž rozdíl mezi jednotlivými výsledky je i několik řádů. Hlavním cílem této práce je rozšířit pole působnosti molekulárních simulací zabývajících se nukleací. Tohoto cíle je dosaženo pomocí paralelizace algoritmů na grafických procesorech (GPGPU), což umožňuje zvýšení rychlosti simulace. Proto je poté možné provádět simulace rozsáhlejších systémů, a tím zlepšit prediktivní schopnost simulací nukleace. Druhým úkolem je výzkum clusterů vytvořených v průběhu simulace. Tyto poznatky jsou využity k vytvoření kritérií vhodných pro klasifikaci metastabilní (přechodně stabilní) kapaliny a páry. Použití těchto kriterií umožňuje detekci nukleace, pomoci níž lze procesu nukleace lépe porozumět. Termofyzikální data shromážděná během před-nukleační fáze simulace mohou být rovněž použita k vytvoření nových stavových rovnic.

Abstract

Theoretical exploration of nucleation as a kind of phase transition process represents a longstanding tradition in thermodynamics, dating from the first observations of droplet nucleation. The research potential in this field has been further propelled by computational tools like Molecular Dynamics (MD), which facilitate a more direct investigation of nucleation processes in contrast to their observation in the natural environment. However, these simulation ambitions encounter computational limitations tied to the scale of the problem. As a result, existing experimental outcomes, simulation outputs, and theoretical models agree only qualitatively, while substantial quantitative disparities spanning multiple orders of magnitude persist. The principal objective of this study is to expand the horizons of simulations pertaining to nucleation. This will be accomplished by harnessing parallelization methodologies within general-purpose programming for graphics processing units (GPGPU), thereby enhancing the speed of calculation. The realization of this objective bears the potential to conduct calculations on larger systems and to enhance the predictive accuracy of nucleation simulations. The secondary goals are to investigate cluster formation during simulation and to develop cluster criteria for metastable (temporary stable) liquid and vapor regions. The establishment of such criteria will enable the classification of nucleating systems, thereby offering profound insights into the underlying processes. The thermophysical data collected during the pre-nucleation phase can be used to construct a novel set of equations of state.

Acknowledgements

This thesis would not be possible without the help, support, and understanding of many people who helped me during my time as a Ph.D. student. I would first like to thank my supervisor, prof. RNDr. Jiří Kolafa CSc. and my supervisor specialist, Ing. Václav Vinš, Ph.D., for their unwavering support, constant motivation, and insightful consultations, which were instrumental in shaping my thesis. I want to thank Ing. Jan Hrubý, CSc. for introducing me to the topic and shaping my scientific path from early bachelor till Ph.D. with his broad expertise that I can't stop envying. I also want to thank doc. Ing. Tomáš Oberhuber, Ph.D., for introducing me to the field of GPGPU. I owe big thanks to Prof. Dr. Roland Span and Dr.-Ing. Monika Thol from RUB who wholeheartedly supported my stay in Bochum and helped with my research on the metastable system as well as the thermophysical package TREND. The following "thank you" is for prof. Dr.-Ing. habil. Jadran Vrabec for his openness, quick and insightful replies, and help with the development of the criteria in *ms*2. I also want to thank my friends and colleagues, in particular Martin Klíma, Sven Pohl, and Tomáš Radnic, who were not only a motivation as fellow Ph.D. students but also of great help in the last months before the thesis submission. Last but definitely not least, I want to dedicate my last thanks to my family, who supported me in their own way long before I had even attempted to do a Ph.D.

Our work related to the topic of this thesis was partially supported by the following projects:

- "Research centre for low-carbon energy technologies", project No. CZ.02.1.01/0.0/0.0-/16_019/0000753, of the OP VVV program of the Ministry of Education, Youth and Sports of the Czech Republic

- "Thermal Energy Storage Materials: Thermophysical Characteristics for the Design of Thermal Batteries", No. GA17-08218S, of the Czech Science Foundation grant

- "Support of the Institute of Thermomechanics of the CAS, v. v. i.", project No. RVO:61388998, of the internal Grant agency of Institute of Thermomechanics

- "Application of analytical mathematical methods in study of physical and biological systems 2", project No. SGS18/189/OHK4/3T/14, of the Student Grant Agency of the Czech Technical University in Prague,

- "Application of rigorous mathematical methods in study of physical models", project No. SGS21/166/OHK4/3T/14, of the Student Grant Agency of the Czech Technical University in Prague

- "Investigation of non-equilibrium steam condensation: A new approach", project No. CEP16-GA0-GA-U/01:1, of the Czech Science Foundation grant

- "Phase behaviour in CCS systems", project No. 7F14466, of the Norwegian Financial Mechanism program of the Ministry of Education, Youth and Sports of the Czech

- "Surface tension of water and water mixtures in equilibrium and metastable state", project No. GJ15-07129Y, GA0/GJ, of the Junior Grant program of Czech Science Foundation grant

- "Droplets, ice and aerosols in silico: combinations of ab initio and classical procedures", project No. 18-16577S, GA0/GA, of the Czech Science Foundation grant

- “Education improvement–priority of UTC in Prague”, project No. CZ.02.2.69/0.0/0.0/16_/0002374, of the Ministry of Education, Youth and Sports of the Czech Republic
- “Laboratories for Excellent Bachelor and Master Degree Programmes”, project No. CZ.02.2.67/0.0/0.0/16_016/0002357, of the Ministry of Education, Youth and Sports of the Czech Republic
- “Formation of droplets in rapid expansions: Between unary, binary, and heterogeneous paths.”, No. GA22-28869S of the Czech Science Foundation grant
- “Aqueous mixtures with salts under extreme conditions – accurate experiments, molecular simulations and modeling”, No. GA22-03380S of the Czech Science Foundation grant

Author’s declaration

I confirm that I have prepared the thesis by my own and having listed all used sources of information in the bibliography.
I further declare that AI tools: ChatGPT [183] Perplexity [191] and QuillBot [196] have been used for review and corrections of the text as English is not the first language of the author. To verify the reasonable use, AI Detector: “Content at scale” [50] have been utilized to verify that text in this work is human generated.

Prague, 30 August 2023

David Celný

Contents

Nomenclature

Acronyms

CCS	Carbon Capture and Storage
CNT	Classical Nucleation Theory
DGT	Density Gradient Theory
DoF	Degree of Freedom
EoS	Equation of State
FCC	Face Centered Cubic
GPU	Graphical Processing Unit
LDF	Local Density Fluctuation
MD	Molecular Dynamics
NN	Nearest Neighbor
SPC/E	Single Point Charge Extended
TRVP	Time Reversible Velocity Predictor
VLE	Vapor Liquid Equilibria

Subscripts

c	critical value (used for temperature, pressure, density)
comp	component
cut	cutoff
eq	equilibrium
hom, inhom	homogeneous, inhomogeneous
kin	kinetic
sot	surface of tension
l	liquid

v	vapor	

Greek letters

α	reduced Helmholtz energy (in the context of EoS)	[-]
δ	reduced density	[-]
ϵ_{LJ}	Lennard-Jones interaction parameter ϵ	[J]
ϵ_0	vacuum permittivity	$8.8541878128 \times 10^{12}$ F/m
Γ	adsorption	[m^{-2} when unscaled]
κ	structural parameter (in the context of PR-EoS)	
κ	friction parameter (in the context thermostats)	
λ	Lagrange multiplier	
μ	mean value (in the context of model based clustering)	
μ_i	chemical potential of component i	[J/mol]
ν	distribution of cluster sizes	
ω	grand potential density (in the context of Phase interface)	[J m^{-3}]
ω	acentric factor (in the context of EoS)	
ϕ	Gaussian density distribution function	
ψ	sphericity	[-]
ρ	density	[kg/m^3]
$\tilde{\rho}$	modified density	[mol m^{-3}]
σ	surface tension	$[mN/m]$
σ_{LJ}	Lennard-Jones interaction parameter σ	[Å]
$\mathbf{\Sigma}$	covariance matrix (in the context of model based clustering)	
τ	reduced inverse temperature	[-]
Θ	scaling factor of expansion	
ξ	coefficient of polytropy	

Latin letters

$\mathcal{A}$	cross-section	
c	coefficient of condensation (in the context of nucleation)	
c	speed of sound (in the context of supersonic expansion)	[m/s]

$c_{i,j}$	influence parameter (in the context of gradient theory)	[J m^5/ mol^2]
CV	coefficient of variance	
d	distance (in the context of clustering)	
e	coefficient of evaporation (in the context of nucleation)	
E	energy	[J]
U	potential energy	[J]
$\mathcal{H}$	Hamiltonian	
$\mathcal{L}$	Lagrangian	
J	nucleation rate	[number of clusters/(s m^3)]
J_w	mass flux	[kg /(m^2 s)]
k_B	Boltzmann constant	$1.38064852 \times 10^{-23}$ m^2 kg/(K s^2)
k_B	universal gas constant	8.3144598 J/(K mol)
N_A	Avogadro constant	6.02219×10^{23}
N	number of molecules	[-]
A	Helmholtz energy	[J]
H	Enthalpy	[J]
Ω	grand potential energy	[J]
p	pressure	[Pa]
p, q	generalized coordinates (in the context of Lagrangian, Hamilton formalism)	
r	radius	[nm,Å]
T	temperature	[K]
v	velocity	[m/s]
v_1	volume of one molecule	[m^3]
V$_m$	molar volume	[m^3/mol]
V	volume	[m^3]
X	artificial variable	[m$^{5/2}$ J$^{-1/2}$]
$\mathcal{Z}$	Zeldovich factor	

Superscripts

CM	center of mass
*	dimensionles properties (used for temperature, pressure, density, ...)

*	critical parameters (used for radius and number of molecules)

Other symbols

$i \in \widehat{n}$	$i \in \{1, \dots, n\}$

Introduction 1

The process of nucleation is a prevalent cornerstone of the physics of condensed matter from the eighteenth century, when the phenomenon of nucleation was first indirectly observed in supercooled water by Fahrenheit [69]. Nucleation was afterwards experimentally and theoretically indirectly investigated [82,184] followed in 1926 by the theoretical work on nucleation published by Volmer and Weber [249] and later extended by Becker and Döring [21] leading to the formulation of classical nucleation theory (CNT). Lately, a number of books have been published [113, 116, 118, 239] due to the wide occurrence of the nucleation phenomenon. The wide applicability is a consequence of nucleation processes close connection with thermophysical properties, atmospheric processes, uses in industry, or interdisciplinary similarities with other processes occurring in nature. The prominent applications are connected with carbon capture and storage (CCS), first outlined in the IPCC report [165] and stressed again in the recent ICCP report regarding the limits of warming [38]. This is an important motivation for the modeling of CO_2 and equations of state research, which leads to technologies for lowering carbon dioxide in the atmosphere. Other applications include methods for the preparation of platinum nanoparticles used in polymer electrolyte membrane (PEM) fuel cells [30, 256].

1.1 State of the art

1.1.1 Phase interface

The initial theoretical investigation of the structure of the phase interface were performed already by van der Waals [234] in 1894. His findings were later extended by Cahn and Hilliard [34,35] who formulated a feasible density gradient theory (DGT). The theory introduced by Cahn and Hilliard describes the free energy of the non-uniform system as an expansion around the uniform system energy. This leads to the expansion of the uniform energy with an additional local disturbance described by so-called gradient energy. The system is modeled in thermodynamic equilibrium; therefore, the investigation of the vapor-liquid equilibrium (VLE) represents an important part of the model. The knowledge of the nonuniform energy is vital in formulating a set of conditions for finding the most probable density distribution within a system containing two fluid phases. Following the method of Cahn and Hilliard [36] coupled with VLE, the investigated system can be solved for the density profile, the work of formation, and interfacial properties such as surface tension. These properties can be further utilized in various nucleation models, e.g., those used in atmospheric research or in engineering applications such as the design of CCS technologies.

The density gradient theory framework presented in this study was initially used to predict the interface properties of pure substances. Predictive models were originally supplied with a simple cubic equation of state (EoS), such as the Peng-Robinson (PR) EoS [51], or the Soave-

Redlich-Kwong (SRK) EoS [210, 271]. For simple systems, such as lower alkanes, cubic EoSs provide reasonable VLE predictions. However, when considering more complex components, cubic EoSs become imprecise, and volume-translated variations [147, 167] or more complex EoSs need to be employed. The predictive capability was enhanced by introducing an association term into EoS [132]. Therefore, the cubic plus association (CPA) EoS [121] or the statistical associating fluid theory (SAFT) equations are preferably used in order to model more complex systems. These equations were used in multiple models [112, 171, 238] aligning well with experimental data. The SAFT-EoSs are commonly used in combination with the DGT to model the density profile and the interfacial properties, as these EoSs provide good prediction for VLE and the saturated liquid density for a large variety of substances, including mixtures [154].

Through the years (2016–2018), when the author focused on the topic, the above mentioned description of the one-component system was extended to binary and multicomponent mixtures, e.g., refs. [145, 155, 170, 171, 189]. In these works, the authors derived the formulas and key equation for the multicomponent problems and provided comparisons with available experimental data. To the best of the authors' knowledge, the multicomponent studies are restricted to the planar phase interface geometry, leading to a simpler form of the solved problem. Some research groups, e.g., refs. [1, 106, 181, 198, 259], applied DGT combined with a convenient EoS on the more complicated spherical interface geometry, i.e., on the description of droplets or bubbles. However, these studies deal with the calculations of pure systems. No studies, including both the mathematical problem formulation and the practical calculations using DGT for the spherical phase geometry in a multicomponent system, were found in the literature in 2018, when this research took place.

1.1.2 Supersonic expansion

Expansion is a process where a vapor transits from a region of high pressure into a region of low pressure. This is usually done through a nozzle, where the expanding vapor cools down while reaching speeds that may reach the speed of sound. This is the case for the supersonic expansion investigated in this work, where the cooling of vapor creates suitable conditions for the nucleation of droplets.

The origins of this field of research can be found in the work of Becker and Döring [21] which laid the basics for the Theory of Nucleation. Theoretical groundwork propelled experimental investigation of expansion in nozzles first conducted by Oswatitsch [185], Kantrowitz [114], and Wegener [254] using both water and wet air as expanding media with pre-1960 technological limitations. Hill [99] and subsequently Wegener [255] investigated the effects of nozzles on the expansion. Continued in 1967 by Stein and Wegener [219], who first used Rayleigh scattering of a beam of helium-neon lasers to measure the concentrations and nucleation rate of expansion. This opened a new era of optical measurements in expansion experiments, which has prevailed to this day.

Molecular simulations of nucleation were first performed for the case of static systems with foundational works from Thompson *et al.* [230], Wolde and Frenkel [228], and Yasuoka and Matsumoto [265, 266]. Simulation of nucleation rates in large systems was continued in work by Diemand *et al.* [59], Tanaka *et al.* [225] and later Angelil [8]. The role of a thermostat and carrier gas was of concern for Wedekind *et al.* [253] and later Halonen *et al.* [95]. Recent experimental investigations focus on cluster temperature with work by Becker *et al.* [20] and pickup cross section of the clusters by Fárník *et al.* [70].

Expansion was only scarcely addressed in a few studies, one by Zhong *et al.* [270] utilizing a long simulation volume with an artificial boundary condition emulating a saturated system on one side and a vacuum on the other. The second was done later by Li *et al.* [143], who used the elipsoid statistical theory to model the water nucleation in an expanding system with the help of molecular simulation.

More recent research by Dingilian *et al.* [60] uses classical nucleation theory to investigate the supersonic expansion of carbon dioxide with carrier gas in comparison with experimental data. A theoretical approach utilizing Navier-Stokes equations was used by Xu *et al.* [263] for direct numerical simulation of supersonic expansion optimized for multiple GPUs.

1.1.3 Properties of metastable systems

Continuing with the literature review of nucleation, CNT, and expansion, the question about the properties of the system is interlaced within. In this review, we focus on the experimental and later simulation investigations of the metastable states, which are the foundation for nucleation to occur.

The first experimental setups included the cloud chamber invented by Wilson [260] in 1897, which achieved supersaturation and enabled measurements utilizing only the resistance thermometry of the employed carrier gas. Other measurements were done on already mentioned expanding systems developed by the group of Wegener [219, 254, 255]. The initial experiments were also performed on co-called diffusion chambers developed by Franck and Hetz [76]. During the later years 1968–1992, measurement of nucleation received much more attention, as reported in the review of nucleation experiments from vapor to liquid by Heist and He [97]. Thirteen new measurement methods were proposed across 60 different fluids, capturing supersaturation rations and nucleation rates for unary as well as binary systems.

For the case of bubble formation, the research begins with the works of Tucker and Ward [232] focusing on the critical size of bubbles in relation to CNT, and Hemmingsen [98] who investigated the supersaturation rates of liquids in contact with walls. The nucleation rate of bubbles was studied by Lubetkin and Backwell [150] which was later continued by Strey *et al.* [224] for both heterogeneous and homogeneous cases using light scattering.

With the development of computers, experimental research shifted in the direction of molecular simulation. Using the Monte Carlo (MC) technique, authors Lee Barker and Abraham [2, 124] were able to contribute to the theoretical investigation of nucleation theory by providing initial estimates for cluster densities, Helmholtz energies, and the most probable cluster shapes. Simultaneously, molecular dynamic studies were performed by McGinty [161] to obtain the radial distribution function of argon droplets, Rusanov and Brodskaya [202] to determine the density profiles of droplets, and Kristensen *et al.* [135] who focused on pair distribution within the droplets.

Simulation investigation was extended with the increased computational power available, allowing for the simulation of larger molecular systems investigating phenomena concerning nucleation. The foundation of large system simulation of nucleation was laid by works from Yasuoka and Matsumoto [265, 266]. Together with large system simulation, alternative methods were proposed, i.e., employing a thermodynamic daemon to investigate nucleation in the steady state by Horsch *et al.* [105]. Cluster growth during nucleation was investigated by Napari *et al.* [173] and later by Ayuba *et al.* [12] including further analysis of kinetics.

For molecular simulations of bubbles, there is a substantially fewer publication on the topic compared to the droplet investigation. Initial works from Kinjo *et al.* [122] extend investigation from droplet to bubble formation observation based on pressure. Later, Wu and Pan [262] investigated bubble size evolution and changes in radial distribution function due to nucleation. Niemark and Vishnyakov [174] utilized MC to look into thermodynamic potential and used Voronoi-Delaunay tessellation to detect the vapor cavity. A more comprehensive study of single bubbles surrounded by saturated liquid was performed by Horsch and Hasse [103]. Cavitation was investigated by Baidakov [14] combining DGT and molecular dynamics.

To our knowledge, the metastability investigation was only briefly mentioned as a side note in most of the works. Direct investigation of metastable regions for Lennard-Jones fluid (LJF) was performed by a few authors, i.e., Linhart *et al.* [148]. This field was most extensively investigated by a group led by Baidakov [13, 15–17].

1.2 Research goals

1.2.1 Chronology of the research

This work is in line with the author's previous research in the field of nucleation and gradient theory [44, 192, 244]. While the gradient theory was used to describe phase interfaces, the obtained results also provide insight into nucleation processes, which was the starting point in 2016.

Research on the phase of interface continued in the 2016–2018 period, supervised by Ing. Jan Hrubý, CSc. in the Institute of Thermomechanics of the CAS, v. v. i. The primary focus was on potentially generalizing across various phase interface geometries and identifying any issues with the current model. During the year 2018, the author closed the topic of phase interface and decided to pursue questions related to nucleation, which were discovered during the previous research.

The first problem identified is the effect of small droplets and their shape evolution during nucleation. This topic was investigated in the by a group of Statistical thermodynamics and simulation at the department of physical chemistry led by prof. RNDr. Jiří Kolafa CSc., who supervised this portion of the research. The primary task was the development of a more effective, parallel molecular simulation package for heterogeneous systems, particularly supersonic expansion simulation. The author was intensively working on this task in 2018 and later in the 2019–2021 period, developing the tools necessary to perform the supersonic expansion simulation on a sufficiently large scale. Focusing on the initial phases of nucleation, this line of research currently offers insights into nucleation, in particular the cluster characteristics modeled in non-equilibrium conditions.

The second big question is the onset of nucleation. More specifically, the condition of the system before the nucleation process even takes place. This research was founded in cooperation with a team from Ruhr Universität Bochum, Lehrstuhl für Thermodynamik, led by prof. Dr-Ing. Roland Span. This group focuses on multiparametric equations of state including both experiments and modeling. The initial investigation of the problem was done as part of the student exchange program, which took place from September 2018 until September 2019. After the program finished, the author have returned to the topic of supersonic expansion. Nucleation onset research was resumed at the end of 2021 and continues until 2023. The knowledge of small droplets from supersonic expansion research was utilized to develop a method for nucleation event detection for both droplets and bubbles. This is an interesting area that opens the door for researching metastable conditions for use in EoS development. It is currently being prepared for publication.

The research on these questions is not considered finished, and the tools are not only maintained but extended in functionality to allow for a future work on our understanding of pre-nucleation and the initial phase of nucleation. This required a certain level of parallelism not only on the part of the algorithms used but from the author as well.

1.2.2 List of tasks

Tasks received during the span of the thesis are collected into the following list.

1. **Phase interface research**
    - Generalize the model of phase interface calculation into a unified form applicable to two main types of interface geometries for multiple components.
    - Develop and maintain the PC(P)-SAFT equation of state in FORTRAN 95, primarily oriented on speed.
    - Apply the developed tools for equilibrium calculation and equation fitting, to model CO_2 mixtures.
    - Validate the model prediction on experimental measurements and density functional theory. Discuss the effect of the developed equation of state.
2. **Supersonic expansion research**
    - Understand the model for supersonic expansion. Identify key requirements for parallelization.
    - Based on these requirements, design and implement an independent parallel solution utilizing the supersonic expansion model that will achieve at around ten-fold speedup against the currently used implementation and enable large single-component simulations.
        - Develop an independent molecular dynamics simulation package tailored for GPUs.
        - Implement mechanisms for high energy conservation.
        - Implement the expansion model in the package.
        - Certify that the developed package speedup has been achieved.
    - Simulate expansion with system with 2024,4048 molecules over periods of 1ţ s
    - Analyze the cluster cross sections and compare the results with the experiment. Discuss the results.
3. **Metastable properties research**
    - Design and develop a method to obtain the properties of metastable fluids.
        - Research the cluster definition and clustering algorithms.
        - Study the available software used by the group (TREND, *ms*2)
        - Develop criteria to identify both cluster and void structures.
        - Adapt the method for runtime parallel execution and implement it into the molecular simulation package *ms*2
    - Calculate the metastable properties of Lennard-Jones fluids using the proposed method.
        - Design and develop an automatic tool for sampling metastable conditions from the equation of state
        - Design and develop tools for managing large amounts of simulation used for local machines as well as supercomputers.
    - Compare the calculated data with the equation of state as well as the data available in the literature.
    - Discuss the achieved results in the context of improving the Lenard-Jones multiparameter equation of state.

1.3 Content of the thesis

The thesis is organized as follows:

- Chapter 1 provides an overview of the research, discussing the current state of the art, research objectives, outline of the thesis content, and research outlook.

- Chapter 2 introduces essential thermodynamic concepts, including potentials, phase transitions, and equations of state. Nucleation and related concepts are introduced in more detail, culminating in derivation of CNT for a single component case. The chapter also touches on the Cahn-Hilliard gradient theory.

- Chapter 3 introduces molecular simulation techniques, including a brief overview of Monte Carlo and a more elaborate introduction to molecular dynamics. This chapter provides the theoretical background required to create custom molecular simulation software. This includes a description of integration schemes, constraints, and temperature control mechanisms, as well as a general form of a solver procedure.

- Chapter 4 explores clustering methods, explaining criteria used in unsupervised and physical clustering, with a focus on the properties of the algorithms and relations between them.

- Chapter 5 presents how nucleation can be investigated from an equilibrium perspective using the Cahn-Hilliard gradient theory. A unified method for modeling interfacial geometry is formulated. Solved problems for planar and spherical geometries are shown together with comparison to literature data for three CO_2-rich binary mixtures.

- Chapter 6 investigates the molecular dynamics simulation of supersonic expansion. The problem formulation is presented and later extended for the use of parallel execution on GPUs. The solution includes a more detailed explanation of the optimization performed and the structure of the parallel solution. Results discuss the validation of the parallel solution, the relation of supersonic expansion with classical nucleation theory assumptions, and comparisons with experimental data.

- Chapter 7 delves into metastable system properties. The chapter focuses on outlining challenges for the problem and proposes a general solution that is adjusted for runtime execution within molecular dynamics software. Parametrization of the method and properties of metastable simulation are discussed in detail. Results include the application of the method to both metastable regions and comparisons with equation of state and simulation results from the literature.

- Chapter 8 summarizes the key research findings.

1.4 Achieved results

This is a summary of the achieved results including author's contribution in impacted publications and student thesis.

1. **Phase Interface Research**

   - A unified model of interface has been developed. The proposed method enables the calculation of surface tension and related properties for multicomponent mixtures. The model shows a good agreement with the experimental data for three investigated CO_2-rich binary mixtures. The effects of polarity are well captured by the PCP-SAFT EoS developed alongside the performed research. The achieved results were summarized in Celný *et al.* [43].
     - David Celný was the leading author, developed the model, implemented the method and EoS, and produced the main body of the text and figures.
   - The developed PC(P)-SAFT EoS was further utilized in the research of more complicated hydrofluoroethers, as presented in a bachelor thesis by Spilková [218].
     - David Celný was the consultant responsible for the PC(P)-SAFT EoS, Python interface, and equilibrium calculation methods, and provided consultations related to programming and EoS.
   - Equation and surface tension modeling for HFE are also present in the work by Vinš *et al.* [243].
     - David Celný was a co-author, contributed the equation of state, supported the surface tension calculation, and wrote the portions of the text concerning EoS and density gradient theory.

2. **Supersonic Expansion Research**

   - An independent molecular dynamics package was successfully developed, primarily designed for GPU utilization. An existing model from Klíma and Kolafa [124] was extended and optimized for the MD package. The GPU implementation is 10× faster than the previous conventional CPU code. This acceleration was achieved for the double precision calculation, guaranteeing sufficient energy conservation of the developed model. The results showing simulations for systems with 2048 and 4096 molecules over 1 µs as well as the comparison with experiments have been presented in the work by Celný *et al.* [41].
     - David Celný was the leading author, extended the model for use on GPUs, implemented most of the method, performed part of the simulations, and produced most of the text and figures.
   - The algorithms proposed were further optimized into the form presented in this work, leading to a substantial speed-up forty times faster than the software used in [124]. This has allowed us to simulate systems with 10240 molecules and enabled further research with a focus on cluster temperature evolution and the effect of pickup cross-sections presented in the work by Klíma *et al.* [123].
     - David Celný was a co-author, contributed the optimized parallel method, implemented most of the parallel method, helped with the simulation, and wrote the portion of the text concerning the use of GPUs together with one figure.

3. **Metastable Properties Research**

    - The problem of obtaining properties of metastable states has been thoroughly investigated, leading to the development of universal grid criteria for the detection of voids and clusters. This allowed efficient detection of the onset of nucleation during the runtime of the molecular dynamics simulation, keeping the simulated system within metastable conditions. The developed criteria were implemented into parallel simulation software *ms*2. The behavior of the initial version of criteria was examined in the publication by Fingerhut *et al.* [73].
        - David Celný was a co-author, contributed the cluster criteria, implemented the whole cluster criteria method into the package, helped with *ms*2 package development, performed all simulations related to cluster criteria, and wrote the portion of the text concerning the cluster criteria with related figures.
    - The developed criteria were integrated with the thermodynamic property package TREND and the simulation package *ms*2. This enabled the sampling of valid metastable state conditions. With the refinement of criteria parametrization, both metastable vapor and more scarcely investigated metastable liquid were successfully sampled. Results produced from the designed method were compared with EoS and literature data for the case of LJF. This part of the research is submitted by Celný *et al.* in [42].
        - David Celný was the leading author, developed the model, implemented the complete method, helped with the development of TREND, performed all the simulations, and produced the main body of the text and all figures except the deviation graphs where he provided the data.

1.5 Future Research

Beyond this thesis and its results, there exist related intriguing problems that require attention. The tools developed for phase interface research are already being applied to study more complex substances. In light of this, we propose areas concerning supersonic expansion and metastable properties research.

1. **Supersonic Expansion Research**

    - According to our preliminary investigation and the currently available experimental data, there is a need for development of efficient parallel implementation of the method for multicomponent systems. This presents a challenge, given that certain optimizations were built with the assumption of a single component system. Nevertheless, achieving this will be an accomplishment, which will enable comparisons with recent studies and lead to collaboration with researchers utilizing different apparatus designs.
    - Further research is also required to provide a full explanation of the creation mechanism behind the clusters with high sphericity from the work of Lengyel *et al.* [142].
    - There is also potential for researching the assumptions made by CNT. This regards the bouncing phenomenon and the merging of larger clusters. This research area may hold the answers to modifications of CNT necessary for better alignment of nucleation theory with molecular simulation and experiments.

2. **Metastable Properties Research**

   - In the area of metastable properties research, a concern for the immediate future involves expanding the methodology to encompass more complex real fluids.
   - Furthermore, there is potential for investigating spinodal decomposition and further extension allowing for the simulation at a wider range of metastable conditions.
   - The third direction concerns the inclusion of the fitting algorithm for the EoS. This would enable the development of an overarching method for improving the EoS. An improvement in metastable regions would have a direct impact on mixture modeling and phase interface modeling.

Thermodynamics of nucleation 2

Thermodynamics is a branch of physics with extension into chemistry that focuses on the relationships between heat, work, temperature, and energy. This is by no means limited in scope, as effects related to heat and temperature constitute a prominent portion of the physical world. Because of this, thermodynamics finds its use in many scientific and engineering fields, including but not limited to chemistry, mechanical engineering, and space research.

With a long history spanning from Sadi Carnot and Lord Kelvin, thermodynamics developed into a multidisciplinary field area encompassing multiple branches of thermodynamics, e.g., classical, statistical, chemical, and biological. Given this wide scope, thermodynamics is no longer a purely physics discipline but relies on mathematical modeling and molecular simulations to tackle the current problems. One of these problems is the process of droplet formation, a common occurrence in nature with profound implications in multiple areas like atmosphere modeling, power generation, surface treatment, and the modern topic of carbon capture and storage. The process of droplet formation has a long tradition, with initial observations dating back to Fahrenheit [69], followed by investigations from Gibbs [82] over the currently utilized kinetic theory by Volmer and Weber [249] into modern day, where scientists are trying to understand the precise working principle and conciliate experimental measurements with theoretical prediction.

Scientific interest in nucleation persists, driving further theoretical investigations that currently reach over $2,400,000$ results on Google Scholar. One specific area of focus is the extension of understanding homogeneous nucleation principles in fluids, which is also the field of our study.

During the research of phase interfaces and nucleation, it became apparent that further research of metastability is required. This led us into our current state, where metastability is presented as a precursor for nucleation.

The aim of this chapter is to provide a review of theoretical concepts required for understanding metastability and nucleation research presented in later parts of this thesis. We start with an introduction of the thermodynamic formalism, which is then used to summarize the knowledge about thermodynamic potentials, equations of state, and phase. These concepts are utilized to introduce the effects of phase interface, metastability, and finally homogeneous nucleation with its primary theory, classical nucleation theory (CNT).

2.1 Concepts of thermodynamics

The fundamentals start with the formalization of the term *system*. In most generic thermodynamic sense, *system* is a portion of physical space separated from its surroundings by a boundary. With a permeable boundary, the system is called *open system*, and conversely, an impervious boundary to any flow of energy or matter results in *isolated system.*

In thermodynamics, an independent set of *state variables* is used to describe *system state* in an instance of time. Examples of state variables include experimentally well-known properties like *temperature* T with unit of [K], *pressure* p [Pa] or *volume* V [m^3]. Volume is often expressed as molar quantity: *molar volume* V_m [m^3/mol], where the Avogadro constant is used to define the number of molecules contained in one mol. The *state function* relates multiple state variables to provide further characterization of the system state, e.g., internal energy U, entropy S, and entalphy H. All of the state functions are independent of the path the system has taken, contrary to *path functions*, which are dependent, like mechanical work W and heat Q. This distinction is particularly important for the characterization of static and dynamic processes.

For any closed system or open system, where boundary conditions are constant for long enough time without any macroscopical change of matter and energy, the system will settle in the entropy maximum according to the maximum entropy principle. System in this conditions and sometimes only the conditions are then called *thermodynamic equilibrium.*

Another important concept is that of *phase* which denotes a region with the same uniform chemical and physical properties. There are confusion with state of matter which are physically distinctive (like gas, liquid, solid, and plasma). But phases are more general terms with patterns, that can be microscopically observable, leading to the meaning of phase used in this thesis as a distinct microscopically observable uniform structure.

It is important to realize that system states can be described by different combinations of state variables and state functions. The number of independent thermodynamics degrees of freedom can be formulated in the case of equilibrium systems with the *Gibbs rule*, relating the degrees of freedom count to the number of components and phases as follows:

$$\mathrm{DoF} = n_\mathrm{comp} - n_\mathrm{phase} + 2 \tag{2.1}$$

Lastly the concept of *thermodynamic ensemble* is formalized as a collection of a large number of systems in different states with common macroscopic attributes. The thermodynamic ensemble is used to describe the statistical behavior of a system, and it is characterized by a set of macroscopic variables. Examples of include microcanonical (NVE), canonical (NVT), and grand canonical (μVT) ensemble, where variables in parentheses are the fixed properties like number of molecules N, total energy E, and chemical potential μ.

2.1.1 Thermodynamic potentials

In this section, we will start from the knowledge of the laws of thermodynamics following [37,131] with the formulation of internal energy differential for system with n_comp components where only work considered is the pressure-volume work:

$$dU = TdS - PdV + \sum_{i=1}^{n_\mathrm{comp}} \mu_i dN_i \tag{2.2}$$

where internal energy is a function of temperature, pressure, and the chemical potentials of individual components μ_i. These variables are also called conjugate to the natural variables of internal energy which are entropy, volume and number of molecules of each component $i \in \widehat{n_\mathrm{comp}}$. This provides the first thermodynamic function (further called thermodynamic potential), which relates entropy, volume, and number of molecules to produce the total energy of the system. With the aforementioned redundancy, we can use the Legendre transformation to change variables and produce a set of known thermodynamic potentials

$$A(T,V,\boldsymbol{N}) = U - TS \qquad \rightarrow \qquad dA = -SdT - PdV + \sum_{i=1}^{n_{\text{comp}}} \mu_i dN_i \tag{2.3}$$

$$H(S,p,\boldsymbol{N}) = U + pV \qquad \rightarrow \qquad dH = TdS + Vdp + \sum_{i=1}^{n_{\text{comp}}} \mu_i dN_i \tag{2.4}$$

$$G(T,p,\boldsymbol{N}) = U - TS + pV \qquad \rightarrow \qquad dG = -SdT + Vdp + \sum_{i=1}^{n_{\text{comp}}} \mu_i dN_i \tag{2.5}$$

$$\Omega(T,V,\boldsymbol{\mu}) = U - TS - \sum_{i=1}^{n} \mu_i N_i \qquad \rightarrow \qquad d\Omega = -SdT - pdV - \sum_{i=1}^{n_{\text{comp}}} N_i d\mu_i, \tag{2.6}$$

where A is the Helmholtz energy, H is the enthalpy, G is the Gibbs energy and Ω is the grand potential.

Performing two types of Legendre transformation is the maximum number possible, leading to a new potential. In the case where all properties are transformed and the differential $dU = TdS - pdV + \sum_{i=1}^{n} \mu_i dN_i$ is applied, the following equations are obtained:

$$U = TS - pV + \sum_{i=1}^{n_{\text{comp}}} \mu_i N_i \tag{2.7}$$

$$dU = TdS + SdT - pdV - Vdp + \sum_{i=1}^{n_{\text{comp}}} \mu_i dN_i + \sum_{i=1}^{n_{\text{comp}}} N_i d\mu_i \tag{2.8}$$

$$\sum_{i=1}^{n_{\text{comp}}} N_i d\mu_i = -SdT + Vdp \tag{2.9}$$

The last equation is called the Gibbs-Duhem equation and relates the chemical potential changes to temperature and pressure changes.

Continuing with the derivation of thermodynamic relations, the form of potentials with new variables leads to the partial derivatives relations, which can be formulated with respect to the natural variables as shown for the example of Helmholtz energy (analogous can be performed for other potentials):

$$S(T,V,\boldsymbol{N}) = -\frac{\partial A}{\partial T}(T,V,\boldsymbol{N}) \tag{2.10}$$

$$p(T,V,\boldsymbol{N}) = -\frac{\partial A}{\partial V}(T,V,\boldsymbol{N}) \tag{2.11}$$

$$\mu_i(T,V,\boldsymbol{N}) = \frac{\partial A}{\partial N_i}(T,V,\boldsymbol{N}) \tag{2.12}$$

Notice eq. (2.11), which provides a formula for the pressure of the system from the Helmholtz energy with is further used in Helmholtz formalism.

These equations are not the only relations; from the comparison of inter-changeable second partial derivatives, Maxwell relations are produced. We show only a selection of all $(n_{\text{pot}}(n_{\text{pot}}-1)/2)$ possible combinations:

$$-\frac{\partial p}{\partial S}(S,V,\boldsymbol{N}) = \frac{\partial T}{\partial V}(S,V,\boldsymbol{N}) \tag{2.13}$$

$$\frac{\partial S}{\partial V}(T,V,\boldsymbol{N}) = \frac{\partial P}{\partial T}(T,V,\boldsymbol{N}) \tag{2.14}$$

$$\frac{\partial S}{\partial P}(T,P,\boldsymbol{N}) = \frac{\partial V}{\partial T}(T,P,\boldsymbol{N}) \tag{2.15}$$

We notice again some useful equations based on partial derivatives that allow the calculation of potentially problematically measured entropic properties. A more generic approach is presented next in section 2.1.1.1.

2.1.1.1 Helmholtz energy formalism

Similar observations led to the realization that system information can be obtained from partial differentiation of the thermodynamic potentials. This concept has strong theoretical meaning, but for practical utilization, the caloric property derivatives are hard to perform. With the previous section and eq. (2.10), an alternative formulation using Helmholtz energy is used instead, introducing the following notation generally accepted in the EoS fitting community:

$$A_{i,j} = \frac{\partial^{i+j} A}{\partial^i T \partial^j V} \tag{2.16}$$

In this formalism, thermodynamic properties can be written as a combination of $A_{i,j}$ terms, assuming that Helmholtz energy is a sufficiently differentiable function and $i, j \in \mathcal{N}_0$. In the case when index is zero, the particular partial derivative is avoided, which means the original Helmholtz energy is $A_{0,0}$. With this formalism, the formulas from the previous section for pressure, entropy, internal energy, enthalpy, and Gibbs energy are easily written as:

$$p = -A_{0,1} \tag{2.17}$$

$$S = -A_{1,0} \tag{2.18}$$

$$U = A_{0,0} + TA_{1,0} \tag{2.19}$$

$$H = A_{0,0} + TA_{1,0} - VA_{0,1} \tag{2.20}$$

$$G = A_{0,0} - VA_{0,1} \tag{2.21}$$

Lustig formalism [151] then provides connection with molecular simulation, allowing to express Helmholtz energy partial derivatives from simulated ensemble properties. Combination of both formalisms is a powerful tool widely used in the field of equation of state fitting and thermodynamic properties modeling.

The comprehensive list of thermodynamics relations using Helmholtz energy formalism can be found in book by Spaan [215]. In the reference reduced temperature, eq. (2.35), reduced density, eq. (2.36), as well as separation of Helmholtz energy into ideal and the residual part, eq. (2.37), were used to list the thermodynamic properties and formulas using the formalism.

2.1.2 Equation of state

The equation of state (EoS) is a function that relates state variables for a given set of physical conditions (thermodynamic state). An easy example was already shown in eq. (2.11), which gives a formula for pressure from the knowledge of the Helmholtz energy. This and similar formulations are used to describe the properties of pure substances and mixtures in liquids and gases, generally denoted as fluids. Equations of state have a wide range of applications and uses in many fields, including fluid modeling, a wide range of engineering and industrial applications (any systems operating with fluids), materials science, and chemistry. It is important to note that a wide variety of equations of state are available, each of which is intended to address a particular scenario. In this thesis, we show, cubic equation, SAFT (Statistical Associating Fluid Theory) equation, and the multiparameter equation. All of the introduced equations have been used in this thesis.

2.1.2.1 Van der Walls EoS

The van der Waals equation of state (vdW EoS) is an equation that extends the ideal gas law to account for the effects of intermolecular forces and the finite volume of gas molecules. It was developed by Johannes Diderik van der Waals in 1873 [236].

$$p = \frac{RT}{V_\mathrm{m} - b} - \frac{a}{V_\mathrm{m}^2}, \tag{2.22}$$

where R is gas constant and pressure p is expressed terms of temperature T, and molar volume V_m. This equation established the family of cubic equation of state, because of cubic polynomial form of $0 = f(p, V, T)$. This is easily seen when fractions are removed from the equation and everything is transferred to right hand side. There are alos two main substance–specific parameters a, b in the VdW equation. Their purpose is to represent the substance behavior with the attraction parameter a and can be related to the critical point properties of the liquid with

$$a = \frac{27(RT_\mathrm{c})^2}{64p_\mathrm{c}}. \tag{2.23}$$

Second is the repulsion parameter b creating the effect of a finite volume of gas, expressed again in therms of critical temperature and pressure as

$$b = \frac{RT_\mathrm{c}}{8p_\mathrm{c}}. \tag{2.24}$$

This equation holds strong theoretical meaning because many thermodynamic properties are analytically calculable with a, b as parameters because of its simple shape. VdW EoS is also used as an approximation of non-ideal gas behavior. The equation has a known deficiency in prediction in high pressure and temperature regions as well as liquid phase prediction. But its beauty is in the simplicity and ease of calculation.

2.1.2.2 Peng Robinson EoS

The Peng-Robinson equation of state (PR EOS) is a cubic equation of state developed by Ding-Yu Peng and Donald Robinson in 1976 [188]. The standard form of PR EoS is:

$$p = \frac{RT}{V_\mathrm{m} - b} - \frac{a\kappa}{V_\mathrm{m}^2 + 2bV_\mathrm{m} - b^2} \tag{2.25}$$

This equation has two main thermodynamical parameters, a and b, similar in meaning to VdW EoS. Parameters are also adjusted to the critical properties of the modeled fluid. An additional structural parameter, κ is related to the acentric factor, ω[1] which characterizes the non-spehricity of the molecule of the modeled fluid.

$$a = 0.45724\frac{(RT_\mathrm{c})^2}{p_\mathrm{c}} \tag{2.26}$$

$$b = 0.07780\frac{RT_\mathrm{c}}{p_\mathrm{c}} \tag{2.27}$$

$$\kappa = \left(1 + \omega\left(1 - \sqrt{\frac{T}{T_\mathrm{c}}}\right)\right)^2 \tag{2.28}$$

[1] The factor is defined as $\omega = -\log(p_\mathrm{eq}/p_\mathrm{c})$ calculated at the reduced termperature equal to $T/T_\mathrm{c} = 0.7$

In this form, PR is a versatile EoS that one can use as a good approximation or valid model based on knowledge of critical parameters and the molecular structure of the targeted fluid. The structure parameter in particular enhances the predictive capabilities of the VdW.

The simple structure enables further extension: Important to mention is the volume translation of PR (VTPR) proposed by Péneloux *et al.* [187], which improves the prediction of liquid density.

This relatively simple model is good for the calculation of non-polar pure fluid properties. The equation provide good prediction for the critical point and is even sufficient for vapor-liquid equilibrium calculations when VTPR is used. This makes it a suitable equation for alkane modelling [153].

Adding simple mixing rules for the parameters a, b

$$a = \sum_i \sum_j x_i x_j a_{i,j} \tag{2.29}$$

$$b = \sum_i \sum_j x_i x_j b_{i,j}, \tag{2.30}$$

where x_i, x_j are the mole fractions of individual components and $a_{i,j}, b_{i,j}$ follows the van der Waals combining rules [235] with optionally one or two additional adjustable mixture pair–specific parameters $k_{i,j}, l_{i,j}$.

$$a_{i,j} = \sqrt{a_i a_j}(1 - k_{i,j}) \tag{2.31}$$

$$b_{i,j} = \frac{b_i + b_j}{2}(1 - l_{i,j}) \tag{2.32}$$

enables the calculation of mixture properties based only on the pure substance knowledge, optionally adjusted with $k_{i,j}, l_{i,j}$ parameters fitted to avaliable experimental data. The simple equation allows for fast calculation, enabling evaluation of multicomponent mixtures, which is the case in the petroleum industry, where the equation still finds its main use [2, 10, 169]. Other areas PR is actively used are investigated by Shardt *et al.* [213].

2.1.2.3 PC-SAFT EoS

The Perturbed-Chain Statistical Associating Fluid Theory (PC-SAFT) equation of state is a member of the SAFT family of EoS developed by Chapman *et al.* [47]. Perturbation theory has been used to quantify the relationship between well-defined site-site interactions of bulk liquid behavior [257]. The Helmholtz energy has been expanded into a series of integrals of the molecular distribution function and the associated potential, which, with the use of perturbation theory, was simplified into the following form based on the individual contributions:

$$A = A^{\mathrm{ig}} + A^{\mathrm{res}} = A^{\mathrm{ig}} + A^{\mathrm{seg}} + A^{\mathrm{chain}} + A^{\mathrm{assoc}} \tag{2.33}$$

Here the first split of the Helmholtz energy into ideal and residual parts is performed, which is then continued for the residual part according to Chapman [48] into terms for segments, chains, and association interactions (relevant for substances exhibiting association, e.g., alcohols).

For PC-SAFT developed by Gross and Sadowski, [87] Wertheim's perturbation theory was used on the hard sphere reference fluid equation from Boublík [28] and Mansoori [157] and rewrote the Helmholtz energy as expansion around the hard sphere fluid:

$$A = A^{\mathrm{ig}} + A^{\mathrm{hs}} + A^{\mathrm{hc}} + A^{\mathrm{disp}} \tag{2.34}$$

Here the terms represent the hard sphere reference, the hard chain, and the dispersion term containing the rest of the perturbation. Where the original integration was either solved exactly, i.e., the hard sphere and hard chain terms, or approximated, as is the case for the dispersion term. The individual formulations are provided in the appendix of the original paper from Gross and Sadowski [87].

This approach is also readily extensible, as new interactions can be incorporated in terms of Helmholtz energy contribution. In this way, dipole interaction A^{dpol} by Gross and Vrabec [90], quadrupole interaction A^{qpol} by Vrabec and Gross [250] and association interaction A^{assoc} by Gross and Sadowski [89] were all introduced into the equation. The SAFT equation accounitng for polarity is often called PCP-SAFT

This construction allows for a large quantity of fluids to be modeled by SAFT type equations, where equation building relies only partially on experimental data to adjust the empirical parameters within the equation. In the end, a relatively low amount[2] of parameters is required to capture the behavior of very complicated fluids like coolants (for example Hydrofluoroethers investigated by Vinš *et al.* [243]). Given the complicated structure of the numerical evaluation of some integrals, the equation's performance is its weak point. There is also reliance on experimental data for adjustable parameter evaluation, which may be an issue with scarcely measured substances. On the other hand, the equation provides very accurate results for pure fluid and viscous liquid equilibria for many classes of interaction, which makes it popular in both the scientific and engineering areas, as certified by thermodynamic packages implementing the equation like Coolprop [22] or TREND [216].

During the research on nucleation, I wrote my own implementation of PC-SAFT in FORTRAN 95, aiming for fast execution by utilizing internal equation state switching to prevent unnecessary recalculation of the coefficients. The equation has been utilized for substance modeling and as a reference equation of state for experiments performed at the Laboratory of Thermophysical Properties of the Institute of Thermomechanics of the CAS .

2.1.2.4 Multiparameter EoS

During the introduction of the previous equation, theoretical knowledge was primarily used to construct the mathematical formulation that would closely represent the data obtained from experimental measurements and computer simulation. But this is not the only option, as the equation can be constructed purely to reflect the experimental data and adjusted later to account for known behavior. In this way, no precise molecular model is required, as the heavy lifting is performed by a wide range of experimental data used for reference of the behavior of the examined fluid.

For the following description, we will adhere to the formalism of critically scaled temperature and density with consequent application on Helmholtz energy

$$\tau = \frac{T_{\mathrm{c}}}{T} \tag{2.35}$$

$$\delta = \frac{\rho}{\rho_{\mathrm{c}}} \tag{2.36}$$

$$\alpha(\tau,\delta) = \frac{A(T,\rho)}{RT} = \frac{A^{\mathrm{ig}}(T,\rho) + A^{\mathrm{res}}(T,\rho)}{RT}. \tag{2.37}$$

Notice that τ is an inverted reduced temperature. This helps in simplificaiton of notation in eq. (2.38) where all terms are expressed as multiplicative factors.

The equation construction is based on data collection from all available experiments, where the particular focus is given to VLE data, heat capacities, and speed of sound measurements. All of these data sets aid in fitting to obtain Highly accurate equation. Vapor-liquid equilibrium

[2] ten parameters are used to represent a non associating polar fluid

measurements capture equilibrium properties and the relationship between vapor and liquid. In particular, pressure-density information helps in fixing the interfacial curve prediction. Heat capacities are important because of their relation to Helmholtz energy as a second derivative, which provides information about the curvature of the Helmholtz function. The speed of sound is then an important verification parameter, combining multiple derivatives of Helmholtz energy, which makes it very sensitive to errors in the Helmholtz energy function.

The equation then combines the data into a weighted set, on which an optimization algorithm searches for the best choice of adjustable parameters. The Helmholtz energy is split according to eq. (2.37) into an ideal gas part and a residual part.

For category of simple fluids, the ideal gas part A^{ig} can be analytically obtained or obtained by the use of quantum simulation and incorporated into the equation independently as a fluid parameter. For complex fluids, one can use the approximation with a rigid harmonic oscillator modeled according to Span [215]. This model offers a way how to calculate parametrization from the ideal gas part form isobaric heat capacity experimental data. For lower temperatures, this ideal gas part model provides a sufficiently accurate prediction. In higher temperatures, further corrections are required [215]. This construction allows independent inclusion of the ideal gas part of Helmholtz's energy without the need for unified fitting of whole Helmholtz energy.

To remove the shape optimization problem from the fitting procedure, the from of the residual part of the equation is fixed to contain n_{pol} polynomial terms, n_{exp} exponential terms, and n_{Gauss} Gaussian terms:

$$A_{\mathrm{r}}^{\mathrm{res}}(\tau,\delta) = \sum_{i=1}^{n_{\mathrm{pol}}} n_i \tau^{t_i} \delta^{d_i} + \sum_{i=1}^{n_{\mathrm{exp}}} n_i \tau^{t_i} \delta^{d_i} \exp(-\delta^{l_i}) + \sum_{i=1}^{n_{\mathrm{Gauss}}} n_i \tau^{t_i} \delta^{d_i} \exp\left[-\mu_i(\delta-\epsilon_i)^2 - \beta_i(\tau-\gamma_i)^2\right] \tag{2.38}$$

where the set of parameters $d_i, t_i, l_i, \mu_i, \epsilon_i, \beta_i, \gamma_i$ is to be determined for the specified choice of the number of individual terms, i.e., $n_{\mathrm{pol}} = 6, n_{\mathrm{exp}} = 6, n_{\mathrm{Gauss}} = 11$ used in the equation for Lennard-Jones fluid by Thol *et al.* [229]. For this task, an optimization algorithm looks for a sufficiently low local minimum in the high-dimensional parametric space. Not much is known beforehand about the error function being minimized, which makes it an arduous process, where the human insight and experience play an important role in constraining the parametric space. Expertise in fitting leads to the creation of an equation of state that represents physically measured behavior while agreeing with the theoretically known properties of the equation of state.

When all terms are tabulated, the evaluation of the equation is fast and can provide predictions over a wide ranges of conditions for which experimental data are available. The primary feature of multiparameter equations of state is their high precision, which is of particular importance for a wide range of practical applications, making these equation particularly interesting for industrial utilization. Therefore, a lot of effort is put into making these equation not only precise in regions of data availability but also to extrapolate well outside of measured data.

Multiple van der Waals loops This brings the discussion to the multiple van der Waals (vdW) loops issue of Multiparameter EoS. The term van der Waals loop or alternatively Maxwell loop refers to the behavior of the equation of state inside binodal region of the phase diagram, where the calculated isotherm exhibits a single sinusoidal swing like in the case of vdW EOS. When the isotherm exhibit more than one of these swings or even discontinuities we call it equation with multiple van der Waals loops [215]. Example of multiple vdW loops is shown in fig. 7.4. It can be argued that multiple loops are the consequence of the employed fitting technique [258] or may be the result of a lack of knowledge in the metastable region. The underlying change in nature from liquid and vapor behavior is therefore not captured well enough, leading to erratic swings in the two-phase region. This issue can be safely neglected in the case of pure fluid evaluation, but in the case of mixtures, it is not necessarily the case [258]. Therefore, discrepancies around the phase equilibrium line can lead to the incorrect prediction of the phase equilibrium of mixtures. This is currently viewed as drawback of many multiparametric equations of state motivating researchers to develop methods to diminish this issue of multiple van der Waals loops like Pohl *et al.* [194].

2.2 Concepts of phase

The phase introduced in section 2.1 is here related to thermodynamic potential to derive the globally observed process. In particular we are interested in phase transitions. The internal structure, or microstate, is not of particular concern in this description, as the focus is on the static systems found in equilibrium.

We start the description from phase equilibrium, illustrating the driving force behind the transition and the related description of stability. In latter part of the section the effect of curvature is introduced.

2.2.1 Phase transition

Let us first start with a system in phase equilibrium at fixed pressure and temperature. Following the reasoning of [94], this system is best described with the Gibbs potential which for the equilibrium should occupy its minimum according to the minimum principle [212]. For the system setting, the total Gibbs potential can be written as the sum of individual contributions for all n_{phase} for n_{comp} as:

$$G = \sum_{i=1}^{n_{\text{phase}}} G^{(i)} = \sum_{i=1}^{n_{\text{phase}}} \sum_{j=1}^{n_{\text{comp}}} \mu_j^{(i)} N_j^{(i)} \tag{2.39}$$

Because the number of particles of one component is not changing, it is always possible to express one selected j' as a complement to the rest of the components. Utilizing this together with the Gibbs potential alternative to eq. (2.12), the condition for n_{eq} phases in equilibrium reads as:

$$\mu_j = \left(\frac{\partial G}{\partial N_j}\right)_{p,T,N_{j'\neq j}} = \left(\frac{\partial G}{\partial N_j^{(i)}}\right)_{p,T,N_{j'\neq j},N_j^{(i'\neq i)}} = \mu_j^{(i)} \tag{2.40}$$

given that the choice of j is arbitrary, we obtain the equality of chemical potentials of a given component across all n_{eq} coexisting phases. This has importance not only for the equilibrium characterization but also for the process leading to equilibrium.

Let us now restrict ourselves to the simple case of two-phase coexistence. With the knowledge of the Gibbs phase rule in eq. (2.1), we know that for a single component and two phases, only one degree of freedom remains, i.e., one independent property describes the phase coexistence. The system temperature and pressure are important controlling factors, and it is also possible that under some parameters, no coexistence of phases is possible. This is illustrated in fig. 2.1, where the jump in chemical potential is visible between the gas and liquid regions of the curve. The case where no coexistence is observed is visible with point c.

Figure 2.1 illustrates three scenarios of transitions. The continuously differentiable path (a) where no phase transition occurs because the system is found in the supercritical region. The critical case with critical point (c) where first derivative is continuous but there is discontinuity in second derivative shown in the change in curvature. Last example is first order transition commonly known as phase transition with path through points (b,d,e,b). In this case there is

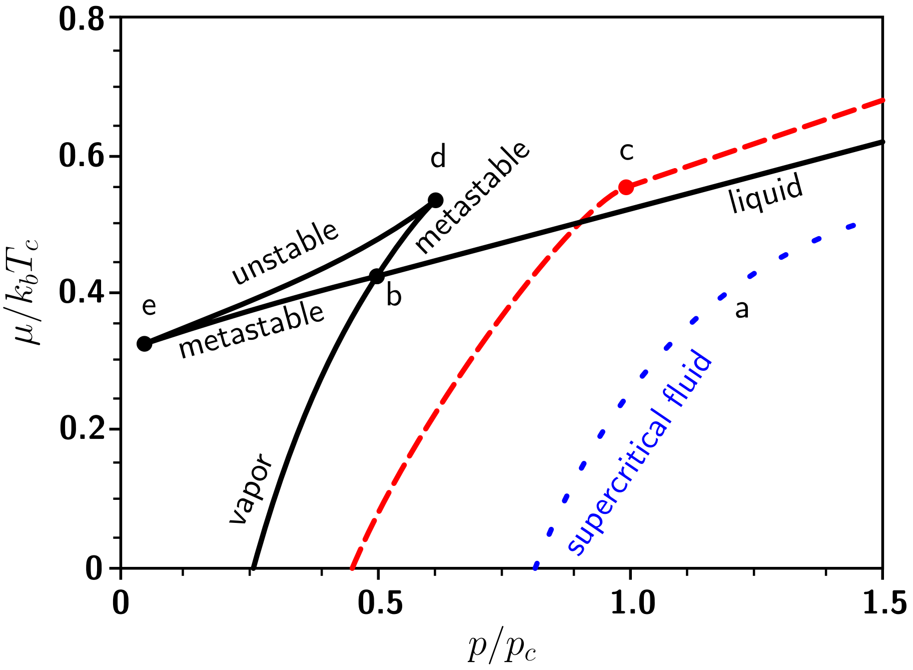


Figure 2.1: Shape of the chemical potential for changing reduced pressure. Van der Waals EoS is used to show supercritical fluid $T/T_\mathrm{c} > 1$ (blue dotted line), critical fluid $T/T_\mathrm{c} = 1$ (red dashed line) and below critical fluid $T/T_\mathrm{c} = 0.85$ (black line) transitions. The imporant crossing points are marked with letters as well as regions related to the stability of the fluid.

discontinuity also in the first derivative shown with the point (b) that is being crossed twice. The first crossing is from the vapor branch of the curve, which extends until point (d) and the second crossing is from the liquid branch, which extends into point (e). For the the last case a stability regions were also denoted, which becomes relevant in section 2.3.1 where the meaning of points (b,d,e) is explained.

2.2.2 Driving force

To answer the question of what force drives the establishment of phase equilibrium, system potential like the ones introduced in section 2.1.1 needs to be chosen. For the case of constant pressure and temperature, the Gibbs energy from eq. (2.5) is a good candidate.

$$\Delta G = G_\mathrm{old} - G_\mathrm{new} = (\mu_\mathrm{old} - \mu_\mathrm{new})N = N\Delta\mu \tag{2.41}$$

Here the subscripts are phase-independent, denoting the direction of transition. For a constant number of particles, we see that the chemical potential difference is the driving force of the transition. For further derivation, we need to specify the phases. In this study, we focus on fluids and consider the transition from gas (also called vapor) into liquid and vice versa. These transitions are further denoted with arrows, i.e., vapor→liquid.

The assumption of ideal gas behavior of the vapor phase and incompressible liquid model [249] can now be used to formulate a substance–agnostic approximation of the chemical potential difference. This is one of the accepted approximations for the driving force. In situation where the assumptions are not fulfilled ideal gas equation can be replaced with a more precise EoS from section 2.1.2. An alternative formulation can also be made using fugacities, but for the purposes of our introduction, the commonly used formulation is sufficient. The driving force

$$\Delta\mu \approx k_\mathrm{B}T\ln(p/p_\mathrm{eq}) - v_1(p - p_\mathrm{eq}), \tag{2.42}$$

where $p_\mathrm{eq} = p_\mathrm{eq}(T)$ is the temperature dependent equilibrium pressure and v_1 is volume of single molecule. Further simplification can be made for cases $v_1 p_\mathrm{eq} \ll k_\mathrm{B}T$, where the second term can be removed. This leads to a form one can usually see in literature [116, 118, 239], which reads

$$\Delta\mu_\mathrm{v} = k_\mathrm{B}T\ln(p/p_\mathrm{eq}) = k_\mathrm{B}T\ln(S). \tag{2.43}$$

Here S is called saturation, a dimensionless number used to specify how far from equilibrium the system resides. This is also indirectly related to the time frame in which phase transitions occur.

In the case $S = 1$, the chemical potential difference is zero and no phase change occurs. In the case of $S < 1$, the system does not even have phase coexistence as there is no promoting force to create it with. This case has a theoretically negative chemical potential difference. In the last case where $S > 1$, we talk about supersaturated system, in which phase transition is promoted. The magnitude of the supersaturation increases until a condition of instability is reached. In the region of instability, the term supersaturation loses its physical meaning, as discussed in the next section.

2.2.3 Phase equilibrium considering interface curvature

The phase equilibrium in general imposes a set of equalities of the chemical potential of each component in the liquid and vapour phases. The simplest case is the planar interface, for which the pressures of both phases have to be equal, i.e., $p_\mathrm{v} = p_\mathrm{l}$. The set of algebraic equations can be iteratively solved for thermodynamic conditions of phase equilibrium at a given temperature and total pressure in this case.

In case of spherical interface, the Young-Laplace equation replaces the equality of pressures. The pressure equation is formulated with the help of Laplace pressure Δp describing the mechanical equilibrium of a surface separating two phases

$$\Delta p = 2\sigma/r_\mathrm{curvature}\,. \tag{2.44}$$

Here, $r_\mathrm{curvature}$ is the radius of curvature and σ denotes the surface tension. The Laplace pressure is a difference between the pressure inside a spherical cluster, i.e., a droplet or a bubble, and within the surrounding phase $\Delta p = p_\mathrm{l} - p_\mathrm{v}$. The Laplace pressure can be used for the definition of the supersaturation of the metastable bulk phase. This means the supersaturation provides some sort of information about how far from binodal the system is.

From the model perspective, the Laplace pressure enables to form a description of curved phase interface. Therefore, using the Laplace pressure formulation for equilibrium calculation, a generalized set of equations applicable on both the planar and the spherical phase interfaces is expressed as follows

$$\begin{aligned} \mu_1^\mathrm{l} - \mu_1^\mathrm{v} &= 0 \\ &\vdots \\ \mu_\mathrm{n}^\mathrm{l} - \mu_\mathrm{n}^\mathrm{v} &= 0 \\ ap_\mathrm{l} + bp_\mathrm{v} - c &= 0, \end{aligned} \tag{2.45}$$

where the pressure equation is parametrized by three parameters allowing for a universal definition of the solved profile geometry. Parameters a and b are dimensionless. Parameter c represents the

Laplace pressure, which equals zero for the planar case with an infinite radius of the interface curvature. Values of parameters a and b vary according to the definition of input quantities; e.g., for spherical interface and prescribed p_l, the parameters have values of $a = 1$, $b = -1$, and $c = \Delta p$.

2.3 Concepts of nucleation

Nucleation is the process of formation of a new thermodynamic phase within an existing phase of a different structure. This process is naturally occurring in every day life, for example, raindrops forming in the clouds or bubbles created in carbonated drinks.

In thermodynamic terms, nucleation is the process during first-order phase transition. The investigated case in this study restricts the transition between vapor↔liquid phases.

Building upon the concepts of phase transition, the introduction into system stability is given with illustration of construction of phase diagram and important stability curves. We also derive the work of formation as well as the critical cluster properties. All these concepts are utilized in section 2.4, where the derivation of the commonly used form of the classical nucleation theory (CNT) is performed.

2.3.1 Stability

Another important aspect of phase transition is the path taken to reach the desired state. In the case of the here investigated vapor↔liquid phase transition, the process is not immediate but undergoes several stages of progression. We can already see hints of this in fig. 2.1, where the gas phase line continues over point (b) until point (d) from which the transition through unstable region is made into liquid phase at (e), which then continues on the liquid branch through (b) into the stable liquid. If the process was immediate, everything would happen at point (b) without any detours like it can be seen for critical case illustrated with point (c). In this case a smooth transition is made between undetermined phases.

This phenomenon arises due to the presence of an energy barrier for the transition, as illustrated in fig. 2.2. For metastable states, the system resides in the locally convex portion of the graph, characterized by a higher attributed Gibbs energy. For metastable states the perturbations of system configuration may crossing the energy barrier which transforms the system into stable phase. The process of the transformation is called nucleation. Metastable system's behavior is determined by the magnitude of perturbation in comparison with the size of the energy barrier. This energy barrier is, in end effect, the energy required to form a transition structure from which the system can grow into the target phase (see configuration 3 in fig. 2.2).

Stability curve Analyzing the barrier height, we can restrict the metastable state to the conditions of the existence of an energy barrier. In this sense, the first limiting case is when no barrier exists because the system is already in phase equilibrium. This line is called correspondence line, and the condition for it is provided by eq. (2.40). For the system with a single component and two phases found in equilibrium, a fixed temperature results in a single free parameter. Binodal is then characterized, as a pressure curve $p_\mathrm{eq}(T)$.

Evaluation of binodal is usually performed by solving the alternative condition of equality of pressures $p_\mathrm{v}(T, \rho_\mathrm{v}) = p_\mathrm{l}(T, \rho_\mathrm{l})$ rather than chemical potentials, as the former is more readily available from the equation of state. With this search done across a range of temperatures, the two branches of the binodal are obtained. From this construction, it is obvious that the curves meet at the critical point where differences between phases vanish. The final shape is visible in fig. 2.3, shown as a black solid line connected at the critical point.

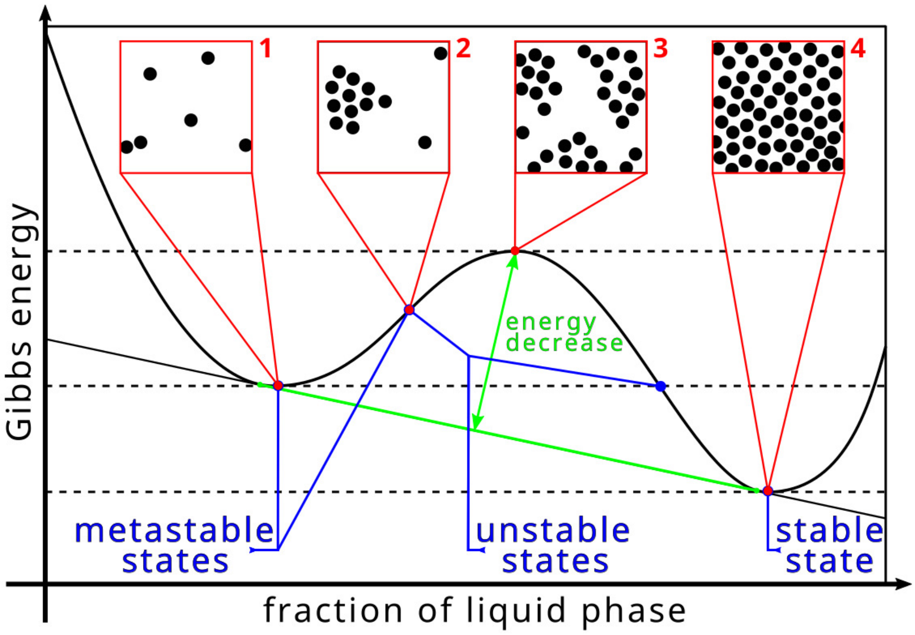


Figure 2.2: Illustration of system stability related to system configuration displayed in graph of Gibbs energy over mole fraction of liquid. Transition to a more favorable state in terms of Gibbs energy means decreasing Gibbs energy to the green tangent. For this transition the energy barrier has to be traversed. The figure also shows several configuration representing vapor system (1), metastable system with cluster (2), unstable system with phase transition occurring in whole volume (3) and finally the liquid system (4).

Instability curve The curve separating the metastable and unstable states is called instability line or spinodal. For its localization we have to consider situations, where the energy barrier is lowered to the point it vanishes. There are no obstruction for even the smallest perturbation to induce for transition which generates a chaotic non-localized phase transitions throughout whole system. In terms of underlying Gibbs potential, the previous local minimum becomes an inflection point while the global minimum is preserved. With the Gibbs potential as a function of pressure at a fixed temperature and a constant number of particles, the original condition can be transformed with the knowledge of the Gibbs potential differential, eq. (2.5), into:

$$0 = \frac{\mathrm{d}^2 G}{\mathrm{d}\rho^2} = \frac{\mathrm{d}p}{\mathrm{d}\rho}. \tag{2.46}$$

Therefore, examining the local extrema of the isotherm $p(T, V)$ can be used to construct both of the spinodal curve branches. The procedure for finding the extrema is done using stable vapor and liquid densities and finding the first extremum point, as illustrated in fig. 2.3. In this way, the artifacts of EoS with multiple Maxwell loops are avoided. The spinodal curves (black dashed line in fig. 2.3) are also joined at a critical point where there is no barrier separating the coexisting phases, as shown in fig. 2.1 by point (c).

In this way, the phase diagram in reduced pressure and reduced density can be constructed from isotherms and consequently separated into stable, metastable, and unstable regions like in fig. 2.3. The binodal, spinodal, and critical isotherm $p(T_\mathrm{c}, \rho)$ are used as separation curves. The vapor region is then restricted by the vapor binodal and critical isotherm. The stable liquid region is found between critical isotherm and the liquid binodal. In this case, the solid phase

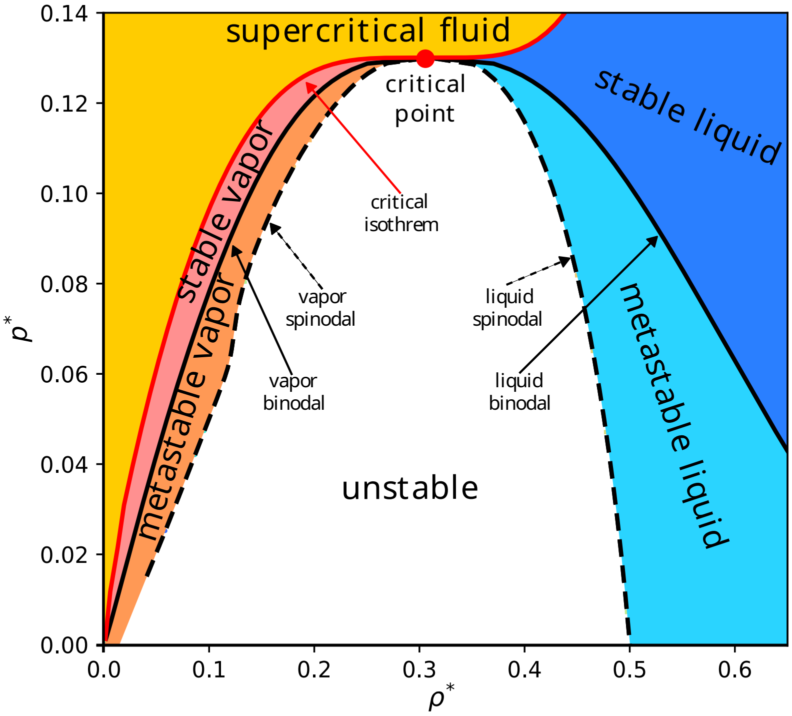


Figure 2.3: Phase diagram of the Lennard-Jones fluid (EoS of Thol *et al.* [229]) in dimensionless variables appendix A.1: pressure over density for varying temperatures. Regions are separated by solid black binodal and dashed black spinodal curves. The red circle shows the critical point. Critical isotherm is shown in red line corresponding to $T^* = 1.321$.

transition or coexistence is not considered. The region above the critical isotherm is called supercritical fluid, which does not factor much into our considerations. The metastable regions are by definition localized in the area between the binodal and spinodal curves and are of prime concern in this study. In the middle of the phase diagram, an unstable region is localized. This region is not of concern as very chaotic changes are occurring there and modeling is heavily restricted.

2.3.2 Cluster

In the previous text, the phrase *groups of molecules* was used without any closer explanation of how the particular grouping is constructed. We will now formalize the group with the term *cluster* which we specify in thermodynamic terms with the choice of dividing surface (interface) for the density function capturing the change between phases. Assuming the shape of the cluster is spherical, the density function is constructed from the center of the sphere, and the dividing surface is the radius of the sphere separating the cluster from the surroundings.

In fig. 2.4, the choice of interface has been selected at $r_{\text{equimolar}}$ in a way that minimizes the excess number of molecules (equimolar dividing surface). This means A_{in} and A_{out} in the figure are of same size.

A practical formulation of the cluster can be constructed in terms of distance with the Euclidean metric in three dimensional physical space. Under this assumption, we can present the following formalization of clusters according to the construction method:

Definition 2.1. For set of positions $r_1, \ldots, r_k$ where $k \in \mathcal{N}$ the nearest neighbor graph is

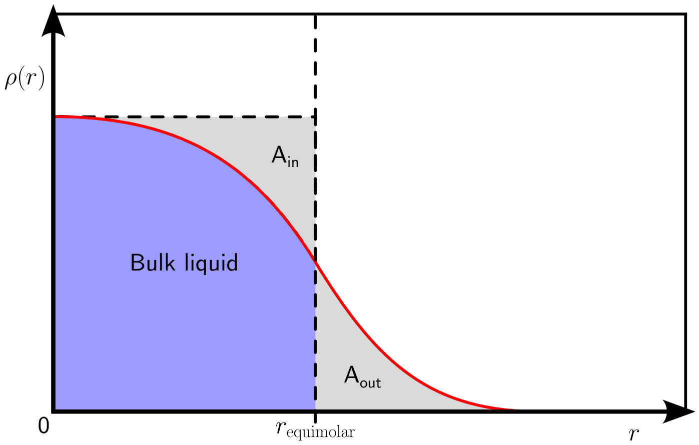


Figure 2.4: Density profile of equimolar dividing surface for spherical cluster.

constructed such that each position r_i is connected to its nearest position r_j and the edge value is equal to the euclidean distance between the points $|r_i - r_j|$

Definition 2.2. Set of $k \in \mathcal{N}$ molecules $c = \{m_1, \dots, m_k\}$ with position vectors $r_1, \dots, r_k$ is called cluster with respect to the neighbor distance r_c when the nearest neighbor graph constructed over position vectors is connected and have all edges smaller than r_c.

The cluster is then identified with the recursive procedure of constructing the neighbor tree and checking the fulfillment of the neighbor distance criterion for each edge of the neighbor graph. This cluster criterion is called Stillinger [223], an alternative formulation of cluster criteria, and therefore different variants of clusters are possible.

2.3.3 Work of formation

The nucleation process in fig. 2.2 is also illustrated with an example of 2D microstates of the nucleating system. The individual stages are shown for vapor→liquid transition, during which the naturally occurring small clusters are continuously created and dissolved until a large enough cluster appears within the system. This particular microstate configuration with a cluster of molecules is an important crossing point in the nucleation process. We will now focus the derivation on finding out how much energy this molecular group contains, which will be equal to the height of the energy barrier, as can be seen from fig. 2.2.

We can start by distinguishing two states of the system: the first is the starting state with homogeneous vapor (configuration (1) in fig. 2.2), while the second is the system with the cluster (configuration (2) in fig. 2.2). The difference in Gibbs potential characterizing the system is then the amount of work the system put into creating the cluster with N molecules. This is called the *work of formation* and for a single component system, it is expressed as:

$$\begin{aligned}
W(N_2) &= G_2 - G_1 = \Delta G && (2.47)\\
&= \Delta U + \sigma A_{\mathrm{ds}} - T\Delta S + p_1 \Delta V && (2.48)\\
&= [(p_1 - p_2)V_2 - p_1\Delta V + T\Delta S + N_2(\mu_2(p_2, T) - \mu_1(p_1, T))] + \sigma A_{\mathrm{ds}} - T\Delta S + p_1\Delta V && (2.49)\\
&= (p_1 - p_2)V_2 + N_2(\mu_2(p_2, T) - \mu_1(p_1, T)) + \sigma A_{\mathrm{ds}}. && (2.50)
\end{aligned}$$

In this derivation, we have the internal energy difference and also the surface contribution system energy that was expended to create the interface of area A_{ds} with surface tension σ. This surface tension holds the difference of pressures inside p_2 and outside p_1 of the cluster and also contains an entropic effect of molecule distribution at the surface, which makes it unsuitable to be included in the internal energy difference. We have also used the knowledge that temperature is constant throughout states (1) to (2) in fig. 2.2, as no energy is exchanged with the surroundings. We can now use the Gibbs-Duhem equation eq. (2.9) for the single component case with a fixed temperature to get the following relation for the chemical potential difference of the cluster:

$$N d\mu = V dP \tag{2.51}$$

$$N_2(\mu_2(p_2, T) - \mu_2(p_1, T)) = V_2(p_2 - p_1) \tag{2.52}$$

Substituting eq. (2.52) into eq. (2.50) for the pressure difference, the chemical potential difference cancels as follows:

$$W(N_2) = N_2(\mu_2(p_1, T) - \mu_1(p_1, T)) + \sigma A_{\mathrm{ds}} = N_2 \Delta\mu + \sigma A_{\mathrm{ds}} \tag{2.53}$$

We now know that the energy barrier depends on the difference in chemical potentials and is promoted by the formation of a dividing interface. When a system with different fixed state variables is assumed, different potentials need to be used for the derivation. Denoting the number of molecules in the cluster as N we get the desired formula for the work of formation of the cluster:

$$W(N) = N\Delta\mu + \sigma A_{\mathrm{ds}} \tag{2.54}$$

A quick analysis of the formula leads to an interesting conclusion. First, we notice the interface term σA_{ds} is always positive because both area and surface tension are positive numbers. This means that when nucleation is favored (small work of formation), the potential term $\Delta\mu$ needs to be negative. We can further analyze the formula for the limiting cases of small and large spherical clusters. For this purpose, the volume of the cluster is expressed with the number of molecules as:

$$V = \frac{4}{3}\pi r^3 = N v_1 \tag{2.55}$$

$$r = \left(\frac{3 N v_1}{4\pi}\right)^{1/3} \tag{2.56}$$

$$A = 4\pi \left(\frac{3 N v_1}{4\pi}\right)^{2/3} = (36\pi)^{1/3} (N v_1)^{2/3} \tag{2.57}$$

The work of formation for a spherical cluster is then:

$$W(N) = N\Delta\mu + (36\pi)^{1/3} (N v_1)^{2/3} \sigma \tag{2.58}$$

There is an important factor for evaluating the probability of finding the cluster of size N in the system as it relates to the probability of finding a cluster of size $P(N)$ as:

$$P(N) \propto \exp\left(-\frac{W(N)}{k_{\mathrm{B}} T}\right) \tag{2.59}$$

Which directly relate to the size distribution within the well relaxed equilibrium system ν_{eq} usually taken as density expressed in [mol/m^3]:

$$\nu_{\mathrm{eq}}(N) = \nu(1) \exp\left(-\frac{W(N)}{k_{\mathrm{B}} T}\right) \tag{2.60}$$

This is the result of the work of the formation construct, which is later applied to the steady state system. But there are some notable issues with this construction, such as the need for a limit case of monomer, which should equal to the initial distribution (in this sense also supply) of monomers $\nu(1)$.

Analyzing eq. (2.58) in the limit of small clusters, the first term vanishes, and the process is dominated by the $N^{2/3}$ decrease of the surface term $A(N)\sigma$. Conversely, for the limit of a large cluster, the potential term dominates the process as it scales with N.

$$\lim_{N\to 1^+} W(N) = \lim_{N\to 1^+} (36\pi)^{1/3} (Nv_1)^{2/3} = (36\pi)^{1/3} (v_1)^{2/3} \tag{2.61}$$

$$\lim_{N\to\infty} W(N) = \lim_{N\to\infty} N = \infty \tag{2.62}$$

This means the work of formation as a function of N has a local maximum where the roles of surface term and potential term are equal.

Critical cluster Searching for first local maximum of work of formation W with respect to N we obtain:

$$N^* = \frac{32}{3}\pi v_1^2 \left(\frac{\sigma}{\Delta\mu}\right)^3 \tag{2.63}$$

This cluster is called critical, and it has an equal probability of growth and shrinking caused by the equal contribution from both terms. The N^* is then the number of molecules in the critical cluster.

To obtain the critical radius, we use the expression eq. (2.56) and translate the previous equation into

$$r^* = \frac{2v_1\sigma}{\Delta\mu}, \tag{2.64}$$

which gives a formula for the radius of a spherical cluster.

2.4 Homogeneous nucleation kinetics

The term homogeneous nucleation means that no external forces or system impurities (external agents) are introduced into the system which would modify the work of formation barrier. However, the real world examples are mostly cases of heterogeneous nucleation, like for example the nucleation of water droplets on dust particles. Understanding of homogeneous nucleation is required first before more complicated process can be reliably modeled. For brevity of the following description, the transition from vapor→liquid, i.e., the formation of droplets, is considered unless stated otherwise.

2.4.1 Steady state nucleation

In this and the following sections, the processes of cluster growth and shrinking are examined. From the evaluation, one can then calculate the rates of the respective transitions and search for the equilibrium value, which leads to the nucleation rate. For this reason, we adopt the ideas of Volmer and Weber [249] who assumed that the processes of cluster growth (condensation) and shrinking (evaporation) were only mediated by single molecule clusters. The authors argued that the probability of collision with a dimer and bigger clusters is small enough to be neglected. In their terminology as well as our definition, definition 2.2, a monomer is considered a cluster of size $N = 1$. Further assumptions include that collision of a monomer with a cluster is the same as monomer condensation, i.e., no bouncing of the monomer from the cluster is considered.

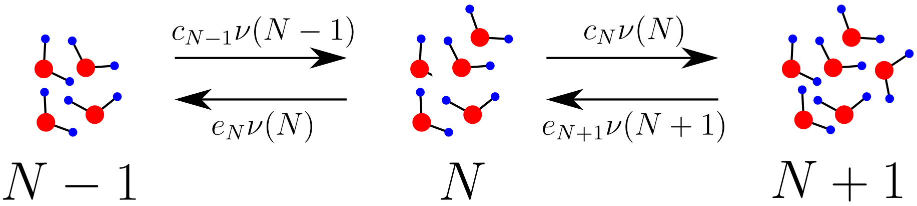


Figure 2.5: Steady state nucleation kinetics illustration for water molecules as system particles. The growth and shrink probabilities are dependent only on cluster sizes.

The final point is the independence of individual condensation and evaporation events from each other. This collectively means the process described is a Markov chain in section 3.3.3. The growth and shrinking can then be conceptualized as a reversible reaction:

$$C_N + C_1 \underset{e_{N+1}}{\overset{c_N}{\rightleftharpoons}} C_{N+1}, \tag{2.65}$$

where C_N is a cluster of size N, with rate of condensation c_N leading to cluster growth and rate of evaporation e_{N+1} leading to shrinking of the cluster as shown in fig. 2.5. With a cluster size distribution function $\nu(N,t)$ we can construct the cluster flux of a particular size N by comparing both condensation and evaporation rates as time independent parameters.

$$J(N,t) = c_N\nu(N,t) - e_{N+1}\nu(N+1,t) \tag{2.66}$$

This leads to a differential equation for cluster size distribution:

$$\frac{\partial\nu(N,t)}{\partial t} = J(N-1,t) - J(N,t) = c_{N-1}\nu(N-1,t) - (e_N + c_N)\nu(N,t) + e_{N+1}\nu(N+1,t) \tag{2.67}$$

The master equation in this form relates the flux difference to the sum of all condensation and evaporation events. To solve it we require to assume a *steady state system*, where individual fluxes are equal. We can imagine the steady state where nucleation takes place creating bigger clusters, but wia means of Maxwell daemon the clusters of certain size are decomposed back into monomers keeping a steady supply. When equilibrium is reached in the system the individual fluxes are equal (because of finite number of molecules), and one universal flux $J = J(i), i \in \widehat{N}$ can be identified as *nucleation rate*. For steady state system the left hand side in eq. (2.67) is equal to zero. This allows to determine the nucleation rate from knowledge of evaporation and condensation rates presented in next section.

2.4.2 Classical nucleation theory for dropplet condensation

Following the Becker and Döring [21] approach, we focus on cluster condensation first, which is easier to model, and later use the detailed balance for the evaluation of condensation rates [113].

We have already assumed that clusters are created by the consequent condensation of monomers, where single collision depends on the surface of the cluster and the probability of approaching monomers from the vapor phase. This can be expressed with:

$$c(N) = \nu A(N) \tag{2.68}$$

where ν is the impingement rate and $A(N)$ is the area determined from equation eq. (2.57). This quantity is determined from gas kinetics [137] with the integration of Maxwell velocity distribution, resulting in the commonly used formulation:

$$\nu = \frac{p_\mathrm{v}}{\sqrt{2\pi m_1 k_\mathrm{B} T}} \tag{2.69}$$

where m_1 is the mass of a single molecule. The impingement rate is here expressed in relation to the vapor pressure p_v under the assumption that the spherical cluster has similar surface behavior as the flat surface [118].

Now the result of the steady state system is applied to eq. (2.65), where for the system in thermodynamic equilibrium condition no nucleation takes place and fluxes are equal to zero. This means that the coefficients of evaporation can be expressed in terms of the rate of condensation as:

$$e_{N+1} = c_n \frac{\nu_\mathrm{eq}(N)}{\nu_\mathrm{eq}(N+1)}. \tag{2.70}$$

In this way, the nucleation rate $J(N)$ from eq. (2.66) is written as:

$$\begin{aligned} J(N) &= c_n \nu(N) - c_n \frac{\nu(N+1)\nu_\mathrm{eq}(N)}{\nu_\mathrm{eq}(N+1)} \\ \frac{J(N)}{c_n \nu_\mathrm{eq}(N)} &= \frac{\nu(N)}{\nu_\mathrm{eq}(N)} - \frac{\nu(N+1)}{\nu_\mathrm{eq}(N+1)} \qquad (2.71) \\ & \qquad (2.72) \end{aligned}$$

Summation over sizes up to a high enough cluster size m, utilizes the feature of the equation where consecutive terms cancel out. Steady state condition of equal nucleation rates, then introduce one universal nucleation rate J as:

$$J \sum_{N=1}^{m-1} \frac{1}{c_n \nu_\mathrm{eq}(N)} = \frac{\nu(1)}{\nu_\mathrm{eq}(1)} - \frac{N(m)}{\nu_\mathrm{eq}(m)} \tag{2.73}$$

The known limiting behavior for small clusters, eq. (2.61), means that the first term on the right-hand side becomes equal to one. For the limit case for large enough $m \to \infty$, the second term approaches zero. Here the knowledge of eq. (2.62) is used for equilibrium distribution, while the real distribution has to remain finite because of the finite supply of the monomers as well as finite time. Rearranging the equation with the known limit behavior yields:

$$J = \left[\sum_{N=1}^{\infty} \frac{1}{c_n \nu_\mathrm{eq}(N)} \right]^{-1} \tag{2.74}$$

Analyzing the sum, we notice that most contributions to the sum are made by cluster sizes around N^*. From eq. (2.60) $\nu_\mathrm{eq}(N)|N = N^*$ has its minimum because of the local maximum of $W(N)|N = N^*$. Consequently, the inverse value $1/\nu_\mathrm{eq}(N)|N = N^*$ will be maximal, explaining the reasoning. The exponential dependency $\exp(-N)$ also makes the contribution vanishing quickly for increasing N. This makes the calculation of this convergent series more convenient by replacing the sum eq. (2.74) by an integral

$$J = \left[\int_1^{\infty} \frac{dN}{c_n \nu_\mathrm{eq}(N)} \right]^{-1}. \tag{2.75}$$

The main issue in this integral is the evaluation of $\nu_\mathrm{eq}(N)$ because the condensation rate can be replaced with the condensation rate of the critical cluster without making much error in the near-critical clusters. In this sense, we can expand the work of formation around the critical cluster size, leading to the following relations for the work of formation:

$$W(N) \approx W(N^*) + 0 + \frac{(N-N^*)^2}{2}\left(\frac{\partial^2 W(N)}{\partial N^2}\right)_{N=N^*} = W(N^*) - \frac{(N-N^*)^2}{2}\widetilde{W}_{N^*} \tag{2.76}$$

With the notation, the second partial derivative of formation work at critical cluster size is $-\widetilde{W}_{N^*}$. From an already performed evaluation of curvature, we know that $\widetilde{W}_{N^*} > 0$. Combining the approximation into an equilibrium size distribution $\nu_{\mathrm{eq}}(N)$ leads to:

$$\nu_{\mathrm{eq}}(N) \approx \nu(1)\nu_{\mathrm{eq}}(N^*)\exp\left(-\frac{(N-N^*)^2}{2k_{\mathrm{B}}T}\widetilde{W}_{N^*}\right) \tag{2.77}$$

This form of the equilibrium size distribution is now conveniently prepared for the integration of eq. (2.75). To obtain the desired Gaussian integral, a continuous extension into negative numbers for ν_{eq} is performed. Because of the localized support around the critical cluster, there is no issue with the extension into negative values, even though it has no physical representation. With the critical condensation rate approximation $c_n \approx c_{N^*}$, we finally get:

$$\begin{aligned} J &= \left[\frac{1}{c_{N^*}\nu(1)\nu_{\mathrm{eq}}(N^*)}\int_{-\infty}^{\infty}\frac{dN}{\exp\left(-\frac{(N-N^*)^2}{2k_{\mathrm{B}}T}\widetilde{W}_{N^*}\right)}\right]^{-1} \\ &= \left[\frac{1}{c_{N^*}\nu(1)\nu_{\mathrm{eq}}(N^*)}\sqrt{\frac{2\pi k_{\mathrm{B}}T}{\widetilde{W}_{N^*}}}\right]^{-1} \\ &= c_{N^*}\nu(1)\nu_{\mathrm{eq}}(N^*)\sqrt{\frac{\widetilde{W}_{N^*}}{2\pi k_{\mathrm{B}}T}} \\ &= \mathcal{Z}\, c_{N^*}\nu(1)\nu_{\mathrm{eq}}(N^*) \end{aligned} \tag{2.78}$$

where presented nucleation rate introduces the Zeldovich factor $\mathcal{Z}$ [267]. This factor is obtained from evaluating the $\widetilde{W}_{N^*}$, see eq. (2.58). Following derivation by Vehkamäki [239], the chain rule $\frac{\partial W(N)}{\partial N} = \frac{\partial W(r)}{\partial r}\frac{\partial r(N)}{\partial N}$ can be used to express the formulas in terms of r^*

$$\widetilde{W}_{N^*} = -\left(\frac{\partial^2 W(N)}{\partial N^2}\right)_{N=N^*} = \frac{v_1^2\sigma}{2\pi(r^*)^4} \tag{2.79}$$

$$\mathcal{Z} = \sqrt{\frac{\widetilde{W}_{N^*}}{2\pi k_{\mathrm{B}}T}} = \frac{v_1}{2\pi(r^*)^2}\sqrt{\frac{\sigma}{k_{\mathrm{B}}T}}. \tag{2.80}$$

The most compact form is obtained from combining the equilibrium size distribution from eq. (2.60) with the condensation rate in eq. (2.68) and replacing the initial size distribution $\nu(1) \approx \rho^{\mathrm{v}}$. The work of formation is usaually expressed in exponential form giving the nucleation rate as

$$\begin{aligned} J &= \mathcal{Z}\nu A(N^*)\rho^{\mathrm{v}}\exp\left(-\frac{W(N^*)}{k_{\mathrm{B}}T}\right) \\ &= J_0\exp\left(-\frac{W(N^*)}{k_{\mathrm{B}}T}\right). \end{aligned} \tag{2.81}$$

This is the most commonly used formula with the preexponential factor J_0 for the nucleation rate of a steady state system, according to the assumptions of the classical nucleation theory. Work of formation used in the expression is given in eq. (2.58). We acknowledge there are alternative formulations of work of formation that lead to slightly different formulation, i.e., for systems with different fixed state variables or different cluster structures.

2.4.3 Known issues of CNT

The developed classical nucleation theory is built around a few key assumptions about the form of the investigated phenomenon and introduces multiple approximations and simplifications to derive the universal nucleation rate. Experimental investigations of nucleation rate over the years have shown significant discrepancies between the model and experiment, motivating further investigation and analysis of possible model deficiencies. This process is still not finished, as there is no model in complete agreement with the experiments to the author's knowledge. In this section, we discuss the three assumptions and approximations suspected to account for the limited applicability of the CNT model.

- **Capillarity approximation**
  CNT assumes that the nucleus interior is a bulk, incompressible fluid and ascribes to the nucleus surface the macroscopic interfacial tension, even though it is not clear that such macroscopic equilibrium properties apply to a typical nucleus.

- **Small cluster description**
  CNT assumes the shape of small clusters is spherical, which is a poor approximation. This results in a misrepresentation of the smallest clusters, which are important for the nucleation process.

- **Choice of dividing surface**
  Closely related is the choice of description of the cluster, which does not reflect the nonuniform nature of the surface of a cluster of any size.

In summary, the classical nucleation theory has limitations that compromise its applicability and cause inconsistencies with experimental results. But the CNT is the most widely used model of the kinetics of nucleation, with important applications in science and industry.

2.5 Cahn-Hilliard gradient theory

The density gradient theory (DGT) was first proposed by Cahn and Hilliard in 1958 [35] as a tool for obtaining the Helmholtz free energy of an inhomogeneous system. The authors postulated that the Helmholtz free energy of a non-homogeneous system can be written as Taylor expansion of the homogeneous one, [35].

The main advantage of this approach is the computational speed and the overall simplicity compared to the more general Density Functional Theory (DFT) or molecular simulations. The simplicity of the approach comes at the cost of lowered accuracy in regions with large density gradients.

DGT formulates the work of formation in terms of grand potential difference $\Delta\Omega$. The potential difference is used to describe the optimal density profile, which is understood as the most probable profile for the considered multicomponent system conditions. The work of formation is defined as the difference between the energy of the homogeneous system and the non-homogeneous system, where the phase interface effects are accounted for. We note that in case of a spherical cluster with the critical radius r^* and known surface tension σ, the work of formation can be determined as follows

$$\Delta\Omega = \frac{4}{3}\pi\sigma(r^*)^2. \tag{2.82}$$

Similar formulations for work of formation $\Delta\Omega$ can be derived using different thermodynamic potentials, but for the case of multicomponent mixtures considered in the later part of the thesis the grand potential Ω is the most suitable one

$$\Delta\Omega(\rho(s)) = \Omega_{\mathrm{inhom}}(\rho(s)) - \Omega_{\mathrm{hom}}(\rho(s)). \tag{2.83}$$

Grand potential Ω depends on the vector of concentration ρ, which can be understood as a density profile $\rho(s)$. In an arbitrary system, density profile is a function of general system coordinates $\rho = \rho(s_1, s_2, s_3)$. This formulation leads to the three-dimensional problem of rather great complexity. Therefore, it is convenient to assume that the system is non-uniform only in one coordinate denoted further as general coordinate s. Note that this assumption is valid for both investigated interface geometries in section 2.2.3. In DGT, the Helmholtz energy is further assumed to be a functional of the density profile $\rho(s)$, i.e. a density function, and its respective gradients. For the situation investigated further, the gradients are simple derivatives, which is of direct consequence of the assumed uniformity.

The Helmholtz energy density of the inhomogeneous system can then be expressed as the Taylor expansion around the homogeneous Helmholtz energy density with higher order terms omitted

$$f_{\mathrm{inhom}}(\rho) = f_{\mathrm{hom}}(\rho) + C_1 \cdot \nabla^2\rho + \frac{1}{2} C_2 \cdot (\nabla\rho)^2 \ldots \tag{2.84}$$

where C_1 and C_2 are the independent Taylor coefficients and f_{hom} denotes the Helmholtz energy density of a hypothetical homogeneous system without the phase interface or a cluster of the new phase. Helmholtz energy density f can be then integrated over the system volume to produce the total Helmholtz energy F. Using same approach as Cahn and Hilliard [35] the Helmholtz energy is obtained

$$F_{\mathrm{inhom}} = \int_s \left[f_{\mathrm{hom}}(\rho) + \frac{1}{2} C_3 \left(\frac{\partial\rho}{\partial s}\right)^2 \right] S\mathrm{d}s. \tag{2.85}$$

Coefficient C_3 denotes the collected coefficients C_1 and C_2 from eq. (2.84). The specific form for relation between the coefficients is related to the direct correlation function [54].

The integration over the remaining coordinates s_2, s_3 is denoted as the surface element S, which is in direct consequence of the uniformity assumption. The Helmholtz energy from eq. (2.85) is now used to express the grand potential of an inhomogeneous system. Substituting the inhomogeneous form obtained here into eq. (2.83), the work of formation can be expressed as follows

$$\Delta\Omega = \int_s \left[\Delta\omega\left(\rho\left(s\right)\right) + \frac{1}{2} C_3 \left(\frac{\partial\rho}{\partial s}\right)^2 \right] S\mathrm{d}s, \tag{2.86}$$

where $\Delta\omega$ is the grand potential density that may be now written as

$$\Delta\omega\left(\rho\right) = f_{\mathrm{hom}}\left(\rho\right) - \sum_{i=1}^{n_{\mathrm{comp}}} \mu_{\mathrm{v},i}\, \rho_i + p_{\mathrm{v}}. \tag{2.87}$$

In eq. (2.87), ρ_i marks the partial molar concentration $\rho_i = \rho x_i$.

Notice that the work of formation $\Delta\Omega$ defined in eq. (2.86) was derived for a generalized type of interface geometry parametrized with general coordinate s and surface element S. This alows for simple specification of the geometry according to the choice of the coordinate system best describing the interface geometry. In accordance with section 2.2.3 the two main cases are: planar geometry best described with Cartesian coordinates and the spherical geometry capturing spherical droplets or bubbles. This is best illustrated in fig. 2.6 where a comparison of both considered geometries is given in context of underalying density profile representation as well as microscopical structure of the interface obtained from molecular simulation. In the figure a choice of general coordinate s is shown as well underlying density profile description which is being searched for by the DGT.

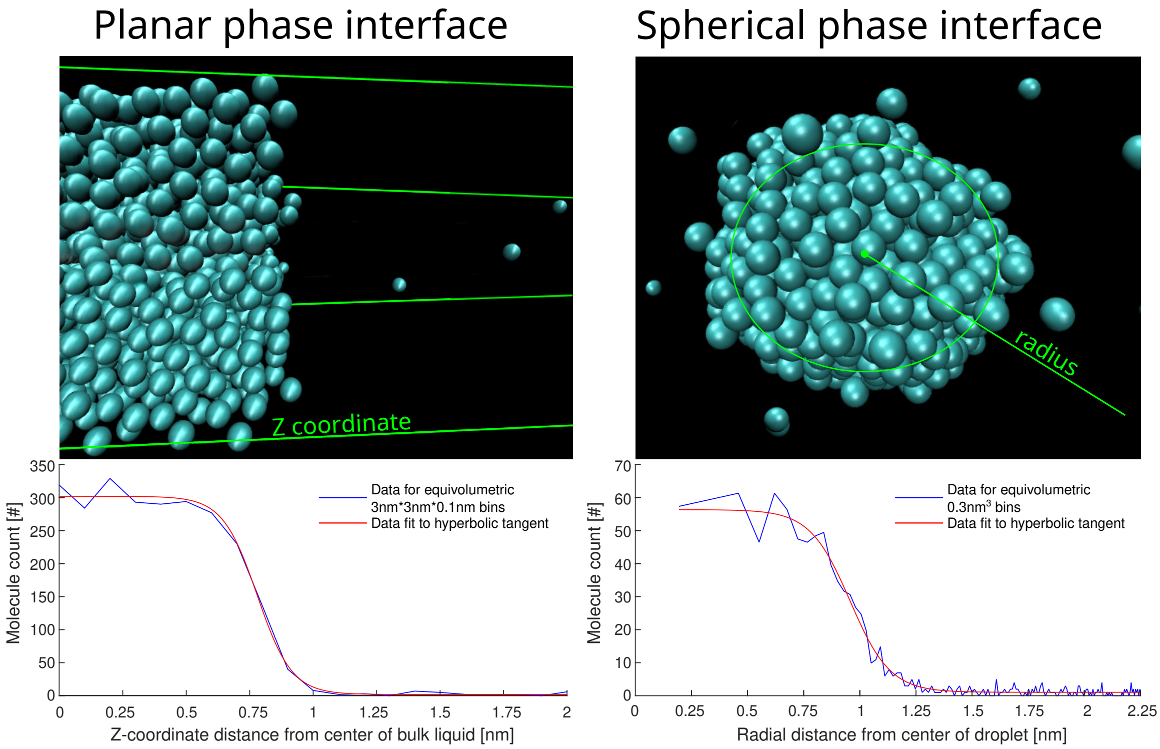


Figure 2.6: Visualization of planar and spherical phase interface geometries, with depicted governing coordinate in which the system is nonuniform. The molecular picture from simulation of argon is supplied with the corresponding density distribution graph. The graph shows comparison between theoretically relevant hyperbolic tangent profile and density histogram using bins of same volumes (which is important consideration for spherical case).

Following paragraphs provide the final formulas for both types of these interface geometries.

For the **planar phase interface**, the interface parameter s becomes coordinate z in the Cartesian coordinate perpendicular to the phase interface and surface element S is replaced with surface area A, which is orthogonal to z. Work of formation then becomes

$$\Delta\Omega = \int_{z_{\text{min}}}^{z_{\text{max}}} \left[\Delta\omega\left(\rho\left(z\right)\right) + \frac{1}{2} \sum_{i,j=1}^{n_{\text{comp}}} c_{i,j} \left(\frac{\partial \rho_i}{\partial z}\right) \left(\frac{\partial \rho_j}{\partial z}\right) \right] A\text{d}z. \tag{2.88}$$

In eq. (2.88), integration over the whole z-dimension can be truncated to the interval $\langle z_{\text{min}}, z_{\text{max}} \rangle$, which has to envelop the interface region resulting in only negligible error of domain truncation.

For the **spherical interface** geometry, the general parameter s is defined as radius r of a sphere, as the system non-uniformity exhibits along this coordinate. Rewriting the integral eq. (2.86) in spherical coordinates yields the following formula with domain truncated again to the large enough radius R_{max} containing homogeneous mother phase, i.e. the bulk phase.

$$\Delta\Omega = \int_{0}^{r_{\text{max}}} \left[\Delta\omega\left(\rho\left(r\right)\right) + \frac{1}{2} \sum_{i,j=1}^{n_{\text{comp}}} c_{i,j} \left(\frac{\partial \rho_i}{\partial r}\right) \left(\frac{\partial \rho_j}{\partial r}\right) \right] 4\pi r^2 \text{d}r. \tag{2.89}$$

The problem formulated in eqs. (2.88) and (2.89) lies in finding the density distribution function $\rho(s)$, which has a proper physical meaning and provides an optimal work of formation. The method for finding the optimum is described in solution section section 5.3.

Molecular simulation 3

3.1 Introduction

Molecular simulation (MS) is a computational technique utilizing mostly classical statistical physics. It has become a useful tool in many branches of science, from engineering with the heavy use of simulation of processes and calculation of thermophysical properties to polymers and colloids to complex biological systems. In recent years, we have observed an increasing demand for simulations of large systems (millions of atoms), particularly in biochemistry but also in material science, polymer science, etc. These efforts have been accentuated by big technological companies such as NVIDIA, leading to their support in the field of biological simulation. The ongoing acceleration of molecular simulations replaces the classical CPU-based parallelism with a GPGPU approach, which enables landmark achievements in molecular simulations to simulate a whole cell [222].

The basic principle of molecular simulation is to approximate the behavior of systems composed of many particles on an atomistic scale. We can thus get the thermodynamic and statistical properties of the system and their time evolution. Simulations provide insight into atom-scale structure not (always) available from an experiment, help interpret the experiment, and provide observations and predictions of experimentally challenging or even unreachable phenomena (high pressure). In this sense, simulations are "theory". In contrast, in many classical statistical-thermodynamical theories, one starts with a molecular model and then obtains statistical-thermodynamical results by applying various approximate theoretical procedures (e.g., integral equations for the structure of liquids). Then, simulations can provide an "experiment" (pseudoexperiment) on an exactly defined model and, in turn, solve the problem of whether the inaccurate results are caused by a bad molecular model or the wrong theory.

Molecular simulation methods can be divided into two main categories: (i) Monte Carlo (MC) methods utilizing stochastic sampling as a means to calculate ensemble averages (we leave aside kinetic Monte Carlo based on events and their apriori known probabilities), and (ii) molecular dynamics (MD) methods based on Newton's laws of motion relying on trajectory to generate the system description. In this thesis, MD is favored over MC because of the need for trajectory generation and the dynamic nature of the observed phenomena. The Monte Carlo method is employed only on a minor scale to improve the properties of initial configurations.

An important point has to be made that the simulations as tools rely on a model (as the potential energy surface described by a force field) utilizing certain simplifications as opposed to the experiment reflecting nature. In turn, model selection is an important factor in the presented results. Simulation cannot fully replace experiment but in many cases it is the best choice we currently have.

We begin with a description of the modeled molecules and the construction of interaction potentials, which serve as the foundation for any simulation. Then, MC methods are presented although they are utilized in this thesis only as a form of preconditioning. Lastly, molecular dynamics is introduced with all the necessary concepts required for the research performed in this thesis.

3.2 Interaction potential

Any modeling is based on approximations. The field of atomistic modeling relies on the Born–Oppenheimer approximation. It says that the motion of electrons can be separated from the motion of much heavier and slower nuclei. In turn, one can uniquely[1] express the energy of the system as a function $U(r^N)$ of positions of atom nuclei, $r^N = \{\boldsymbol{r}_i\}_{i=1}^N$, where N is the number of atoms. Such a function is called the *potential energy surface* (PES). Since there is no time dependence, the corresponding forces are conservative.

The potential energy surface can be obtained by an (approximate) solution of the electronic Schrödinger equation. This approach is called *ab initio molecular dynamics* and is computationally demanding, allowing one to simulate hundreds of atoms at most at a picosecond scale. A modern compromise approach uses a neural network trained on smaller systems (clusters) calculated by good quantum methods [134] to reach a considerable speedup. The classical and most efficient approach (millisecond scale for proteins at the best computers) is based on expressing function $U(r^N)$ by a formula composed as a sum of many terms. This formula (with the corresponding forces) is called *force field* in chemistry, although in physics, term (interatomic) *potential* or just (potential) *model* are common. The potential here is considered in the physical sense i.e. the force is force $F = -\nabla U$.

in this section the interaction potentials are introduced, formalizing the basic properties. Then we introduce the compounds simulated in this study in relation to the interactions they exhibit as well as the structure of the fluids. In the latter part, the potential field development principles are briefly explained. The aim is to explain how the potentials are obtained for simulation rather than rigorous derivation, which would exceed the focus of this thesis.

3.2.1 Types of interaction potentials

The intermolecular forces are traditionally classified into bonded and nonbonded. With bonded forces molecular structure is preserved, be it bonds, bond angles, torsion etc. Nonbonded forces represent the atom–atom interaction involving repulsion and attraction including Coulombic forces.

3.2.1.1 Neutral atom interaction

In order to describe the interactions between neutral atoms, it is important to consider both the short-range and long-range contributions and the forces involved. The short-range contribution is characterized by a strong repulsion occurring for atoms in close proximity. This is the consequence of Pauli's exclusive principle and the fact that atoms are not point-like entities. Quantum calculations indicate an exponential dependence of potential on the interatomic distance r as follows:

$$u(r) \approx e^{-c_{\text{rep}} r} \tag{3.1}$$

With c_{rep} is a constant to be determined.

[1] One may consider two different PESs corresponding to two different electron states, e.g., singlet and triplet in the case of phosphorescence. We will not consider such cases here.

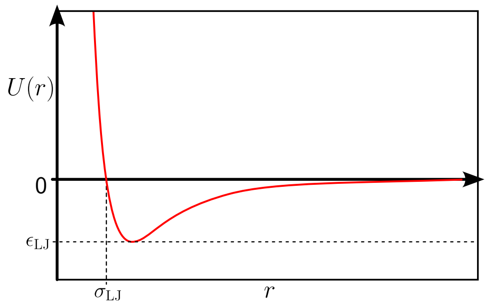


Figure 3.1: The LJ potential describes both the attraction and repulsion between particles with illustrated role of parameters $\epsilon_{\mathrm{LJ}}, \sigma_{\mathrm{LJ}}$.

The second effect concerns the attraction of neutral atoms over longer distances. A specific example is the London force, where instantaneous fluctuations in the electron distribution of an atom induce temporary opposite polarization, resulting in momentary attractive forces. This belongs to the van der Waals type of forces, which are in chemistry [11] defined by a potential decaying with distance as $1/r^6$. The accepted approximation is:

$$u(r) \approx c_{\mathrm{att},6} r^{-6} + c_{\mathrm{att},8} r^{-8} + \dots \tag{3.2}$$

The combined effects of short-range repulsion and longer-range attraction in neutral atom interactions can be represented by various mathematical functions. Examples include the Square Well potential [18] or Buckingham potential [32]. However, for the practical purposes of this study, the Lennard-Jones(LJ) potential is considered sufficient to accurately model the desired behavior of neutral atom interactions.

The Lennard-Jones potential shown in fig. 3.1 is mathematically expressed as a function of the interatomic distance r, involving two parameters: the well depth ϵ_{LJ}, which determines the strength of the interaction, and σ, which is the distance where the potential first reaches zero. Parameter σ is related to the nearest distance between particles, resulting in the transformation of both eqs. (3.1) and (3.2) into a formula:

$$u_{\mathrm{LJ}}(r) = 4\epsilon_{\mathrm{LJ}} \left[\left(\frac{\sigma_{\mathrm{LJ}}}{r} \right)^{12} - \left(\frac{\sigma_{\mathrm{LJ}}}{r} \right)^{6} \right] \tag{3.3}$$

where the exponential has been replaced by the power of the twelfth order. This potential realization has proven to be very versatile in modeling the behavior of atom interactions in various scientific fields.

3.2.1.2 Charged particle interaction

It is a common first approximation if one excludes the effects of polarizability and consider the charge as localized at a single point. The charge is then typically positioned at the center of the atom, although some exceptions may apply. The visualization of Coulomb potential in fig. 3.2 shows both attractive and repulsive interactions, which are determined from the combined charges of interacting particles $Q = q_1 q_2$ and the interparticle distance r according to the Coulomb potential.

$$u_{\mathrm{ele}} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r} = \frac{k_e Q}{r} \tag{3.4}$$

where ϵ_0 is permitivity of a vacuum combined with the other prefactors into the Coulomb constant k_e.

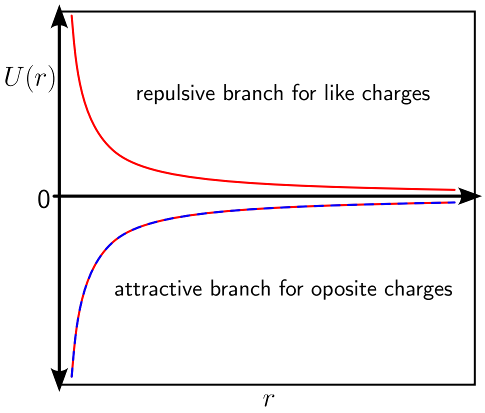


Figure 3.2: Coulomb potential for case of like and unlike pair of charged particles.

An important characteristic of the Coulomb potential is its distance dependence, which follows an inverse relationship represented by r^{-1}. This means that Coulombic forces do not diminish rapidly like those in the Lennard-Jones potential. Instead, they extend over significant distances, similar to the way gravitational forces behave. In other words, the influence of Coulombic forces can be felt even at relatively large separations between charged particles.

The long-range nature of the Coulomb potential poses technical challenges in molecular simulation. When simulating systems with charged particles, such as polar molecules, the interaction between distant charges needs to be sufficiently accounted for. This can be achieved by either including long-range interactions with a long cutoff or employing specialized techniques like Ewald–Komfeld summation, which account for longer-range interactions. In either case, a balance needs to be achieved, ensuring both accurate potential evaluation and efficient management of computational resources during the simulation.

3.2.2 Building interaction potential

To mitigate the computational cost associated with evaluating potentials over their full range, modifications are often employed to improve efficiency while maintaining accuracy in potential energy and force calculations. Modification include potential truncation and or shifting. Alternatively the splines can be used to represent the final potential to further diminish the computational cost.

3.2.2.1 Truncated and shifted potential

Full-length evaluations of potentials involve considering interactions over wider ranges, resulting in numerous interactions that only marginally contribute to the total potential energy. As a solution, shortening of the effective range of potential is proposed with potential function truncation. To perform this operation rigorously would require the potential function to have finite support, but due to the quickly vanishing terms, an approximation of the original potential is created instead, where the associated error is deemed acceptable. In practice, nearly all potentials are truncated at some distance r_{cut} leading to the alternative potential functions of form:

$$u_{\text{LJ,cut}}(r) = \begin{cases} u_{\text{LJ}}(r), & r \leq r_{\text{cut}} \\ 0, & r > r_{\text{cut}} \end{cases} \tag{3.5}$$

While the error made in potential evaluation may be negligible, the change in potential strongly influences the force calculated from it. Because of the differential relation between potential and

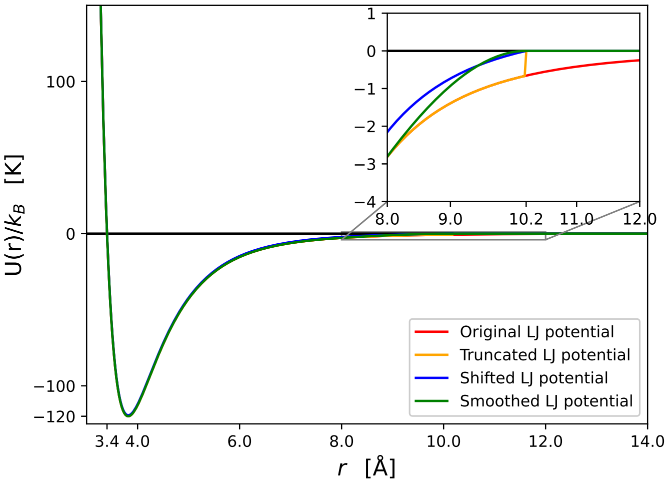


Figure 3.3: Comparison of original potential with the truncated, shifted and smoothed modifications. The used potential parametrization is comparable to argon with $\sigma_{\mathrm{LJ}} = 3.4$ and $\epsilon_{\mathrm{LJ}}/k_{\mathrm{B}} = 120$ and cutoff distance is $r_{\mathrm{cut}} = 3\sigma_{\mathrm{LJ}}$.

force, $F = -\nabla U$, any jump introduced into potential will incur infinite force at the jump point. This kind of potential cannot be used in molecular dynamics simulation and the jump at cutoff distance needs to be corrected. A common correction is performing a shift of the whole function such that the jump is aligned and the function remains continuous at r_{cut}. This moves the whole potential upwards by a constant, as shown in fig. 3.3 changing the absolute value of the potential itself into:

$$u_{\mathrm{LJ,cut}}(r) = \begin{cases} u_{\mathrm{LJ}}(r) - u_{\mathrm{LJ}}(r_{\mathrm{cut}}), & r \leq r_{\mathrm{cut}} \\ 0, & r > r_{\mathrm{cut}} \end{cases} \tag{3.6}$$

In this way the potential can be used in the simulation even with the systematic difference in potential energy, because the forces and kinetic behavior remain the same. When higher precision of energy or energy dependent variables is required, a correction for added energy is performed, effectively subtracting some predetermined value from the final potential energy value as show in section 3.2.2.3.

3.2.2.2 Truncated and smoothed potential

An alternative correction is a smooth connection of potential in the region around the jump point. This is done by means of a crossover function in the area before the cutoff distance r_{cut}. One example is:

$$u_{\mathrm{LJ,smooth}}(r) = \begin{cases} u_{\mathrm{LJ}}(r), & r < r_{\mathrm{smooth}} \\ C_1 \cdot (r^2 - r_{\mathrm{cut}}^2)^2, & r \in [r_{\mathrm{smooth}}, r_{\mathrm{cut}}] \\ 0, & r > r_{\mathrm{cut}} \end{cases} \tag{3.7}$$

where constant C_1 and smoothing start radius r_{smooth} are determined from the condition of continuity of first derivatives at the borders where the crossover function is applied. A comparison of all modifications is shown in fig. 3.3. If the Ewald summation is not practical for the Coulomb interactions, a carefully designed short-range approximation can still be an acceptatble option.

3.2.2.3 Error correction

In the previous section, modifications to the potential function were introduced, inevitably changing the potential energy of the system and associated properties (i.e., pressure). In some cases, the error incurred is negligible and can be safely ignored, as is the case for cutoff distances of $r_{\text{cut}} \geq 6\,\sigma_{\text{LJ}}$ (usually $3\,\sigma_{\text{LJ}}$ is considered enough for gaseous systems). If the system is homogeneous beyond the cutoff distance, one can include a correction which for the case of pure fluid reads as:

In other cases, a correction to total energy is performed to account for the contributions of molecules further away than cutoff molecules would make within the system. Adding over all interaction between molecules i, j in the system, the energy lost due to the use of modified potential instead of full can be evaluated by integrating the difference in the used potentials $\Delta u_{i,j}(r) = u_{i,j} - u_{\text{cut,i,j}}$.

$$\Delta U = \sum_{i<j} \Delta U_{\text{i,j}} = \sum_{i<j} 4\pi\rho \int_{r_{\text{cut}}}^{\infty} \Delta u_{i,j} dr \tag{3.8}$$

The fundamental problem is that the system utilizing the modified potential is used by a small, finite number of entities (molecules in the system), and the exact calculation of error is of the same complexity as the original problem. Therefore, the generally accepted mean field approximation is used, which states that at a sufficient distance, the system already behaves homogeneously. This approximation yields good results in homogeneous simulation and holds reasonably well for the bulk portion of the system. Therefore, density can be assumed to be a constant equal to the number density of all sites in the system $\rho = N/V$. In this way, the energy correction is obtained as ΔUV.

3.2.3 Considered molecules

The main objective of this work is to investigate the phenomenon of nucleation. To ensure clarity and focus on the phenomenon itself, a class of simple fluids is chosen for research purposes, avoiding any complexities that could potentially obscure the investigation. In this section, substances of increasing complexity will be discussed, beginning with an artificial Lennard-Jones fluid and progressing to the examination of water. In the context of this thesis, water is referred to by the model name in order to emphasize the variations the model introduces into the description of the substance.

Lennard Jones fluid The Lennard-Jones fluid (LJF) is a simplified model of a monoatomic electro-neutral fluid used in a wide range of applications. The primary research method is molecular simulation or, alternatively, extrapolations from experimentally obtained properties of noble gases.

The LJF has garnered considerable attention due to its general simplicity and low computational demands resulting from the specific form of its interaction potential. By utilizing the inter-particle interaction from eq. (3.3) and setting the parameters set to unity $\sigma_{\text{LJ}} = 1.0, \epsilon_{\text{LJ}} = 1.0$, the LJF fluid in dimensionless units is obtained. The outcomes of these evaluations find practical use in the research of more complex phenomena, often serving as a reference fluid for perturbation theories of liquids, a basis for model extensions, or a validation fluid within thermodynamically related software. The extensive research conducted on LJF covers a wide range of properties, spanning over 35,000 data points obtained from simulations. Rather than highlighting individual studies, we refer to the comprehensive summary by Stephan *et al.* in [221]. A common modification of LJF is the truncation and shift of the potential, referred to as LJTS, which has also been thoroughly investigated by Stephan in [220].

Table 3.1: Default paramterization of argon model used within the MACSIMUS, Mac_module, and *ms*2 software.

	molar mass [g/mol]	σ_{LJ} [Å]	$\epsilon_{\mathrm{LJ}}/k_{\mathrm{B}}$ [K]
MACSIMUS	39.948	3.405	119.8
Mac_module	39.948	3.4	120
*ms*2	39.948	3.3952	116.79

Argon In the context of this study, argon can be viewed as a real fluid alternative to LJF. Being a noble gas of simple monoatomic structure, the LJ parametric model can be immediately used for its description. Three parametrizations used by packages utilized in this study are provided in table 3.1. MACSIMUS [126] and *ms*2 [73] are standalone molecular simulation packages and Mac_module is software implemented for the research of supersonic nucleation as a part of this study. The availability of accurate experimental data for argon further enhances its utility as a model system for a wide range of thermodynamic studies. The uses include a reference system for testing simulation methodologies, assessing force fields, and validating computational techniques. The inert nature of argon allows its use as a moderator of heat transfer medium in mixtures, further stressing the need for accurate thermophysical properties and associated modeling of argon. Because of its close relationship with LJF, improvements made in the understanding of either argon or LJF lead to better force fields for other fluids.

Nitrogen Nitrogen is a fundamental component of the Earth's atmosphere, and because of its low reactivity and stability, it is a favored gas in many experiments. This motivates both the measurement [108] and modeling [217] of the thermophysical properties of nitrogen. From a modeling perspective, nitrogen is a diatomic molecule with a rigid bond between atoms, with structure shown in fig. 3.4. The linear nature of the molecule also limits the degrees of freedom, making it an interesting test substance in molecular simulation tools, in particular for selective testing of bond preservation mechanisms and verifying proper transformation of kinetic energy into rotation of molecules.

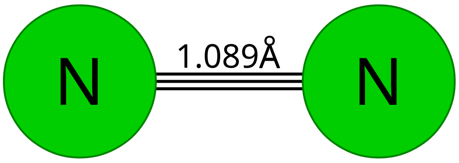


Figure 3.4: Visualization of molecule of nitrogen with the bond length.

Model parametrization using LJ potential and general structural properties of nitrogen are presented in table 3.2. Nitrogen parameterization for both MACSIMUS and Mac_module follows values proposed by Fisher *et al.* [74]. Distinction between atomic and molecular properties is made in subscript.

SPC/E water model Water poses significant challenges for molecular modeling due to its complex nature and the multitude of properties it exhibits. It is complicated to accurately include interactions caused by hydrogen bonding that influence basic properties like density, viscosity, or heat capacity. Water also exhibits solid-liquid-vapor in a relatively narrow temperature and pressure range, leading to rapid changes in the underlying structure of hydrogen bonds. Moreover, water poses several anomalous properties, which complicate the matter of modeling even further.

Table 3.2: Default parametrization of nitrogen model withing different simulation software.

	bond length [Å]	molar mass$_{\mathrm{N_2}}$ [g/mol]	$\sigma_{\mathrm{LJ,N}}$ [Å]	$\epsilon_{\mathrm{LJ,N}}/k_{\mathrm{B}}$ [K]
MACSIMUS	1.08892776	28.0134	3.3078	36.6727
Mac_module	1.08892776	28.0134	3.3078	36.6727
*ms*2	1.0464	28.014	3.3211	34.897

Table 3.3: Default parametrization of SPC/E model withing different simulation software.

	OH bond length [Å]	HOH angle [rad]	molar mass$_{\mathrm{SPC/E}}$ [g/mol]	$\sigma_{\mathrm{LJ,O}}$ [Å]	$\epsilon_{\mathrm{LJ,O}}/k_{\mathrm{B}}$ [K]	δ_{H} [e]
MACSIMUS	1.0	$\arccos(-1/3)$	18.0154	3.1655	78.1974	0.4238
Mac_module	1.0	$\arccos(-1/3)$	18.0154	3.1655	78.1974	0.4238
*ms*2	-	-	18.0154	-	-	-

There is a known maximum of density around 4 °C maxima in heat capacity, speed of sound and other properties. Another anomalous points are suspected in the super-cooled liquid region of water as investigated by Vinš *et al.* [245].

In response to these complications, multiple models have been developed, trying to address specific temperature ranges and selected thermophysical properties that they could model with a higher precision while inevitably avoiding the rest. It is therefore crucial to be aware of the strengths of the selected model and account for its inaccuracies.

We will now address a predominately used Simple Point Charge Extended (SPC/E) water model [23] a successor of Simple Point Charge [24]. The original SPC and SPC/E follow the geometrical shape of water, which corresponds to the tetrahedral structure given by sp3 hybridization. The structure with bond lengths and angles is presented in fig. 3.5. The model contains three atoms with rigid bonds and angles. Molecules interact through the LJ potential, and because of partial charges present on oxygen and both hydrogen atoms, electrostatic interaction is also required. There is no LJ term on hydrogen. The details of the parametrization for MACSIMUS and Mac_module are presented in the table 3.3, for *ms*2 user have to supply own parametrization for specified water model.

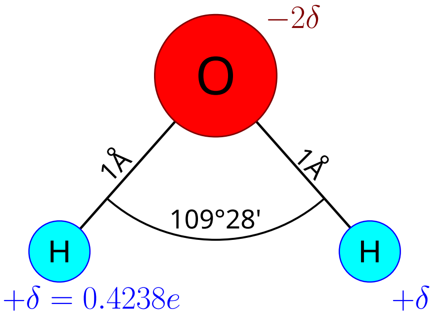


Figure 3.5: Visualization of the SPC/E water model with bond and angle specification. Partial charge on atoms is shown using $\delta\pm$.

Even with low complexity of SPC/E inter-molecular (SPC/E–SPC/E) interaction result in a one LJ evaluation and 3^2 electrostatic evaluation for all combination of charged sites. The Coulomb potential for the interaction is parametrized for H–H interaction with elementary charge e.

The main advantage of SPC/E is its simplicity and, therefore, the speed of simulation. The model performs well in the vapor region of the phase diagram but is not suitable for VLE calculation, where the original SPC model performs better, and is completely out of scope for solid prediction.

3.3 Monte Carlo simulation

3.3.1 Working principle

The goal of the methods is to evaluate a thermodynamic property (expectation value), $\langle X \rangle$, as an ensemble average of a quantity of interest, X, which is a function of configuration.

In the context of equilibrium thermodynamics, the system is described by a Hamiltonian defined in a phase space with position and momenta as variables. If we are interested only in time-independent quantities, the integration over momenta can be performed in advance, resulting in generalized position dependence, i.e., configuration space. Therefore, the target property is in the form of a mean over a configuration space, which in the canonical (NVT) ensemble reads as:

$$\langle X(\boldsymbol{r}^N) \rangle = \int X(\boldsymbol{r}^N) \rho(r^N) dr^N = \frac{\int X(\boldsymbol{r}^N) e^{-\beta U(\boldsymbol{r}^N)} d\boldsymbol{r}^N}{\int e^{-\beta U(\boldsymbol{r}^N)} d\boldsymbol{r}^N} \tag{3.9}$$

The integration runs over $3N$ degrees of freedom, i.e., N vectors $\boldsymbol{r}^N$. For β we follow the notation from the thermodynamic chapter. The denominator is the configuration integral. In general, the analytical solution of the integrals is prohibitively complex for increasing N. Therefore, a numerical solution is sought using the sampling method. We will focus on a description of a widely utilized extension of this idea first described by Metropolis [164] as this is the entry point in the Monte Carlo simulation of thermodynamic properties that is used in this thesis. For the working principle description, we will now introduce two concepts utilized in generic Monte Carlo methods, of which the Metropolis algorithm is no exception.

3.3.2 Ergodicity

A mathematically rigorous definition of ergodicity [65, 199] is too complex for our purposes. Instead, we will focus on two generally accepted physically-based formulations: [72, 197, 233].

Definition 3.1. Let us consider a system with prescribed energy with microstates forming a hypersurface in the phase space. If such a system fulfills the condition that a neighborhood of any state is reachable from any state in finite time, then the system is called *ergodic*.

This definition can be formulated alternatively [233] as:

Definition 3.2. Dynamic system evolving according to Hamilton's equations is called *ergodic*, when it visits all points on the constant energy hypersurface in an infinite time.

Note here the use of infinite time for the whole system as opposed to finite time reachability in the case of two microstates. The reason is the uncountable amount of microstates within the same energy hypersurface. The important consequence of these definitions is that the hypersurface does not have isolated or unreachable subregions.

Ergodicity is an important condition for molecular simulations. In the case of Monte Carlo, ergodicity makes the replacement of the integral for a sum meaningful because the whole system can be explored from the sampled states [72]. From the practical point of view, this assures the convergence of simulation (MD or MC) approximate evaluation of quantity of interest, i.e., for every $\epsilon > 0$ there exists simulation time or number of steps so that the standard error of the simulation averages is better than ϵ.

In general, the ergodic condition is assumed as always fulfilled. However, when studying energy barriers, especially in small systems, this assumption can be less clear. The ergodic condition often works within thermodynamic limit introduced by statistical mechanics. The limit effectively changes the perspective from microscopic to macroscopic by the means of using the central limit theorem in conjunction with increasingly large system where $N \to \infty$,$V \to \infty$ but density is held constant. In sufficiently big system the fluctuation become negligible because their magnitude decrease with $1/\sqrt{N}$ and the phenomenological Thermodynamics is obtained in form introduced in chapter 2.

3.3.3 Markov Chain

To introduce the Markov process in full would again extend beyond the scope of this study. We therefore limit ourselves to discrete case of Markov chain an introduce important concepts required for explanation of nucleation in fig. 2.5 and the Metropolis algorithm, which is primarily utilized throughout this study. Let us first note the definition of Markov Chain by Brooks *et al.* [31]:

Definition 3.3. A sequence $X_1, X_2, \ldots$ of random variables of some finite set is a Markov chain if the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$ depends on X_n only. The set in which the X_i take values is called the state space of the Markov chain.

Notation and properties used in Monte Carlo methods are provided below.

Definition 3.4. The onditional distribution of X_{n+1} given X_n will be called the transition probability distribution and denoted as (X_{n+1}, X_n).

Another common terms for transition probability are stochastic matrix or transition matrix.

Definition 3.5. The transition probability distribution is called reversible with respect to an initial marginal distribution of X_1 if $(X_k, X_{k+1}) = (X_{k+1}, X_k)$ for the Markov chain $X_1, X_2, \ldots$ and $k \in \mathrm{N}$.

Definition 3.6. A Markov chain is stationary if for $k \in \mathrm{n}$ the conditional distribution of $(X_{n+2}, \ldots, X_{n+k})$ given X_{n+1} does not depend on n.

Now we will formulate an important theorem that allows replacing integration with summation for the integral, as in the case of eq. (3.9). The proof can be found in [31].

Theorem 3.7 (strong law of large numbers)**.** *If* $\langle |X| \rangle < \infty$*, then* $\frac{1}{n} \sum_{k=0}^{n-1} X_k \to \langle X \rangle$ *almost surely as* $k \to \infty$*.*

3.3.4 Metropolis algorithm

Following the reasoning of section 3.3.1, the investigated ergodic system and reversible Markov chain with stationary transition distribution section 3.3.3 transforms the integral in eq. (3.9) into a sum of the generated samples.

$$\langle X(\boldsymbol{r}^N)\rangle = \frac{\sum_{k=1}^{\mathrm{k_{samples}}} X_k(\boldsymbol{r}^N)e^{-\beta U_k(\boldsymbol{r}^N)}}{\sum_{k=1}^{\mathrm{k_{samples}}} e^{-\beta U_k(\boldsymbol{r}^N)}} \tag{3.10}$$

The Markovian stationary and reversible sequence of sampled configurations $\boldsymbol{r}^N$ is constructed according to definition 3.3. The Metropolis algorithm described below ensures that the constructed Markov chain fulfills the reversibility and stationarity conditions and that the definition 3.7 is valid. For more complex analysis, these conditions can be used in error estimation done within the Monte Carlo method. Error estimation in Monte Carlo is not of concern in this study as molecular dynamics is utilized for the presented result data calculation.

We can finally proceed to the explanation of the method proposed by Metropolis in 1953 [164] and still in use today. First, a trial configuration, $\boldsymbol{r}^N_{\mathrm{trial}}$, is generated by a small change from $\boldsymbol{r}^N$ so that the conditional probability density of the opposite move is the same. The new configuration (step $k+1$) is

$$\boldsymbol{r}^N_{k+1} = \begin{cases} \boldsymbol{r}^N_{\mathrm{trial}}, & \frac{\rho(\boldsymbol{r}^N_k)}{\rho(\boldsymbol{r}^N_{\mathrm{trial}})} > \phi \\ \boldsymbol{r}^N_k, & \text{otherwise} \end{cases} \tag{3.11}$$

where $\phi \in [0,1)$ is a uniformly distributed (pseudo)random number. Note that if the probability of the trial configuration is greater than that of the old one, it is always accepted; otherwise, the trial configuration is rejected when it is unlikely and $\boldsymbol{r}^N_{k+1} = \boldsymbol{r}^N_k$.

3.4 Molecular dynamics

The second simulation method able to calculate the expectation values of thermodynamic quantities from a microscopic description is molecular dynamics. A sample set of configurations (so called trajectory) is obtained from an (in principle deterministic) time evolution instead of from stochastic sampling described in section 3.3. The fact that both methods give the same expectation values is the result of the Ergodic hypothesis section 3.3.2. Moreover, the time evolution of the trajectory is generated, which is often of interest.

We start with a general description of molecular dynamics in the frame of classical (analytical) mechanics, where important properties of the system are shown. Then, the most popular integration methods (Verlet and Gear) are covered, discussing the error propagation and stability of the schemas. Afterwards, constrained mechanics is introduced, and the method of preserving the bonds lengths is explained with the SHAKE algorithm. Methods required for the simulation of more complicated ensembles are also sketched. The focus is given to thermostats; methods that control the temperature of the system. A more detailed explanation of the core principles and related methods used in MD can be found in reference material [77, 197, 206]. The fundamentals presented here provide a mathematical outlook on molecular dynamics and provide insight required for introducing the GPU paradigm into the simulation, which will be the task in later chapter.

To simplify the description, the universal term *element* will be used for a molecule or atom. In this way, monoatomic argon and triatomic water molecules can be interchanged in the description without loss of generality in the first few sections before structural concerns are answered with constrained mechanics. The word *system* is used here for a set of interacting particles under prescribed force field and initial conditions.

3.4.1 Practical formulation of the problem

Here we first introduce the general case of molecular dynamics of system in Cartesian coordinates with the usually employed assumptions. The purpose is to give reader unfamiliar with MD a quick overview in simple terms, which is later formalized.

The modeled system of N molecules gets its mechanical behavior from classical physics with interaction potential in the form presented in section 3.2. In this framework, the force field is traditionally composed of *bonded forces* and *non-bonded forces*,

$$U = U_{\mathrm{b}} + U_{\mathrm{nb}}. \tag{3.12}$$

The bonded forces, U_{b}, include chemical bonds, bond angles, and torsions. The non-bonded forces include interactions of atoms at two different molecules for simple fluids considered here [2]. External forces affecting the system may also be included in the non-bonded contribution.

For the common case without external forces, the non-bonded forces are most often approximated by the *pairwise-additive* or *two-body* approximation as

$$U_{\mathrm{nb}}(r^N) = \sum_{i=1}^{N}\sum_{j=1}^{i-1} u^P(r_{ij}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} u^P(\boldsymbol{r}_i, \boldsymbol{r}_j) \tag{3.13}$$

where $r_{ij} = |\boldsymbol{r}_j - \boldsymbol{r}_i|$ is the interatomic distance, and the sum is over all atoms except for bonded ones (1–2 and 1–3). In practice, attempts for a more precise description beyond pair additivity are rare because of large computational demands. They may include polarizability, short-range three-body forces (as Axilrod–Teller potential), etc. We remind in passing that the second form of eq. (3.13) has its application in GPU-based implementation.

When all potential energy contributions are known, the force can be directly evaluated as a negative gradient. The force acting on particle i then becomes

$$f_i = -\frac{\partial}{\partial r_i} U(r^N) = -\sum_{i\neq j} u'(r_i, r_j)\frac{r_{ij}}{|r_{ij}|} \tag{3.14}$$

with the derivative analytically available. Molecular dynamics is then based on the numerical integration of Newton's (or equivalent) equations of motion using the forces calculated in this manner.

3.4.2 Hamiltonian and its properties

In this section, the general working principle from section 3.4 is formalized using the Hamiltonian formalism [233]. The use of the formalism allows for inspection of error propagation and derivation of important properties like conservation laws used for validation of the simulation. The Hamilton's equations of motion read as:

$$\dot{q}_\alpha = \frac{\partial \mathcal{H}}{\partial p_\alpha} \tag{3.15}$$

$$\dot{p}_\alpha = -\frac{\partial \mathcal{H}}{\partial q_\alpha} \tag{3.16}$$

where q and p denote the generalized coordinates and momenta, respectively, and the subscript α denotes the elements of the vector variables q and p. The Hamiltonian $\mathcal{H}$ is formulated using

[2] For more complicated substances, the self-interaction can't be neglected. Examples are hydrocarbons where interactions are separated by more than three bonds (the case of three bonds, i.e., 1–4, can include both torsions and rescaled non-bonded terms)

Lagrangian $\mathcal{L}$, which is typically expressed in the Carthesian system predominately used in MD. The Hamiltonian

$$\mathcal{H}(\boldsymbol{q},\boldsymbol{p}) = \sum_{\alpha=1}^{N} p_\alpha \dot{q}_\alpha - \mathcal{L} = \sum_{\alpha=1}^{N} \frac{\dot{q}_\alpha^2}{2m_\alpha} + U(\boldsymbol{q}) \tag{3.17}$$

contains the kinetic and potential contribution. Following from the formulation of $\mathcal{H}$ in eq. (3.17), fundamental conservation laws are derived next. For this purpose, Noether's theorem [176] and its formulations can be applied when the underlaying symmetries are known. This leads to three conservation laws with a direct application in simulation itself.

3.4.2.1 Conservation of energy

The first symmetry of the Hamiltonian is with respect to time. The conservation can be proven in general (independent of the coordinate system used) by taking the time-derivative of the Hamiltonian,

$$\frac{d\mathcal{H}}{dt} = \sum_{\alpha=1}^{N} \frac{\partial \mathcal{H}}{\partial q_\alpha}\dot{q}_\alpha + \frac{\partial \mathcal{H}}{\partial p_\alpha}\dot{p}_\alpha = \sum_{\alpha=1}^{N} \frac{\partial \mathcal{H}}{\partial q_\alpha}\frac{\partial \mathcal{H}}{\partial p_\alpha} - \frac{\partial \mathcal{H}}{\partial p_\alpha}\frac{\partial \mathcal{H}}{\partial q_\alpha} = 0 \tag{3.18}$$

where the properties of eq. (3.17) and symmetry of second derivatives of the Hamiltonian were used to show the conservation of the Hamiltonian.

From a practical point of view, total energy, or Hamiltonian conservation, is an important test of numerical integration methods (the integrator or propagator) for stability, noise, and secular energy drift. This is especially important in our study, where adiabatic nucleation is considered.

3.4.2.2 Conservation of total momentum

For other conservation laws, one can either know the conserved property and prove and verify it, like it has been done in section 3.4.2.1, or discover the conserved property from the knowledge of the invariant. Here, Noether's theorem is reformulated using perturbation theory [208]. Following the notation of perturbation or infinitesimal transformation of the time $t' = t + \delta t(t, \boldsymbol{q})$ and coordinates $q'_\alpha = q_\alpha + \delta q_\alpha(t, \boldsymbol{q})$ states that for each Noether's perturbation and some perturbed motion function $f(t, \boldsymbol{q})$ [3] term $C_\mathrm{m}(t, q, \dot{q})$ corresponds to the constant of motion

$$C_\mathrm{m}(t, \boldsymbol{q}, \dot{\boldsymbol{q}}) = \left(\frac{\partial \mathcal{L}}{\partial \dot{q}_\alpha}\dot{q}_\alpha - \mathcal{L}\right)\delta t(t, \boldsymbol{q}) - \frac{\partial \mathcal{L}}{\partial \dot{q}_\alpha}\delta q_\alpha(t, \boldsymbol{q}) + f(t, \boldsymbol{q}) \tag{3.19}$$

In the context of this study, it can be understood as a sort of blueprint for converting known invariants into perturbation formalism and, with eq. (3.19) into conserved properties. For a more detailed explanation, refer to chapter 6.3 in the book by Lovelock and Rund [149].

To apply the shown procedure to total momentum, we start with the invariance of the system with respect to the translation group. This particular invariance is a direct consequence of the absence of any external field, as discussed in section 3.4.1. Continuing the perturbation formulation, we note that time is not affected by the transformation, removing the time perturbation from the formula. Nonetheless, the perturbation of the generalized coordinate reads as $q' = q + \delta q$. Due to the translational invariance the perturbation made can be understood as if the resulting perturbations do not matter and could be set to unity, i.e. $q' = q + 1$. Substituting into the

[3] the function needs to be differentiable and satisfy the intragral condition in equation (8) from Sarlet [208] but for here considered case it is equal to zero as only subclass of perturbation is used.

perturbation eq. (3.19) and with a motion function equal to zero, we find out that a system with translational invariance leads to the conservation of generalized momentum

$$\frac{\partial \mathcal{L}}{\partial \dot{q}_\alpha} = p_\alpha \tag{3.20}$$

Similarly to the conservation of total energy (the value of the Hamiltonian in time), the conservation of total momentum P represents a test of a numerical simulation algorithm (the *integrator* or *propagator*). Note that the translational symmetry is also preserved when the periodic boundary conditions are used.

3.4.2.3 Conservation of angular momentum

For a system with the Lagrangian insensitive to a small rotation $\delta\varphi$ around arbitrary axis a, we can formulate another conservation law. Utilizing the formalism from section 3.4.2.2, the change incurred by the rotation only affects the Cartesian coordinates, resulting in removed time perturbation and leaving coordinate perturbation in the form $q'_\alpha = q_\alpha + (a \times q)\delta\varphi$. By substituting into eq. (3.19) and utilizing circular shift property of triple product, $\boldsymbol{a} \cdot (\boldsymbol{b} \times \boldsymbol{c}) = \boldsymbol{b} \cdot (\boldsymbol{c} \times \boldsymbol{a}) = \boldsymbol{c} \cdot (\boldsymbol{a} \times \boldsymbol{b})$, we obtain the conserved property of angular momentum L_α with a similar choice of motion function:

$$\frac{\partial \mathcal{L}}{\partial \dot{q}_\alpha} \cdot (a \times q_\alpha) = p_\alpha \cdot (a \times q) = a \cdot (q_\alpha \times p_\alpha) = a \cdot L_\alpha \tag{3.21}$$

Analogously to the total momentum, the angular momentum observation is an important property for validation of the integrator in the free space (vacuum) boundary condition.

3.4.2.4 Time reversibility

When talking about time reversibility, the Hamiltonian formalism is not even required, yet it clearly illustrates that changing the direction of time has no effect on the form of the equation generated. Consider eq. (3.17), where time t is replaced with $-t$; therefore, the first time derivative also switches sign:

$$\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}(-t)) = \sum_{\alpha=1}^{N} \frac{(-\dot{q}_\alpha)^2}{2m_\alpha} + U(\boldsymbol{q}) = \mathcal{H}(\boldsymbol{q}, \boldsymbol{p}(t)) \tag{3.22}$$

This property is closely connected to energy conservation. A numerical finite-difference method for integration of the equations of motion that is time-reversible cannot exhibit a secular drift of total energy for a system in equilibrium. One would generally expect that small numerical errors are independent random variables (let us say δE in one time step of length Δt); then, the expected energy conservation error is $\delta E\sqrt{t/\Delta t}$ and can be both positive and negative. In the section 3.4.3 better methods will be shown.

3.4.2.5 Symplectic property

For this particular property, the Hamiltonian presented in section 3.4.2 needs to be transformed into matrix formalism. First, let us agglomerate generalized coordinates and momenta under one vector and transcribe the set of eq. (3.17) into vector form

$$\boldsymbol{x} = (\boldsymbol{q}, \boldsymbol{p}) = (q_1, \ldots, q_{3N}, p_1, \ldots, p_{3N}) \tag{3.23}$$

$$\dot{\boldsymbol{x}} = \left(\frac{\partial \mathcal{H}}{\partial p_1}, \ldots, \frac{\partial \mathcal{H}}{\partial p_{3N}}, -\frac{\partial \mathcal{H}}{\partial q_1}, \ldots, -\frac{\partial \mathcal{H}}{\partial q_{3N}}\right) \tag{3.24}$$

In concise block matrix notation this reads

$$\dot{\boldsymbol{x}} = \boldsymbol{M}\frac{\partial \mathcal{H}}{\partial \boldsymbol{x}} \tag{3.25}$$

$$\boldsymbol{M} = \begin{pmatrix} \mathbf{0} & \boldsymbol{I} \\ -\boldsymbol{I} & \mathbf{0} \end{pmatrix}, \tag{3.26}$$

where I is the square identity matrix of rank $3N$. We can see that $\boldsymbol{M}$ is regular, its transposition equals its inverse and its negative: $\boldsymbol{M}^{\mathrm{T}} = \boldsymbol{M}^{-1} = -\boldsymbol{M}$; i.e., it is orthogonal. The initial value problem, eqs. (3.25) and (3.26), leads to a unique solution which also means that a unique transformation exists, giving solution x_0 from x_t. The Jacobian elements for this transformation are:

$$J_{k,l} = \frac{\partial x_t^k}{\partial x_0^l}. \tag{3.27}$$

The properties of the Jacobian can be used to obtain information about the errors of the associated solution transformation and their propagation. For the Jacobian matrix eq. (3.27) of the described system eq. (3.25), we show that the following property $\boldsymbol{J}^T\boldsymbol{M}\boldsymbol{J} = \boldsymbol{M}$ is satisfied. For this reason, we apply the Hamilton formalism and search for the solution of eqs. (3.15) and (3.16) which can be rewritten for our conservative system as:

$$\dot{q}_\alpha = \frac{p_\alpha}{m_\alpha} \tag{3.28}$$

$$\dot{p}_\alpha = F_\alpha(\mathbf{q}) \tag{3.29}$$

As expected from section 3.4.1, we obtain the Newton equation of motion for which the solution is integrated up to the order of $O(t^2)$ as:

$$q_\alpha(t) = \frac{F_\alpha(\mathbf{q})t^2}{2m_\alpha} + \frac{p_\alpha(0)t}{2m_\alpha} + q_\alpha(0) \tag{3.30}$$

$$p_\alpha(t) = F_\alpha(\mathbf{q})t + p_\alpha(0), \tag{3.31}$$

where the formula for momentum was substituted into the final formula for position. The solution is then used to express the partial derivatives using the Kronecker delta notation:

$$\frac{\partial q_\alpha(t)}{\partial q_\beta(0)} = \delta_{\alpha,\beta}\ ,\ \frac{\partial q_\alpha(t)}{\partial p_\beta(0)} = \frac{t}{2m_\alpha}\delta_{\alpha,\beta} \tag{3.32}$$

$$\frac{\partial p_\alpha(t)}{\partial q_\beta(0)} = 0\ ,\ \frac{\partial p_\alpha(t)}{\partial p_\beta(0)} = \delta_{\alpha,\beta}. \tag{3.33}$$

Finaly, the Jacobian matrix is constructed from the partial derivatives in the block shape of the same dimensions as matrix $\boldsymbol{M}$.

$$\boldsymbol{J} = \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ \frac{t}{2m_\alpha}\cdot\boldsymbol{I} & \boldsymbol{I} \end{pmatrix} \tag{3.34}$$

The proof of the property $\boldsymbol{M} = \boldsymbol{J}^T\boldsymbol{M}\boldsymbol{J}$ now follows in a straightforward manner using blockwise multiplication

$$\boldsymbol{J}^T\cdot\boldsymbol{M}\cdot\boldsymbol{J} = \begin{pmatrix} \boldsymbol{I} & \frac{t}{2m_\alpha}\cdot\boldsymbol{I} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix}\cdot\begin{pmatrix} \mathbf{0} & \boldsymbol{I} \\ -\boldsymbol{I} & \mathbf{0} \end{pmatrix}\cdot\begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ \frac{t}{2m_\alpha}\cdot\boldsymbol{I} & \boldsymbol{I} \end{pmatrix} =$$

$$= \begin{pmatrix} \boldsymbol{I} & \frac{t}{2m_\alpha}\cdot\boldsymbol{I} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix}\cdot\begin{pmatrix} \frac{t}{2m_\alpha}\cdot\boldsymbol{I} & \boldsymbol{I} \\ -\boldsymbol{I} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{pmatrix} = \boldsymbol{M} \tag{3.35}$$

With this key property in eq. (3.35), the symplectic property is formalized in the following definition using the obtained result for $\boldsymbol{M}$ and $\boldsymbol{J}$.

Definition 3.8. Square matrix S of size $2n$ is called symplectic if and only if it satifies condition $\boldsymbol{S}^T \cdot \boldsymbol{M} \cdot \boldsymbol{S} = \boldsymbol{M}$, where $\boldsymbol{M}$ is the nonsingular square matrix with block form $\begin{pmatrix} \mathbf{0} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{pmatrix}$.

For simplicity, we refer to the Hamiltonian of the system as symplectic when the corresponding matrix $\boldsymbol{J}$ is symplectic. Similarly, a finite-difference integrator may also be symplectic, which in some cases leads to the existence of a perturbed Hamiltonian; in turn, the total energy differs from the exact solution by a bounded function, as discussed further in section 3.4.3.4.

3.4.3 Integration schema

For a pure molecular system (i.e., not a mixture), the problem is formulated as a set of equations of motion consisting of $N \cdot m$ second-order differential equations. Here, N represents the number of molecules, and m represents the number of atoms per molecule, as each atom's position needs to be calculated.

As discussed in the previous section on the properties of the Hamiltonian (Section 3.4.2), the precision of predictions can be enhanced, and error propagation can be mitigated within acceptable limits when the schema fulfills reversibility and symplectic properties. Therefore, the design of schemas that meet these conditions plays a pivotal role in the solution process.

With the aforementioned conditions in mind, a section on integration schemas will be presented with Verlet's and Gear's algorithms. The overall structure of an iterative schema is shown, along with further information such as bond preservation and temperature control. Furthermore, the explicit solution method for preserving constraints imposed by bonds and fixed angles within individual molecules will be demonstrated using the SHAKE algorithm.

3.4.3.1 Verlet

Let us start with the simple integration schema called the Verlet method [241]. The desired update of position q_α can be derived from the Taylor expansion around t by $\pm\Delta t$ (backward and forward),

$$q_\alpha(t \pm \Delta t) = q_\alpha(t) \pm \dot{q}_\alpha(t)\Delta t + \frac{1}{2}\ddot{q}_\alpha(t)\Delta t^2 \pm \frac{1}{6}\dddot{q}_\alpha(t)\Delta t^3 + \mathcal{O}(\Delta t^4), \tag{3.36}$$

where $\ddot{q}_\alpha(t) = F_\alpha(\boldsymbol{q}(t))/m_\alpha$ (and $\dot{q}_\alpha = p_\alpha/m_\alpha$ if needed). The velocity term can be removed from the equations by adding the two variants together:

$$q_\alpha(t + \Delta t) + q_\alpha(t - \Delta t) = 2q_\alpha(t) + \frac{F_\alpha(\boldsymbol{q}(t))}{m_\alpha}\Delta t^2 + \mathcal{O}(\Delta t^3) \tag{3.37}$$

This formula contains an error of fourth order and can be rearranged into the shape of a solution schema used in molecular simulation. This form provides an explicit formula for the calculation of the next time step,

$$q_\alpha(t + \Delta t) = 2q_\alpha(t) - q_\alpha(t - \Delta t) + \left[\frac{F_\alpha(\boldsymbol{q}(t))}{m_\alpha} + \mathcal{O}(\Delta t^2)\right]\Delta t^2, \tag{3.38}$$

from which it is apparent that the method is of second order if $\mathcal{O}(\Delta t^2)$ is neglected. A single force evaluation is required per time step.

The algorithm requires knowledge of $q_\alpha(0)$ and $q_\alpha(-\Delta t)$ for a start at time $t = 0$. The latter value can be obtained from the usual initial conditions, $q_\alpha(0)$ and $\dot{q}_\alpha(0)$, using Taylor expansion, eq. (3.36), to the second order; often, the first order suffices.

Even though the *schema* is simple, it fulfills the requirements for time reversibility, see section 3.4.2.4. The symplectic property (described for the Hamiltonian in section 3.4.2.5) is also preserved.

In the same spirit, velocities at time t are not know at the time the forces are evaluated. They can be evaluated after a step is finished using the second-order formula

$$\dot{q}_\alpha(t) = \frac{q_\alpha(t + \Delta t) - q_\alpha(t - \Delta t)}{2}. \tag{3.39}$$

The equivalent set of equations called "Velocity Verlet" is defined by two steps containing the velocity explicitly written here in momenta $p_\alpha = m_\alpha \dot{q}_\alpha$ as:

$$q_\alpha(t + \Delta t) = q_\alpha(t) + \frac{p_\alpha(t)}{m_\alpha}\Delta t + \frac{F_\alpha(t)}{m_\alpha}\frac{\Delta t^2}{2} \tag{3.40}$$

$$p_\alpha(t + \Delta t) = p_\alpha(t) + F_\alpha(t) + F_\alpha(t + \Delta t)\frac{\Delta t}{2} \tag{3.41}$$

Although the formulas contain the velocity explicitly, velocity $q_\alpha(t + \Delta t)$ is still available *after* the evaluation of $F(t + \Delta t)$.

3.4.3.2 Leap-frog

Another schema giving trajectory identical to both algorithms is called Leap-frog:

$$p_\alpha\left(t + \frac{\Delta t}{2}\right) = p_\alpha\left(t - \frac{\Delta t}{2}\right) + F_\alpha(\boldsymbol{q}(t))\Delta t \tag{3.42}$$

$$q_\alpha\left(t + \Delta t\right) = q_\alpha(t)m_\alpha + p_\alpha\left(t + \frac{\Delta t}{2}\right)\Delta t \tag{3.43}$$

In this approach, the update of momenta is carried out half a timestep before the update of positions. The position update, in turn, utilizes the obtained momenta, which makes the order of execution particularly important to preserve the beneficial properties of the schema. The velocity of eqs. (3.39) and (3.41) is equivalent to

$$p_\alpha(t) = \frac{p_\alpha(t - \frac{\Delta t}{2}) + p_\alpha(t + \frac{\Delta t}{2})}{2} \tag{3.44}$$

Although the trajectories of all the above methods are identical (with the appropriate initial conditions), there is a difference in the per element kinetic energies. The "Velocity Verlet" (VV) kinetic energy E_{LF} at time t is based directly on velocities eqs. (3.39) and (3.41), whereas the "Leap-frog" (LF) kinetic energy $E_{\mathrm{kin,LF}}$ is the arithmetic average of the energies calculated at half-times:

$$E_{\mathrm{kin,VV}}(t) = \frac{p_\alpha(t)^2}{2m_\alpha} \tag{3.45}$$

$$E_{\mathrm{kin,LF}}(t) = \frac{p_\alpha(t - \frac{\Delta t}{2})^2 + p_\alpha(t + \frac{\Delta t}{2})^2}{4m_\alpha}. \tag{3.46}$$

The total value is obtained as a sum over all elements in the system. Both versions of the kinetic energy calculation differ by $\mathcal{O}(\Delta t^2)$. In case when a thermostat is included (see below), the trajectories no longer differ.

According to Kolafa and Lísal [127], the LF schema is slightly more accurate for most quantities of interest. The LF schema is equivalent to the Verlet method, which means that the desirable properties of time reversibility and symplecticity are retained. This modified schema is favored for its fast performance and explicit nature, offering a balance between computational efficiency and precision.

3.4.3.3 Gear

An alternative to the direct propagator introduced by Verlet relies on the utilization of past positions and momenta (the "history"). In this way, previously calculated configurations are used to construct a multiple steps predictor-corrector method. One of such methods was originally introduced by Curtiss [52] and later formalized and named after Gear [80]. Because of the computational demand posed by the evaluation of forces, only a subclass of all schemas with a single RHS evaluation is of interest for use in MD. Depending on the number of time steps kept in the history, different order schemas can be designed. In this study, the Gear schema of the fourth order is employed; therefore, the following description focuses on this case.

History can be kept in the form of past timesteps $t, t-\Delta t, t-2\Delta t, t-3\Delta t$ or alternatively in form of numerical differences expressed in time t as:

$$Q_\alpha(t) = \begin{pmatrix} q_\alpha(t) \\ \dot{q}_\alpha(t)\Delta t \\ \ddot{q}_\alpha(t)\frac{\Delta t^2}{2} \\ \dddot{q}_\alpha(t)\frac{\Delta t^3}{6} \end{pmatrix}. \tag{3.47}$$

This description is equivalent, because the sum of the elements with arbitrary Δt gives the polynomial spanning of the aforementioned points. With this history vector, the predictor schema is constructed similarly, utilizing the Taylor expansion, here up to the fourth order. Collecting the terms into a matrix yields an easy-to-read formula for the predictor:

$$Q_{\alpha,\text{predictor}}(t+\Delta t) = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix} Q_\alpha(t) \tag{3.48}$$

Predicted positions can be used to calculate forces and evaluate the RHS of equation eq. (3.52), which can alternatively be calculated from the third row containing predictions for the forces. The difference can be interpreted as the error of the predictor:

$$Q_{\alpha,\text{error}}(t+\Delta(t)) = \frac{F_\alpha(\boldsymbol{q}(t+\Delta t))}{2m_\alpha}\Delta t^2 - Q_{\alpha,\text{predictor}}(t+\Delta t) \tag{3.49}$$

The corrector is:

$$Q(t+\Delta t) = Q_{\alpha,\text{predictor}}(t+\Delta t) + \begin{pmatrix} a_0 = 1/6 \\ a_1 = 5/6 \\ a_2 = 1 \\ a_3 = 1/3 \end{pmatrix} \cdot Q_{\alpha,\text{error}}(t+\Delta(t)) \tag{3.50}$$

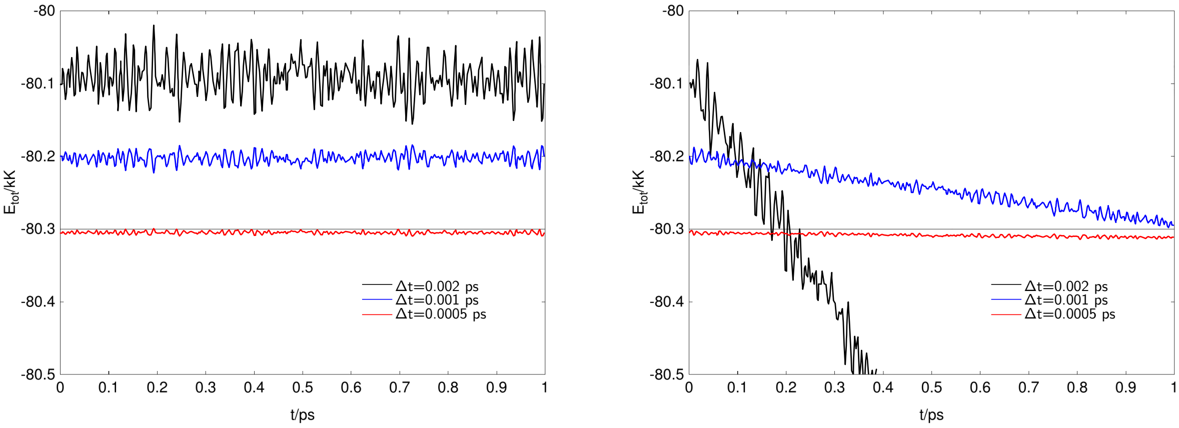


Figure 3.6: Leap-frog integrator

Figure 3.7: Gear integrator of fourth order

Figure 3.8: Energy conservation for decreasing integration step Δt. Energy evolutions are shifted by the same value for each integration step.

where constants $\{a_0, a_1, a_2, a_3\}$ are determined [80] using the conditions of maximum stability of the iterations. To be more precise: By linearizing the predictor-corrector schema, we get a set of linear recurrence equations for error propagation. The explicit solution of such a set contains the roots of the characteristic equation. If any of the roots were (on absolute value) greater than 1, the method would be unstable because any error would exponentially grow. The above set of constants guarantees that the roots are zero or unity (the latter correspond to the errors in the initial conditions which cannot be avoided).

The Gear method offers an effective integration schema with a higher order of accuracy which makes accurate MD simulations more efficient (longer time step). However, it's important to note that the Gear method lacks time reversibility (hence, it is not symplectic) due to its asymmetric nature. It necessitates the implementation of additional control mechanisms to maintain energy balance within the system. Despite these limitations, the Gear method is particularly valuable for NVT ensembles and also for rotating hard bodies with the equations of motion expressed by quaternions. Recently Janek and Kolafa [111] have reexamined the coefficients to improve time reversibility at the expense of the order of the method. For the above method, $\boldsymbol{a} = (0, 2/3, 1, 1/3)$.

3.4.3.4 Properties of the schema and their effects on precision

In section 3.4.2, the fundamental properties have been discussed without the context of their immediate use. Here we discuss several particular examples to elucidate the properties of various integrators.

The invariance to the change of the direction of time was shown for the system of section 3.4.2.4. In the numerical integration methods, this is also an important property, ensuring the accuracy of prediction and the propagation of error. It can be easily seen that any version of the Verlet method is time-reversible (it is even symplectic). In contrast, the Gear schema, like many other predictor-corrector schemas, does not fulfill the time reversibility condition [4]. The reason is the difference of result after application of the schema in the following forward-backward sequence: $t \to t + \Delta t \to t$. The associated error propagation can be observed in fig. 3.8, where various integration schemas for error propagation are compared. The symplectic property of the system section 3.4.2.5 is related to the preservation of the phase space. Because of this, if the schema

[4] The second-order Gear method is an exception because it is equivalent to Verlet and Leap-frog.

also fulfills the symplectic condition, then the generated trajectory will remain bound in the phase space packet. This means that the associated error will be bound as well. Therefore, the symplectic property ensures that error does not drift out of proportion, as illustrated by Leap-frog integrator in fig. 3.6. In comparison Gear schema is example of integrator not fulfilling time reversibility leading to rapid drift of energy shown in fig. 3.7.

3.4.4 Constraints

In our previous discussions, we have focused on an unconstrained system of interacting particles, referring to them as elements rather than atoms. However, it is important to introduce the distinction between atoms and molecules in this section. A molecule is a collection of atoms that possess a specific structure. While the concept of structure has not been significant in the earlier discussions, it becomes important when modeling more complicated substances. In this study, we will specifically focus on a narrower aspect of constrained molecular dynamics, addressing the following points of interest:

1. Modeling chemical bonds by a harmonic or other continuous function introduces high-frequency vibrations, especially for light hydrogen atoms, which require short time steps to integrate properly. This loss of efficiency can be to some extend alleviated using multiple-timesteps schemas.

2. Moreover, these fast vibrations are weakly coupled to the rest of the system. From the perspective of quantum theory, such vibrations are rarely excited anyway. Thus, vibrating hydrogen bonds are often replaced by fixed ones. In consequence, no energy is stored in the bonds.

3. The same argument applies for small molecules (such as water), which can be replaced by a rigid body. Although special equations for rigid body dynamics can be applied (e.g., based on quaternions), using constraints is a valid option.

4. It is often the case that positions or distances in a molecule need to be fixed; hence, the force is also of interest. Some temperature, pressure, or polarizability control mechanisms [26] may impose constraints on the generalized coordinates $(\boldsymbol{q}, \boldsymbol{p})$.

5. For model molecules, the angle constraints can be transformed into bond constraints. For example, in the SPC/E water model section 3.2.3, the angle $\mathrm{H-O-H}$ is preserved by fixing the H–H distance with an artificial bond.

In this section, the theory of holonomic constraints is used to derive mechanisms for bond preservation. These constraints are applied to maintain the fixed bonds and angles between atoms within a molecule during the simulation. To accomplish this during simulation, an algorithm called SHAKE (or its variants) is employed.

3.4.4.1 Holonomic constraints for rigid molecules

First, the definition of holonomic constraints, which are independent of the generalized momenta is written as follows:

$$c_i(\boldsymbol{q}) = 0, i \in \widehat{o} \tag{3.51}$$

The rigid bond in the nitrogen molecule fig. 3.4, can be used as the simplest example

With the set of o constraints, the task of bond preservation can be transformed into a problem of Langrangian multiplier using the Hamilton formalism, eqs. (3.28) and (3.29):

$$\dot{q}_\alpha = \frac{p_\alpha}{m_\alpha} \tag{3.52}$$

$$\dot{p}_\alpha = F_\alpha(\boldsymbol{q}) - \sum_{\beta=1}^{o} \lambda_\beta \frac{\partial c_\beta(\boldsymbol{q})}{\partial q_\alpha} \tag{3.53}$$

In this formulation, the Hamilton equations depend on $3N - o$ generalized coordinates and o constraints of the form of eq. (3.51). This can be formalized into a new changed Hamiltonian H_c where a point transformation from original coordinates to the modified ones generates a canonical transformation [83]. In theory, using the Hamilton formulation is preferred as opposed to Lagrange because it is possible to transform the system with constraints into the standard form of Hamilton equations eqs. (3.15) and (3.16) with a constrained Hamiltonian H_c. This is by no means straightforward as demonstrated by Ciccotti and Ferrario [49]. In the field of MD simulations, it is possible to use both the Lagrange and Hamilton formalism [55]; however, the Hamilton formalism leads to a loss of numerical precision because of the linear growth of the Lagrange multipliers in time.

Therefore, in the following section section 3.4.4.2, a more direct iterative approach is introduced using the Lagrange formalism [204].

3.4.4.2 Constrained dynamics

The constrained dynamics algorithm can be formulated specifically for the integration method like the Verlet method, which is often utilized in practical applications. But here a more general approach is preferred that illustrates the working mechanism of the algorithm as introduced in the previous section.

Let us start from an extended Lagrangian with Lagrange multipliers accounting for o holonomic constraints:

$$\mathcal{L} = \sum_{\alpha}^{N} E_{\text{kin}} - U(\boldsymbol{q}) + \sum_{\beta=1}^{o} \lambda_\beta \frac{\partial c_\beta(\boldsymbol{q})}{\partial q_\alpha} \tag{3.54}$$

Continuing in the derivation of the modified equations of motion from the Lagrangian, the following second order differential equation with unknown parameters λ_β is obtained:

$$\ddot{\boldsymbol{q}}_\alpha = \frac{F_\alpha}{m_\alpha} + \frac{1}{m_\alpha} \sum_{\beta=1}^{o} \lambda_\beta \frac{\partial c_\beta(\boldsymbol{q})}{\partial q_\alpha} \tag{3.55}$$

Solving for λ_β we can utilize the knowledge that any constraints $c_\gamma(\boldsymbol{q})$ are constants in time, therefore their second time derivatives are zero.

$$\frac{\mathrm{d}^2 c_\gamma(\boldsymbol{q})}{\mathrm{d}t^2} = \sum_{i}^{N} \ddot{\boldsymbol{q}}_i \frac{\partial c_\gamma(\boldsymbol{q})}{\partial q_i} + \sum_{i}^{N} \sum_{j=1}^{N} \dot{\boldsymbol{q}}_i \frac{\partial c_\gamma(\boldsymbol{q})}{\partial q_i \partial q_j} \dot{\boldsymbol{q}}_i = 0 \tag{3.56}$$

The second derivative contains terms similar to those in the modified equation of motion. Substituting them from eq. (3.55) into eq. (3.56) with the change to i, j indices notation is generated:

$$\sum_{i=1}^{N} \frac{F_i}{m_i} \frac{\partial c_\gamma(\boldsymbol{q})}{\partial q_i} + \sum_{i=1}^{N} \frac{1}{m_i} \sum_{\beta=1}^{o} \lambda_\beta \frac{\partial c_\beta(\boldsymbol{q})}{\partial q_i} \frac{\partial c_\gamma(\boldsymbol{q})}{\partial q_i} + \sum_{i=1}^{N} \sum_{j=1}^{N} \dot{\boldsymbol{q}}_i \frac{\partial c_\gamma(\boldsymbol{q})}{\partial q_i \partial q_j} \dot{\boldsymbol{q}}_i = 0 \tag{3.57}$$

Rearranging the terms and shortening notations with mass matrix $\boldsymbol{M}_{\beta,\gamma}$ and force term F_γ, the final set of equations can be written in simple form:

$$\sum_{\beta=1}^{o} \boldsymbol{M}_{\beta,\gamma}\lambda_\beta + F_\gamma = 0 \tag{3.58}$$

$$M_{\beta,\gamma} = \sum_{i=1}^{N} \frac{1}{m_i}\frac{\partial c_\beta(\boldsymbol{q})}{\partial q_i}\frac{\partial c_\gamma(\boldsymbol{q})}{\partial q_i} \tag{3.59}$$

$$F_\gamma = \sum_{i=1}^{N} \frac{F_i}{m_i}\frac{\partial c_\gamma(\boldsymbol{q})}{\partial q_i} + \sum_{i=1}^{N}\sum_{j=1}^{N} \dot{\boldsymbol{q}}_i \frac{\partial c_\gamma(\boldsymbol{q})}{\partial q_i \partial q_j}\dot{\boldsymbol{q}}_i \tag{3.60}$$

Solving the set of o unknown Lagrange multiplies from eq. (3.58) is the final step in the algorithm. Depending on the structure of the molecule, this procedure may require a sparse matrix solver like the conjugated gradients method in the case of large protein molecules. The constraint forces equal the Lagrange multipliers. In addition, the algorithm guarantees the constraints with the integrator precision only. Such errors would explosively grow in time; therefore, another step (fortunately with the same matrix) must be added to guarantee the constraints.

SHAKE In contrast, the SHAKE algorithm merges both steps to one. While constrained dynamics solves the matrix for all Lagrange coefficients at once, SHAKE iterates through individual constraints.

SHAKE algorithm is closely related to the Verlet (or equivalent) method building upon already derived position update, eqs. (3.40) and (3.43). The basic principle of the algorithm is:

1. For past configuration $\boldsymbol{q}(t)$, where constraints are satisfied, perform the update of position into $\boldsymbol{q}(t+\Delta t)$ without considering constraints.

2. Correct the positions of two involved particles q_j, q_k to preserve the constraint c_i.

3. If $i > 1$, repeat step 2 for all constraints $c_i, i \in \widehat{o}$, until they are all satisfied. Note that multiple iterations may be required as correcting one constraint may break other.

The correction is done with a fictive force acting in direction $q_{jk} = q_k - q_j$, and the scale of this force is acaled with λ (converges to the Lagrange multiplicator). This parameter is evaluated from the condition that constraint $[q_{jk}(t+\Delta t) + \lambda q_{jk}(t)]^2 = l^2$ where l is bond lenght of the desired q_{jk}. The approximate result neglecting terms of higher order $\mathcal{O}(\Delta t^2)$ is:

$$\lambda = \frac{|q_{jk}(t+\Delta t)|^2 - |q_{jk}(t)|^2}{2q_{jk}(t+\Delta t)\cdot q_{jk}(t)} = \frac{|q_{jk}(t+\Delta t)|^2 - l^2}{2q_{jk}(t+\Delta t)\cdot q_{jk}(t)} \tag{3.61}$$

The positions are then updated taking into account individual masses of the elements.

$$q_{\mathrm{j,SHAKE}}(t+\Delta t) = q(t+\Delta t) + \lambda\frac{1/m_j}{1/m_j + 1/m_k}q_{jk} \tag{3.62}$$

$$q_{\mathrm{k,SHAKE}}(t+\Delta t) = q(t+\Delta t) - \lambda\frac{1/m_k}{1/m_k + 1/m_j}q_{jk} \tag{3.63}$$

The symmetrical nature of this change ensures preservation of the center of mass and momenta. An important benefit of the method is time reversibility, when constructed with already time reversible method of a Verlet kind.

3.4.5 Non-microcanonical ensemble consideration

Until this point, most of the explanations have required mathematical principles and have not included much of the physical nature of the problem. Terms like thermodynamic ensembles were not incorporated into the derivation of the method's principles. Our assumptions – constant energy (the Hamiltonian is the integral of motion), constant number of molecules, and constant volume (or a system of finite size in a vacuum) correspond to the microcanonical ensemble (for definition see 2.1), also referred to as NVE[5]. It is more practical to fix temperature, as it is done in most experiments. For this sake, the canonical ensemble is denoted as NVT.

3.4.5.1 NVT ensemble

In theoretical thermodynamics, the transformation from constant-entropy (or thermal energy in equilibrium) to constant temperature is achieved with the Legendre transformation. If a MD simulation is allowed to exchange energy with some external thermal bath, the total energy will fluctuate. It can be shown that under specific conditions, some of these methods generate samples of the NVT ensemble or do so in the thermodynamic limit. Three main examples are discussed next.

Andersen thermostat Andersen [7] proposed the selected particle velocity to be sampled from the Maxwell-Boltzmann distribution. In this way, the method is similar to Monte Carlo, and the generation of a canonical ensemble average is produced, as Andersen demonstrated in ref. [7]. Interaction time and molecule(s) affected by the thermostat are selected in accordance with the Poisson process. A normal distribution is then utilized to obtain the new velocity with the pseudo-random number generator.

In practice, the Andersen thermostat is not often used because of its slow convergence and discontinuities in trajectories.

Berendsen thermostat Another approach to adjusting the temperature during molecular dynamics (MD) simulations is to utilize the relation that the instantaneous temperature is in equilibrium given by the total kinetic energy of molecules in the system via the equipartition principle. By scaling the velocities of the molecules according to the desired temperature, the temperature can be effectively controlled. The scaling factor is [175]:

$$\sqrt{\frac{T}{T_{\mathrm{kin}}}} = \sqrt{\frac{T \cdot \mathrm{DoF} \cdot k_{\mathrm{B}}}{\sum_{i=1}^{N} \frac{p_i^2}{2m_i}}} \tag{3.64}$$

Here the kinetic temperature is calculated from the equipartition theorem, where DoF denotes the number of (ergodic) degrees of freedom. In practice, a single update is enough to adjust the temperature to the desired value and may only be required occasionally. For continuous rescaling in each timestep, only a subset of molecules is selected to provide more gradual temperature control. This "velocity-rescaling" method is a primitive way of controlling temperature. The produced ensemble is not canonical.

A more flexible version of velocity rescaling is the common Berendsen method. In the most usual implementation, the rescaling is embedded to the equations of motion. Using the

[5] There is a small inaccuracy in this statement. The usual definition of the microcanonical ensemble does not fix other integrals of motion except the total energy, whereas momentum and angular momentum (depending on the symmetry of the external potential) are conserved in the "MD-NVE" and do not fluctuate.

linearization of eq. (3.64), we get the equivalent form:

$$\ddot{\boldsymbol{r}}_\alpha = \frac{\boldsymbol{F}_\alpha}{m_\alpha} - \kappa(T_{\text{kin}} - T)\dot{\boldsymbol{r}}_\alpha, \alpha \in \widehat{N} \tag{3.65}$$

where the "friction parameter" κ can be adjusted to set the temperature relaxation time.

This thermostat offers greater control over the speed of convergence, but once again, the generated ensemble is not canonical. The benefit is generally its fast (exponential) convergence, as heat is quickly removed from fast moving molecules in the system. Both the Berendsen and direct velocity-scaling thermostats can lead to the artifact called the "flying ice cube". In such scenario, a group of molecules in a fixed structure referred to as an ice cube is rotating to balance out the targeted kinetic temperature requirement.

Canonical Sampling through Velocity Rescaling In this method by Bussi et al. [33], the friction parameter κ is a random variable with the probability distribution selected in a way that the canonical ensemble is obtained. The thermostat can still relax exponentially to the target temperature as in previous case. This method is called Canonical Sampling through Velocity Rescaling (CSVR) and it is commonly used in computational biochemistry.

Nosé-Hoover thermostat Extending upon the versatility of the Berendsen method, another well-known modification of the equation of motion was proposed by Nosé [177], where the system is coupled with an artificial thermostat by means of another variable. The method belong to a group of methods called "Extended Lagrangian".

In his first article, Nosé added a pair of variables $s = (q_s, \dot{q}_s)$ in the Lagrange formalism to link the system with the thermal bath. Variable q_s is the scaling factor of velocities; in addition, potential energy $\text{DoF}_{\text{Nose}} k_{\text{B}} T \ln(q_s)$ and kinetic energy $Q{\dot{q}_s}^2/2$ are added to the Lagrangian:

$$\mathcal{L}(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \sum_{\alpha=1}^{N} \frac{m_\alpha \dot{q}_\alpha^2}{2q_s^2} + \frac{Q\dot{q}_s^2}{2} - U(\boldsymbol{q}) - \text{DoF}_{\text{Nose}} k_{\text{B}} T \ln(q_s) \tag{3.66}$$

where DoF_{Nose} is the number of degrees of freedom (conserved properties subtracted) plus one for s. Parameter Q affects the timescale of the interaction with the heat bath; it is sometimes called "thermostat mass" because of its place in the kinetic energy term. The Hamiltonian is easily obtained from the Lagrangian using eq. (3.17) as:

$$\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) = \sum_{\alpha=1}^{N} \frac{p_\alpha^2}{2m_\alpha q_s^2} + U(\boldsymbol{q}) + \text{DoF}_{\text{Nose}} k_{\text{B}} T \ln(q_s) + \frac{p_s^2}{2Q} \tag{3.67}$$

with conjugate momenta $p_\alpha = m_\alpha \dot{q}_\alpha / q_s^2, p_s = Q\dot{q}_s$. By integrating the partition function (or the mean value of any observable) over the whole hypersurface of the constant value Hamiltonian, Nosé [178] demonstrated that the Hamiltonian in the eq. (3.67) form indeed generates the canonical distribution. The equations of motion from the modified Nosé Hamiltonian then read:

$$\dot{q}_\alpha = \frac{p_\alpha}{m_\alpha q_s^2} \tag{3.68}$$

$$\dot{p}_\alpha = F_\alpha \tag{3.69}$$

$$\dot{q}_s = \frac{p_s}{Q} \tag{3.70}$$

$$\dot{p}_s = \sum_\alpha \frac{p_\alpha^2}{m_\alpha q_s^3} - \frac{\text{DoF}_{\text{Nose}} k_{\text{B}} T}{q_s} \tag{3.71}$$

Since scaling q_s in the kinetic energy causes practical problems if variable q_s drifts out of unity for any reason (integrator errors, changes caused by cooling), Hoover [102] later proposed the following non-canonical variables: $dt = dt_{\mathrm{old}} = q_s dt_{\mathrm{new}} = q_s d\tilde{t}$ transforming the previous set of variables into unscaled real time here denoted with a tilde:

$$d\tilde{t} = \frac{dt}{q_s} \tag{3.72}$$

$$\tilde{q}_\alpha = q_\alpha, \;\; \tilde{p}_\alpha = \frac{p_\alpha}{q_s} \tag{3.73}$$

$$\tilde{q}_s = q_s, \;\; \tilde{p}_s = \frac{p_s}{q_s} \tag{3.74}$$

This unscaling returns to a more straightforward interpretation of the quantities, where problems with time drift from the original Nosé method are alleviated and time-step does not need to be adjusted as it was the case in Nosé formulation.

Finally, it is more convenient according to Nezbeda *et al.* [175] to transform $s = (q_s, \dot{q}_s)$ into $\ln s = \eta$. Applying this transformation to eqs. (3.68) to (3.71), we get the practical Nose-Hoover equation of motion used in molecular dynamics as follows:

$$\dot{q}_\alpha = \frac{p_\alpha}{m_\alpha} \tag{3.75}$$

$$\dot{p}_\alpha = F_\alpha - \frac{p_\eta}{Q} p_\alpha \tag{3.76}$$

$$\dot{\eta} = \frac{p_\eta}{Q} \tag{3.77}$$

$$\dot{p}_\eta = \sum_\alpha \frac{p_\alpha^2}{m_\alpha} - \mathrm{DoF}_{\mathrm{Nose}} k_{\mathrm{B}} T \tag{3.78}$$

Since non-canonical transformations have been used, the conserved property calculated from these equations is not a Hamiltonian. The Nosé–Hoover thermostat is sometimes extended to several added variables which improves egodicity. For use in this thesis, the thermostat in the form of eqs. (3.75) to (3.78) is sufficient.

3.4.6 Solver procedure

In this section, we discuss more technical issues of the implementation of the above theory to systems of molecules interacting via a force field.

Let us first assume that software for translating the user input into molecular structure, simulation parameters including the force field, and placement of the molecules into the initial structure within the system volume are all available. For simplicity, we further assume that the equilibration of the initial configuration is performed in a similar manner to the solver procedure described now. This effectively means that the initial conditions (starting positions and starting velocities of all atoms in the system) have been relaxed into a more probable configuration, whose properties can now be calculated in the so-called production run. During production, the simulation is performed with the evaluation of system properties like pressure, and either saved or written to the console. This is repeated until the requested simulation time is reached. The description now proceeds with individual tasks performed during a single timestep. For clarity, one of possible solver structure sequence is given in the form of a list:

I. **Initialization or update of interaction structure**
For more efficient force evaluation, it is common to use an interaction storage structure which reduces the $\mathcal{O}(N^2)$ complexity of the naive treatment of pairwise-additive forces

by two nested loops. Two main approaches are the Verlet neighbor list and the domain decomposition, which preserve information about close molecules that have a chance for interaction. This structure is updated during this step to reflect the configuration of the system.

II. **Force evaluation**
In this step, individual interactions between atoms in the system are evaluated from the holding structure. This is the most expensive operation within the solver procedure. The forces $F(t)$ calculated at time t are accumulated into the force vector used in the integration schema.

III. **Integrator step**
Following the chosen integration schema, previous positions, velocities, and currently calculated forces are used in the prediction of the next positions. In this section, Leap-frog schema is chosen as an example of the internal structure.

1. **Velocity update**
   Calculate velocities at time $t + \Delta t/2$, eq. (3.42).
2. **Position update**
   Calculate the new positions at time $t + \Delta t$, eq. (3.43)
3. **Constraints**
   If there are constraint bonds, run the SHAKE algoritm until converged. Positions at time $t + \Delta t$ are updated.
4. **Velocity correction**
   If SHAKE has been used, calculate new velocities at time $t + \Delta t/2$ from positions at $t + \Delta t$ and t. (This and the previous two steps can be merged together into the algorithm called RATTLE.)
5. **Velocity rescaling**
   Velocity-rescaling thermostat can be implemented here.

Note that the Nosé–Hoover thermostat and similarly a barostat, and also the algorithm used in the supersonic expansion simulation, require a knowledge of velocities at the **II force evaluation** stage, where they are not known. This issue can be solved by several means. The above steps excluding force evaluation may be iterated; if fully converged, the algorithm is time-reversible (but not symplectic). There are tailored exactly time-reversible methods using half and quarter-steps [160]; again, the most demanding forces are evaluated only once. Finally, it is possible to use predicted velocities [111].

IV. **Properties evaluation**
Evaluate the total kinetic and potential energy as well as other requested thermodynamic properties and outputs.

Clustering 4

Clustering is the technique of separating a set of objects or data points into similarity groups called clusters. The similarity within groups is determined with criteria or classifiers based on the property or feature of individual data points. When no previous characteristic or external information is employed and the criterion operates on unrefined data, the process is called unsupervised classification. In this way, clustering may reveal previously unknown structures within the data, making it a very useful technique with a wide range of applications. These include machine learning, data mining, image analysis, bioinformatics, and research on nucleation.

One of the key characteristics of clustering is the ambiguity of the cluster definition, which varies according to the criteria used. In consequence there are many variants of clusters that do not need to be equivalent in characteristics (amount of points,shape,cross section). This is especially the case for different families of criteria. For example; clustering used in the machine learning detects more spread cluster patterns as opposed to the compact ones produced from molecular systems. Both fields also pose their own challenges for the criteria used, such as the data points with tenths of dimensions or the large amounts of data. To account for the ambiguity of cluster definitions in the text, in cases where the type of cluster is not clear from the context, a specification about the criterion is supplied, i.e., the Stillinger cluster.

Cluster analysis is a challenging problem not only because of the definition ambiguity but also because of the general need to consider multiple factors like similarity measures, criterion complexity, efficiency, and stability of detection with respect to the initial conditions. This creates a complicated field to navigate generally, and specific clustering methods need to be specially designed based on problem formulation to account for the mentioned aspects. Furthermore, it is well known [133] that no clustering method is capable of handling every type of cluster structure (including shape, size, and density). And the solution using combination of clustering methods is still a challenging problem [133]. This motivates the design problem specific cluster criteria.

Following sections introduce various criteria for clustering, starting with a review of criteria used for unsupervised clustering. Unsupervised clustering is usually employed in clustering applications and data analysis tools. The main characteristic for these use-cases is the higher dimensionality (see definition 4.2) of the space in which the data are located. The general introduction is then continued into the criteria used for physical cluster detection, applicated primarily to the spatial data generated from molecular simulation. This difference in the data is also major reason why unsupervised and physical clustering are not mentioned together in the literature [109, 133, 182, 264] creating separation between these categories.

It is the opinion of the author that there is a lot of inspiration to be obtained from comparing those two fields. The signs of the inspiration are shown in the text as some of the techniques naturally permeate through the research fields.

4.1 Unsupervised clustering algorithms

Cluster analysis is the organization of a set of observations, usually represented as a vector of measurements. The vector alternatively represents a point in a multidimensional space. The clustering algorithm is a procedure that assigns a label to each point in space based on the point's location with respect to other surrounding points. The procedure often relies on point similarity (closeness to other points) to make the decision, but no other apriori information is usually provided.

In this section, the used mathematical notation is formalized. Afterwards, a brief summary of the main categories of clustering algorithms is presented. The categories follow naming used by Kotsiantis and Pintelas [133], who designed category names based on the main concept of the criteria in the category. In this section, focus is given to introducing the general idea and properties of the criteria family with an illustration of the family member algorithm. For a more detailed review, please refer to the summary paper by Xu and Tian [264].

4.1.1 Terminology of clustering

To be able to capture the general nature of the solved problem, the following notation is introduced based on the publication by Jain *et al.* [109]. The first concepts to introduce are observation and attribute.

Definition 4.1. An observation is the vector $\boldsymbol{x} = (x_1, \ldots, x_o)$ of o attributes (measurements) x_i.

where the vector of atributes is called an Observation. Then for given observation

Definition 4.2. The total number of attributes x_i, denoted as o is the dimensionality of given observation x.

Following the notation a set of n observations is $\mathcal{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Individual attribute j of the observation i in the set is denoted using double indexing $x_{i,j}$. In this way, the set of observations can be understood as a rectangular matrix of attributes with rank n, o. A clustering is then defined for the attribute matrix as follows:

Definition 4.3. For a set of observations $\mathcal{X}$ clustering is mapping of observations to integer labeled clusters $l_i \in \widehat{k}$. Integer number k is then the total number of clusters (labels).

The clustering, which is also a function, is then called *hard clustering* and allows for labels to be unified with clusters. This situation occurs when outliners (or noise) are not assigned to any cluster during the classification. The most common case in unsupervised hard clustering is to use the bijective clustering function. The case where mapping is not injective is called *fuzzy clustering.*

The individual attributes of observation can be both quantitative and qualitative entities like numbers, intervals, or enumerations of properties (e.g., colors). Depending on the attributes quantifiability, a metric can be constructed to compare the similarity of two observations. In further text, observations located in parametric (feature) space are denoted as points, and the metric is denoted as *distance* with the notation $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

The most common distance is the Minkowski metric:

$$d_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{k=1}^{o} |x_{i,k} - x_{j,k}|^p \right)^{1/p} . \tag{4.1}$$

With the Euclidean metric for $p = 2$ and the Manhattan metric for $p = 1$. These are simple examples that rely only on observations. More advanced distance evaluation can also include the surrounding points of the examined $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$, introducing the attribute matrix as an additional input $d(\boldsymbol{x}_i, \boldsymbol{x}_j, \mathcal{X})$. Refer to [109], for more examples of metric.

4.1.2 Density based clustering

Since proximity plays a key role in an intuitive notion of a cluster, methods utilizing the nearest neighbor (NN) classification based on the minimum distance from all points in the surrounding area are the first set of algorithms naturally considered candidates for clustering. The NN is also essential portion of more complicated algorithms presented in further sections. One of the first iterative procedures of this kind was proposed by Lu and Fu [79], where the NN technique with threshold was used to categorize graphemes [1]. In a sense density based clustering algorithms are constructed as a generalization of the nearest neighbor approach. Clusters are then based on the spatial coverage of data points in a region. The key idea of density-based clustering is that for each instance of a cluster, the neighborhood of a given radius r has to contain at least a minimum number of instances $n_{\mathrm{min,r}}$, which directly relate to n_{min} NN identification.

DBSCAN algorithm The DBSCAN stands for Density Based Spatial Clustering of Applications with Noise and was introduced by Ester *et al.* [67]. DBSCAN uses the previous notation of distance to construct clusters based on the closeness of the points, utilizing again principle of NN for the identification. The identification is formalized into the following direct density reachability:

Definition 4.4. A point $\boldsymbol{x}$ is directly density reachable from point $\boldsymbol{y}$ when distance separating them is smaller than radius parameter of the algorithm $d(\boldsymbol{x}, \boldsymbol{x}) < r_{\mathrm{DBSCAN}}$ and number of points in neighborhood of $\boldsymbol{x}$ is greater than $n_{\mathrm{min,r}}$.

The neighborhood is a generally accepted set of all points with a distance lower than the neighborhood radius. In principle this is a single step of the NN algorithm with additional condition specifying the minimal number of neighbors. To follow this parallel, density reachability is defined as:

Definition 4.5. A point $\boldsymbol{x}$ is density reachable from point $\boldsymbol{y}$, when there exists chain of direct density reachable points $\boldsymbol{z}_i i \in \widehat{n}$ connecting both points $\boldsymbol{x} = \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n = \boldsymbol{y}$.

By this approach, the notion of the nearest neighbor chain is extended to include sufficient density until the last element, which may not fulfill the second condition from definition 4.4. Such construction allows clustering of high density point structures as clusters while avoiding paths of loose conection in between. The DBSCAN therefore differentiates between three sets of points: core points fulfilling density and connectivity conditions; border points only fulfilling the connection; and noise points not fulfilling any of the conditions.

Algorithm is then initiated from an arbitrary point and iteratively extended using the condition of definition 4.4 to classify the points into the three cathegories. When there are no further inclusions possible from the initially selected point (to either core cluster or border), the algorithm proceeds to the next unclassified point. Points that do not fulfill any condition during the procedure are classified as noise.

[1] written characters representing letters

Properties of density based algorithms

- Has medium level of complexity $\mathcal{O}(n \log n)$ where n is number of points.
- Exhibits high clustering efficiency of arbitrary shape of data, but is sensitive to systems with nonhomogeneous density.
- Contains implicit noise filtering.
- Has increased sensitivity to parametrization of criterion. This is especially important with regards to density of investigated system.

4.1.3 Partitioning based clustering

A distinct feature of the partitional clustering algorithm is the construction of a partitioning structure, which is a set of dividing hyperplanes in o dimensional space. This division is usually created by optimizing the error distance function to accommodate the initially specified number of clusters. During the method operation, the partitioning is constructed around the centeroid, which is an artificially created point representing the center of the cluster. Finding an optimal placement for the centroids then defines the clustering. This is done iteratively constructing a partitioning structure around the centroids during the iteration.

An alternative construction of the partitioning introduced by Kaufman and Rousseeuw [117] can be done using medoids. Artificially added centroids are replaced with actual points called medoids, which are used as representatives of a cluster. For more detail, please refer to the authors book [117], which introduces the algorithm in the context of its implementation.

k-means algorithm The best-known method of utilizing centroids is the k-means algorithm [163]. At the start of the k-means, n_{clust} initial centroids are randomly placed in the parameter space. Then all the remaining points are assigned to the closest centroid.

$$f_{\text{k}-\text{means}}(x_i) = p \iff ||\boldsymbol{x}_i - \boldsymbol{c}_p||^2 \leq ||\boldsymbol{x}_i - \boldsymbol{c}_j||^2 \; \forall \in j \in \widehat{n_{\text{clust}}} \tag{4.2}$$

Then a search for the optimal placement of centroids is initiated, where the value of the following error function

$$\text{err}(\mathcal{X}, \mathcal{L}) = \left(\sum_{k=1}^{n_{\text{clust}}} \sum_{k=1}^{n_j} ||\boldsymbol{x}_i^{(j)} - \boldsymbol{c}_j||^2 \right)^{1/2} \tag{4.3}$$

is minimized. The Euclidean metric is used here to compare the distance between all n_j points of j-th cluster $\boldsymbol{x}_i^(j)$ with the cluster's centroid c_j. This minimization is done by updating the centroids positions as the mean of all points within the cluster

$$c_j = \frac{1}{n_j} \sum_{k=1}^{n_j} x_j, \tag{4.4}$$

and calculating the update point assignment with eq. (4.2) until all position of the centroids converge.

Properties of k-means algorithm

- Is efficient in processing large data sets with complexity of $\mathcal{O}(n_{\text{clust}} n t)$, where n is number of points and t is number of iterations.

- Can become trapped at the local optimum. More elaborate initial centroid placement were developed to ammend this issue.

- Is oriented towards finding spherical–shaped clusters.

- It is sensitive to noise and all data need to be available at the start.

Relation to Voronoi tesselation Of note is the relation of the k-mean algorithm to Voronoi tessellation, visible from the point assignment in eq. (4.2). We can illustrate this with a model case where each point of Voronoi tessellation is chosen to represent a centroid for k-means. The resulting Voronoi cells then exactly corresponds to the clustering constructed by the k-means in that particular iteration.

4.1.4 Hierarchical based clustering

These methods group data instances into a tree of clusters using two main approaches: agglomerative and divisive. The agglomerative methods build clusters incrementally until one big cluster is created at the root. The divisive methods split the space until only monomer clusters remain at the leaf nodes. The tree terminology was used intentionally because both approaches yield tree data structures called dendrograms.

The main difference between the realizations of the hierarchical algorithms is the different formulation of the similarity metric. The metric is this family of algorithms called proximity, including single-link, complete-link, and average-link definitions.

Single-link similarity considers the similarity between the two most similar instances, one from each cluster. It handles non-elliptical shapes well but is sensitive to noise and outliers.

Complete-link similarity, on the other hand, measures the similarity between the two most dissimilar instances from different clusters. It is less affected by noise and outliers but may break large clusters and struggle with convex shapes.

Average-link similarity strikes a compromise between the two, allowing for the balance of features from the previous two cases. This approach is used to tailor solution specific to the solved problem.

Single link agglomerative algorithm A comprehensive description of the algorithm is given by Sneath and Sokal [214], where the single link algorithm is a context of complete and average linking with different underlying metrics used.

The basic operational principle of the algorithm is to construct a tree from the list of links produced by proximity metric. At first, each point is assigned a leaf node of tree [2] with distinct labeling. This creates bottom level representing the primitive monomer clustering. Then a list of links between points is constructed, where each link contains participating points and the associated proximity value. In the single-link case, proximity can be the nearest neighbor Euclidean distance. Sorting this list by proximities in ascending order provides a construction manual for the dendrogram. Each new level of the tree then includes links with a step-wise increase in the considered proximity cutoff d_k. For floor k, then all links with proximity less than d_k are already represented by edges in the dendrogram. The sorted list labels are updated during the process, allowing reuse of the ordered list until all the points contain the same label, which completes the dendrogram construction.

[2] the node of tree that is only connected to parent but not child

Properties of hierarchical clustering algorithms

- Is fast to construct and has complexity of $\mathcal{O}(n)$.
- There is no apriori need for cluster specification.
- Porides an easy visualization via the dendrogram. Multiple cluster partitionings can be created by cutting the dendrogram at different levels.
- Proximity definition can be constructed to fawor certain shapes of clusters.

Realation to partitioning algorithm Considering divisive algorithms, a similarity with the space partitioning algorithms, like binary splitting tree, can be observed. In both types of algorithms a tree of volumes (clusters) are constructed. The tree structures are also similar in the representation of the root[3] node as all-encompassing volume (cluster) and the leaf nodes as volumes (clusters) containing individual monomers. The notable difference is the used proximity definition, which, in the case of a binary splitting tree, is a simple halving of volume. In the case of a k-d tree, the median of points contained in the volume is chosen as the dividing structure[4].

Further examples of hierarchical algorithm Some of the more sophisticated hierarchical clustering algorithms, as reviewed by Xu and Tian [264] include: Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [269], Clustering Using REpresentatives (CURE) [92] and (CHAMELEON) [115] which relies on graph partitioning.

4.1.5 Grid based clustering

Expanding on the concepts of the density-based clustering, this approach introduces the concept of a grid. The quantized space is divided into a finite number of cells (hyper-rectangles), and necessary operations are performed within this artificial grid. Thus, many operations can be simplified, leading to more efficient algorithms. For example, distance can be determined purely from the cell indexes instead of the original attribute vectors (i.e., by using the Manhattan metric). The typical algorithms based on grids are STatistical INformation Grid-based method (STING) [252] and CLustering In QUEst (CLIQUE) [4].

The core idea of STING is to use the rectangular grid as partitioning and construct hierarchical structure like in section 4.1.4. The data within single levels is then iteratively clustered, navigating the tree structure. This also means only voxel-like clusters can be created, connected solely through the rectangular neighboring faces.

CLIQUE takes advantage of the grid to follow the construction of density-based clustering section 4.1.2. The algorithm identifies dense cells in lower dimensions and expands into higher dimensions, utilizing the identified dense cells as guesses. This is particularly beneficial for data with high dimensionality, or data where principal components are hard to identify. When all dense cells are identified, CLIQUE forms a graph of dense cells and performs a depth-first search to find all connected components of the graph. These components are then labeled as the resulting clusters.

Properties of grid based clustering algorithm

- Have low time complexity of $\mathcal{O}(n)$ as well as high efficiency of calculation.

[3] node of the tree that has no parent but only child connection

[4] This is again a choice that can be linked to the idea of the k-means algorithm.

- Show a high potential for parallelization with good scalability on the size of the grid.
- The detection is sensitive to mesh granularity. With similar concern for the accuracy of the cluster detection.
- The choice of grid edge parameter has to be adjusted to detected objects for best results. And there results are represented with voxels.

4.1.6 Model based clustering

The basic idea is to select models for each cluster and find the best fit for the model. The particular model can be, for example, a mixture of Gaussian distributions where individual clusters are located by the individual Gaussian distributions. The parameters of the model (Gaussian mixture model) are estimated from the points used as data.

There are two main types of model-based clustering classes: statistical learning (e.g., EM, COBWEB and GMM) and neural network learning (e.g., SOM and ART) which have received a lot of attention in recent years.

The concept of the model based clustering is introduced with one of the first examples of statistical learning: the expectation-maximization (EM) algorithm by Dempster *et al.* [56] using a mixture model.

Expectation maximization The input data for the algorithm are observations y_i, $i \in \widehat{n}$ with dimensionality o which are understood as random realizations from some unknown parametrization of the model distribution. The algorithm then adjusts the parametrization to fit the data with maximum confidence. The selected model, in this case, is the multivariate Gaussian density mixture. Because of the underlying clustering task, the model is multinodal and composed of n_{gm} density distributions, where a single distribution has the following structure:

$$\phi_k(\boldsymbol{y_i}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma_k}) = \frac{1}{\sqrt{(2\pi)^o \det(\boldsymbol{\Sigma}_k)}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_k)\right). \tag{4.5}$$

The single multivariate distribution is parametrized with a means vector μ_k of size o and a square covariance matrix $\boldsymbol{\Sigma}_k$ of rank o. By manipulating the covariance matrix, the model can be adjusted from general shaped cluster detection to, i.e., targeting only spherical clusters with $\boldsymbol{\Sigma}_k = \lambda \boldsymbol{I}$. To shorten further notation, the parametrization of a single density distribution is often collected under a single parametric structure denoted with θ. This means that the mixture is expressed as

$$\sum_{k=1}^{n_{\text{gm}}} \tau_k \phi_k(\boldsymbol{y}_i|\theta_k), \tag{4.6}$$

with weights τ_k stating how likely it is the observation $\boldsymbol{y}_i$ belong to the k-th distribution. This means the weights are probabilities and need to satisfy $\tau_k \leq 0$ and $\Sigma_{k=1}^{n_{\text{gm}}} \tau_k = 1$.

The likelihood that the given parametric structure $\boldsymbol{\theta}$ and weights $\boldsymbol{\tau}$ correspond to the set of observations $\mathcal{Y}$ is given by

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathcal{Y}) = \prod_{i=1}^{n} \sum_{k=1}^{n_{\text{gm}}} \tau_k \phi_k(\boldsymbol{y}_i|\theta_k). \tag{4.7}$$

The EM algorithm operates on the observation as if it was an incomplete set sampled from the full data set $\mathcal{X}$, where the subset $\mathcal{Z}$ contains the unobserved elements. Assuming the $\mathcal{X}$ data set is independent and with the same distribution [75] as the observed $\mathcal{Y}$ the observed likelihood, which is searched by the EM algorithm, is specified as:

$$\mathcal{L}_o\left(\mathcal{Y}|\boldsymbol{\theta},\boldsymbol{\tau}\right) = \int \mathcal{L}\left(\mathcal{X}|\boldsymbol{\theta},\boldsymbol{\tau}\right) dz, \tag{4.8}$$

and the complete data log-likelihood of the observed data y_i based on z_i is

$$l(\boldsymbol{\theta},\boldsymbol{\tau},z_{i,k}|\mathcal{X}) = \sum_{i=1}^{n}\sum_{k=1}^{n_{\mathrm{gm}}} z_{ik}\log\tau_k\phi_k(\boldsymbol{y}_i|\theta_k), \tag{4.9}$$

where coefficients z_{ik} identify whether x_i belongs to nodal group k.

With the specified observed log-likelihood, the EM algorithm [56] contains two steps: the expectation step (E-step) and the maximalization step (M-step). Expected value is then constructed [75] as follows:

$$\tilde{z}_{ik} = \frac{\tilde{\tau}_k\tilde{\phi}_k(\boldsymbol{y_i}|\tilde{\theta}_k)}{\sum_{j=1}^{n_{\mathrm{gm}}}\tilde{\tau}_j\tilde{\phi}_j(\boldsymbol{y}_i|\tilde{\theta}_j)}. \tag{4.10}$$

The tilde used for parameters represents the expected parameterization of the following iteration of the algorithm. In the maximization step, the optimal parametrization is searched, giving the maximal log-likelihood eq. (4.9). For a multivariate normal mixture, the estimate for the M step can be expressed as:

$$\tilde{\tau}_k = \frac{n_k}{n} \tag{4.11}$$

$$\tilde{\mu}_k = \frac{\sum_{i=1}^{n}\tilde{z}_{ik}\boldsymbol{y}_i}{n_k} \tag{4.12}$$

$$n_k = \sum_{i=1}^{n}\tilde{z}_{ik}, \tag{4.13}$$

where covariance matrix $\tilde{\Sigma}_k$ depends on the type of parametrization used. Specific shapes are given shown in the work by Celeux *et al.* [39].

The complete algorithm gives the means and covariance matrices of the model, which localize the clusters. Choosing the confidence level, the isosurface boundary can be drawn to representing the clusters.

Properties of model based clustering algorithm

- Has generally high time complexity reaching up to $\mathcal{O}(n^2)$ where n is number of points. This complexity primarily depends on the choice of model and the optional use of heuristics.
- There is a wide range of models to select from. Moreover, a custom model can be designed to fit the particular problem.
- Have issues to solve problems without apriori knowledge. For case where number of clusters is unknown another model needs to be used for estimation.
- Have a high accuracy of prediction with attributed confidenece.

4.2 Physical clustering algorithms

From general case of the unsupervised clustering, the physical clustering first adjusts the supposed data dimensionality, number of data points, and metric.

Observations are based on data obtained mostly by molecular simulation and can be comprised of spatial data, velocity data, and, optionally, potential energy. The dimension of the space is reduced to 3, which results in low dimensionality for the observations. On the other hand, the number of observations is equal to the number of molecules in the system, which can make the datasets substantially large. The metric used for spacial and velocity data is the Euclidean distance, and for the case of potential energy, the absolute difference is usually sufficient.

The previous specification would still fit within the framework of unsupervised clustering but there are thermodynamic considerations that need to be included. The algorithms need to identify the clusters and not classify noise as a separate cluster, meaning that only hard clustering algorithms with outliers removed are considered (see section 4.1 for specifics).

From the theory of physical cluster [140] follows another important condition, which is the unique division. The condition states in simple terms that physically observed distinct groups are also classified as two clusters. In consequence, the loose contention of groups of molecules[5] cannot be classified as a single cluster even though it is a single group.

Another concern is related to the non-uniformity of the detected objects. During the nucleation, various sizes and, more importantly, shapes of clusters are present in the system. This is further considered for purposes of CNT, where similar configurations are deemed equivalent. Fulfilling these conditions enables the development of a nucleation theory from the physical cluster definition [140]. For the physical clusters, a certain balance needs to be struck between proper cluster identification and individual cluster configuration.

In the following text, the term criterion is used for physical based clustering algorithms to differentiate from the algorithm introduced for unsupervised clustering. Some of the most commonly used formulations of cluster criteria operating on the vapor→liquid transition are presented. The criteria are mostly based only on the spatial data [211] but attempts have been made to introduce potential energy into the detection,i.e., in the work by Hill [100].

4.2.1 Naive criterion

A close relationship between the physical clustering and the density based clustering is leveraged in this cluster criterion operational principle. The physical cluster is understood as a close grouping of molecules, relying on density as the main factor for the classification.

One of the simplest approaches, called “naive” is mainly used for illustrative purposes. The idea of the criterion is to evaluate the density on the neighborhood with radius r_{naive} around each molecule. The clusters are identified with a number density value, deciding whether the molecule is densely or sparsely surrounded. Clusters can be identified as densely surrounded molecules. This, in principle, can work well when the homogeneous distribution of uniform clusters is being investigated and only the molecules at the center of the molecular groups are identified, as visualized in fig. 4.1. Arrows from atom centers are used to represente belonging to the cluster illustrated in the fig. 4.1.

This is the “naive” assumption, which is rarely fulfilled. In consequence, this criterion makes it difficult to parametrize the choice for $r_{\text{naive}} \in (2\sigma_{\text{LJ}}, 10\sigma_{\text{LJ}})$. Furthermore, when heterogeneous cluster sizes are present in the system, the criterion reports more clusters than there are physically present in the system (see red molecule in fig. 4.1). This overcounting contradicts the requirement of unique cluster division, meaning nucleation theory cannot be developed.

[5] Can be imagined as Christmas lights where lights are the clusters and loose conection is the electric cable.

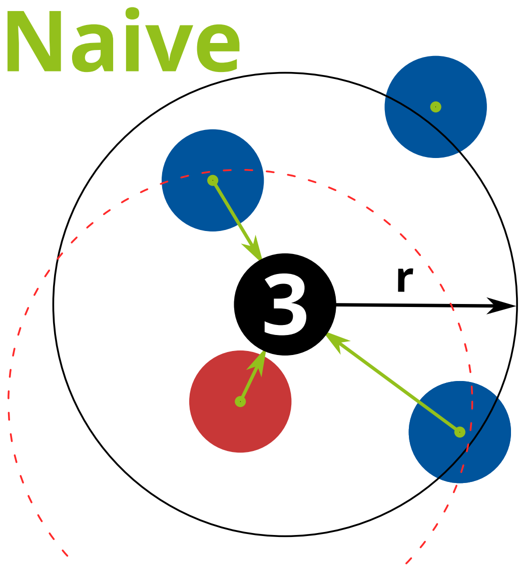


Figure 4.1: Example of naive criteria cluster classification with $N = 4$ molecules. Black molecule was used to detect the cluster with three neighbors within radius r. Red molecule report the same cluster as black.

To correct the overcounting issue of the naive approach, the following modification is proposed. In the labeling step, an agglomeration of the densely surrounded molecules removes the extra reporting. The agglomeration can be done hierarchically or, like in the case of the density algorithm with the nearest neighbor approach. The agglomeration is described in section 4.1.4 and section 4.1.2 respectively.

4.2.2 Stillinger criterion

One of the most widely used definitions of a physical cluster is the one presented by Stillinger [223]. In this criterion, a cluster is defined as groups of molecules where each is reachable from another. The corner molecules are reached using a chain of interconnected molecules no further apart than some apriori set radius $r_{\text{stillinger}}$. The operation principle is illustrated in fig. 4.2, where the neighboring range is visualized using circles and arrows that point to the point centers used for the detection.

For reference, a more rigorous definition of the criterion was used in definition 2.2. There is also a relation with the nearest neighbor technique as well as the formulation of denity reachability from section 4.1.2. The Stillinger criterion is then implemented using recursion as a way to grow the clusters from the monomers used for initialization. This iterative approach is reason why the criterion is sometimes referred to as the Stillinger chain.

Addressing the conditions required for physical cluster to construct a nucleation theory we start with the unique separation. Given the choice of $r_{\text{stillinger}}$ a clear separation between clusters can achieved. The usual starting choice is $r_{\text{stillinger}} = 1.5\sigma_{\text{LJ}}$ of underlying LJ potential for the investigated substance or the first peak of radial distribution function. The bigger concern is the role of the configuration. Stillinger has formulated an expression for the excluded volume of the system based on a specific cluster of length only. There is an argument to be made [140] that for the Stillinger cluster, the configuration of molecules in the chain also plays a significant role, not only the chain length. This has led to improvements extending the Stillinger criterion with N/V volume concepts section 4.2.3 leading to the N/V-Stillinger criterion.

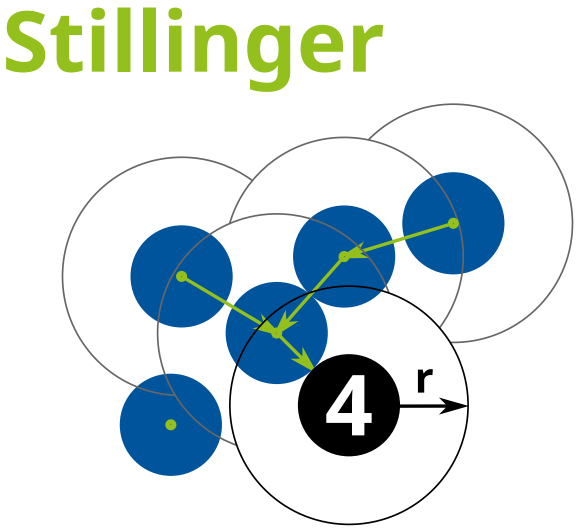


Figure 4.2: Example of Stillinger criterion cluster identification with $N = 5$. Detection radius r is shown for performed recursion with arrows denoting the parent molecule.

4.2.3 i/V and N/V criteria

This criterion builds upon the foundation idea of the i/V criterion proposed by Lee *et al.* [140]. They considered a cluster of i particles, that are confined to the spherical volume V. The enclosure sphere is constructed around the center of mass of the involved i particles, as illustrated in section 4.2.3 left. This approach enabled the calculation of some properties of a cluster, like its free energy or radial distribution function, but for use in the nucleation theory, it was deemed not too useful. The reason being the arbitrary choice of confining volume, which made the cluster definition not unique. The history of further modification is presented in the work of Ellerby *et al.* [66], with a more detailed analysis of the confining volume.

One way to resolve this confinement issue is presented by Schaaf *et al.* [209]. They proposed an energy barrier condition where molecules within the cluster require a total potential energy contribution from other molecules in the cluster to be below an apriori set bound α. The outside molecules are classified as those with the energy contribution above α. To achieve the desired behavior the potential is modified into:

$$u_{\mathrm{N/V}}(r) = \begin{cases} u(r), & r > r_{\mathrm{min}} \\ -u(r_{\mathrm{min}}), & r \leq r_{\mathrm{min}} \end{cases} \tag{4.14}$$

This resembles a proximity criterion in the form of interaction energy, but the high energy particles are excluded (classified as part of the vapor phase). This allows the particles of the clusters to be determined, allowing the calculation of the center of mass. A sphere can be then constructed with a radius equal to the distance of the furthest molecule in cluster from the center of mass. This way, $N - 1$ molecules are part of the cluster interior, and one molecules is removed to the cluster shell. Cluster shell is still counted as part of the cluster, as illustrated in section 4.2.3.

The presented construction ensures unique division. Furthermore, each cluster represents a class of configuration sharing the same center of mass, volume and energy barrier α.

This enables the creation of an equilibrium cluster, forming the basis for the work of formation derivation. Nucleation theory based on the N/V cluster is shown in the work of Schaaf [209].

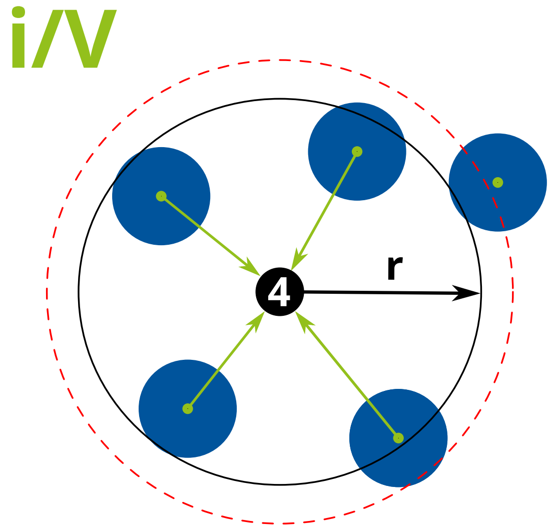


Figure 4.3: Example configuration of i /V cluster with $N = 4$ shown in the center of mass. The issue with arbitrary bounding volume choice is illustrated by red dashed line.

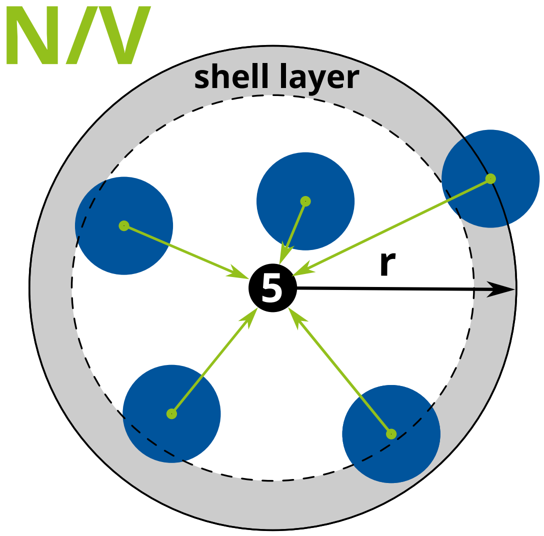


Figure 4.4: Example configuration of N /V cluster with $N = 5$ shown in the center of mass. Shell layer is visualized in gray with the furthest molecule at the border of the bounding volume.

Nucleation from the equilibrium view 5

5.1 Introduction

Accurate modelling of vapour-liquid phase interfaces is important for better understanding of various natural processes, such as formation of droplets in atmosphere or bubbles in liquid water, but also for an improved design of many engineering applications, e.g., in the purification of natural gas, design of combustion processes, modelling of cavitation, or in development of new carbon capture and storage (CCS) technologies. In particular, the prediction of a non-equilibrium phase transition, connected with formation of a new phase out of a mother phase in a metastable state, requires a detailed knowledge of the phase interface and its properties. The phase interface can primarily be defined by its shape, which is in a basic case planar or in case of nucleation of droplets and bubbles spherical. Even though the phase interfaces play a crucial role in phase transitions their description is still far from being satisfactory. This work focuses on semi-empirical modelling of phase interfaces that can be used in engineering applications rather than on more comprehensive and computationally demanding approaches such as molecular simulations.

This research task represents a culmination of a long-term research of group at the Institute of Thermomechanics of the Czech Academy of Sciences. Modelling of a planar phase interface of a multicomponent system was in the department of Thermodynamics first investigated by Vinš [247] which was later reexamined for the spherical interface geometry of one component system [106, 193]. The main aim of this work was to develop solution method of GT combined with SAFT-type EoS for the spherical phase interface geometry of a multicomponent systems with initially outlined in Celný *et. al.* [45]. The method is extended into derivation of a computational approach for more general interface geometry in multicomponent system including additional important properties with experimental relevance.

A brief introduction to the gradient theory was already given in sections 2.2.3 and 2.5, where both planar and spherical geometry is shown. This allows for the derivation of the related core problem in section 5.2. The core problem is subsequently solved in section 5.3 with a so-called split scheme, which divides the problem into an algebraic and a differential part. These parts are described in section 5.3.1 and section 5.3.2, respectively. The solution occurs as a fixed sequence according to constrains imposed by the core problem formulation. Output quantities of the model are primarily the concentration profiles and other quantities describing the phase interface, i.e. the surface tension, the work of formation and the surface adsorption. The important formulas used for the computation of additional properties are described in section 5.3.3. In section 5.4, the results are discussed in detail and compared with the available experimental data for the examined CO_2 relevant mixtures with n–butane and SF_6 and with the predictions of a more general Density Functional Theory by Ebner and Evans [64,68]. The theoretical results were validated on experimental measurements performed in the department by Vinš *et al.* [248] and available in the literature.

5.2 Problem formulation

The problem is formulated for general case of vapor→liquid phase transition of multicomponent system. Starting from the theoretical investigation of work of formation the condition of most probable density profile representing the system can also understood as a problem of finding the saddle point of the work of formation functional $\Delta\Omega$. Type of this saddle point present in the work of formation functional space is further influenced by underlying interface geometry. The riding-like saddle point is characteristic for the spherical phase interface, while the ravine-like saddle point is characteristic for the planar phase interface.

The variational calculus provides a reliable framework for this kind of problem definition. In this context the criterion for the optimal density profile is in terms of variation of work of formation

$$\delta\Delta\Omega\left[\rho(s)\right]_{\rho=\rho_0} = 0. \tag{5.1}$$

The variation is performed over density profiles $\rho(s)$ around some profile ρ_0 with general coordinate s. We have employed notation from section 2.5 and utilized the assumption of homogeneity of the system in the other two spatial coordinates leaving only s where the density inhomogeneity across the phase interface is expressed. The condition eq. (5.1) for extremum is then found as a solution of the Euler-Lagrange equations written for a system with n_{comp} components

$$S\,\frac{\partial\Delta\omega\left(\rho\left(s\right)\right)}{\partial\rho_k} + \frac{S}{2}\sum_{i,j=1}^{n_{\mathrm{comp}}}\frac{\partial c_{i,j}}{\partial\rho_k}\left(\frac{\partial\rho_i}{\partial s}\right)\left(\frac{\partial\rho_j}{\partial s}\right) - \frac{\mathrm{d}}{\mathrm{d}s}S\sum_{i=1}^{n_{\mathrm{comp}}}c_{i,k}\left(\frac{\partial\rho_i}{\partial s}\right) = 0,\ \ k\in\widehat{n_{\mathrm{comp}}}. \tag{5.2}$$

In the formula S denotes the saturation for the vapor→liquid transition and grand potential density difference of homogeneous and inhomogeneous systems $\Delta\omega$ is used. This set of n_{comp} equations [1] is further simplified rewriting the partial derivative to capture the meaning behind the driving force of the transition with

$$\frac{\partial\Delta\omega\left(\rho\left(s\right)\right)}{\partial\rho_k} = \Delta\mu_k. \tag{5.3}$$

The eq. (5.2) contains a new variable $c_{i,j}$ called the influence parameter, which in essence tries to quantify the effect of an individual component on the phase interface. The influence parameter for a pure component $c_{i,i}$ can be obtained from models or most commonly from the correlation of experimental data for the surface tension. In case of multicomponent systems no straightforward option is available and the mixed influence parameter $c_{i,j}$ with non-equal subscripts $i\neq j$ is therefore evaluated mostly as a geometric mean

$$c_{i,k} \doteq \sqrt{c_{i,i}\cdot c_{k,k}}. \tag{5.4}$$

Consequently, the influence parameters in mixtures are expected to be similar to those of one-component system. However for some systems, this assumption is not sufficient and eq. (5.4) is multiplied by $(1-\gamma_{ij})$, where γ_{ij} is an adjustable binary parameter. In this study we can omit the possible extension as it does not influence the developed mathematical model , i.e. $\gamma_{ij}=0$ is assumed in further calculations. We note that the influence parameter shows in general a certain

[1] hat notation $\hat{n}$ represent the set of integers $1,\ldots,n$

density dependence [53]. Nevertheless in most cases, the density dependence of $c_{i,j}$ is rather small and can be neglected leading to

$$\frac{\partial c_{i,j}}{\partial \rho_k} = 0,\ \ k \in \widehat{n_{\text{comp}}}, \tag{5.5}$$

which is also the case in this work. Combining eqs. (5.3) to (5.5) together with derivative notation $\rho_i' = \partial\rho_i/\partial s$ the set of Euler-Lagrange equations eq. (5.2) is transformed into

$$\Delta\mu_k - \sum_{i=1}^{n_{\text{comp}}} \sqrt{c_{i,i}\cdot c_{k,k}} \left(\frac{1}{S}\frac{dS}{ds}\rho_i' + \rho_i''\right) = 0,\ \ k \in \widehat{n_{\text{comp}}}. \tag{5.6}$$

The eq. (5.6) represents the core problem of the mathematical model formulated as a set of n_{comp} second order differential equations. The equations are supplied with Dirichlet boundary conditions obtained from the phase equilibrium calculation.

5.3 Problem solution

Term $\Delta\mu_k$ in eq. (5.6) is in general analytically non-integrable due to the fact that it is computed from complex EoS. No explicit relation for this term can be usually obtained as most od the EoSs do not lead to the analytically expression for $\Delta\mu_k$. Additionally, the left hand side in eq. (5.6) contains factor dS/ds dependent on the interface geometry. To overcome both problems simultaneously an extension to authors previous work [40] was performed and an unified numerical method for arbitrary geometry was developed.

Reorganizing the shape of eq. (5.6) to move the non-integrable term to the right hand side(RHS).

$$\sum_{i=1}^{n_{\text{comp}}} \sqrt{c_{i,i}} \left(\frac{1}{S}\frac{dS}{ds}\rho_i' + \rho_i''\right) = \frac{\Delta\mu_k}{\sqrt{c_{k,k}}}, k \in \widehat{n_{\text{comp}}}. \tag{5.7}$$

Solution of the core problem given by eq. (5.7) is possible in this form, however a considerable computational effort would be required. The defined problem require to solve n_{comp} mutually interconnected differential equations. Moreover, problem complexity greatly increase with the number of components making a direct solution prohibitivly expensive for $n_{\text{comp}} > 2$.

It is therefore favourable to further modify the form of the core problem eq. (5.7) to overcome the cumbersome fully differential solution. In this study, a similar approach as in refs. [51, 136, 166, 167] was used to transform the original set of equations into an algebraic problem and a simplified differential problem. A similar approach is to take the structure in eq. (5.7) where all elements with component index k were transferred to RHS. This allows to substract the first equation $k = 1$ from the remaining equations $k \in 2\dots n_{\text{comp}}$. Because all the equations have the same sum term at the left hand sized, the original set eq. (5.7) is transformed into

$$\sum_{i=1}^{n_{\text{comp}}} \sqrt{c_{i,i}} \left(\frac{1}{S}\frac{dS}{ds}\rho_i' + \rho_i''\right) = \frac{\Delta\mu_1}{\sqrt{c_{1,1}}} \tag{5.8}$$

$$\frac{\Delta\mu_k}{\sqrt{c_{k,k}}} = \frac{\Delta\mu_1}{\sqrt{c_{1,1}}},\ \ k \in 2\dots n_{\text{comp}}, \tag{5.9}$$

consisting of a single differential equation in eq. (5.8) and a set of algebraic equations in eq. (5.9). The differential equation is the first equation from the original problem and the algebraic part comprises of $n_{\text{comp}} - 1$ nonlinear equations with identical RHS. This new formulation allows for a so-called split solution. Both parts are to be solved in a fixed order beginning with the algebraic and continuing to the differential part. Continuity between both parts is facilitated by the RHS of the equation which is being passed between solvers. The order of execution cannot be changed so the new formulation given in eqs. (5.8) and (5.9) preserves the connections between original differential formulation in eq. (5.7).

As a part of the proposed solution the the following notation with artificial variable X as the linking value for both types of solved sub-problems is employed.

$$X = \frac{\Delta\mu_1}{\sqrt{c_{1,1}}} \tag{5.10}$$

In addition to variable X, the partial molar densities also need to be treated. Following modification is inspired by the problem of a monotonous density. In the original formulation of DGT by Cahn and Hilliard [34], multiple component systems required at least one partial density having a monotonous character along the main coordinate axis s. The *monotonous condition* implies that the remaining molar concentrations have to be expressed as functions of the selected monotonous density. Therefore, a problem with the selection of mixture components and their common relations arises. The order of components influence the solution method and in worse case even the obtained results. Moreover, the partial density of the first component is expressed in terms of the dimensional coordinate, while the other components are given as functions of the first component density. These issues leads to cumbersome description that have consequences in usability of the method and creates a preferential (unequal) treatment of some of the component densities.

We note that the proposed approach is similar to the work of Liang *et. al.* [146]. However, both studies were carried out independently. Major focus of this work has been the modelling of spherical phase interfaces of multicomponent systems. Both studies introduce a modified density $\tilde{\rho}$, which allows for softening the problem with the monotonous density selection as well as processing all partial densities in the same manner. In this case, all partial densities are described as functions of $\tilde{\rho}$ in the following way

$$\tilde{\rho} = \frac{\sum_{i=1}^{n_{\text{comp}}} \sqrt{c_{i,i}}\rho_i}{\sum_{i=1}^{n_{\text{comp}}} \sqrt{c_{i,i}}}, \tag{5.11}$$

where n_{comp} is the number of components in a mixture and $c_{i,i}$ is the influence parameter of the i-th pure component. The modified density can be understood as a weighted mean density, where the weights are the influence parameters of individual components. The idea of this type of density is motivated by the form of eq. (5.7).

The monotonous character is justified by the *monotonous condition* mandatining the existence of monotonous component with high value of the influence parameter. Visualization in fig. 5.1 shows an example for a binary mixture with c_1 of the *monotonous* component being almost 7-times higher than c_2 of the second component. In spite of the influence parameter involvement the benefit of modified density lies in its shape. It can be easily proven that the shape of the modified density remains the same for the first and the second coordinate derivatives. The proof relies on fact that the influence parameters are independent of dimensional coordinate.

With the density modified in terms of eq. (5.11) and variable X given by eq. (5.10), the differential part of eq. (5.8) can be written in the following way

$$\frac{1}{S}\frac{dS}{ds}\tilde{\rho}' + \tilde{\rho}'' = \frac{X}{\sum_{i=1}^{n_{\text{comp}}} \sqrt{c_{i,i}}}. \tag{5.12}$$

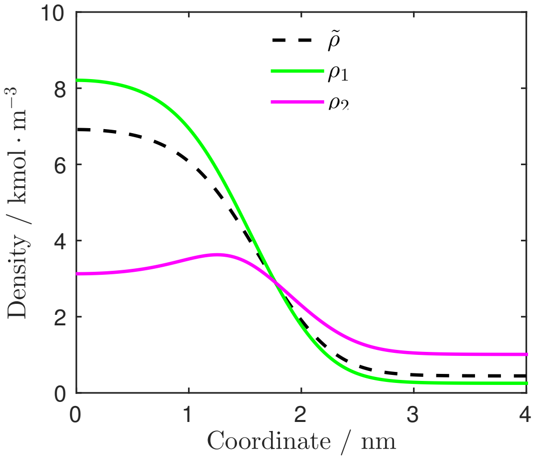


Figure 5.1: Illustration of the modified density profile over the planar phase interface. Influence parameters are $c_1 \doteq 1.608 \cdot 10^{-19}\,\mathrm{J\,m^5\,mol^{-2}}$ for the *monotonous component* and $c_2 \doteq 2.398 \cdot 10^{-20}\,\mathrm{J\,m^5\,mol^{-2}}$ for the second component.

The eq. (5.12) is valid for the arbitrary geometry, i.e. shape of the phase interface for the solved problem see the examples illustrated in fig. 2.6. Hence a generalized solver can be used for the both geometry-types. When numerical efficiency and speed are of concern a specialized solvers can be developed for each geometry allowing for further optimization. For example, when the factor dS/ds equals 1, the problem can be numerically integrated, which corresponds to the the planar phase interface case.

The algebraic part given by eq. (5.9) is also treated using notations eq. (5.10) and eq. (5.11). Consequently, the equation for the modified density $\tilde{\rho}$ has to be added to the algebraic system, which subsequently composes of n_{comp} nonlinear equations

$$\begin{aligned}\frac{\Delta\mu_2}{\sqrt{c_{2,2}}} &= -X\\ &\vdots\\ \frac{\Delta\mu_{n_{\mathrm{comp}}}}{\sqrt{c_{n_{\mathrm{comp}},n_{\mathrm{comp}}}}} &= -X\\ \sum_{i=1}^{n_{\mathrm{comp}}}\sqrt{c_{i,i}}\rho_i &= \tilde{\rho}\sum_{i=1}^{n_{\mathrm{comp}}}\sqrt{c_{i,i}}.\end{aligned} \tag{5.13}$$

It is easily observed that the algebraic system of equations in eq. (5.13) does not depend on the type of the interface geometry as no direct reference to parameter s or S is present. This feature permits the independent solution of problem regardless of the geometry type which can be further used for optimization in an simultaneous solution of multiple algebraic parts. The algebraic system is computed at first as variable X is the link between the algebraic part and the subsequently solved in the differential part of the problem.

5.3.1 Algebraic system solution

Given the nonlinear character of the algebraic part eq. (5.13) the solution is based on the Newton-Raphson. System properties can be further improved by rearranging the algebraic equations into form that yields a symmetric Jacobian Matrix. This is a direct consequence of the partial derivatives interchangeability shown previously by Liang *et. al.* [146]. The form used for solution is then

$$\begin{aligned}
\Delta\mu_2 + X\sqrt{c_{2,2}} &= 0 \\
&\vdots \\
\Delta\mu_{n_{\mathrm{comp}}} + X\sqrt{c_{n_{\mathrm{comp}},n_{\mathrm{comp}}}} &= 0 \\
\sum_{i=1}^{n_{\mathrm{comp}}} \sqrt{c_{i,i}}\rho_i - \tilde{\rho}\sum_{i=1}^{n_{\mathrm{comp}}} \sqrt{c_{i,i}} &= 0
\end{aligned} \tag{5.14}$$

The computational procedure is developed around the Newton-Raphson iterator utilizing the Jacobian inversion method.

The whole procedure operates across the *modified density* discretization. The schema is stepping over the individual modified densities for which the solutions are found. The discretization also helps for the initial estimate choice required by the solver. While the first density relies on equilibrium estimate each consequent sampled density uses the previous solution as an estimate. In this way a number of iterations are significantly decreased further enhancing the speed of solution.

The solution values are coupled into the data structure with modified density sampling $\tilde{\rho}^{(1)}, \tilde{\rho}^{(2)}, \ldots \tilde{\rho}^{(\mathrm{disc})}$ that can be understood as a table of properties sampled at the discretization points.

$$\begin{array}{cccc}
\tilde{\rho}^{(1)} & \tilde{\rho}^{(2)} & \ldots & \tilde{\rho}^{(\mathrm{disc})} \\
\hline
\rho_1^{(1)}\left(\tilde{\rho}^{(1)}\right) & \rho_1^{(2)}\left(\tilde{\rho}^{(2)}\right) & \ldots & \rho_1^{(\mathrm{disc})}\left(\tilde{\rho}^{(\mathrm{disc})}\right) \\
\rho_2^{(1)}\left(\tilde{\rho}^{(1)}\right) & \rho_2^{(2)}\left(\tilde{\rho}^{(2)}\right) & \ldots & \rho_2^{(\mathrm{disc})}\left(\tilde{\rho}^{(\mathrm{disc})}\right) \\
\vdots & \vdots & \ldots & \vdots \\
\rho_{n_{\mathrm{comp}}}^{(1)}\left(\tilde{\rho}^{(1)}\right) & \rho_{n_{\mathrm{comp}}}^{(2)}\left(\tilde{\rho}^{(2)}\right) & \ldots & \rho_{n_{\mathrm{comp}}}^{(\mathrm{disc})}\left(\tilde{\rho}^{(\mathrm{disc})}\right) \\
X^{(1)}\left(\tilde{\rho}^{(1)}\right) & X^{(2)}\left(\tilde{\rho}^{(2)}\right) & \ldots & X^{(\mathrm{disc})}\left(\tilde{\rho}^{(\mathrm{disc})}\right)
\end{array} \tag{5.15}$$

For the data structure shown here the superscripts are used to denote the discretization index of the points instead of the component numbers used in the rest of the chapter.

This data structure in eq. (5.15) represents the basis for a piecewise cubic interpolation used in the subsequent step. The piecewise cubic interpolation brings two important benefits: (a) fewer discretization points are required to hold comparable level of accuracy, (b) brevity and overall simplification of the differential solver. It often happens that the differential solver requests a value not present in the discretized sampling in eq. (5.15). With the interpolation method, such value can easily be computed from the data structure saving more expensive iterations of the algebraic solver.

The intermediate interpolation step can be therefore understood as a transformation of the sampled data structure in eq. (5.15) into set of functions $\rho_1\left(\tilde{\rho}\right), \rho_2\left(\tilde{\rho}\right), \ldots, \rho_{n_{\mathrm{comp}}}\left(\tilde{\rho}\right), X\left(\tilde{\rho}\right)$. The function dependencies are shown in fig. 5.2 for one example of such algebraic solution procedure. The last benefit of interpolation is that simpler density dependencies can be modeled sufficiently with lesser amount of points while retain the same numerical accuracy of conseqent differencial step.

The main reason behind the Newton-Raphson iterator in the algebraic solver is the overall simplicity and solution speed of the solver. One can argue that with the use of interpolation method other more sophisticated solvers can also be employed. However, this could result in performance constrain directed on the discretization, which is better to be avoided.

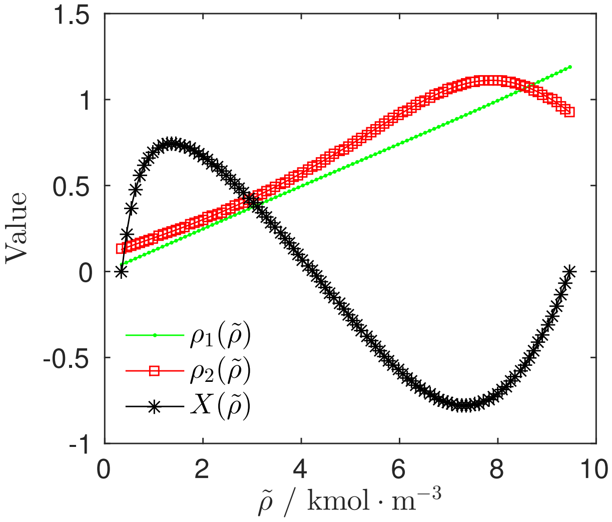


Figure 5.2: Example of the interpolated results from the algebraic solver, i.e. eq. (5.14). For comparison purposes, the quantities X, ρ_1, and ρ_2 are shown dimensionless (see appendix A.2). The *monotonous* component 1 of the binary mixture is represented by green pointed line −●−, *non-monotonous* component 2 is depicted by red squared line −□−, and the binary variable X is represented as black stared line −∗−.

One could also object against the initial estimate quality used in the iterator. The worst case can lead to the over-iteration, which would reduce the speed benefits but due to the employed stepping approach this is no longer the issue. This approach ensures quality of the estimate for sufficiently fine discretization which is not so hard requirement as the usage of more sophisticated solvers would be.

Our analysis of solver behavior has revealed that refinement of discretization brings notable benefits for first few refinements. It was empirically found that about 100 discretization points are sufficient in most cases. It is reasonable to expect that finer discretizations do not extend the knowledge of eq. (5.15) more than the interpolation.

In this way the data structure provides a description of densities and X dependencies required for the differential equation solver.

5.3.2 Differential equation solution

The algebraic solution is followed by the differential solver, in which an interpolation of variable $X(\tilde{\rho})$ is primarily used as it forms RHS of the differential equation eq. (5.12). Utilising the previous knowledge of the selected interface geometry, a specialized differential is developed. This feature is especially useful for planar geometry case, where the solution can be found even analytically allowing for higher precision to be achieved. The analytical solution is presented in section 5.3.3. The more general solution method, primarily used for the spherical phase geometry, is described here. The differential equation in eq. (5.12) with the general factor dS/ds can be modified for the spherical phase interface as follows

$$\frac{2}{r}\tilde{\rho}' + \tilde{\rho}'' = \frac{X(\tilde{\rho})}{\sum_{i=1}^{n_{\mathrm{comp}}} \sqrt{c_{i,i}}}, \tag{5.16}$$

where the general coordinate s was replaced for the radius r.

A numerical solution method is required for eq. (5.16). The general analysis determined that the mathematical problem exhibits three attractors to which the solution tends to evolve. From the thermodynamic perspective there should be only two attractors corresponding to systems of bulk liquid phase, bulk vapor phase. The third atractor is consequence of asssumption of static system without any disturbances where a hypothetical third attractor corresponds to critical density in between of the bulk phases. This is better shown in fig. 5.3, where red profile aligns to bulk liquid attractor, black (desired) profile to bulk vapor attractor and teal profile is skewed by the third attractor corresponding to the unstable system.

Considering this behavior of three attractors, the solution method has to be robust enough to overcome this obstacle. The testing of the solution method, often via the trial and error, pointed out that the shooting method coupled with the predictor–corrector type of solver provides a convenient and robust solution. Wide range of methods were tested and deemed not to be applicable due to the widespread convergence issues caused mainly by the physical nature of the problem. Unfortunately, in some cases the selected method even failed to converge. The possible reasons are discussed below in more detail.

The chosen shooting method translates the original boundary value problem into the initial value problem which is easier to solve in this setting. Main aim of the shooting method is to find the initial density in the centre of a new phase, i.e. a droplet, that yields the density profile aligning with apriori known mother phase density, i.e. vapor. An important condition for the shooting method is that the density derivative in the center of a droplet equals zero, which is a required by assumed spherical symmetry of the droplet.

Proposed scheme utilizes the shooting method coupled with the MATLAB® implementation of the predictor–corrector differential solver called "ODE45". This solver was developed from the method proposed by Dormand and Prince [63] for Runge-Kutta type of iterative solvers.

The shooting method is further supplied with a decision criterion responsible for the selection of the next value of the shooting parameter α_{next}. It was found that a bisection method constructs a reliable criterion and is capable to cope with the steep nature of the investigated searching task for the optimal shooting parameter α_{optimal} and other attractors of the problem. In theory this criterion is guaranteed to converge to the optimum when the optimum is located within the starting interval used for bisection.

The shooting method criterion is based on a minimum value search for a difference between the vapour density at the end of the droplet density profile and the bulk density of the surrounding mother phase. In this way we specifically target at the correct attractor in the density of the mother phase. If an incorrect attractor is reached, the shooting parameter of the following step is chosen accordingly. The illustration of the density profiles for various values of the shooting parameter α leading to different attractors of the problem are shown in fig. 5.3. In this figure we can notice that the attractors at the vapour and liquid density exhibit only one sided attraction (within the vapour-liquid density band). Outside this band the profiles are unphysical and strong repulsion is observed instead.

A droplet in a supersaturated vapour, forming from the mother phase, is handled similarly. The initial shooting parameter α corresponds to the liquid density within the droplet centre in this case. For complete information the method also work from the opposite transition where bubbles form within liquid. In such situation the criterion remains same while supplying the new meaning for the mother phase now liquid and initial density estimate now bulk vapor.

5.3.3 Calculation of density profile and related properties

The developed solving procedure allows the calculation of the density profile for the targeted interface geometry. The profile gives a information about the properties of the interface itself like surface tension, but also indicate whether adsorption takes place in case of multiple component mixture.

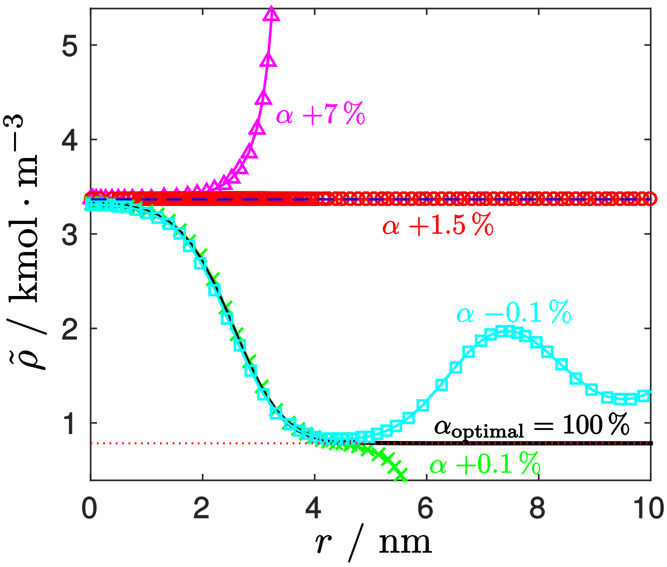


Figure 5.3: Density profiles for binary mixture of CO_2 + n-nonane for shooting parameters $\alpha = \tilde{\rho}_0$. Selected case provides discernible differences of profiles for large variations of shooting parameter.

As already mentioned for the **planar phase interface**, the core problem can be also solved analytically [112,147,167]. The planar case results in the form of following integral for the density profile expressed with the axial coordinate z as

$$z(\rho) = z_0 + \int_{\rho_0}^{\rho} \sqrt{\frac{\sum_{i,j=1}^{n_{\text{comp}}} c_{i,j} \left(\frac{\partial \rho_i}{\partial z}\right) \left(\frac{\partial \rho_j}{\partial z}\right)}{2\Delta\omega}}\, d\rho, \tag{5.17}$$

where ρ_0 stands for the initial density of integration at the spatial coordinate z_0. The initial values are chosen based on the equilibrium calculation setting the ρ_0 to bulk density and consequently z_0 to location in the interfacial region, where the bulk density is reached. It is important to note that the choice of ρ_0 and z_0 determines the profile orientation based the location of the starting point, e.g. starting from bulk density of mother phase.

In eq. (5.17), only the partial densities are left to be determined. These values are readily available from the inverted data structure in eq. (5.15). Performing a backward conversion from the modified density $\tilde{\rho}$ in terms of eq. (5.11) the partial densities are obtained. This method is reliable but the condition of injectivity of transformation has to be fulfilled, e.g. the modified density is monotonous. In cases where the modified density would not be monotonous the backward transformation would then becomes ambiguous.

For the **spherical phase interface**, the differential solver produces results in a form of the modified density as a function of radius $\tilde{\rho}(r)$, which is further transformed to the partial densities $\rho_i(r)$. The transformation is performed with the piecewise cubic interpolation functions $\rho_i(\tilde{\rho})$ applied on the discretized points along the phase interface data in form of eq. (5.15).

The **surface tension** in case of the vapour→liquid phase interface, represents another important property describing the forces exerted onto the dividing interface that hold the phases separate from each other. For the systems with the planar interface geometry, the following expression for the surface tension σ can be used [166, 168, 271]

$$\sigma_{\text{planar}} = \int_{\rho^{\text{V}}}^{\rho^{\text{L}}} \sqrt{2\,\Delta\omega \sum_{i,j=1}^{n_{\text{comp}}} c_{i,j} \left(\frac{\partial \rho_i}{\partial z}\right)\left(\frac{\partial \rho_j}{\partial z}\right)}\, d\rho. \tag{5.18}$$

According to Liang *et. al.* [146], the integration of eq. (5.18) can be performed using the modified density $\tilde{\rho}(r)$, which slightly improves the accuracy, as no additional backward transformation is required.

When considering the spherical case, the direct integration is not possible for the evaluation of the surface tension. But the curved surface can be described by the Young-Laplace equation in eq. (2.44) relating the surface tension to the curvature and the Laplace pressure Δp. After a simple treatment [246], an equation for the surface tension on a spherical phase interface can be obtained

$$\sigma_{\text{spherical}} = \sqrt[3]{\frac{3\Delta\Omega\Delta p^2}{16\pi}}, \tag{5.19}$$

where $\Delta\Omega$ denotes the work of formation determined from eq. (2.89).

Another important property related to the phase interface research is the **surface of tension**[2]. The surface of tension interprets the DGT results in terms of the Gibbsian surface thermodynamics, in which the relevant quantities are defined with a reference model of a simple sphere with radius r_{sot}. For a curved interface, the surface of tension is a surface at which the surface tension acts according to the Laplace pressure Δp given by eq. (2.44). For the planar geometry, the surface of tension can be understood as the centre of force for the surface tension similarly as the centre of mass is for the gravity. The governing equation for the surface of tension over a planar phase interface is given as follows.

$$z_{\text{sot}} = \int_{-\infty}^{\infty} \left[\sum_{i,j=1}^{n_{\text{comp}}} c_{i,j} \left(\frac{\partial \rho_i}{\partial z}\right)\left(\frac{\partial \rho_j}{\partial z}\right)\right] z\, dz. \tag{5.20}$$

Here the integration bounds are expressed as infinity, however the real computation is performed only for sufficiently large interval to envelop the region of interest of the computed density profile.

Similarly to the surface tension σ, the surface of tension for the spherical phase interface can also be expressed from the Young-Laplace equation and the work of formation $\Delta\Omega$, i.e. from eq. (2.44) and eq. (2.82).

$$r_{\text{sot}} = \sqrt[3]{\frac{3\Delta\Omega}{2\pi\Delta p}} \tag{5.21}$$

where $\Delta\Omega$ can be obtained from eq. (2.89).

While the surface of tension is thought concept it gives the option of computing the properties originally introduced by the CNT such as the excess number of molecules ΔN and the adsorption Γ. The **excess number of molecules** gives the information about the excess amount of molecules compared to the homogeneous gaseous system. For illustration see fig. 2.4 where the excess number of molecules is used for choice of equimolar dividing surface. The **adsorption** or also the surface excess number of molecules connects the excess number of molecules with a surface area to provide information about the amount of molecules localized at the surface. The amount of adsorbed molecules is usually computed in moles and thus the results of following

[2] Not to confuse with surface tension

formulas are scaled with Avogadro constant (not expressed in defining formulas). The following integration is performed for the general spatial coordinate s.

$$\Delta N_i = \int_{s_{\mathrm{min}}}^{s_{\mathrm{max}}} \left[\rho_i(s) - \rho_i^{\mathrm{V}}\right] ds \tag{5.22}$$

Similarly to the surface of tension, the coordinate limits should be set to envelop the whole computed density profile in a way that omitted area has negligible influence on the final result accuracy.

The adsorption can be then again expressed in an unified manner for the planar and spherical phase interfaces using a general coordinate s as follows.

$$\Gamma_i = \int_{s_{\mathrm{min}}}^{s_{\mathrm{sot}}} \left[\rho_i(s) - \rho_i^{\mathrm{L}}\right] S ds + \int_{s_{\mathrm{sot}}}^{s_{\mathrm{max}}} \left[\rho_i(s) - \rho_i^{\mathrm{V}}\right] S ds \tag{5.23}$$

The formulation eq. (5.23) shows the integration is divided into two regions: one preceding and other following the location of the surface of tension s_{sot}. Depending on the orientation of the density profile, the bounds of integration are flipped to correspond with the profile calculated direction.

For the spherical geometry case $s_{\mathrm{min}} = 0$, $s_{\mathrm{max}} \gg r_{\mathrm{sot}}$, the formula reflects the choice of a spherical dividing surface (r_{sot}) and the adsorption transforms into following formula

$$\Gamma_i = \frac{\Delta N_i}{4\pi r_{\mathrm{sot}}^2} - \frac{r_{\mathrm{sot}}(\rho^{\mathrm{L}} - \rho^{\mathrm{V}})}{3}. \tag{5.24}$$

5.4 Results

In the previous work of the author in [40] focus was given spherical interface case applied illustrated with mixture of nitrogen and carbon dioxide. The primary goal of the study was calculation of the density profiles. In this extension a broader view is employed with the generalized interfacial geometry approach applied on two mixtures relevant for CCS: n-butane + CO_2 and SF_6 + CO_2. The topic is extended into other properties including surface tension, surface of tension and adsorption. We also provide comparison with more general theory and experimental data to verify the correct operation of the proposed method.

5.4.1 Artificial variable X

An overview of the computed results obtained with the described model for both phase interface geometries is provided here. The variable X , playing an important role in the developed solution, is discussed at first. In figs. 5.4 and 5.5, the dependency of X on the modified density $\tilde{\rho}$ is shown for various values of the Laplace pressure Δp for mixtures SF_6 + CO_2 and n-butane + CO_2, respectively. The base shape of variable X corresponding to the planar geometry with $\Delta p = 0\,\mathrm{Pa}$ and further evolves with the increasing Laplace pressure preserving the boundary values. These end values of $X(\tilde{\rho})$ are equal to zero, which is a direct consequence of the equilibrium condition and the definition of X according to eq. (5.10).

It should be noted that even though the variable X changes for different systems, its shape remains wave-like function, however with different amplitudes. In figs. 5.4 and 5.5, the greyscale lines decreasing hue corresponding to increasing values of the Laplace pressure. The black dashed line, representing the planar geometry (i.e. the non-supersaturated system), is added for comparison.

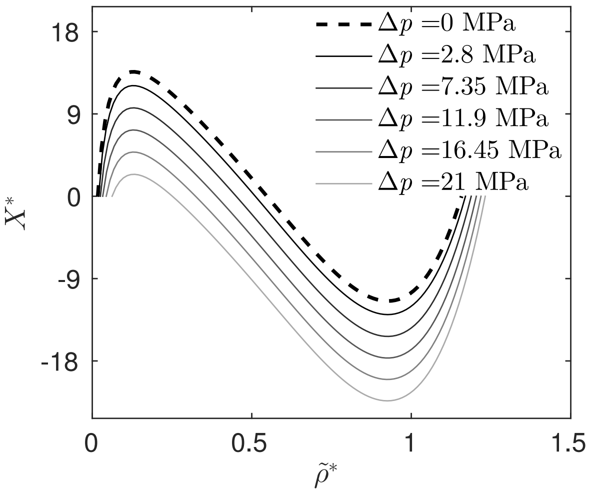


Figure 5.4: Variable X scaled to dimensionless shape by eq. (A.11) for SF_6 + CO_2 mixture at $T = 230\,\mathrm{K}$

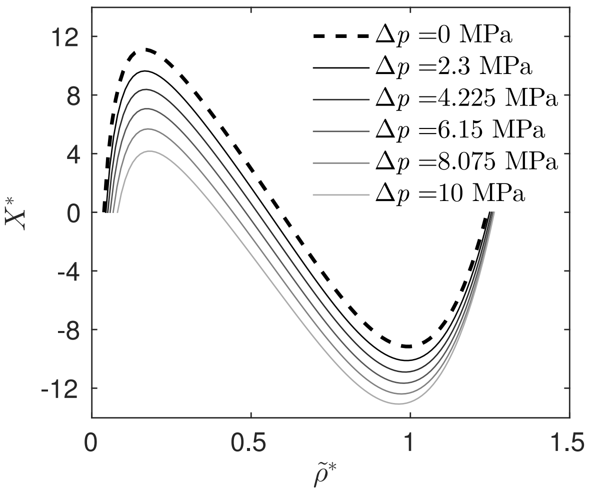


Figure 5.5: Variable X scaled to dimensionless shape by eq. (A.11) for n-butane + CO_2 mixture at $T = 300\,\mathrm{K}$

5.4.2 Density profile

Density profiles describing the phase interfaces are computed either directly from eq. (5.17) for the planar geometry or according to the general method described in section 5.3.1 and section 5.3.2. Examples of the density profiles for n-butane + CO_2 mixture are shown in figs. 5.6 and 5.7 for the planar and the spherical type of geometry, respectively. The total density ρ, the modified density $\tilde{\rho}$, the partial densities of individual components ρ_i, and the surface of tension s_{sot} , that would be employed in CNT, are depicted in the figures. Distinct feature of both profiles is the substantial adsorption of CO_2 within the interface predicted by the PCP-SAFT + DGT model. The adsorption becomes more pronounced with increasing Δp as can be seen from the comparison of selected cases with $\Delta p = 0$ and $\Delta p = 5.53\,\mathrm{MPa}$. We note that the profiles are computed until the stop criteria of the solution method is reached, which in the spherical case results in rather long vapour part of the profile. Due to a direct computation of the planar case, the vapour part is considerably shorter in fig. 5.6. The planar geometry has an arbitrary selected onset of computation at $z = 0$. This point is used only as a reference for the interface thickness, while in case of the spherical geometry the radial distance is directly related to the droplet size.

Graphs depicted in figs. 5.8 and 5.9 show the density profiles in terms of the the modified density $\tilde{\rho}$ for two binary mixtures n-butane + CO_2 and SF_6 + CO_2 at various values of the Laplace pressure Δp. The profiles are aligned according to the surface of tension set to $s = 0\,\mathrm{nm}$.

It can be seen in fig. 5.8 that for n-butane + CO_2 mixture the density profiles are closely attached to the planar profile corresponding to $\Delta p = 0$. The only considerable difference between the profiles can be seen on the vapour side where the bulk density increases with higher supersaturation, i.e. with higher Δp. Results in fig. 5.9 computed for SF_6 + CO_2 mixture show similar behaviour except for one misaligned profile for the highest Laplace pressure of 21 MPa. This is an illustration of a small size droplet of a limited dimension. At high supersaturations, the droplet shrinks in diameter and the density in its centre becomes significantly smaller than the bulk liquid density. Consequently, the small size droplet is solely composed of the phase interface and the associated surface of tension cannot be easily compared to larger droplets. The research of behavior of small droplets is a challanging topic that is not well aligned with assumptions made by the here develop solution method.

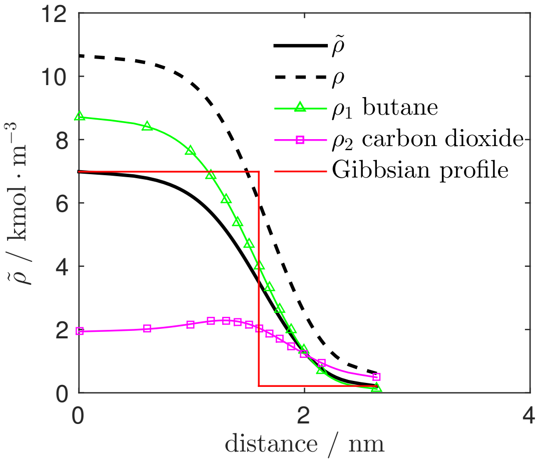


Figure 5.6: Planar density profiles of n-butane + CO_2 mixture at $T = 300\,\mathrm{K}$, $p = 1.36\,\mathrm{MPa}$, and $\Delta p = 0\,\mathrm{MPa}$ predicted with PCP-SAFT + DGT. Gibbsian comparison corresponds to location of the surface of tension z_{sot}.

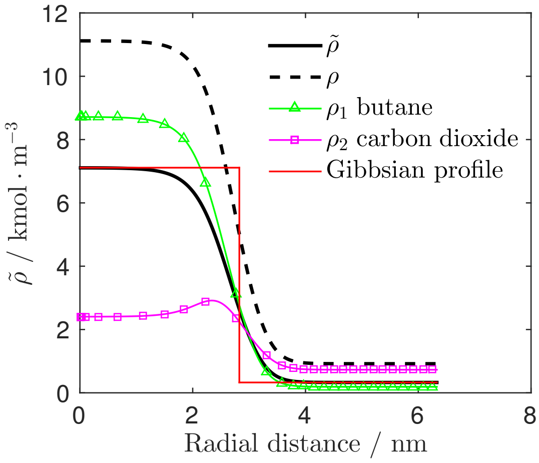


Figure 5.7: Spherical density profiles of n-butane + CO_2 mixture at $T = 300\,\mathrm{K}$, $p = 1.85\,\mathrm{MPa}$, and $\Delta p = 5.53\,\mathrm{MPa}$ predicted with PCP-SAFT + DGT. Gibbsian comparison corresponds to location of the surface of tension r_{sot}.

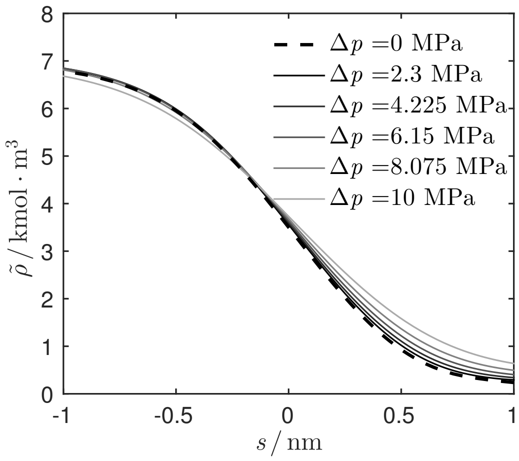


Figure 5.8: Comparison of the density profiles for n-butane + CO_2 mixture at $T = 300\,\mathrm{K}$ and various values of the Laplace pressure Δp calculated with the PCP-SAFT + DGT model. Surface of tension is placed at $s = 0$ for each profile.

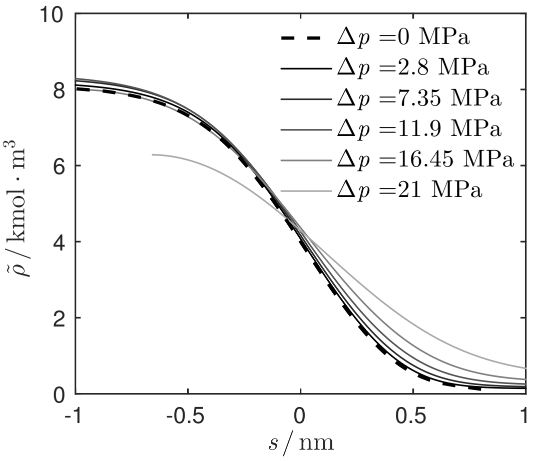


Figure 5.9: Comparison of the density profiles for SF_6 + CO_2 mixture at $T = 230\,\mathrm{K}$ and various values of the Laplace pressure Δp calculated with the PCP-SAFT + DGT model. Surface of tension is placed at $s = 0$ for each profile.

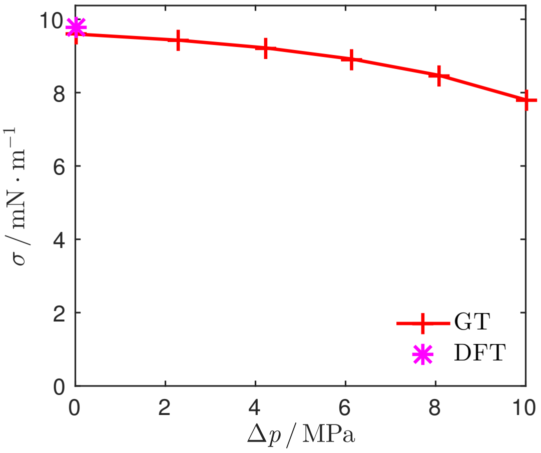


Figure 5.10: Surface tension of n-butane + CO_2 mixture depending on Laplace pressure Δp obtained with the PCP-SAFT + DGT model compared with the DFT prediction for the planar geometry at $T = 300\,\mathrm{K}$.

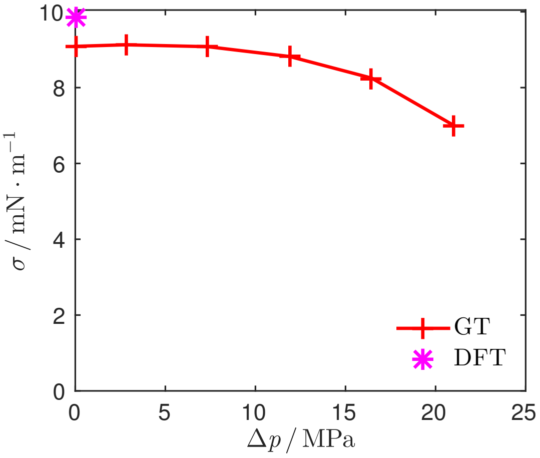


Figure 5.11: Surface tension of SF_6 + CO_2 mixture depending on Laplace pressure Δp obtained with the PCP-SAFT + DGT model compared with the DFT prediction for the planar geometry at $T = 230\,\mathrm{K}$.

5.4.3 Surface tension

Graphics in figs. 5.10 and 5.11 show the surface tension depending on the Laplace pressure corresponding to the density profiles provided in figs. 5.8 and 5.9, respectively. The computations abide the approach given by eqs. (5.18) and (5.19). The PCP-SAFT + DGT results are compared with a more general Density Functional Theory (DFT) proposed by Ebner [64] and further developed by Evans [68]. DFT does not fundamentally require the gradient assumption and provides more general results at a cost of substantially slower computation. From the comparisons in figs. 5.10 and 5.11, it can be seen that the presented DGT-model slightly under-predicts the surface tension compared to DFT. However, the difference lies within a reasonably small range. In fig. 5.11, the difference between DGT and DFT becomes more pronounced due to a rather problematic nature of the SF_6 + CO_2 mixture as the two components have quite close critical temperatures of $T_c(SF_6) = 318.7\,\mathrm{K}$ and $T_c(CO_2) = 304.1\,\mathrm{K}$. The SF_6 + CO_2 system was selected in order to test the robustness of the developed model. The solution algorithm can treat this mixture in the same manner as the system with a clear *monotonous* component such as the n-butane + CO_2 mixture.

As can be seen in figs. 5.10 and 5.11, the model behaves according to the theory with a characteristic arc-shaped dependency of the surface tension on the increasing Δp. This trend can be well observed in one-component systems [106, 193], which can be solved even for the miniscule droplets typical of a rapid drop in the surface tension; see , e.g., figure 15 in ref [193]. The results provided in figs. 5.10 and 5.11 could also be calculated at higher values of Δp however with rather significant error caused by the miniscule concentrations of both phases, which would lead to a badly conditioned problem both for the equilibrium and the algebraic solvers.

5.4.4 Adsorption

In fig. 5.12 adsorption Γ is shown representing another important property of the phase interface, depending on Laplace pressure Δp . The example is provided for n-butane + CO_2 mixture at a constant temperature of 300 K. As can be seen, CO_2 shows relatively high adsorption within the

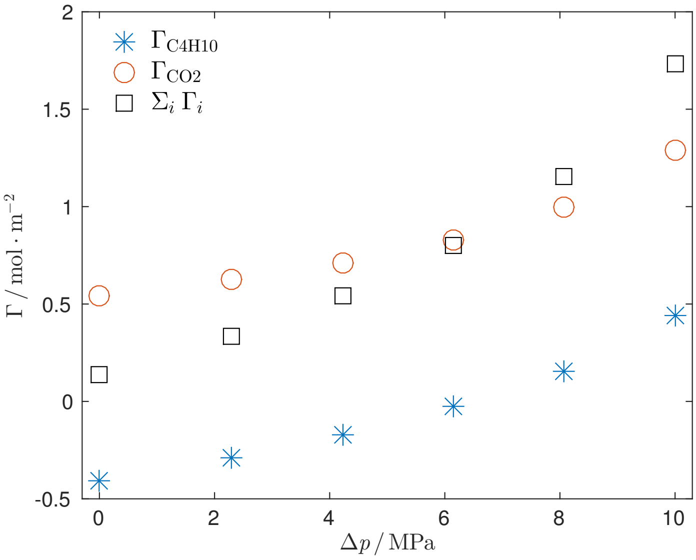


Figure 5.12: Adsorption of individual components of n-butane + CO_2 mixture depending on Laplace pressure Δp at a constant temperature $T = 300\,\mathrm{K}$.

interface already at $\Delta p = 0$, i.e. for the planar phase interface, which is further continuously increasing with increasing Δp. The increase of $\Gamma(CO_2)$ can also be seen from the comparison of figs. 5.6 and 5.7 showing the density profiles at two different values of Δp at the same temperature of 300 K. It is notable that with increasing Δp, the adsorption of butane is also increasing from the negative values for the planar and large spherical interfaces to the positive values for smaller droplets. An increase of Γ(butane) is a direct consequence of the higher supersaturation of the system.

5.4.5 Comparison with experiments

Theoretical validation of the presented model and comparison with the DFT theory was showed in figs. 5.10 to 5.12. A brief comparison with the experimental data for the surface tension over the planar phase interfaces of CO_2 mixtures is provided further. The depicted results were obtained by DGT combined with the PCP-SAFT equation of state accounting for quadrupolar CO_2. All mixtures were modelled with a zero binary interaction parameter k_{ij} in the Lorentz-Berthelot combining rule for the PC-SAFT energy parameter [88].

Data for the surface tension of n-butane + CO_2 mixture by Brauer and Haugh [29] and Hsu *et. al.* [107], are shown in figs. 5.13 and 5.14. A relatively good agreement between the experimental data by Brauer and Haugh [29] and the model can be seen in fig. 5.13. The DGT + PCP-SAFT model is qualitatively well aligned to the data with a constant over-prediction under 10 %. At temperatures 310.93 and 319.26 K, the calculation was aborted at pressures above approximately 0.63 and 0.73 MPa, respectively. The computation was terminated prematurely at these points due to an inaccurate prediction of the phase equilibrium. The predicted phase equilibrium for n-butane + CO_2 mixture does not correspond to the experimental data at the given temperatures and pressures. This discrepancy can be partially improved by adjusting the binary interaction parameter k_{ij} that shall be set to 0.125 for PC-SAFT and to 0.036 for PCP-SAFT in this case [86]. However as the main purpose of this work is to introduce a new DGT-based model, a constant $k_{ij} = 0$ is considered out of demonstrative reasons.

In the second comparison for n-butane + CO_2 mixture shown in fig. 5.14, three experimental

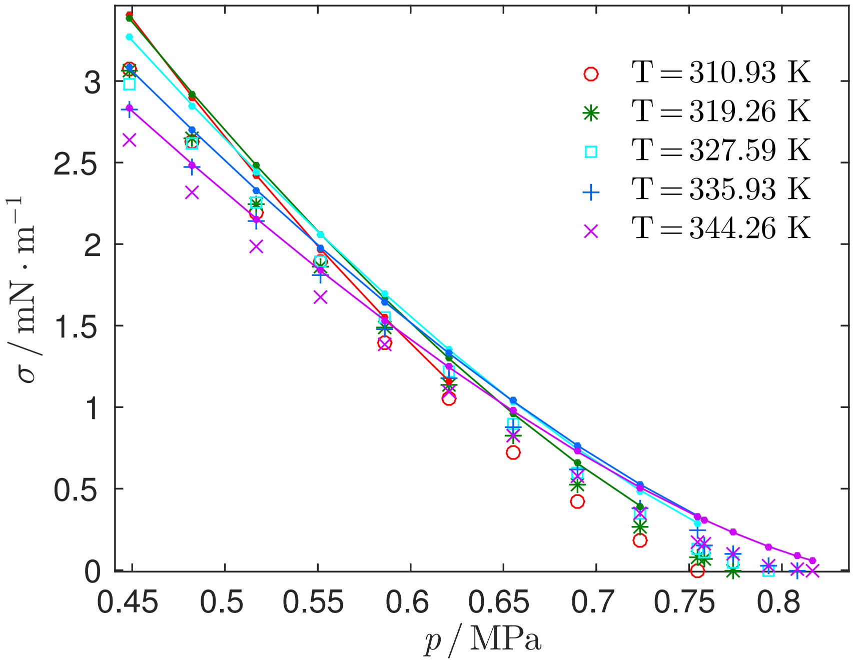


Figure 5.13: Comparison of the PCP-SAFT + GT model with the experimental data [29] for the surface tension of n-butane + CO_2 mixture.

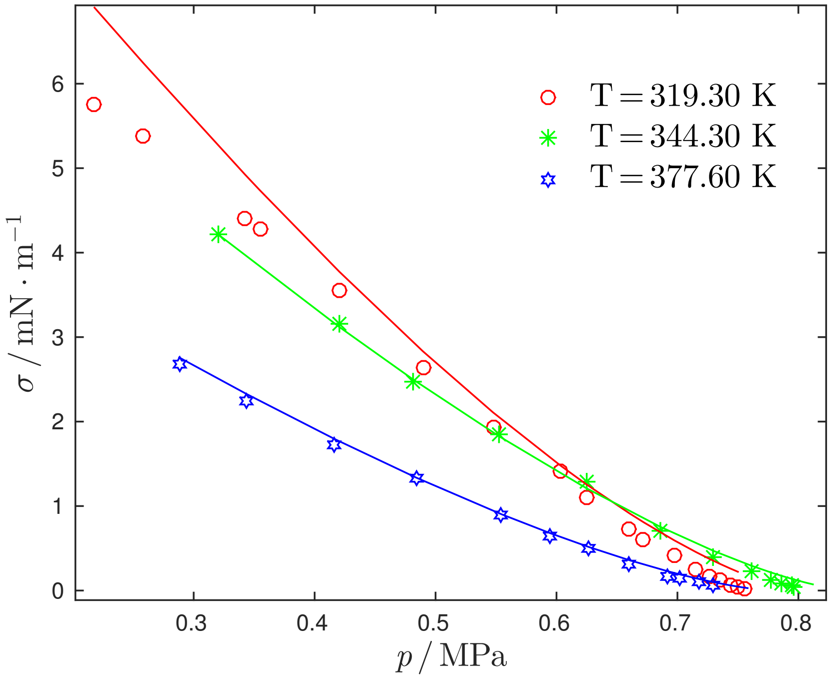


Figure 5.14: Comparison of the PCP-SAFT + GT model and the experimental data [107] for the surface tension of n-butane + CO_2 mixture.

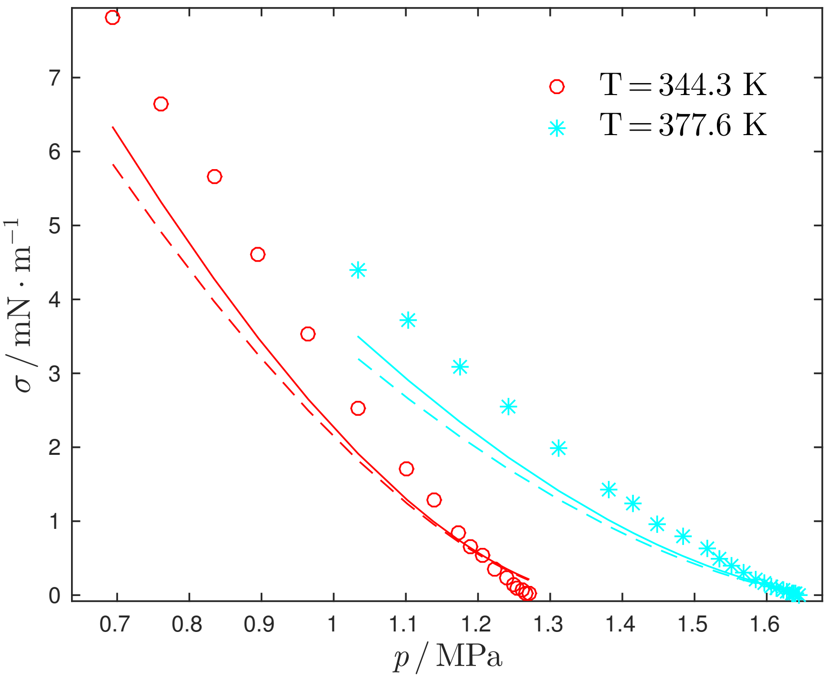


Figure 5.15: Surface tension of n–decane + CO$_2$ by Nagarajan and Robinson [172] compared to the PCP-SAFT + DGT model (solid lines) and the PC-SAFT + DGT model (dashed lines) both with $k_{ij} = 0$.

datasets by Hsu et al. [107] are presented. System conditions are well reproduced by the PCP-SAFT + DGT model with a very good precision over wide pressure range even with the binary interaction parameter k_{ij} equal to zero. However at the lowest temperature of 319.3 K, the model slightly deviates from the experimental data in the low pressure region below approximately 0.35 MPa. Slight discrepancies between the model and the experimental data for n-butane + CO$_2$ can be attributed to the quadrupolarity of CO$_2$ and to a significant adsorption of CO$_2$ within the phase interface.

Comparison of the experimental data by Nagarajan and Robinson [172] for n–decane + CO$_2$ with the PCP-SAFT + DGT model and the PC-SAFT + DGT model is shown in fig. 5.15. In both cases a zero binary interaction parameter k_{ij} is considered. The model with PCP-SAFT EoS provides slightly better prediction. The average deviation (AD) and the average absolute deviation (AAD) from the experimental data, given as

$$\mathrm{AD} = \frac{1}{N} \sum_{i=1}^{N} \left(\sigma_{i,\mathrm{exp}} - \sigma_{i,\mathrm{mod}} \right) , \tag{5.25}$$

$$\mathrm{AAD} = \frac{1}{N} \sum_{i=1}^{N} \left| \sigma_{i,\mathrm{exp}} - \sigma_{i,\mathrm{mod}} \right| , \tag{5.26}$$

have values of 0.29 mN m^{-1} and 0.36,mN m^{-1} for PCP-SAFT + DGT and 0.38 mN m^{-1} and 0.44,mN m^{-1} for PC-SAFT + DGT, respectively. However, as can be seen both models underpredict the experimental surface tension especially at lower pressures. The model predictions can be improved by considering a non-zero binary interaction leading to a better prediction of the phase equilibrium, i.e. more accurate phase compositions at given T and p; see, e.g., ref [246], where the non-zero values for k_{ij} were employed.

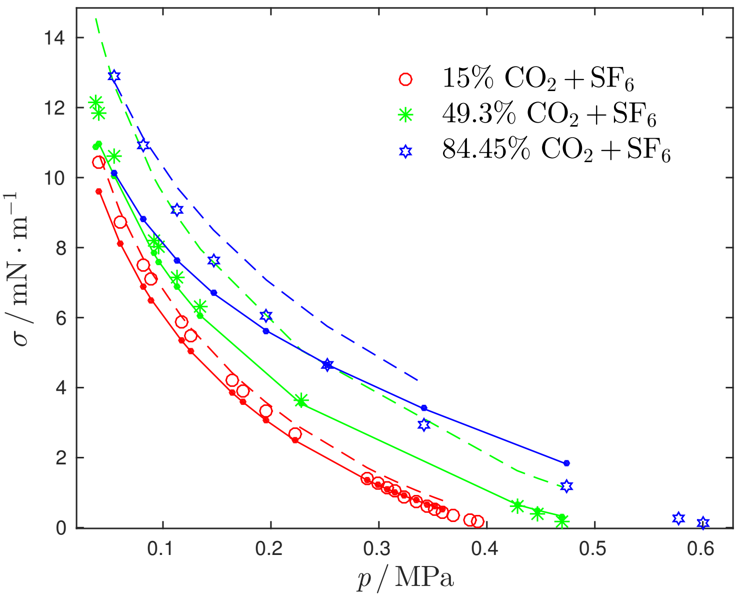


Figure 5.16: Surface tension for $SF_6 + CO_2$ mixture predicted by the PCP-SAFT + DGT model (solid lines) and the PC-SAFT + DGT model (dashed lines) compared with the experimental data by Do and Straub [62] at three constant mixture compositions. Temperature of the data lies between 210 and 310 K.

Data for the surface tension of $SF_6 + CO_2$ system measured by Do and Straub [62] for three different mixtures of a given overall composition are shown in fig. 5.16. We note that the data were taken from diagrams in the article and can therefore suffer of slightly higher uncertainty. The PCP-SAFT + DGT model shows quite good agreement also with this system, which is, unlike n-butane + CO_2, typical of rather close critical temperatures of both components. The only discrepancy between the data and the model can be seen for the mixture with the highest content of CO_2 of around 85 % at lower pressures. Results of the introduced PCP-SAFT + DGT model can in general be treated as quite reliable also for $SF_6 + CO_2$ system as no additional parameter, i.e. the binary interaction parameter k_{ij} or binary parameter γ_{ij} for the influence parameter $c_{i,j}$, was adjusted to the mixture data. In fig. 5.16 results are shown for the PC-SAFT + DGT, which does not consider quadrupolar CO_2. The PC-SAFT + DGT model overpredicts the surface tension compared both to the experimental data and the PCP-SAFT + DGT model. The largest difference can be seen for the middle composition of 49.3 % CO_2. The average deviation and the absolute average deviation are 0.40 mN m^{-1} and 0.48 mN m^{-1} for the PCP-SAFT + DGT model and −0.77 mN m^{-1} and 0.78 mN m^{-1} for the PC-SAFT + DGT model, respectively.

5.5 Conclusions

A new model for multicomponent mixtures with various phase interface geometries was developed in this work. The model is based on the density gradient theory (DGT) describing the phase interface and the PC-SAFT equation of state employed on the thermophysical properties computations. The DGT formulation was modified to allow for solution of both planar and spherical phase interface geometries. This approach enabled a unified derivation of the core problem described in section 5.2. The core problem of the density profile solution in its plain form would require substantial computational resources therefore an innovative solution approach was proposed in section 5.3. The new solution approach utilizes the shape of the core problem and divides it into an algebraic part and a differential part. Both parts are solved with a help of a transition variable according to the shape of the prescribed interface geometry. Unlike for the spherical interface, a direct solution can be used in case of the planar interface geometry.

The formulas for properties describing the phase interface , i.e. the density profile, the adsorption and the surface tension, were also derived. All the formulas were analyzed from the theoretical background and compared with other theories, i.e. Gibbsian thermodynamics for the interface profiles and DFT for the surface tension prediction. The DGT-based model was found to provide comparable results to DFT for the surface tension at the planar phase interface. The predictive ability of the introduced model was investigated on selected mixtures with CO_2. The model predictions are in good agreement with the experimental data for the surface tension over relatively wide ranges, even though no additional parameter was correlated to the mixture data. A considerable influence of the improved PCP-SAFT EoS accounting for quadrupolarity compared to original PC-SAFT EoS was detected when combined with the DGT-based model. The PCP-SAFT + DGT model provides significantly better predictions especially for the mixture of SF_6 + CO_2.

The developed model is being used in the planning and evaluation of experimental measurements of droplet nucleation in CO_2-rich systems. Elements of this research are further utilized in investigation of the thermophysical properties of Hydrofluoroethers (HFE) [6,218,243]. These fluids nonflamable, dielectric and low Global warming potential fluids are an interesting replacement for currently used coolants and working fluids across wide range of applications like electronic cooling, cascade refrigeration, heat transfer fluids, lubricant carriers or rinsing agents.

Simulation of supersonic expansion 6

6.1 Introduction

Water clusters ready for photochemical and other experiments can be obtained via supersonic expansion [27]. During the expansion, saturated water vapor expands through a nozzle into a vacuum, reaches supersonic speed, cools down, and nucleates into clusters. These clusters are similar to atmospheric aerosol particles; therefore, their properties have a great meaning in the chemistry of the atmosphere [9].

The cross-section of water clusters can be estimated [71, 141] by their capability to pickup other particles on their surface. After considering polarization and other phenomena that usually cause the pickup cross-section to be larger than the geometric cross-section, it has been deduced that there is still an unexplained increase for large clusters, which has been attributed to clusters of irregular shape [71, 142] rather than spherical. This is, to some extent, counterintuitive and indicative of complex phenomena that may take place. One possibility is the aggregation of smaller frozen clusters later in the apparatus as a consequence of turbulent flow and/or sonic boom [124]. Similarly, twin xenon clusters were identified in supersonic expansion in very narrow and long nozzles [201]. Most experiments detect frozen clusters [156], which may serve as precursors for aggregates such as snowflakes [19].

A number of experiments have determined the nucleation rates during expansion [81, 120, 226]. These measurements were performed typically in Laval nozzles where smooth walls best guarantee isentropic expansion. Conical nozzles were considered in other theoretical work [101].

Atomistic molecular simulations have been applied to homogeneous nucleation for years [8, 59, 61, 144, 190]. These works focused mostly on testing theoretical predictions of the nucleation rate. Other works dealt with the characterization of developing particles. Typically, a constant volume simulation cell with some cooling protocol (with or without carrier gas) is applied in these works [152, 205]. Such computer experiments as well as the classical nucleation theory are based on the assumption of almost equilibrium during cluster growth [95], at least since the stage of a critical cluster. Such clusters are naturally spherical (which does not rule out the possible aggregation of frozen clusters later in their flight). This assumption is to some extent valid in wide and long Laval nozzles, but is inappropriate in micrometer-wide nozzles of millimeter lengths with microsecond time of flight.

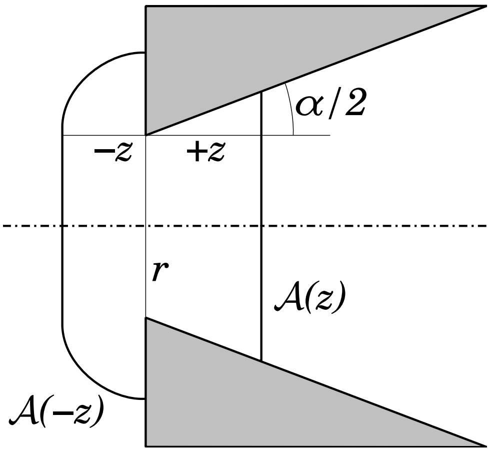


Figure 6.1: Geometry of the nozzle with diameter r, opening angle α and cross-section $\mathcal{A}$. Direction of expansion is shown with $+z$. Figure is adapted from [124].

6.2 Problem formulation

The problem formulation was motivated by the experiments performed on a medium expanding through a nozzle. With the theoretical knowledge of thermodynamics and fluid dynamics, a model is designed that can be evaluated using molecular dynamics simulation. Because of the nature of the solved problem, the common simulation software is not suitable for the problem, as discussed after the conditions for the simulated system are introduced. The novel nature (in standards for the year 2018) of the implementation is further pronounced with the use of graphics processing units as the main computational resource. To explain our reasoning for the presented solution, a brief overview of GPGPU is provided at the end of the problem formulation.

6.2.1 System conditions

The experimental setup is based on two chambers with different pressures (typically one is saturated vapor and the other is vacuum) separated by a nozzle. When the nozzle is opened, vapor flows from the storage chamber to a vacuum. The geometry of the nozzle, fig. 6.1, allows for gas to expand, filling the available volume. This consequently cools down the vapor crossing to the metastable region, where the formation of droplets is stimulated.

The particular shape of the nozzle is designed for adiabatic expansion, meaning no heat is removed from the system. This is due to the rapid expansion, where the effects of thermal motion on the walls are assumed to be without friction. The shape of the nozzle cross-section $\mathcal{A}(z)$ can be formalized as

$$\mathcal{A}(z) = \begin{cases} \pi(r^2 - \pi r z + 2z^2) & z < 0 \\ \pi(r + z\tan(\alpha/2))^2 & z \geq 0 \end{cases}. \tag{6.1}$$

The vapor is first assumed to behave like ideal gas, allowing for the model to be developed, but for real vapor behavior, a correction mechanism is proposed in the later part of the study.

$$p(z) = \frac{\rho(z)RT(z)}{M}, \tag{6.2}$$

where R is gas constant, M is molar mass, and ρ is density (not molar). Ideal gas allows us to use the Mayer formula to relate the heat capacities C_V and C_p as $C_V = C_p - R$. Further, because the process is also adiabatic, it can be described using a polytrope with a coefficient

$\xi = C_p/C_V$[1]. For the purposes of later derivation, the known adiabatic conditions with the polytropic coefficient ξ are transformed into the differential form through differentiation of the logarithm of the original condition (the shown formulas are equivalent):

$$p\rho^{-\xi} = \text{const} \quad \rightarrow \quad \frac{\partial p}{\partial z}\frac{1}{p} = \xi \frac{\partial \rho}{\partial z}\frac{1}{\rho} \tag{6.3}$$

$$T\rho^{1-\xi} = \text{const} \quad \rightarrow \quad \frac{\partial T}{\partial z}\frac{1}{T} = (\xi - 1)\frac{\partial \rho}{\partial z}\frac{1}{\rho} \tag{6.4}$$

$$T^{\xi}p^{1-\xi} = \text{const} \quad \rightarrow \quad \frac{\partial T}{\partial z}\frac{1}{T} = (\gamma - 1)\frac{\partial p}{\partial z}\frac{1}{p} \tag{6.5}$$

Related to the adiabatic expansion is the conservation of total mass flux. This means no particles are lost, e.g., due to deposition on walls. This means that the equation of continuity is satisfied in the form

$$J_w(z) = v(z)\rho(z)\mathcal{A}(z), \tag{6.6}$$

where the mass flux J_w depends on the velocity v, density ρ and cross-section $\mathcal{A}$.

The experiment is designed to allow for a continuous stream without interruptions, and because there are no barriers in the flow, it is considered non-turbulent when in the nozzle. To anchor the problem in fluid dynamics, the flow is considered to be a compressible stationary flow. Neglecting further the role of elevation, the Bernoulli equation in the form used by Landau and Lifshitz [139] as:

$$0 = \frac{\partial}{\partial z}\left(\frac{v^2}{2} + H\right) = \frac{\partial}{\partial z}\left(\frac{v^2}{2} + \frac{p}{\rho}\right) \tag{6.7}$$

Where the original enthalpy $H = U + p/\rho$ is simplified because of the constant internal energy U leading to the final form of the equation. The Bernoulli equation expresses the preservation of energy, which is comprised of the kinetic part $v^2/2$ and the enthalpic part. Integrating eq. (6.7) yields the energy of the simulated system moving through the nozzle with velocity v. This is useful for verification as well as the derivation of formulas in the next section.

This completes the problem description, where four functions $T(z)$, $p(z)$, $\rho(z)$, and $v(z)$, are tied together with four eqs. (6.2) and (6.5) to (6.7), which enables the problem solution.

6.2.2 Design aim

The typical nozzle lengths used in this type of experiment are millimeters, and speeds reach 1 km/s (for nitrogen expanding from 500K [138]); therefore, trajectories several µs long are needed. However, common parallel implementations of molecular dynamics (MD) based on domain decomposition are inefficient for inhomogeneous systems containing a few clusters in dilute gas. Most highly efficient codes are oriented on biological systems (e.g., Gromacs) and are almost always performed at constant temperature; therefore, the single precision arithmetic is sufficient, and the algorithm may often forget a small interaction close to the potential cutoff because the thermostat easily fixes such disruptions. In contrast, the expansion is an adiabatic process, meaning long-term energy conservation is very important. In the previous work of Klíma and Kolafa [124], that only 1000 molecules were used. This was not sufficient for a good statistical representation for clusters of sizes up to several hundred molecules.

Therefore, the implicitly parallel MD software developed specifically for highly inhomogeneous systems was requested. Considering the computational resources available and imposed demands on the simulation, the Nvidia Graphics processors were selected for the task, partly also because

[1] Instead of commonly used γ to not confuse with the surface tension

of the availability of the well-documented CUDA framework [179] at the beginning of the development.

At the start of the development cycle in 2018, the primary philosophy of the Mac_module was established. The aim was to reach at least tenfold wall time speedup against the original serial (or parallel with bad load balancing) implementation in MACSIMUS [126]. Additionally, the relative energy conservation has to be better than 10^{-3}J/mol ps. The primary focus of the software was on pure systems of water molecules, followed by testing substances such as nitrogen or argon (see section 3.2.3 for parametrizations used). These assumptions allowed for different organization of the data holding structure, potential construction and evaluation, as well as specifics of individual calculation routines (kernels) to be more adjusted for use on GPUs. Solving these problems, which are presented in section 6.4, allowed the author to create software that calculates the given task of supersonic expansion for large heterogeneous systems with a simulated time span of 1 µs.

6.2.3 Architecture consideration

The GPU (graphics processing unit) and GPGPU (general purpose GPU) hardware specification and parallel paradigm have to be reflected in the design of the code; otherwise, the performance gains do not outweigh the invested work. This paradigm is represented with the generally known rules: throughput-first, coalesced memory accesses, GPU saturation, and branch-less code with reasonable kernel launch strategies as advised by the CUDA toolkit documentation [180]. With the basic terminology, here are four specially targeted areas of importance for the solved case. For specific GPU microarchitecture specifications, consult chapter 19.2 of the documentation [180].

6.2.3.1 Thread divergence

There is a significant difference in the employed parallel paradigm that values memory locality more than in the CPU case. The reported theoretical scale of parallelism is when groups up to 32 threads (a group of 32 threads is called a warp) execute the instructions together. This explains the name single instruction multiple threads (SIMT), which is the parallel architecture of the NVIDIA GPU's [180]. This is important when branching in the executed kernel code is encountered. In that case, the branch (if-then-else statement) execution is serialized, leading to a decrease in performance due to some threads waiting. We can call this a divergent case as the condition in code divides threads into two groups that need to be executed.

6.2.3.2 Coalesced access

Another related topic is the coalescing of memory reads and writes. Quoting the CUDA-programming guide 12.2 [180]: "Global memory resides in device memory, and device memory is accessed via 32-, 64-, or 128-byte memory transactions." This means that misaligned data or data scattered across the memory require more memory loads than just the total size of memory requested divided by 32 (or 64, or 128). Before the data are available, the warp is asleep or trying to evaluate other warps for which the data were available. Proper data access pattern, so-called coalesced access, is a significant factor during optimization. The programmer is tasked with designing the kernels (or the whole algorithm) in a way that allows threads to read and write the data into memory in sequential 32, 64, and 128-bit chunks. As will be described later, this requirement poses a significant issue for any many-body problem as the interaction between bodies is determined by the Euclidean distance of moving elements, resulting in a scattered access pattern.

6.2.3.3 Load balancing

A programatically controled collection of threads is called block. During the execution, these blocks are allocated to the chip on the different physical Streaming Multiprocessor (SM) where they are executed. This leads to a heterogeneous nature of execution on the GPU, where the programmer does not have a direct control over which blocks are executed first. This changing order of execution of blocks from run to run explains why direct error propagation investigation is a very complicated problem. Given the essentially random distribution of blocks on SMs, the task of achieving the desired performance demands that the work is equally distributed among the available SMs. This is called load balancing, and because of the lack of direct control over block allocation, it is the task for the programmer to designing blocks with a similar amount of work in them to fulfill this load balancing requirement.

6.2.3.4 Computational units

One of the downsides of the GPU processor that is relevant to the implementation in this study is the disparity between the different processing units available. The observed throughput for double precision operations is up to 64 times less than for single precision ones for low-end (e.g., GeForce desktop series but also for the recent RTX series) GPUs that are widely available. Even GPUs dedicated to scientific calculation (e.g., the TESLA series) still have half the double precision throughput as compared to floating operations. The disparity in precision calculation is a direct consequence of the original purpose of the GPU, and it has to be accounted for in any software that requires high numerical precision. The scenario is much worse when special function units are responsible for the evaluation of trigonometric, exponential, logarithmic, and square root functions. The programmer has to pay close attention to this disparity when higher precision is required because most of the functions are approximated to the accuracy of single precision.

6.3 Problem solution

The initial problem solution was developed in the group led by the author's supervisor, Jiří Kolafa, and coleague Martin Klíma during the years 2017–2018 [124]. We will now formulate the original solution, which was later extended and modified for the purpose of the parallel implementation by the author. The goal was to remove the low efficiency issue reported by Klíma and Kolafa in [124]. Higher efficiency enables the simulation of larger systems for a longer period of time.

The proposed idea for molecular simulation was to design a protocol allowing the simulation of a system within the expanding medium. The simulated system does not know about the rest of the medium, while its properties are externally controlled during the expansion. The primary parameter is the expansion axis coordinate z later related to time. The controlled properties like the system density (volume) are then evolved based on the simulation time. The novel approach is the departure from the NVT ensemble, where effects caused by the use of thermostat cannot be neglected, allowing only the study of the system under equilibrium. The proposed approach instead enables to perform adiabatic simulations similar to NVE[2], where the expansion is governed in such a way that the adiabatic evolution of the system is preserved. This expansion protocol then allows for the study of non-equilibrium evolution.

We will now develop the equation necessary for the design of such an expansion protocol. Relying on the publication by Klíma and Kolafa in [124]. The author's extension primarily

[2] The energy conserved is in form of sum of Enthalpy and kinetic energy of the moving system.

concerns the parallelization and consequent implementation of the protocol into standalone molecular dynamics simulation software using GPGPU, as presented in section 6.4.

6.3.1 Expansion of ideal gas

The goal of the solution is to obtain an evolution equation for important properties like temperature, density, and velocity. Using the velocity functions, the coordinates can then be related to time dependence. In this way, the simulation time determines where the simulation box is located in the nozzle and what the density of the system should be.

The first step is to determine the characteristics of the governing functions to verify that the used model corresponds to the observed experimental knowledge. For that reason, the Bernoulli equation for compressible flow in eq. (6.7) is formulated for the ideal gas case

$$0 = \frac{\partial}{\partial z}\left[\frac{v^2}{2} + \frac{RT\xi}{M(\xi-1)}\right]. \tag{6.8}$$

which is differentiated into:

$$\begin{aligned} 0 &= v\frac{\partial v}{\partial z} + \frac{R\xi}{M(\xi-1)}\frac{\partial T}{\partial z} \qquad (6.9)\\ &= v\frac{\partial v}{\partial z} + \frac{RT\xi}{M}\frac{\partial \rho}{\partial z}\frac{1}{\rho}, \qquad (6.10)\end{aligned}$$

where the adiabatic condition from eq. (6.4) was used to transform the differencial from temperature to density.

Taking the continuity eq. (6.6) and transforming it also into the differential form in the same way as the adiabatic condition gives expression for the density derivative

$$\frac{\partial \rho}{\partial z}\frac{1}{\rho} = -\frac{\partial v}{\partial z}\frac{1}{v} - \frac{\partial \mathcal{A}}{\partial z}\frac{1}{\mathcal{A}}. \tag{6.11}$$

Inserting eq. (6.11) into eq. (6.10) leads to

$$0 = v\frac{\partial v}{\partial z} - \frac{\xi RT}{M}\left(\frac{\partial v}{\partial z}\frac{1}{v} + \frac{\partial \mathcal{A}}{\partial z}\frac{1}{\mathcal{A}}\right) = v\frac{\partial v}{\partial z} - c^2\left(\frac{\partial v}{\partial z}\frac{1}{v} + \frac{\partial \mathcal{A}}{\partial z}\frac{1}{\mathcal{A}}\right), \tag{6.12}$$

where $c = \sqrt{\xi RT/M}$ represents the speed of sound.

Rewriting the derivatives as logarithms collects the prefactors into the following:

$$\begin{aligned} 0 &= v^2\frac{\partial \ln v}{\partial z} - c^2\frac{\partial \ln v}{\partial z} - c^2\frac{\partial \ln \mathcal{A}}{\partial z}\\ \frac{\partial \ln v}{\partial z} &= \frac{1}{(v/c)^2-1}\frac{\partial \ln \mathcal{A}}{\partial z} \qquad (6.13)\end{aligned}$$

where the final formulation allows to examine the behavior in the region before the nozzle $z < 0$ and inside the nozzle $z > 0$. From analysis of eq. (6.13) for $z < 0$ in conjunction with the cross-section term $\partial \ln \mathcal{A}/dz < 0$ it follows that $v < c$. For the eq. (6.13) to be satisfied at $z = 0$, the velocity has to be equal to the speed of sound, $v(0) = c(0)$. This provides an important condition for the flow entering the nozzle. As the flow continues into the nozzle and z increases, the temperature decreases and the fraction $v/c > 1$. This means the flow reaches supersonic velocities. In this way, the ideal gas model agrees with the physical idea of the supersonic expansion process.

With the condition $v(0) = c(0)$, the equation for temperature evolution can be derived. Integrating eq. (6.8) gives the content of the bracket on the left-hand side. Where the initial enthalpic term for the ideal gas is expressed on the right hand.

$$\frac{v^2}{2} + \frac{RT\xi}{M(\xi-1)} = \frac{RT_0\xi}{M(\xi-1)} \tag{6.14}$$

Where T_0 is the initial temperature equal to the vapor medium before the valve is released. Inserting the condition $v(0) = c(0)$ with the formula for the speed of sound defined in eq. (6.12) and simultaneously holding $z = 0$ leads to:

$$\frac{\xi RT(0)}{M} + \frac{RT(0)\xi}{M(\xi-1)} = \frac{RT_0\xi}{M(\xi-1)}, \tag{6.15}$$

which can produce an equation for the temperature of the medium at the nozzle throat.

$$T(0) = \frac{2T_0}{\xi+1}. \tag{6.16}$$

The speed of sound at the entry to the nozzle follows immediately after, when $T(0)$ is substituted back into the speed of sound c.

$$c(0) = \sqrt{\frac{2\xi RT_0}{(\xi+1)M}} \tag{6.17}$$

To continue developing the formula for velocity from eq. (6.14), expressing the density and temperature arbitrary z is required. For that reason, the adiabatic condition eq. (6.4) in the integrated form is used to relate densities with the derived entry temperature $T(0)$ and the initial temperature T_0. The same relation can also be expressed from eq. (6.16) leading to the following equalities:

$$\left(\frac{\rho(0)}{\rho_0}\right)^{\xi-1} = \frac{T(0)}{t_0} = \frac{2}{\xi+1}. \tag{6.18}$$

The first equality can also be used to evaluate density at z from the knowledge of the temperature at z.

Inserting eqs. (6.17) and (6.18) into eq. (6.14) leads to an equation for the velocity only in terms of the initial values T_0, ρ_0 and density ρ:

$$\frac{v^2}{2} + \frac{RT_0\xi}{M(\xi-1)}\left(\frac{\rho}{\rho_0}\right)^{\xi-1} = \frac{RT_0\xi}{M(\xi-1)} \tag{6.19}$$

The only property left to be determined is density. For that purpose, the continuity equation in eq. (6.6) can be utilized as $\rho = J_w/v\mathcal{A}$. Remembering the flux J_w is assumed constant, the initial conditions for flux in the combination with eqs. (6.17) and (6.18) can be used again to express the flow only with the means apriori known properties as:

$$J_w = \left(\frac{2}{\xi+1}\right)^{\frac{1}{\xi-1}} \sqrt{\frac{2RT_0\xi}{M(\xi-1)}}\rho_0\mathcal{A}(0) \tag{6.20}$$

The dependence of density on velocity makes the eq. (6.19) transcendental in velocity and explains why no direct solution is povided and the equation is solved numerically. Note that two solutions are obtained from eq. (6.19); the bigger root is used for the system within the nozzle $z > 0$, while the smaller root may be utilized for the system before entering the nozzle $z < 0$.

Information about the expanding system is completed with the relation for the simulation time. The formula follows directly from integration of the obtained velocity

$$t = \int_0^z \frac{1}{v(s)} ds. \tag{6.21}$$

The integrated coordinate s is used from the entry point to the current location z in the nozzle.

In this way, the problem is now fully determined with the necessary formulas for the evaluation of temperature, density, velocity, and time. The primary input for the integration method is the density evolution $\rho(t)$. All the equations are valid in the case of ideal gas expansion which is adjusted for the real gas in next section.

6.3.2 Model schema

Up to this point, the expansion was assumed to behave like ideal gas; this allowed us to formulate the necessary equations for flux as well as others to be able to perform simulations of an expanding system. However, nucleation not only modifies the equation state, but also introduces time factor. We need an expansion protocol, $\rho(t)$, that would be consistent with the equations of fluid dynamics.

Since high-precision energy conservation is required, the symplectic Verlet method is required. We have chosen Leap-frog described in section 3.4.3.2 as an integrator and SHAKE from section 3.4.4.2 for constraints. Using a box expansion protocol, $\rho(t)$, leads to the same requirement as for the Nosé–Hoover thermostat and similar methods, namely that velocities are required before they are calculated. We chose the TRVP method [127] with good time reversibility, although symplecticity is slightly violated.

Following the model formulation in section 6.2.1 for a simulation cell carved out from an expanding beam, it means that the cell enthalpy, H, plus the flow kinetic energy, E_{flow}, of the cell is constant, eq. (6.7). The sum of these parts is the total energy of the system:

$$E_{\text{tot}} = H + E_{\text{flow}} = \text{const}, \tag{6.22}$$

Note that the simulation box itself is periodic, and its momentum is zero. The energy E_{flow} can be obtained by integrating the pressure-volume work using second-order formulas for numerical derivatives to obtain

$$E_{\text{flow}}(t + \Delta t) = E_{\text{flow}}(t) - \frac{V(t) + V(t + \Delta t)}{2} \cdot [p(t + \Delta t) - p(t)] \tag{6.23}$$

It can be seen that for a very fine sampling of the system, using eq. (6.23) is not a cost-effective solution because of the configuration pressure calculation required in the formula. For regular large-scale simulation, this is not suitable, and an alternative approach is proposed that does not incur significant error.

We have performed several numerical tests to verify that the total energy eq. (6.22) is constant with sufficient precision (for example, see section 6.6.1). In turn, we can safely use equation $E_{\text{flow}}(t) = H(0) - H(t)$, where the flow velocity at $t = 0$ is zero.

From eq. (6.7), kinetic energy of the flow can be directly calculated as $E_{\text{flow}} = m\dot{z}^2/2$, where m is the box mass. Expressing the velocity from the two fomulations for $E_{\text{flow}}(t)$ gives

$$v(t)^2 = \frac{2(H(0) - H(t))}{m} \tag{6.24}$$

Form velocity $v(t)$, the position, z, of the simulation box in the nozzle, is gathered as integral

$$z(t) = z_0 + \int_0^t v(\tau) \mathrm{d}\tau. \tag{6.25}$$

With position, the relations for nozzle cross-section $\mathcal{A}(z)$ follow from eq. (6.1) allowing the evaluation of mass flux J_w according to eq. (6.6). According to the assumed steady state flow, the flux has to be constant, which is the basis of the adjustment from the ideal gas case to the real gas represented by the molecular model.

Real gas adjustment The method for the correction was designed by Martin Klíma (for a more detailed explanation, see [123, 125]). In principle, a predictor in the form of an ideal gas approximation is being iteratively corrected with short simulation runs of 200 ps to achieve constant flux J_w.

This iteration is done for smaller systems, where the cost of multiple restarts of the simulation is manageable. In this way, a density evolution $\rho(t)$ for real gas in the form of piecewise linear interpolation sampled at 200 ps intervals is obtained. For production systems, this dependence is scaled to fit the desired system size, making no larger error than ±2.5% in flux conservation. This accuracy is sufficient to consistently bind the microscopic description obtained from MD with fluid dynamics. Because the $\rho(t)$ is precalculated, no expansion model needs to be employed.

6.4 Parallel solution

The bottom-up algorithm design was adopted, with most of the work dedicated to the GPU. For the single component case, where the number of molecules must be multiple of 32. Furthermore, all the data in matrix form are flattened, which is better for coalescing. The molecular structure is hardwired into the software and loaded into constant memory before computation. The user input and output processing are done on the host side and are omitted from the description.

Our MD code customarily (see section 3.4.6) consists of three costly steps. The most time consuming part is force evaluation. Second is the Verlet list construction (or update), which can technically become more demanding but is executed less often. The integration of the equations of motion together with SHAKE follows.

Thanks to using the TRVP predictor, SHAKE is the only remaining iterative method. A fixed number of iterations SHAKE (15) is used with superrelaxation to guarantee sufficient relative precision better than 10^{-10}. Note that in the publication by Klíma and Kolafa [124] the accuracy threshold of only 10^{-7} was used, which led to a (still acceptable) but suboptimal energy drift of 0.4% per µs.

In this section, the key points of our parallel solution implementation are addressed. The underlying data structure is addressed first with a discussion of why double precision is necessary. An explanation of interaction holding structure choice follows next, along with examples of how the Verlet list is being processed.

6.4.1 Data representation

During the simulation, precision and related energy conservation are of great importance (more so than execution speed). The design shown in fig. 6.2 of the holding structure for positions and velocities is therefore designed as a so-called center of mass (CM) scheme with a focus on precision. The CM scheme describes each molecule, i.e., SPC/E water by the x, y, z center of mass vectors of length N and x, y, z relative atoms vectors of length $N\,n_{\text{sites}}$.

This design has several advantages over the commonly employed absolute atomic (x, y, z) values (positions, velocities), namely:

- Verlet list can be constructed at the molecular level, lowering the complexity while preserving coalesced access to data.

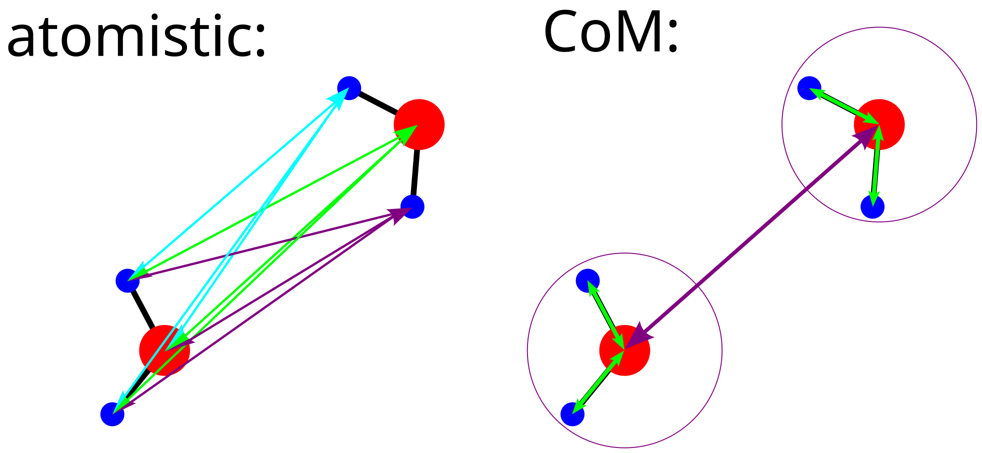


Figure 6.2: Representation of center of mass description in coparison with atomistic description for SPC/E water model. Vectors show the mutual interaction evaluation for both descriptions.

- A preemptive cutoff distance check at the molecular level can be employed, removing the need for further inter-molecular distance during force evaluation.
- Rounding errors of potential calculation are diminished.
- A similar error diminishing is observed for SHAKE constraint calculation.

The last error of calculation is less visible with the use of double precision, but it is necessary and even more so for the mixed single/double approach used in the implementation. To elaborate on the precision concerns, the errors are incurred for the large box sizes present for dilute systems in the latter parts of the expansion. In this scenario, distant molecules are being compared while their intermolecular structure remains the same. This means numbers are compared with different orders of magnitude, which leads to larger rounding errors.

Even for small systems and boxes this is clearly observable when energy conservation is studied in an adiabatic simulation for the same simulation conditions but with different precision and using the CM technique. In fig. 6.3, the case of float precision shows a dramatic difference of four orders of magnitude in the energy drift of cases with and without CM. This example case shows why single precision is unfeasible for the adiabatic simulations. Figure 6.3 also shows that float precision with CM cannot replace double precision because the required energy conservation below 0.5% per µs is not achieved. Quick calculations yield that more than 2000× improvement is required, which is achieved only for double precision. To conclude, double precision is required for adiabatic simulation, where energy conservation is critical. In consequence this condition limits the potential speedup for some GPUs, which can differ in throughput up to 64× between single precision and double precision due to the computation units available.

6.4.2 Interaction structure

With the mentioned requirements for performance and taking into account the overview [200], the initial MD module was designed as a brute force approach to check the efficiency of the available parallelism and tackle the potential for parallelization available within interaction evaluation. In the case, where the CM approach is used and only molecular centers are evaluated, this results in an N^2 interaction calculation for the force update during a single iteration.

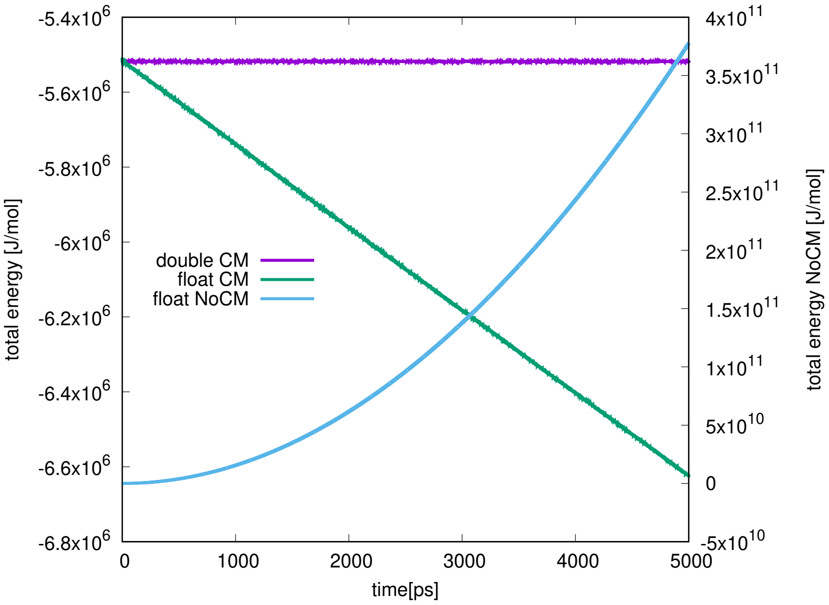


Figure 6.3: Total energy conservation graphs from NVE ensemble simulations of bulk SPC/E water with same conditions $N = 2048$, $T_0 = 350\,\mathrm{K}$, $\rho_0 = 997$ kg/m^3 and different combination of precision and scheme settings all using same 1fs timestep . Notice that the single precision NoCM is drawn with the right-hand side axis

In this sense it can be also thought that brute force is a good fit for a GPU with enough parallelism and good access. But the available computation resources quickly become insufficient with increasing N making an optimization of interaction evaluation a necessity. There is also the issue of the overwhelming number of zero interactions (zero force and zero potential energy), as shown in fig. 6.4. Memory throughput is therefore being wasted on interactions that do not contribute to the evolution of the system. The figure shows this effect for two systems with different densities, one chosen as vapor and the other as liquid for the SPC/E water model. The examined clustered case is located between the two shown border cases.

There are two popular approaches to mitigate the calculation of zero interactions. Methods based on domain decomposition (as the linked-cell list method) that divide the system volume with an artificial grid and evaluate the interaction between molecules based on cell identity. This is a popular method for homogeneous systems where similar cell occupancy can be guaranteed. For an inhomogeneous system, domain decomposition becomes inefficient, and additional operations to preserve load balancing are necessary. Most of the load balancing constitutes sorting algorithms, which impose additional complexity. Moreover for expanding systems there is a need for increasing number of domains significantly decreasing the effectivity of this approach.

We have therefore chosen the approach using the Verlet list [5, 78]. This has already been tried with homogeneous system simulation on GPU [84, 195]. The Verlet list preserves the information of the nearest neighbors for each particle in the system. This is similar to the working principle of the naive criterion described in section 4.2.1. The choice of the Verlet list reach is calculated as the potential cutoff plus padding. The padding is added to reduce the need to update the list because its construction complexity is $\mathcal{O}(n^2)$. This temporal nature and the consequent need for reconstruction is the known downside of the Verlet list, but it can be amortized when performed only when necessary and with low frequency. But for expanding system the Verlet list preserve its structure.

In fig. 6.4 the effectivity of the padded Verlet list in removing the zero interaction is shown in

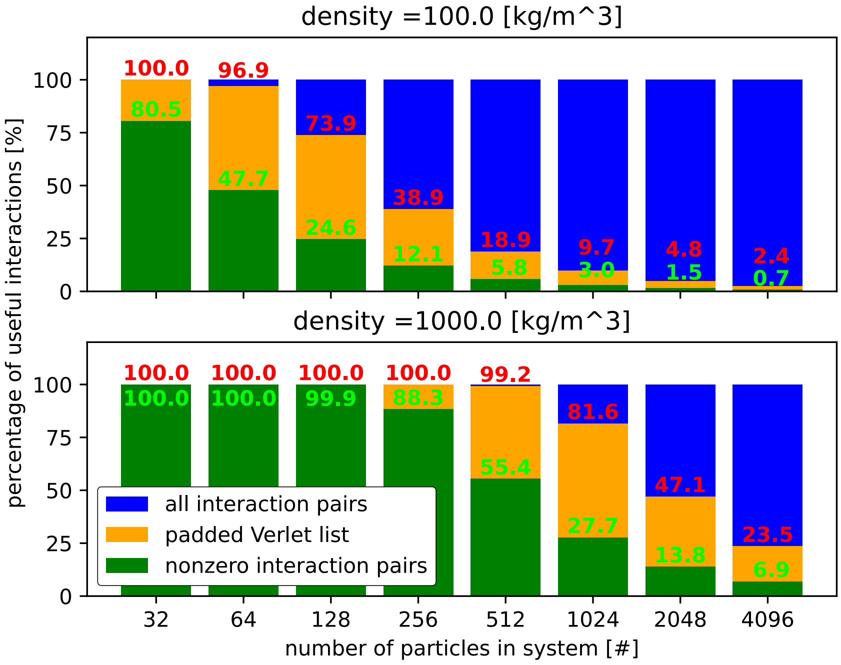


Figure 6.4: Visualization of force interactions percentage in different systems sizes and two densities corresponding to homogeneous vapor and liquid and with cutoffs defined vide infra. Comparison of strictly necessary interactions (green) with considered interactions in case of full interaction matrix (blue) and padded Verlet list (orange)

comparison with the actual nonzero interaction. The cutoff padding is chosen as a compromise between the lifetime of the Verlet list and the zero interactions included in the padded list, and in practice, it can reach up to 50% of the potential cutoff.

The construction and subsequent reconstructions are performed on GPUs, which eliminates the need for exchanging position vectors between the host and device. The Verlet list can be viewed as a sparse matrix of j indexes with a jagged structure, i.e., only the first elements of each row are valid interaction indices. Index i is inferred from the row during the accession of the structure. Verlet list structure row lengths remain uneven as molecules have varying number of neighbors, which is the consequence of system being inhomogeneous. Even with the uneven structure Verlet list is better candidate for expanding systems. Because as the system expands and distances between molecules widen the Verlet list contains on average less nonzero interaction that need to be evaluated. This means the simulation is sped as time progresses up instead of slowing down as is the case for link cell list method.

For list construction, the original two-loop iteration over all i, j molecule combinations is replaced by kernel launch in a 2D grid of 1D blocks with the number of threads adjusted for specific hardware. As shown in listing 1, the block shares molecules j between all threads while each thread evaluates interaction with its own molecule i. This allows coalescing of the global memory reads, but the primary reason is the reuse of the loaded memory. For memory writes the situation is not as simple, because the pair interaction depends on the condition of the nolecular distance being below the `verlet_cutoff`. From the performance analysis and with testing of the different access patterns, the use of atomic operations was chosen as the best option for incrementing the interaction indices into the Verlet structure.

Notice that only the blocks operating on the same row i can be in conflict during an atomic operation. Moreover, all interactions are investigated for Verlet list construction, and not only half like is the common practice on CPUs. This removes the scatter pattern during writing into memory. Another reason is the reduction of conflicts during atomic access to memory. As a consequence, the same molecule interaction with itself is resolved in the condition dr > 0.0 on line 36 in listing 1.

```
 1 __global__ void verlet_reconstruct (value_t* d_position_x,
 2                                     value_t* d_position_y,
 3                                     value_t* d_position_z,
 4                                     unsigned int* d_verlet,
 5                                     unsigned int* d_verlet_occupancy)
 6 {
 7   extern __shared__ value3_t j_positions[];
 8   value3_t i_position;
 9   value_t dx,dy,dz,dr2;
10   int i;
11   int row_id = blockIdx.x * blockDim.x + threadIdx.x;
12   int coll_id = blockIdx.y * blockDim.x + threadIdx.x;
13 
14   // all threads load position_i
15   i_position.x = d_position_x[row_id];
16   i_position.y = d_position_y[row_id];
17   i_position.z = d_position_z[row_id];
18 
19   // load the data into shared memory
20   j_positions[threadIdx.x].x = d_position_x[coll_id];
21   j_positions[threadIdx.x].y = d_position_y[coll_id];
22   j_positions[threadIdx.x].z = d_position_z[coll_id];
23   __syncthreads();
24 
25   for (i = 0; i < blockDim.x; i++)
26   {
27     dx = i_position.x - j_positions[i].x;
28     dx -= d_lx*round(dx/d_lx);
29     dy = i_position.y - j_positions[i].y;
30     dy -= d_ly*round(dy/d_ly);
31     dz = i_position.z - j_positions[i].z;
32     dz -= d_lz*round(dz/d_lz);
33 
34     dr = dx*dx+dy*dy+dz*dz;
35 
36     if(dr2 < d_cutoff_verlet*d_cutoff_verlet && dr2 > 0.0)
37     {
38       // save the index the actual position save the index
39       d_verlet[row_id*VERLET_WIDTH+atomicInc(&d_verlet_occupancy[row_id],
       VERLET_WIDTH+1)] = blockIdx.y * blockDim.x+i;
40     }
41   }
42 }
```

Listing 1: Verlet reconstruction kernel for general precision value_t.

Please note that the kernel shown in listing 1 is an example to elucidate the operation principle, and as such, it is without optimization used in our final simulation code. Further optimization include: shared memory is also utilized to collect the produced interaction indexes, leading to a single write into the Verlet structure. To diminish the register intensity of the kernel, local variables are reused, data types are reduced based on maximal limits, and loops are unrolled.

paragraph name A more elaborate optimization technique is aimed at the size of the Verlet list and its sparsity. Because of the discussed effect the expanding system has on the structure, it becomes progressively more beneficial to convert the Verlet list into its dense form. This means the empty tails of the rows are removed. This can be achieved by precalculating or approximating the size of the Verlet list, which is allocated and then filled with interactions. In this case, the padding is increased up to 50% to extend the lifetime of the list that is essentially constructed twice (first is calculating the row lengths, and the second fills the interactions). To lessen the computational strain, an approximation of the size is used based on the previous list, and expensive recalculations are done less frequently.

The Verlet list construction is an $\mathcal{O}(n^2)$ operation, which is amortized by the number of force evaluations the list remains useful. In our design, we adjust the Verlet list padding so that at least between $15 - 100$ force evaluations can be achieved per one list construction. The list's longevity depends on the maximum distance any molecule from the list has traveled. When this value exceeds the Verlet list padding, the list needs to be reconstructed. This is to prohibit the loss of interaction due to molecules flying from outside the Verlet list reaching the interaction cutoff.

6.4.3 Interaction potential

Although the Lennard-Jones and charge–charge terms in the intermolecular potential seem to be simple enough not to require any special treatment, it is not true if we realize that they must be not only cut off, but also smoothed for use in molecular dynamics see section 3.2.2.2. However, this means that the evaluation kernel would need to switch between the short-range (original), smoothed, and zero part, which slows down the parallel execution. We thus used splines, where conditions are replaced by cast to integer pointing into a lookup table in constant memory.

As a compromise between the sizes of the spline constant tables and accuracy, a quartic splines for the potential were used. Another point to consider is whether to use the atom–atom separation r or its square r^2 (which is directly available in the code) as the independent variable. Since the necessary range to cover is much wider in the latter case, leading to bulky tables, we decided to opt for square root evaluation and to spline the intermolecular potentials as a function of r. The quartic polynomial in each interval is constructed from five conditions: the potential values and derivatives at the interval ends, and the value at the center of the interval.

The Lennard-Jones potential was replaced by splines in the interval $[\Delta r, 3\sigma_{\mathrm{LJ}}]$, where $\Delta r = \sigma_{\mathrm{LJ}}/32$ is the grid and $\sigma_{\mathrm{LJ}} = 3.165558\,\text{Å}$ is the Lennard-Jones collision diameter. The cutoff was $4\sigma_{\mathrm{LJ}}$. In the interval $[3\sigma_{\mathrm{LJ}}, 4\sigma_{\mathrm{LJ}}]$, the potential was replaced by a single quartic polynomial (i.e., all spline coefficients in this interval were identical) with zero potential and force for $4\sigma_{\mathrm{LJ}}$ and the additional condition $u(3.5\sigma_{\mathrm{LJ}}) = 0.2u(3\sigma_{\mathrm{LJ}})$ which gave the optically smoothest decay to zero at the cutoff.

The $1/r$ term in the Coulomb energy is shifted, cut off, and smoothly sewn to zero,

$$\frac{1}{r} \approx \begin{cases} 1/r - S, & \text{for } r < \alpha c, \\ (r - \alpha c)^3 (A + Br), & \text{for } \alpha c < r < c, \\ 0, & \text{for } c < r. \end{cases} \tag{6.26}$$

A previously tested [129] cutoff $c = 12.6875\,\text{Å} \approx 4\sigma_{\mathrm{LJ}}$ and $\alpha = 142/203$ which is close to the recommended $\alpha = 0.7$ was used. Parameters S, A, B are determined so that the potential and forces are continuous. The potential is expressed by quartic splines as above with grid 1 Å/32; again, all polynomials for $\alpha c < r < c$ are the same.

The forces are the minus gradients of the potentials; therefore, they are expressed by cubic splines. Note that these splines do not satisfy the common condition of continuous derivatives. However, with regard to our concern, jumps in the derivatives of the forces do not impact energy conservation. The alternative approach, which uses the same order of splines for both forces and energy, leads to noisy total energy, although this error does not accumulate. This is a less preferable option. In the possible memory bound problem (the evaluation is restricted by memory access), a slightly costlier calculation of energy is not an issue.

6.4.4 Force evaluation

The most expensive part is the force evaluation, performed in full (both i, j and redundant j, i interactions are evaluated) once in each iteration. For a fixed system size N the complexity of the evaluation is governed by the number of interactions, as shown in fig. 6.4. The interactions for liquid and vapor systems substantially differ, and nucleating system are initially located in between (due to the expansion they eventually have less intarection than vapor). This is related to already discussed jagged structure of the Verlet list, particularly when clusters of higher than bulk vapor density are present in the system. To effectively access this Verlet list, a multitude of force evaluation schemes were tried in the initial stages of the development. The results obtained for older GPU were later deemed irrelevant for more powerful GPUs used later, opting for a simpler solution that can be extended to adapt to the different GPU microarchitecture specifications.

In our molecular-based design, the force evaluation proceeds at three conceptual levels. The top level in listing 2 corresponds to the grid and kernel launches, ensuring the forces from the last time-step are cleaned. This way, conditions can be removed from the force evaluation kernel. Top level also calculates grid size for the flattened Verlet structure, with row length being, in the worst-case scenario, the number of molecules. The individual blocks are 2D rectangles of threads, where the x dimension corresponds to a row of the Verlet structure and the y dimension is utilizing the instruction-level parallelism (ILP) technique to increase the amount of work done by a single thread.

In the second level, the interactions between molecules are calculated according to listing 3. The holding structure is accessed to obtain the relevant interaction pair, which calculates the force and potential energy contribution to the i-th (`id_i` in the listings) molecule from the j-th molecule (`id_j` in the listings). This presents a particular challenge with the access pattern. Coalesced access to the i molecule is achieved, but j-th molecules are accesses through indexes in the Verlet list. And while Verlet list access is coalesced, the positions of j-th molecules are scattered. This is not an issue that a different holding structure could correct [85, 195] as molecules are distributed through the system, which is the consequence of thermal motion. To amend this issue, we have considered three options:

- The indexes in the Verlet list may be sorted, which may result in less scattered access.
  - This is only partially effective and brings benefits for systems with a density close to that of liquid.
- The indexes in the Verlet list may be replaced with positions.
  - Best access pattern fully with coalesced read.
  - Is memory demanding because integer index is replaced by $3 * (1 + n_{\text{sites}})$ double values.

```
void force_calculation(value_t *d_position_x_in,
             value_t *d_position_y_in,
             value_t *d_position_z_in,
             value_t *d_position_ref_x_in,
             value_t *d_position_ref_y_in,
             value_t *d_position_ref_z_in,
             value_t *d_force_ref_x_in,
             value_t *d_force_ref_y_in,
             value_t *d_force_ref_z_in,
             value_t *d_epot_in,
             unsigned int* d_verlet)
{
  thread_per_block = 32;
  nblock = sysp.molecule_count_aligned/thread_per_block
  WORKLOAD = 16;

  dim3 thread_p_block(thread_per_block,sysp.molecule_count_aligned/WORKLOAD);

  // clean force vetors form last iteration
  set_zero<<<nblock*N_SITES, thread_per_block>>>(d_force_ref_x_in);
  set_zero<<<nblock*N_SITES, thread_per_block>>>(d_force_ref_y_in);
  set_zero<<<nblock*N_SITES, thread_per_block>>>(d_force_ref_z_in);

  // clean potential energy vetor form last iteration
  set_zero<<<nblock, thread_per_block>>>(d_epot_in);

  // calculate new forces
  n_body_vmk<<<nblock,thread_p_block>>>
      (d_position_x_in, d_position_y_in, d_position_z_in,
       d_position_ref_x_in, d_position_ref_y_in, d_position_ref_z_in,
       d_force_ref_x_in, d_force_ref_y_in, d_force_ref_z_in,
       d_epot_in, d_verlet);

  cudaSafeKernell();
  return;
}
```

Listing 2: Force calculation function for general precision value_t equal to double.

  - Imposes an update of the Verlet list in each timestep which becomes similarly inefective like domain decomposition methods.

- Rely on the caching capabilities of the GPU and ensure that even the scattered access yields useful data.
  - Requires optimization of data structure.
  - Is very hardware dependent (size of cache).
  - Is easiest to implement.

Based on the trial implementation of the three options, the last one was found to be the most cost-effective when applied to the solution in listing 3. The first option was employed partly in the optimization of the Verlet list memory writes.

Memory writing can be again performed in a coalesced manner, because contributions are written only to the molecule i. This means force contributions j are not recycled, as doing so would incur scatter write into memory with multiple conflicts. Therefore, the locally incremented force and potential energy are only attributed to the i-th molecules, and the kernel is launched on the “whole” Verlet list rather than its “half”. In this design, multiple blocks can operate on a single row, requiring the use of atomic operations for writing out the results. The use of atomic operations prevents the data race and is more effective than the synchronization of results into temporary variables with subsequent reduction. The maximum possible conflict is determined by the number of molecules divided by the workload, which can be fine-tuned for the examined system size.

```
__global__ void molecule_interaction (value_t *d_position_x_in,
                  value_t *d_position_y_in,
                  value_t *d_position_z_in,
                  value_t *d_position_ref_x_in,
                  value_t *d_position_ref_y_in,
                  value_t *d_position_ref_z_in,
                  value_t *d_force_x_in,
                  value_t *d_force_y_in,
                  value_t *d_force_z_in,
                  value_t *d_epot_in,
                  unsigned int* d_verlet)
{
  unsigned int id_i = (blockDim.x*blockIdx.x + threadIdx.x);
  unsigned int id_j = VERLET_WIDTH*id_i+threadIdx.y;
  value_t output[3*N_SITES+1] = {0.0}; //store atomistic forces +energy
  value_t dx,dy,dz;

  // processing of sequence j indexes
  for (int ii= 0; ii < blockDim.y; ii++)
  {
    dx = d_position_x_in[id_i] - d_position_x_in[d_verlet[id_j]];
    dx -= d_lx*round(dx/d_lx);
    dy = d_position_y_in[id_i] - d_position_y_in[d_verlet[id_j]];
    dy -= d_ly*round(dy/d_ly);
    dz = d_position_z_in[id_i] - d_position_z_in[d_verlet[id_j]];
    dz -= d_lz*round(dz/d_lz);

    //substance specific interaction evaluation
    mol_interaction(id_i,d_verlet[id_j],
                    dx, dy, dz,
                    d_position_ref_x_in,
                    d_position_ref_y_in,
                    d_position_ref_z_in,
                    output);
    WORKLOAD++; // increment ro preserve coalesced access for id_j
  }

  // increment the atomistic forces for sites of molecule i
  for(int i=0; i<N_SITES; i++)
  {
    atomicAdd(&d_force_x_in[N_SITES*id_i+i], output[1+3*i]);
    atomicAdd(&d_force_y_in[N_SITES*id_i+i], output[2+3*i]);
    atomicAdd(&d_force_z_in[N_SITES*id_i+i], output[3+3*i]);
  }
  // increment potential energy for molecule i
  atomicAdd(&d_epot_in[id_i], output[0]);

  return;
}
```

Listing 3: N-body evaluation kernel for general precision value_t.

The interactions are being accessed in chunks of the y-dimension of the block and processed. For each interaction, the molecular distance is calculated and forwarded into the device function, evaluating interactions between the atoms of the molecules i and j according to the molecular structure section 3.2.3. The case of SPC/E is shown in listing 4. The partial results are accumulated into the local variable output. The output contains the force and potential energy contribution to i from all evaluated j. In the example case, all the conditions for evaluating the empty interaction are left to the generic mol_interaction function. This is possible for the single component case, where the Verlet structure uses the same row index to indicate the invalid interaction. Calculating with the same indices results in zero distance, which results in the zero contribution, as was desired. This is shown in listing 6 on the line 5.

```
__device__ void spce_interaction(unsigned int id_i,
                                 unsigned int id_j,
                                 value_t dx,
                                 value_t dy,
                                 value_t dz,
                                 value_t *d_pos_ref_x,
                                 value_t *d_pos_ref_y,
                                 value_t *d_pos_ref_z,
                                 value_t *output)
{
  value_t dx_atm,dy_atm,dz_atm;

  dx_atm = dx +(d_pos_ref_x[N_SITES*id_i] -d_pos_ref_x[N_SITES*id_j]);
  dy_atm = dy +(d_pos_ref_y[N_SITES*id_i] -d_pos_ref_y[N_SITES*id_j]);
  dz_atm = dz +(d_pos_ref_z[N_SITES*id_i] -d_pos_ref_z[N_SITES*id_j]);

  // LJ interaction between O-O
  lj_spline(dx_atm, dy_atm, dz_atm, 0, output);

  // electrostatic interaction
  for (int ii = 0; ii < N_SITES; ii++)
  {
    for (int jj = 0; jj < N_SITES; jj++)
    {
      dx_atm = dx +(d_pos_ref_x[N_SITES*id_i+ii] -d_pos_ref_x[N_SITES*id_j+jj]);
      dy_atm = dy +(d_pos_ref_y[N_SITES*id_i+ii] -d_pos_ref_y[N_SITES*id_j+jj]);
      dz_atm = dz +(d_pos_ref_z[N_SITES*id_i+ii] -d_pos_ref_z[N_SITES*id_j+jj]);

      lj_elspline(dx_atm, dy_atm, dz_atm,
                  ii,d_mol_charge_template[ii]*d_mol_charge_template[jj],
                  output);
    }
  }
}
```

Listing 4: Evaluation of SPC/E–SPC/E interaction.

The last conceptual level is the forcefield evaluation. At this point, the device functions are written using macros (see listing 6) to contain all the substance–specific properties of the spline designed according to section 6.4.3. The example of a macro block for SPC/E is shown in listing 5. This device function is responsible for accumulating the potential energy and force from the lowest-level function into the output structure.

The lowest level functions of the potential and force perform a lookup in the table containing the appropriate spline. The table is stored in the constant memory loaded at the start of the simulation, along with spline-specific parameters. The grid constant for the potential Δr named in

```
__device__ void lj_spline(value_t dx_in, value_t dy_in, value_t dz_in,
                 int i_in, value_t *output)
{
  value_t dr_fr; // dr later recycled to fr value of force

  // for SPCE version lj_force, lj_pot r is required
  dr_fr = sqrt(dx_in*dx_in +dy_in*dy_in +dz_in*dz_in);

  output[0] += lj_pot(dr_fr);
  dr_fr = lj_for(dr_fr)/(dr_fr);

  output[1+3*i_in] += dr_fr*dx_in;
  output[2+3*i_in] += dr_fr*dy_in;
  output[3+3*i_in] += dr_fr*dz_in;

  return;
}
```

Listing 5: Spline evaluation for case of SPC/E.

the program as POT_WIDTH and the minimal/maximal effective indices MIN_SIZE/MAX_SIZE of the spline. For a single component, this has proven to be the best performing solution, allowing a significant reduction in the number of branches in the program. The end results of the force

```
__device__ value_t lj_pot(value_t x)
{
  unsigned int i;
  // get lookup table index
  i=(unsigned int)(x*POT_WIDTH);
  if (i<MIN_SIZE || i> MAX_SIZE)
  {
    return 0.0;
  }
#if POT_COLL == 3 // case of quadratic spline
  i *= 3;
  return d_pot_spline[i] +x*(d_pot_spline[i+1] +x*(d_pot_spline[i+2]));
#elif POT_COLL == 4 // case of cubic spline
  i *= 4;
  return d_pot_spline[i] +x*(d_pot_spline[i+1] +x*(d_pot_spline[i+2] +x*(
    d_pot_spline[i+3])));
#elif POT_COLL == 5 //case of biquadratic spline
  i *= 5;
  return d_pot_spline[i] +x*(d_pot_spline[i+1] +x*(d_pot_spline[i+2] +x*(
    d_pot_spline[i+3] +x*(d_pot_spline[i+4]))));
#endif
}
```

Listing 6: Interaction potential in form of lookup table interaction.

calculation are x, y, z atomistic force vectors and the potential energy calculated for molecule i. The division shown here represents the code structure, which was designed for feature separation, aiding in cases of issues or applying optimizations. Similarly, like in the case of the Verlet structure, the code sample shown here differs from the routines used in our final calculations lacking further optimizations. The optimization generally makes the program harder to read and understand, which motivated their removal from the samples. Let us briefly mention a few further optimization: register strain reduction with variables recycling (illustrated in listing 5 lines 7 and 10), explicit loop unrolling with (i.e., on lines 21 and 23 in listing 4).

6.4.5 Propagator along given density–time dependence

The symplectic (therefore, time-reversible) Verlet-class propagators (including SHAKE for maintaining bond lengths) are not suitable for integrating the equations of motions with the right-hand side depending on velocities because velocities $v(t)$ at time t can be calculated only *after* the step from t to $t + \Delta t$ has been finished. Such equations appear not only in thermostats and barostats implemented by the extended Lagrangian approach, but also for expansion controlled by a given density–time dependence. There are three groups of methods available to overcome this missing knowledge: (i) time-reversible methods based on the Trotter decomposition on the Liouville operator [159], (ii) iterations, where the step (including SHAKE) starts with the first guess of $v(t)$ and is repeated until accurate enough [231], and (iii) predicting the velocities by the TRVP method [128]. The last method is the simplest and most suitable for GPUs because all steps (including SHAKE) are performed sequentially only once. We are aware of other potentially suitable methods, like modified Gear methods with good time reversibility, examined by Janek and Kolafa in [110]; however, more research is required before they can be applied to models with constraints.

The schema starts with a prediction of molecular velocities at time t using the time-reversible velocity predictor [128] of order $k = 2$ shown in eq. (6.27). The predictor is, in contrast with [124], applied only to the centers of mass because the relative distances of the centers of mass change over time. In turn, there is no product of velocities subtracted in eq. (6.34).

Next, the box rescaling parameters Θ and v_f^{CM} are calculated with eqs. (6.28) and (6.29). The required values of V and $l = V^{1/3}$ are simply obtained from the density–time dependence and not integrated as in isobaric simulations. Forces acting on atoms $\boldsymbol{f}_j$ are calculated in eq. (6.38) using global atom positions $\boldsymbol{r}_j^{\mathrm{g}} = \boldsymbol{r}_i^{\mathrm{CM}} + \boldsymbol{r}_j$, where CM denotes the center of mass of molecule i and $\boldsymbol{r}_j$ are relative to the CM. Further, the classical Leap-frog schema in eqs. (6.34) and (6.35) is performed, and the resulting positions are rescaled by $\Theta(t + \Delta t/2)$ in eq. (6.36). The periodic boundary conditions are applied to the resulting centers of mass. Then, the Leap-frog schema is performed for atomic coordinates in eqs. (6.36) and (6.39). When constraints are present, the SHAKE algorithm is applied to update the relative positions of atoms in eq. (6.41). The update ensures constraints but invalidates the velocities of atoms, which are recalculated as differences in positions according to eq. (3.39). The kinetic energy is then calculated based on updated atomic velocities and molecular velocity, which, because of the preservation of CM by SHAKE, remains unchaged.

In the following algorithm given in the form of equations, index i denotes particles and j interaction sites; $\sum_{j \in i}$ denotes a sum over all sites of molecule i.

$$\boldsymbol{v}_i^{\text{CM}}(t) = \frac{5}{3}\boldsymbol{v}_i^{\text{CM}}(t-\Delta t/2) - \frac{5}{6}\boldsymbol{v}_i^{\text{CM}}(t-3\Delta t/2) + \frac{1}{6}\boldsymbol{v}_i^{\text{CM}}(t-5\Delta t/2) \tag{6.27}$$

$$\Theta(t+\Delta t/2) = \left[\frac{V(t+\Delta t)}{V(t)}\right]^{\frac{1}{3}} \tag{6.28}$$

$$v_f^{\text{CM}}(t) = \frac{1}{\Delta t}\ln\left[\frac{l(t+\Delta t/2)}{l(t-\Delta t/2)}\right] \tag{6.29}$$

$$\boldsymbol{f}_j(t) = -\frac{\partial U}{\partial \boldsymbol{r_j}^{\text{g}}(t)} \tag{6.30}$$

$$\boldsymbol{f}_i^{\text{CM}}(t) = \sum_{j\in i}\boldsymbol{f}_j(t) \tag{6.31}$$

$$\dot{\boldsymbol{v}}_i^{\text{CM}}(t) = \frac{\boldsymbol{f}_i^{\text{CM}}(t)}{m_i} - v_f^{\text{CM}}(t)\cdot\boldsymbol{v}_i^{\text{CM}}(t) \tag{6.32}$$

$$\dot{\boldsymbol{v}}_j(t) = \frac{\boldsymbol{f}_j(t)}{m_j} - \frac{\boldsymbol{f}_i^{\text{CM}}(t)}{m_i} \tag{6.33}$$

$$\boldsymbol{v}_i^{\text{CM}}(t+\Delta t/2) = \boldsymbol{v}_i^{\text{CM}}(t-\Delta t/2) + \Delta t\cdot\dot{\boldsymbol{v}}_i^{\text{CM}}(t) \tag{6.34}$$

$$\boldsymbol{r}_i^{\text{CM}}(t+\Delta t) = \boldsymbol{r}_i^{\text{CM}}(t) + \Delta t\cdot\boldsymbol{v}_i^{\text{CM}}(t+\Delta t/2) \tag{6.35}$$

$$\boldsymbol{r}_i^{\text{CM}}(t+\Delta t) = \boldsymbol{r}_i^{\text{CM}}(t+\Delta t)\cdot\Theta(t+\Delta t/2) \tag{6.36}$$

$$l = l\cdot\Theta(t+\Delta t/2) \tag{6.37}$$

$$\text{Periodic boundary conditions crop} \tag{6.38}$$

$$\boldsymbol{v}_j(t+\Delta t/2) = \boldsymbol{v}_j(t-\Delta t/2) + \Delta t\cdot\dot{\boldsymbol{v}}_j(t) \tag{6.39}$$

$$\boldsymbol{r}_j(t+\Delta t) = \boldsymbol{r}_j(t) + \Delta t\cdot\boldsymbol{v}_j(t+\Delta t/2) \tag{6.40}$$

$$\boldsymbol{r}_j(t+\Delta t) = \text{SHAKE}\left(\boldsymbol{r}_j\right) \tag{6.41}$$

$$\boldsymbol{v}_j(t+\Delta t/2) = (\boldsymbol{r}_j(t+\Delta t) - \boldsymbol{r}_j(t))/\Delta t \tag{6.42}$$

$$E_{\text{kin}}(t+\Delta t/2) \tag{6.43}$$

6.5 Evaluation of characteristics

6.5.1 Cluster temperature

Connected with the system energy is the temperature, which is readily from the kinetic energy. To investigate the temperature of a cluster, we first need to define the co-called cluster temperature

$$T_{cl} = \frac{2(E_{\text{kin,c}} - E_{\text{tr,c}})}{k_{\text{B}}(\text{DoF}_c - 3)}. \tag{6.44}$$

Here the kinetic energy of cluster $E_{\text{kin,c}}$ is divided by the Boltzmann constant and the degrees of freedom of cluster DoF_c from which 3 is subtracted for conserved momenta in the periodic boundary conditions. Because the cluster is also moving, the kinetic energy caused by this movement $E_{\text{tr,c}}$ needs to be subtracted. The movement of the cluster is calculated from the center of mass of the cluster, which is substituted into the kinetic energy formula

$$E_{\text{tr,c}} = \frac{1}{2m_c}\left|\left(\sum_{i=1}^{N_{\text{c}}} m_i\boldsymbol{v}_i\right)^2\right|. \tag{6.45}$$

In this equation, N_{c} is the number of molecules in the cluster, and m_c is the total mass of the cluster obtained from the individual masses of molecules m_i. For SPC/E water, the masses

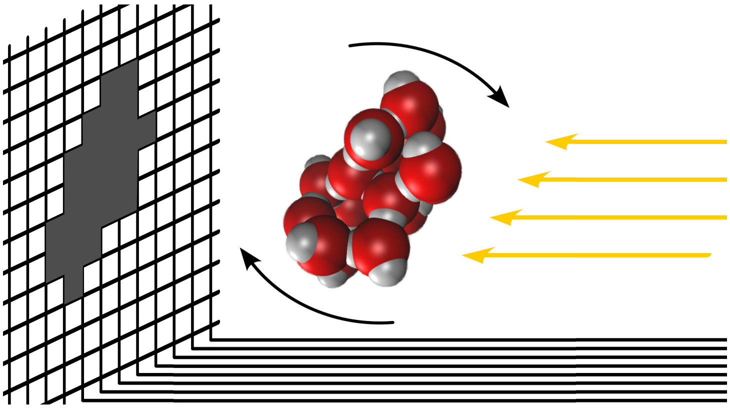


Figure 6.5: Concept of cross-section calculation using a ray and shadow analogy. Rays are shown as yellow arrows and shadow cast by molecular cluster is grey are shown in grid. Cluster systematic rotation is shown as black arrows.

are obtained from table 3.3 and $\mathrm{DoF}_c = N_\mathrm{c}\mathrm{DoF}_{\mathrm{SPC/E}}$. For the rigid model of SPC/E used for calculation, $\mathrm{DoF}_{\mathrm{SPC/E}} = 6$.

6.5.2 Cluster shape characteristics

One of the characteristics of a the cluster obtained from its shape is its (dis)similarity to a sphere. This particular measure, ψ, was defined by Wadell [251] as a ratio of the surface area of the object compared to the surface area of a sphere. This also means that if object is similar to sphere the $\psi \approx 1$ and for for increasing values of ψ the object is more dissimilar to a sphere.

$$\psi = \frac{S_{\mathrm{cluster}}}{S_{\mathrm{sphere}}}, \tag{6.46}$$

Here, the S_{sphere} is the cross-section area of a sphere with the same volume as N water molecules with the water bulk density ($\rho = 997\ \mathrm{kg/m^3}$). Cluster is therefore compared to the ideal sphere formed from the same amount of molecules of ideally bulk density water. No closest packing or interstitial spaces are considered for the sphere surface area.

To evaluate the cross-section of the cluster, which is represented as a set of points with interaction potentials, the average geometric crossection is used. This quantity can be imagined as the average shadow the cluster creates when placed between a screen and light, and consequently, the cluster is rotated as visualized in fig. 6.5.

For the calculation of the cross-section of the simulated object, the Monte Carlo in section 3.3 integration is used. The atoms are replaced by spheres of radii given by the sum of the atom Lennard-Jones collision radius ($\sigma_{\mathrm{LJ}}/2$) plus the helium probe radius ($\sigma_{\mathrm{LJ}}/2 = 1.25\,\mathrm{\AA}$). This set of spheres is projected onto a $0.1\,\mathrm{\AA}\times0.1\,\mathrm{\AA}$ grid at a random initial orientation. This is repeated for another 15 directions according to the dodecahedron-based Gaussian quadrature [162], with the final crossection evaluated as the average of the covered grid faces.

The average geometric cross-section is another important shape characteristic of the cluster. Note that the experiment by Lengyel *et al.* [142] have provided the pickup cross-sections. While quantitative comparison is hindered by the systematic overestimation of pickup cross-section compared to geometric cross-section. The qualitative comparison of trends as well as dimensionless sphericity is valid between both approaches [71].

6.6 Results and Discussion

The development of custom-built software that actively utilizes the power of GPUs for higher precision simulation has posed multiple challenges. In this regard, we present here first the answers to the posed requirements for energy conservation and speedup. After the model validation, the analysis of clusters is provided. This envelops the cluster shapes, sphericities, and growth during the expansion simulation. We further discuss the effect of the merging of clusters and the chance of molecules bouncing from the clusters with regards to CNT. The final part focuses on the comparison of simulated data with experiments by Lengyel *et al.* [141, 142].

6.6.1 Energy conservation

The first area discussed is the achieved energy conservation for expanding the test system with $N = 256$ where energy fluctuations are higher than for the production system sizes. In fig. 6.6 it can be seen that the calculated total energy eq. (6.22) deviates only slightly from the constant value. The linear fit has a slope of $-2.40 \cdot 10^{-4}$ J/(mol ps).

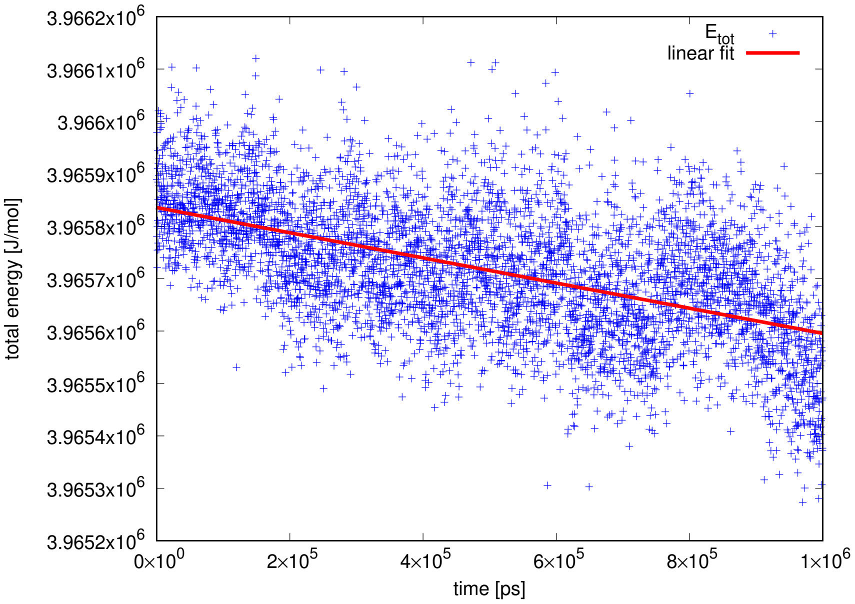


Figure 6.6: Total energy evolution for a test simulation of expansion with 256 SPC/E molecules with double precision. One point is shown for each 200 ps

The total slope value across the needed timescale of 1.0 µs is −1 J/mol per a molecule, which is negligible in comparison with the hydrogen bond energy, 20 to 25 kJ/mol. (per a mole of simulation boxes). We have therefore proved the molecular dynamics of expanding systems utilizing the model in section 6.3. This also certifies that the presented parallel solution is precise enough and that the results generated from it are valid molecular simulation pseudo-experiments. It further enables the solution to proceed as described in section 6.3.2

6.6.1.1 Achieved speedup

The second concern of the software is execution speed. For this sake, we have compared the implementation with the previously used sequential code [125] implemented in MACSIMUS. DL_POLY was added to represent the simulation package used for thermodynamic properties simulation (choice based on the year 2021). To illustrate the improvement achieved during the development, we provide a comparison with an older version of our implementation, which was presented in [41].

Three system configurations were simulated to compare performance with different initial conditions. Systems consisting of bulk water, low density vapor containing clusters, and water vapor were chosen. In this way, method operation is shown for dense systems with the highest amount of interactions as well as the initial vapor conditions from which the simulations are started. The cluster case is taken from a simulation of expansion, showing how individual methods operate for very large simulation boxes with heterogeneous density distributions.

This type of comparison is targeted more at practical use instead of following the definition of speedup. The aim is to answer the question of which implementation to use in a given situation and, more importantly, what speed benefits can be expected by doing so. The question of sequential and parallel comparison is not very relevant for answering this question. We illustrate the notion with a comparison between the older version and the version presented in this work.

In general, the design strategies significantly differ between CPU and GPU oriented code (see the GPU-specific optimization in sections 6.4.2 and 6.4.4), using techniques that are not suited for sequential execution.

The comparison of the different algorithms (optimizations) used by DL_POLY, MACSIMUS[3] and Mac_module for bulk, cluster, and vapor configurations with $N = 2048, 4096, 10240$. The benchmark simulations were performed in the NVE ensemble for 10 000 timesteps with the same initial configuration using SPC/E water molecules. The following hardware was used: CPU Intel(R) Core(TM) i7-8750H 2.20 GHz, GPU NVIDIA GeForce GTX 1060 (Max-Q)[4], NVIDIA GeForce RTX 3090[5], and NVIDIA TESLA V100[6]. The comparisons are listed in table 6.1.

		GPU-module						**MACSIMUS**	**DL_POLY**
#	**System**	**GTX 1060**		**RTX 3090**		**TESLA V100**		**i7-8750H**	**i7-8750H**
		old	new	old	new	old	new	old	old
		time [s]	time [s]	time [s]	time [s]	time [s]	time [s]	time [s]	time [s]
2048	Bulk	364.9	86.0	49.5	23.4	52.4	31.8	650.3	5,237.4
	Clusters	34.8	2.5	8.3	1.1	9.2	1.0	50.6	275.9
	Vapor	14.0	2.5	4.8	1.2	5.4	1.1	15.2	181.0
4096	Bulk	744.5	207.2	101.4	47.7	102.3	57.5	1,295.0	18,716.6
	Clusters	62.0	4.3	14.1	1.5	9.4	1.3	59.8	480.5
	Vapor	48.8	4.4	11.7	1.5	6.8	1.3	29.1	393.3
10240	Bulk	—	508.3	—	139.5	—	153.9	3,242.2	—
	Clusters	—	11.8	—	2.6	—	1.8	184.0	—
	Vapor	—	59.4	—	14.5	—	15.6	73.8	—

Table 6.1: Simulation times (lower is better) in seconds and acceleration of two versions of GPU–module for 10 000 timesteps. Old version reflect the algorithm as presented in [41] and new represent the algorithm as described in this work. Listed values are arithmetic means from 10 runs.

[3] The MACSIMUS software was tested in the single-threaded version utilized in previous work by Klíma and Kolafa [125].

[4] with Intel Core i7-8750H and 16 GB RAM

[5] with AMD Ryzen Threadripper 3960X and 32 GB RAM

[6] with Intel XEON Gold 6130 and 96 GB RAM

It can be seen in table 6.1 that the performance of our algorithm depends on the used GPU hardware and its features. We have compared three GPU processors (GTX1060, V100, and RTX3090) of three chip micro-architectures (Pascal, Volta, and Ampere) with different amounts of SM blocks (10, 80, and 82) achieving varying memory bandwidths (192.2 GB/s, 897.0 GB/s, and 936.2 GB/s) across different bus widths (192 bit, 4096 bit, and 384 bit). The consequences of these differences can be noticed in the table.

As expected, the old version using the lower performance GTX1060 does not achieve the desired performance gain, and for the largest system in the vapor phase, the timing is even worse than in the original MACSIMUS code. This is a direct consequence of over-saturating (memory bandwidth) the card calculation capabilities (SM block count) as opposed to the different optimizations more suited for CPU that are used by MACSIMUS (linked-cell list method). But many of the issues were removed in the new version, and now even the weakest card consistently outperforms MACSIMUS, with speedups up to 24× in the case of cluster systems. The saturation is noticed at $N = 10240$ where MACSIMUS again approaches the GPU performance for the vapor configuration. This means that the benefit of GPU implementation is already attainable on low-end devices, which was a pleasant surprise.

Substantially better results are achieved for the RTX3090 and especially for the A100, with more memory bandwidth and SMs that can be leveraged for larger problems. Bulk liquid simulations were improved from $\approx 12\times$ in the original version by a factor of two. The similar performance field is for vapor configurations, where the speedups for the biggest systems are only 5× in comparison to MACSIMUS. Here we notice that algorithm benefit more from bandwidth, and the available SMs on the RTX3090, showing better performance than older Volta microarchitecture. Interestingly, the amount of double precision units did not factor as relevant at this stage. The difference in the amount of interaction becomes visible for the system with $N = 10240$.

The most noticeable gains are observed for the configurations for which the algorithms are optimized. Here we see the most substantial speedups of over 100× for the largest system, with a consistent 7× improvement versus our previous version of the algorithm.

The comments were focused around MACSIMUS because, with our current understanding, we cannot explain with certainty the reason for the significantly slower single-threaded DL_POLY calculation. We suspect that the role of system inhomogeneity for which the settings were adjusted plays a role here, or there may be an issue with how DL_POLY constructs its holding structure. In contrast, MACSIMUS uses a simple linked-cell list method, with the list newly constructed around each particle at each timestep.

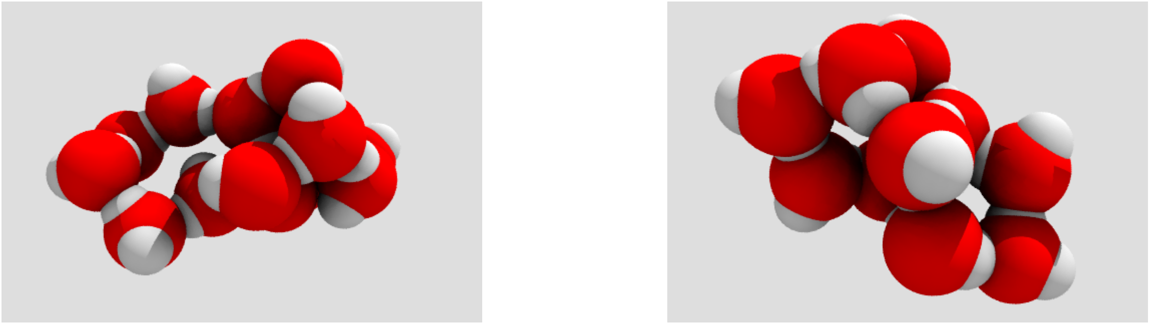
From the analysis of the algorithm behavior, a difference brought about by the newer architecture chip was noticed. Using our code as a benchmark, the V100, representing a TESLA high performance data center type GPU, is being outperformed by the new generation of desktop-based RTX 3090. In the observed case, the new micro-architecture relies more on faster memory and scheduler mechanisms instead of double precision throughput than was originally expected to be more prominent. This finding is important for further investigation and potential improvement of the algorithm.

In conclusion, the new version of the GPU parallel simulation package Mac_module brings significant improvements over its predecessor version and the original CPU implementation. This means that calculating the simulation of expansion is now a matter of days instead of months, as it was previously. We'd like to remind the reader that during production, the overall speedup is based on the performance of the vapor configuration during the initial phases, which gradually transforms into the case of cluster configurations. We believe that the new version makes the expansion simulation significantly more cost-effective compared to what was considered feasible in 2018, when the author contemplated implementing the GPU parallel algorithm.

6.6.2 Cluster Shape

The properties of clusters developed in supersonic expansion were studied in detail. We were interested in the numbers of molecules in clusters, their cross-section, and the sphericity described in section 6.5.2. Molecule connectivity to define clusters was defined by a hydrogen–oxygen distance less than 2.45 Å which corresponds to the first minimum on the radial distribution function H–O in the liquid phase. This distance is used in the Stilliger cluster criterion section 4.2.2. This way, the number of all clusters (as well as monomers) can be obtained in each time frame captured during a simulation. Cluster classification is performed after the simulation is finished.

An example of a few shapes of clusters is shown first to give the general idea of clusters created during the initial phase of the simulation of the expansion. Two of these small below-critical clusters are shown in fig. 6.7. We can see that the shapes of these clusters are not spherical, as is assumed by the CNT.

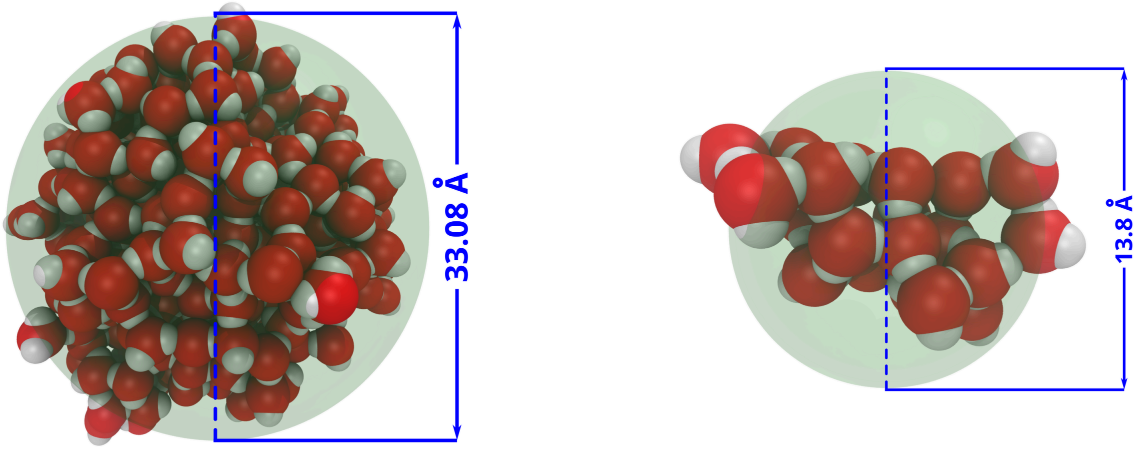


(a) Cluster with 13 SPC/E

(b) Cluster with 11 SPC/E

Figure 6.7: Visualization of two small sized clusters showing the nonspherical shapes.

To further investigate the sphericity described in eq. (6.46), we will use the largest cluster found in the later section of the simulation. Two typical examples are presented in fig. 6.8 illustrating the used sphericity definition. Note that the shown clusters exhibit a sphericity difference of $\Delta\psi = 0.4$.



(a) Cluster of 439 SPC/E molecules with cross-section 859.8 Å^2, the corresponding sphere diameter of 33.1 Å, and sphericity 1.28

(b) Cluster of 21 SPC/E molecules with cross-section 149.5 Å^2, the corresponding sphere diameter of 13.8 Å, and sphericity 1.68

Figure 6.8: Visualization of two differently sized clusters of observably different sphericity sourced from [41].

With regards to CNT, the cluster in fig. 6.8a exhibits spherical characteristics with a sphericity of 1.28, while the cluster in fig. 6.8b deviates from the spherical assumption, displaying a sphericity of 1.68. The observed elongate cluster shapes are prevalent during the initial phase of expansion simulation, suggesting that the spherical assumption is inaccurate. These cluster shapes, which were analyzed in this study, can contribute to a further development of the theory. We acknowledge that while prior research on cluster shapes was performed, that research operated under equilibrium conditions, which may have influenced the observed shapes.

6.6.3 Evolution of clusters

We have conducted several sized simulations numbering 1000, 2048, 4096, and 10240 molecules in two prescribed conditions investigated in the experiment [142]. The required number of molecules is dictated by the maximum size reported in the experiment [142] and the sufficient time is inferred from the time when the cluster no longer grows. The stored trajectories sampled by 10 ps are analyzed, and cluster information is calculated. The development of the maximal cluster in the system of 10240 molecules and 4096 molecules are shown in fig. 6.9 and fig. 6.10 respectively. The nucleation process is demonstrated in its entirety, starting from the vapor phase located before the nozzle throat under the prescribed initial T_0, p_0. The clear sign of this effect is the nucleation onset observed after around 0.1 µs, where the simulation box starts to travel through the nozzle throat. The graphs exhibit noticeable jumps in the sizes around 0.2 µs. More common upward jumps correspond to merging, as is the case in fig. 6.9 for $N = 2048$ at 0.31 µs. In contrast, a downward jump signifies breaking a large cluster into parts, see fig. 6.9 for $N = 4096$ at 0.22 µs. The maximal cluster size is observed to converge at $\sim$ 0.9 µs, after which its size slightly fluctuates.

6.6.3.1 Cluster merging

The cluster merging phenomenon was further investigated from the performed simulation. Namely, the shapes and sizes of the clusters during the process of merging were observed. Sphericity was also evaluated for the three stages before, during, and after merging. Important for future

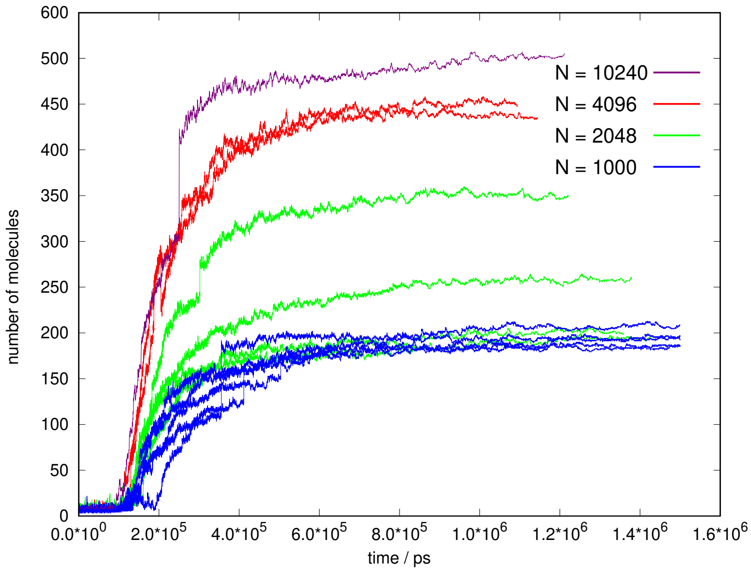


Figure 6.9: Maximal cluster size evolution for $\boldsymbol{T_0 = 409\ \mathrm{K}}$, $\boldsymbol{p_0 = 313.5\ \mathrm{kPa}}$

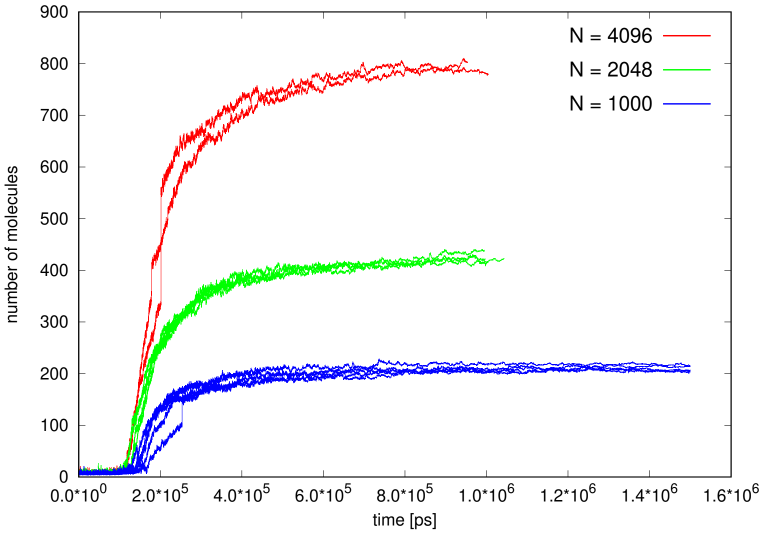


Figure 6.10: Maximal cluster size evolution for $\boldsymbol{T_0 = 424\ \mathrm{K}}$, $\boldsymbol{p_0 = 474.7\ \mathrm{kPa}}$

comparison with the experiment is the change in sphericity between two separate clusters into a final cluster.

The question here is whether this merging phenomenon can explain the rise of sphericity. According to the analysis of multiple occurrences of the merge, this does not seem to be the case. The approaching cluster is fully incorporated and forms a cluster with sphericity corresponding to its size, as shown in fig. 6.11.

Sidestepping the investigated behavior of sphericity, the merging phenomenon is of interest in relation to the CNT. The working assumption of CNT is in direct contradiction, only allowing for single molecule changes to the cluster. But during our simulation of expanding systems, we have seen multiple occurrences of these phenomena, suggesting an unaccounted process in the cluster formation. We are aware that our simulation system employs periodic boundary conditions (PBC), theoretically enabling the collision of clusters that would not meet in a system without PBC. But employing PBC is not only for technical reasons; there is also thermodynamic reasoning

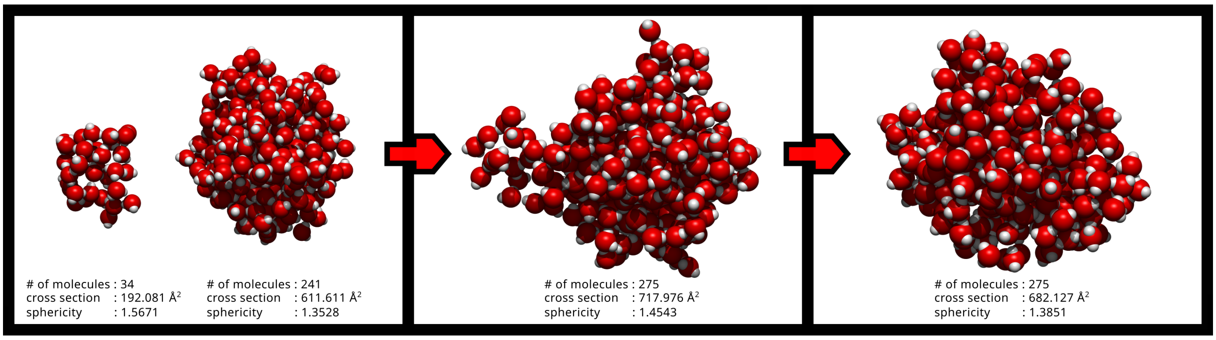


Figure 6.11: Clusters evolution during merging phenomenon with corresponding cluster properties. Snapshots were created with same orientation.

to approximate the behavior of the similarly composed surroundings. Under this argument, we can assume that a cluster crossing through PBC represents a foreign cluster entering the carved expanding cell. On the other hand, there is the factor of expansion to account for because the systems are continuously growing in volume, making the transition through the PBC less likely. It is of interest that the velocity at which the cluster moves is inversely proportional to its size. This means that the larger the cluster, the less likely it is that it will cross the PBC. To conclude this reasoning, there is a potential to point out the possible inconsistency of CNT through our simulations that are performed far from equilibrium. For certainty of this claim, further investigation of clusters crossing PBC is required.

6.6.3.2 Cluster bounce

Another phenomenon that is in conflict with the CNT assumption our observation of molecules bouncing from clusters. During the visualization of simulation configuration evolution, an occurrence suggesting bouncing was observed. Because the investigation of pickup cross-sections is performed in the group, tools to calculate and visualize collisions of single molecules and clusters were developed. The setup is designed by Martin Klíma and operates in the following manner: The cluster taken from the expansion simulation is fixed by its center of mass in the system and cooled to that prevent spontaneous evaporation. This cluster is then bombarded with monomers with varying initial positions and varying initial velocities. At the end, it is examined whether the molecules were picked up by the cluster or not.

During these experiments, there is a subset of rare events where a molecule is directly bouncing from the surface of the cluster, as illustrated in fig. 6.12. In the figure, an agglomeration of multiple frames was performed to show the video playback in one figure, resulting in a distortion of cluster molecules due to the thermal motion of the cluster. In fig. 6.12, we can see the reflection of the molecule path in the plane of observation as well as the decrease in velocity after the contact. The visualization uses the same time step difference between frames, which means the distance traveled by a molecule is indicative of its velocity. For this particular situation, the initial velocity was 17.07 Å/ps, which is 1707 m/s. For cases of colision of static probe molecules with clusters flying in the expanding jet, this is an achievable speed. But the collision inside the simulation box, with relative velocities equivalent to the temperature of the system, would require collisions to be observed for velocities below 1000 m/s.

In conclusion, the bouncing phenomenon remains unclear, requiring further investigation of lower velocity collisions. Cases observed so far correspond to collisions quickly followed by the evaporation of particles.

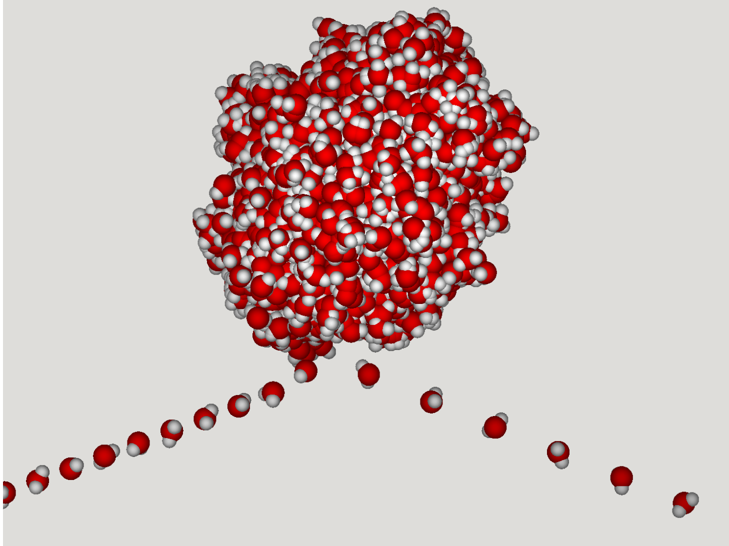


Figure 6.12: Visualization of simulation playback of molecule bouncing from the surface of the cluster. The visualization was performed with “trace” option of *show* command from MACSIMUS [126] utilities.

6.6.4 Evolution of temperature

To determine whether the clusters are frozen, the evolution of the temperature was performed as part of our ongoing research in work by Klíma *et al.* [123]. In fig. 6.13 the temperature distribution of clusters ($N_\mathrm{c} > 10$) during the simulation is shown in comparison with the system temperature. We can see that the average of the cluster temperatures is above the system, which is the consequence of slower evaporative cooling of the clusters. Given the used SPC/E model with a melting point of 215 K [237] we can now prove that most of the clusters (to account for fluctuating cluster temperatures) are not frozen during the simulation, even though the simulation temperature is below the melting line at the end of 12 ps.

Based on previous work by Klíma and Kolafa [124], the clusters are below freezing point (200–215 K), but not sufficiently to nucleate and freeze, as investigated by Bartell and Lennon [19] and later by Manka *et al.* [156]. According to research done by Angelil *et al.* [8] water freezing is very difficult to achieve in simulations (even with better models than SPC/E). To fully understand the aggregation of potentially frozen droplets is left to future study. In the current simulation setup, it is hard to account for freezing with one direct simulation, and more work on the model will be required first.

6.6.5 Comparison with experiments

Tables 6.2 and 6.3 collect characteristics of the simulated clusters for two respective initial conditions, and compare them with the experimentally obtained data (bold in the table caption). In all cases, three categories (main columns) are shown: the geometric mean size corresponding to the expected log-normal distribution of cluster sizes, the biggest, and the second biggest cluster. To answer the research question, three parameters are shown for each category: the number of molecules in cluster N, cross-section S in Å^2, and sphericity ψ. Please note that clusters smaller than 11 molecules are not counted, reflecting the choice in [142] enforced by the used experimental apparatus.

Table 6.2 contains simulation results for lower $T_0 = 409\,\mathrm{K}$ and $p_0 = 313.5\,\mathrm{kPa}$ resulting in a lower density of vapor. Therefore, more of smaller clusters are expected, which is reflected in the

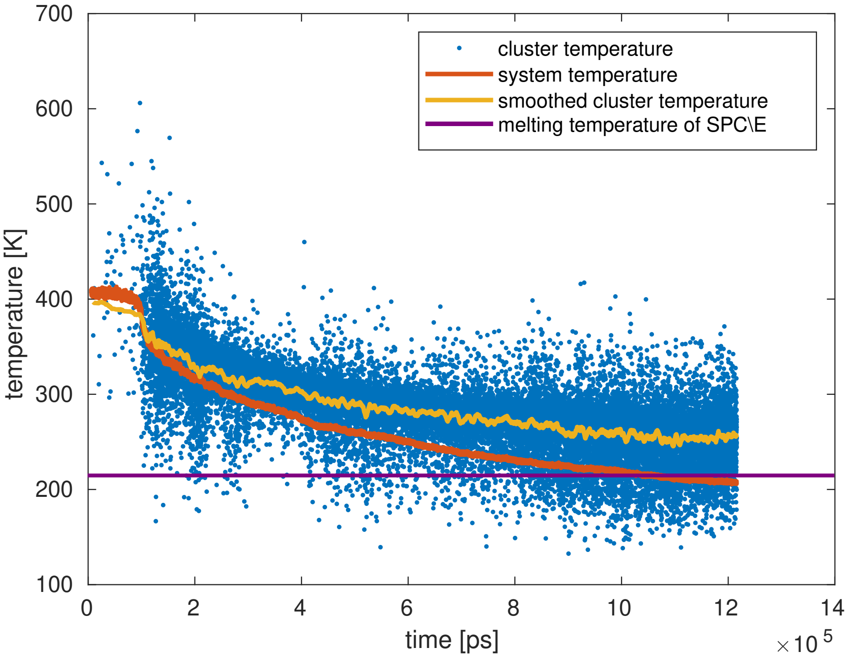


Figure 6.13: Evolution of temperature during the simulation of expansion of 10240 SPC/E molecules with initial temperature $T_0 = 409$ K. Temperatures of cluster with $N_c > 10$ is compared to averaged cluster temperature, system temperature with depicted melting line. Adjusted from [123]

size of the second biggest cluster. Because of a sharp disproportion in the case of $N = 1000$, larger systems were simulated, resulting in much better statistical coverage of cluster sizes. Additionally, more than one simulation with the same starting conditions were performed for smaller systems to gain better statistical coverage, especially for large ($N \geq 150$) clusters.

It is important to note that in the case of 4096 molecules, the largest simulated cluster is bigger than the experimental average cluster, and the second largest cluster is comparable to the first one. This indicates that we are likely close to the sufficient number of molecules needed to reliably describe the expansion. This was validated by simulation for 10240 molecules, where even the second biggest cluster is larger (in the number of molecules) than the one observed experimentally.

As can be seen in both table 6.2 in the geometric mean column and N of the biggest cluster column, the sphericity of the simulated clusters that achieved sizes above the experimental 337 span from 1.32 to 1.52. This is indicated by both the 4096 simulation and one case of the 2048 run. The biggest 10240 run is located in between with 1.44. This is significantly lower than the experimentally obtained sphericity of 2.70. A reader can refer to visualizations figs. 6.8a and 6.8b to see the difference of 0.4 in sphericity, which is approximately half of the difference discussed here. We can also see the trend for sphericity listed in the colums for the biggest and second biggest clusters. The trend observed is that the sphericity of the bigger cluster decreases with increasing size.

The notable discrepancy does not stop at sphericity but is already observable in the associated cross-section, which needed to be adjusted from the experiment to our observed geometric ones.

The results for higher initial density and temperature, $T_0 = 424\,\mathrm{K}$ and $p_0 = 474.7\,\mathrm{kPa}$, are summarized in table 6.3. Here, the higher initial density of vapor leads to significantly larger

	geometric mean			the biggest cluster			the second biggest cluster		
N_{part}	$\langle N\rangle$	$\langle S\rangle$ [Å^2]	$\langle\psi\rangle$	N	S [Å^2]	ψ	N	S [Å^2]	ψ
1000	33	183	1.50	186	513	1.35	18	123	1.53
	47	225	1.48	184	502	1.33	12	100	1.65
	$N \leq 10$	$N \leq 10$	$N \leq 10$	207	534	1.31	$N \leq 10$	$N \leq 10$	$N \leq 10$
	$N \leq 10$	$N \leq 10$	$N \leq 10$	191	501	1.29	$N \leq 10$	$N \leq 10$	$N \leq 10$
	$N \leq 10$	$N \leq 10$	$N \leq 10$	195	497	1.27	$N \leq 10$	$N \leq 10$	$N \leq 10$
2048	185	506	1.33	260	622	1.31	132	412	1.36
	43	212	1.49	350	764	1.32	17	121	1.56
	133	416	1.37	194	517	1.32	162	469	1.35
	86	332	1.45	198	526	1.33	192	524	1.35
4096	102	361	1.42	435	860	1.28	159	473	1.38
	43	221	1.54	448	870	1.27	219	585	1.38
10 240	69	284	1.44	544	979	1.26	416	853	1.31

Table 6.2: Simulation results performed with the nozzle opening angle $\alpha = 30°$ under system conditions: **$T_0 = 409$ K**, **$p_0 = 313.5$ kPa** simulated for 1 µs. Experimental mean cluster properties from [141,142] for these initial conditions are following: **$\langle N\rangle = 337$**, **$\langle S_{\text{geo}}\rangle = 476$ Å^2**, $\langle S_{\text{pick}}\rangle = 1530$ Å^2, **$\langle\psi\rangle = 2.70$**. The data below the experimental cluster size threshold are omitted.

clusters in smaller quantities. Although for the 4096 runs the biggest cluster still exceeds the mean experimental value of 587, the second biggest clusters are significantly smaller. In this case, we get a really narrow cluster size distribution. Nevertheless, the sphericity for large clusters again decreases in direction to unity, which is in disagreement with the experiment.

6.7 Conclusions

This part of the thesis presents molecular simulation software developed for the purpose of calculating nucleation phenomena on GPUs. This task presents its own set of challenges in areas like energy conservation and efficiency for long simulations of highly inhomogeneous systems. We have shown that our software is able to answer these challenges and opens a way to use GPUs even on tasks requiring double precision calculation units. This model and parallel solution fit into the area of simulation of heterogeneous molecular systems – tasks not available in known MD simulation packages.

The original motivation of this study was to elucidate the nature of the shapes of clusters produced in systems expanding through a nozzle. This kind of research has been previously hard to reach because of the computation demands posed by microsecond-scale simulation runs for sufficiently large systems that would capture the experimentally measured mean cluster sizes. We have utilized GPGPU techniques to calculate up to ten times larger systems than in our previous studies across two starting conditions. The simulated times exceeded ~ 1 µs. For statistical evaluation, several runs have been calculated on the GPUs available to our research team. Although the simulated cluster sizes are in moderate agreement with the experiment, the

	geometric mean			the biggest cluster			the second biggest cluster		
N_{part}	$\langle N\rangle$	$\langle S\rangle$ [Å^2]	$\langle\psi\rangle$	N	S [Å^2]	ψ	N	S [Å^2]	ψ
1000	$N\leq 10$	$N\leq 10$	$N\leq 10$	195	510	1.30	$N\leq 10$	$N\leq 10$	$N\leq 10$
	$N\leq 10$	$N\leq 10$	$N\leq 10$	196	512	1.30	$N\leq 10$	$N\leq 10$	$N\leq 10$
	$N\leq 10$	$N\leq 10$	$N\leq 10$	209	545	1.33	$N\leq 10$	$N\leq 10$	$N\leq 10$
	$N\leq 10$	$N\leq 10$	$N\leq 10$	210	535	1.30	$N\leq 10$	$N\leq 10$	$N\leq 10$
	$N\leq 10$	$N\leq 10$	$N\leq 10$	197	521	1.32	$N\leq 10$	$N\leq 10$	$N\leq 10$
2048	58	264	1.51	422	842	1.28	31	180	1.56
	32	185	1.57	423	859	1.31	14	117	1.72
	$N\leq 10$	$N\leq 10$	$N\leq 10$	439	863	1.28	$N\leq 10$	$N\leq 10$	$N\leq 10$
	$N\leq 10$	$N\leq 10$	$N\leq 10$	415	847	1.30	$N\leq 10$	$N\leq 10$	$N\leq 10$
4096	145	445	1.38	780	1231	1.24	27	161	1.53
	$N\leq 10$	$N\leq 10$	$N\leq 10$	803	126	1.25	$N\leq 10$	$N\leq 10$	$N\leq 10$

Table 6.3: Simulation results performed with the nozzle opening angle $\alpha = 30°$ from initial conditions: $\boldsymbol{T_0 = 424}$ **K**, $\boldsymbol{p_0 = 474.7}$ **kPa** simulated for 1 µs. Experimental mean cluster properties from [141, 142] for these initial conditions are following: $\boldsymbol{\langle N\rangle = 587}$, $\boldsymbol{\langle S_{\mathrm{geo}}\rangle = 689}$ **Å^2**, $\langle S_{\mathrm{pick}}\rangle = 3332$ Å^2. The data below the experimental cluster size threshold are omitted.

cross-sections are markedly smaller than the data reported by the experiment. The assumption that clusters are supercooled liquid droplets would require much longer time to cool enough by evaporation sufficient for the clusters to freeze. At this point, it is questionable whether increasing the system size further would bring different results and explain the systematic difference in the experimentally reported values for cross-section and sphericity. Possible ways to explain the suggested aggregation of icicles may include diffusion kinetics of (frozen) clusters at times longer than can be achieved in simulations, as well as performing simulations mimicking directly the pickup experiment with respect to cluster dipole moments. The investigation of pickup is continuing in the research group, with the publication [123] accepted for publication.

We conclude that the developed software opens possibilities for long simulations of large, highly heterogeneous systems. In our future work, we would like to extend the software capabilities and employ further optimization to provide a more pronounced speedup compared to already highly optimized CPU implementations.

7 Investigation of metastable system properties

Nucleation research has evolved over time, starting with insights from Fahrenheit [69], followed by contributions from Gibbs [82] into the kinetic theory by Volmer and Weber [249] that is still in use today. The knowledge related to nucleation finds applications across a wide range of fields, including energy production [25, 130, 261], nucleation of droplets in atmospheric simulation [96, 119, 268], chemical process engineering [207], and surface treatment [240].

Experimental investigation of nucleation faces challenges due to the short-lived, microscopic nature of the phenomenon. Nevertheless, various methods were developed to tackle this issue, as reported by Heist and He [97]. Rising experimental interest spread to water droplets studied by Viisanen *et al.* [242], carbon capture and storage (CCS) relevant mixtures shown in work by Vinš *et al.* [248] and Čenský *et al.* [46].

Molecular dynamics (MD) simulations, leveraging increased computational capabilities, have also emerged as a valuable tool to study the kinetics of nucleation; see, for instance, Neimark and Vishnyakov [174], Horsch and Vrabec [105] or Kalikmaninov [113].

Scientific interest in nucleation persists, driving further theoretical investigations. One specific area of focus is the extension of understanding homogeneous nucleation principles in fluids and the improvement of models used for heterogeneous nucleation.

Our research aims to comprehensively explore the metastable region of the phase diagram, including the liquid side, and develop a universal mechanism applicable to both vapor and liquid metastable states. While investigation of metastable states have received attention, there is still a lack of comprehensive measurements and simulations in this specific area. By bridging this gap, the goal is to enhance the understanding of nucleation and effects at the interface discussed in section 5.4 and to improve multiparameter equations of state [215].

7.1 Problem formulation

The core problem is based on the task of section 1.2.2 of obtaining data in the metastable region of the phase interface diagram. Investigating the metastable region experimentally is challenging at the microscopic level. In particular, the metastable liquid region is even more challenging due to the nature of droplets. The currently available data accumulated in both metastable vapor and metastable liquid, is insufficient for the needs of EoS improvement.

Conversely, the theoretical investigation using CNT to predict the characteristics of metastable regions faces several issues, as discussed in section 2.4.3. This makes the prediction unsuitable for the investigated task. Therefore, an intermediary between theory and experiment in the form of molecular simulation is chosen with the intention of providing pseudo-experiment data and employing the theory in a form that does not introduce the problems of CNT.

The problem transforms with the use of MD into the detection of phase change in the system, more specifically the detection of the onset of nucleation. This understanding is further modified by the request to solve this task with efficiency, allowing runtime execution during the simulation.

First, the problem challenges are presented, which are consequently formalized for the molecular dynamics approach to the solution presented in the next section section 7.2.

7.1.1 Problem challenges

Based on the theoretical elucidation of nucleation from section 2.3 and more specifically stability from section 2.3.1, several challenges concerning the investigation of metastable states are identified.

7.1.1.1 Lifetime of metastable state

The characteristic property of metastability is the lifetime of the state. This lifetime depends on the stochastic behavior of the studied or simulated system and is based on the likelihood of developing the critical cluster. While CNT can be used to formulate the prediction based on the thermodynamic condition of the system using eq. (2.59) with critical cluster size from eq. (2.63), this approach formulation is not precise enough to provide the solution to the problem.

One of the reasons that contributes to the challenge is that the onset of nucleation separating the metastable state does not depend on saturation as featured in the previous formulas (eqs. (2.59) and (2.63)) but also on system size N (as seen in the work of Yasuoka and Matsumoto in [266]). From a stochastic perspective, it is obvious that larger systems have a higher likelihood of generating the configuration necessary to cross the energy barrier. But this has significant consequences for molecular simulation, where system size is an important parameter linked to the predictive capability of the simulation.

7.1.1.2 Cluster and void

Thermodynamic systems under metastable conditions exhibit, given enough time, phase transition precursors. These precursors are called clusters for vapor→liquid[1] transition direction. For the common observer on a macroscopic scale, these are also called droplets. For the opposite liquid→vapor[1] transition, voids are formed instead, macroscopically observed as bubbles.

And while most of the theory in section 2.3 can be reused[2] for the bubble description, the physical characterization is particularly difficult to detect. This is illustrated in fig. 7.1 showing 2D representations of both phase precursors.

In the example metastable vapor molecular system (left in fig. 7.1), the problem consists of detecting the groups of molecules. This is immediately related to clustering, as described in chapter 4. For a metastable liquid system (right in fig. 7.1), a lack of molecules needs to be detected instead. But not complete emptiness, as the content of a bubble is composed of vapor. Considering further that molecular simulation tools represent molecules as a set of points without any volume, the task of detecting empty spaces is further complicated. This can be confirmed by the lesser number of works focusing on metastable liquid states.

[1] Throughout this chapter, the transition direction is used to denote the considered case, e.g., liquid→vapor. In some cases, the system starting conditions are used instead, i.e., a metastable liquid system (which transits from liquid into vapor).

[2] Apart from saturation, which needs to be defined separately.

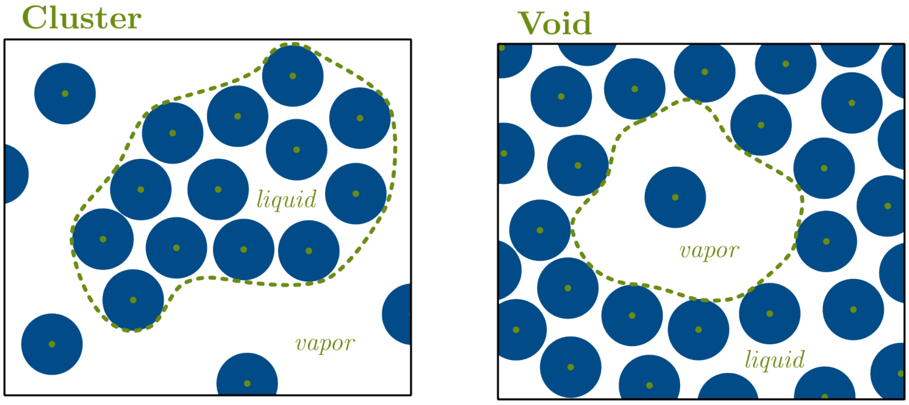


Figure 7.1: Visualization of cluster (left) and void (right) in 2D space with denoted imaginary boundaries of the structure separating the phases.

7.1.1.3 Obtaining the initial conditions

Given the time dependent nature of the metastable state, the initial conditions are of particular importance. There is a strong emphasis on generating initial conditions for the molecular simulation that do not introduce the phase interface precursors into the system while providing an accurate description of the metastable system. This dichotomy needs to be solved within the simulation protocol; otherwise, a systematic error is introduced into the calculated metastable state characteristics.

7.1.1.4 Runtime execution

The task of runtime execution is motivated by resource management for the investigation of the metastable states. Given the not precisely known lifetime of the metastable state, the performed simulation often suffers from state that is no longer valid, which ends up wasting the computational resources without providing the desired result. This can be simply stated as MD software lacking awareness of the change of state in the system.

To amend this, the solution method needs to avoid post–processing (which is the common case) and perform all necessary operations either before or during the simulation execution. This means the complexity of the presented solution has to be below the complexity of the force calculation[1].

The solution should also be incorporated into the parallel design of the software. It is also useful to present a solution with a low impact on the existing implementation.

7.1.2 Mathematical formulation

We first assume the spatial data are provided from the simulation in the form of a vector of the positions of individual atoms located in three dimensional space. Using the molecular structure, the atoms can be grouped into molecules. It can then be safely assumed that the simulation spatial data are also available as molecular positions $\boldsymbol{r}_i, i \in \widehat{N}$ with N as the number of molecules and relative positions of atoms according to the CM approach used in section 6.4.1.

Based on our experience with the phase interface described in chapter 5 and cluster shapes examined in chapter 6, the density was chosen as the main descriptor of the system instead

[1] Force calculation is generally the most expensive step during iteration, as shown in section 3.4.6

of the spatial data directly. This means that a three dimensional density function evolving in time, $\rho = \rho(x, y, z, t)$, is attributed to the particular simulation run. Because our intention in characterizing the nature of the system state is based on extrema examination, the density function needs to be sufficiently smooth.

We therefore propose the following construction for the density function using multivariate distribution from eq. (4.5) for the special case of three dimensional space. The density distributions are then placed at each molecule's center of mass, $\boldsymbol{\mu_i} = \boldsymbol{r}_i$. The spherical molecular structure is achieved with the diagonal covariance matrix, whose elements are based on the interaction Lennard-Jones potential as $\boldsymbol{\Sigma} = \sigma_{\mathrm{LJ}}/2 \cdot \boldsymbol{I}$. This approach gives the smooth density function representation of the spatial data at a single timestep t as a sum over molecules in the system

$$\tilde{\rho}(x, y, z, t) = \sum_{i}^{N} \phi(\boldsymbol{r}_i). \tag{7.1}$$

In this way, the evolution of the spatial data can be transformed into the evolution of the density function, which we are going to use for the characterization and detection of the onset of nucleation.

For this model operation, we define the metastable system condition based on knowledge of the critical cluster or void, which can be obtained without many of the assumptions required for the derivation of CNT.

Definition 7.1. A system with thermodynamic conditions in the metastable region remains metastable if the transition precursors remain below the critical size.

After the detection of the precursor with over critical size, we speak about the onset of the nucleation, and the system is no longer considered metastable. This also defines the lifetime of the metastable state until the first encounter of the overcritical transition precursor.

Condition of continuous simulation The definition 7.1 is constructed for the case of continuous simulation strictly adhering to the runtime execution. This is the case for the simulation software used in this work, where multiple independent simulations are executed in parallel. This means that the simulation cannot return to previous times. This leads to instances where the simulation is terminated before the optimal termination point is reached.

The reason is the stochastic nature of the cluster growth shown in fig. 2.5. Following the critical cluster definition, there is still fifty percent chance of the cluster shrinking. Later, when the maximal cluster shrinks below the critical size mark, the case of premature termination is encountered.

With the density formalism and adjusted definition of metastable state, the original problem of obtaining simulation data for systems in the metastable region is transformed into performing a molecular simulation until the critical precursor occurs in the simulated system. This provides a necessary generalization enveloping both types of precursors under one characteristic given by the density function $\tilde{\rho}$.

7.2 Solution of identification problem

To detect the precursors under the density description, the same transformation needs to be performed.

The first part of the solution is to characterize the detected objects in the density description from eq. (7.1). Using the same transformation, the spatial precursors from fig. 7.1 are converted into the shape shown in fig. 7.2.

Analyzing the density distributions of the precursors reveals characteristic features that can be used in the design of the detection method. These features are illustrated in fig. 7.2 using a solid red line to cut the density distribution with the associated density cross-section provided at the edge of the figure.

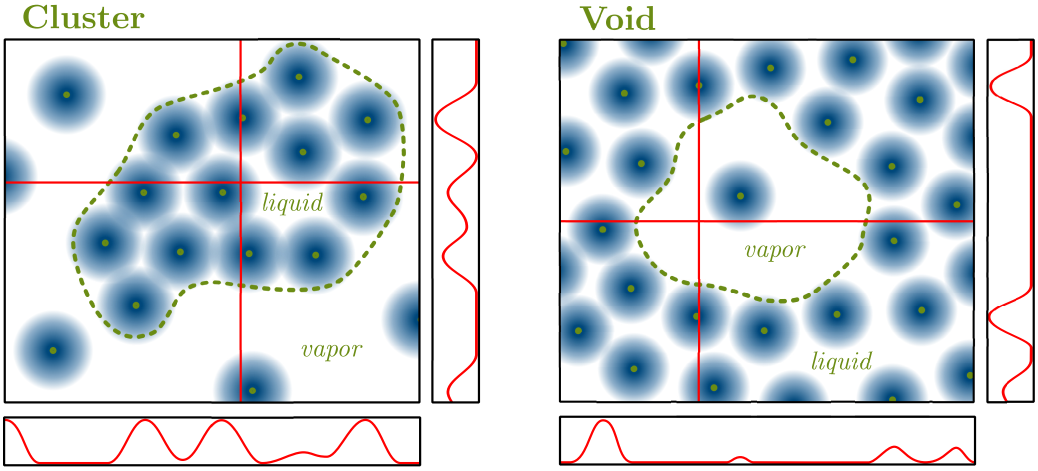


Figure 7.2: Visualization of the density representation of a system with a cluster (left) and a density representation of a system with a void (right) in 2D space. Cross-sections of density distribution in the x and y directions are shown on the sides of the figures, where the cutting planes are depicted as solid red lines.

In the given example, we can already see that the cluster can be represented as having an elevated density compared to the surrounding volume. A void can be represented as a depression in the density of the surrounding volume. This gives the foundation for the design of the runtime criteria.

7.2.1 Designing a criterion for runtime execution

The density approach has yielded a unified description of the transition precursors. But for the runtime execution, replacing each point with a density distribution and evaluating the density is a very costly operation.

To bypass the costly transformation of points into density with multivariate Gaussian, a less precise density evaluation based on a histogram is used. We can utilize the principle of grid based clustering from section 4.1.5 and construct a regular grid on which the density is evaluated.

7.2.1.1 Local density fluctuation

To encompass the similarities between both transitions in an efficient manner using the approximation of density on the grid, the term "local density fluctuation" (LDF), is defined next in detail.

The term LDF incorporates the concept of local system density approximation performed on regular cubic lattice tilling of the system volume (see fig. 7.3). For criteria operation, it is important that all cells are of the same volume, regardless of their position within the system volume. This ensures artifacts are not introduced into the density approximation at the edges.

By prescribing a reference density value for all cells, a universal relative comparison is established. This approach helps in identifying features of the density function, which are classified as density fluctuations. The primary observed characteristic is the presence of fluctuations, rather than the magnitude of the fluctuations.

In practical terms, this allows us to associate negative deviations from reference with a scarcity of molecules (voids) and positive deviations with an abundance of molecules (clusters). By categorizing both cases as fluctuations, a simplified method to capture both transition precursors is constructed. Where appropriate density reference is used to characterize the system density from which the fluctuations deviate.

In this construction, both voids and clusters are classified as LDF. Clusters are identified using an upper bound comparison, while voids are detected using a lower bound comparison. Two different reference values discussed in section 7.2.1.3 can be used to increase the precision of the detection for both criteria cases.

7.2.1.2 Designing the criteria

Following the definition of LDF, the universal grid criterion using three parameters a, b, c is constructed. The following paragraphs elaborate on the role of each parameter and provide theoretical bounds for their choice, allowing for the construction of criteria with different foci. The operation principle with the role of parameters a, b, c is shown in fig. 7.3.

In the text, several terms specific to the grid criteria are used. Consider first a regular cubic tiling of the simulation cell, which is further referred to simply as the grid. Vertices of the grid are called grid points, and cells are the individual volumes of the tiling; each cell is represented by a cube with grid points as its vertices. The term neighborhood of grid point refers to all cells that contain the said grid point as one of their vertices.

The grid is shown in fig. 7.3 with plus signs as grid points, and the neighborhood is visualized as a green square around the grid point with the number three[2].

Grid edge The first value, a, is the *grid edge* size. The choice of a regular cubic grid allows for computation complexity mitigation. Using regular cubic allows for local density evaluation with complexity $\mathcal{O}(n)$. This construction enables the run-time execution of the criterion.

The general restricting values for a are derived from the system size, which in the case of a regular cubic system volume V leads to

$$a \in \left(0, \sqrt[3]{V}\right]. \tag{7.2}$$

In practice, choosing $a = k\sigma_{\mathrm{LJ}},\ k \in \mathcal{N}$ leads to the simplification of consequent formulas as well as the quick transfer of the detection model from parametrization to parametrization as the grid becomes relative to the interaction potential of the simulated substance σ_{LJ}.

In cases where the grid does not align precisely with the system volume, periodic boundary conditions are used to construct the image of the grid cell to ensure all cells have the same volume.

Reference bound The second parameter b is denoted as *reference bound* and used for fluctuation classification. This parameter represents the reference value against which the grid point is classified. For performance purposes, the classification is designed in the simplest possible terms as integer comparison of the number of molecules found within the neighborhood of every grid point (see fig. 7.3).

For considered $\mathcal{R}^3$ space, the construction leads to each molecule being accounted for by eight grid points. This corresponds to averaging performed over the neighboring cells without

[2] The only grid point that has three molecules in the neighborhood

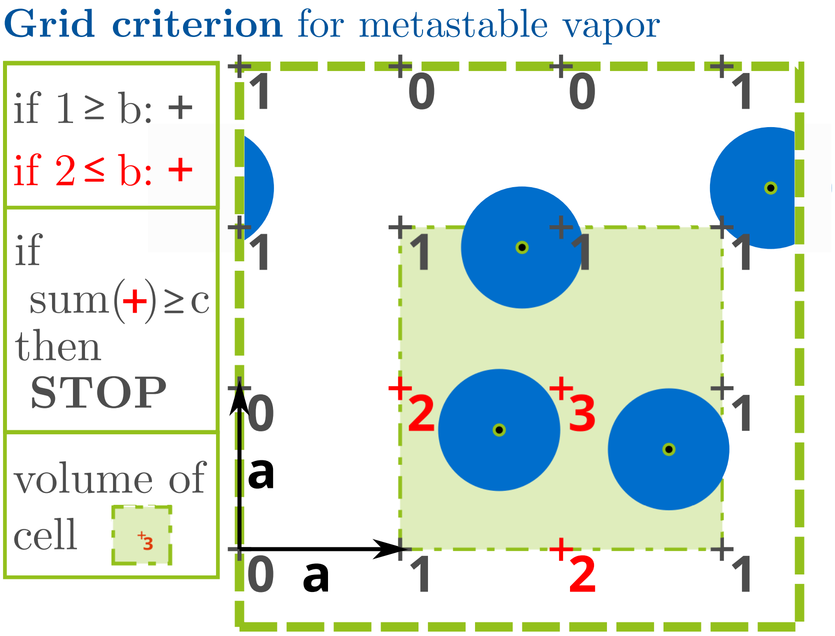


Figure 7.3: 2D visualization of the criteria design and the meaning of the parameters. Grid points are denoted as plus signs, accounting for the presence of surrounding atoms, shown as blue circles with black centers. The cell volume for the contribution is shown as a green–filled rectangle with a dash–dotted border. The system volume is denoted by dashed green lines with an illustration of the periodic boundary condition. The operation principles of parameters b and c adjusted for this illustration are explained in the legend.

the division by the number of neighbors. In consequence of the applied averaging, the detection abilities are enhanced for cases of nonspherical entities as well as entities not aligned with the detection grid[3].

The parameter b is physically constrained by the empty cell as its minimum and the fully occupied cell as its maximum. While the minimal occupancy is straightforward, for the maximal occupancy estimate, the $n^{\max}_{\text{mol,gp}}(a, r)$ function is required.

$$b(r) \in [0, n^{\max}_{\text{mol,gp}}(a, r)] \tag{7.3}$$

The function $n^{\max}_{\text{mol,gp}}(a, r)$ depends on the parameter a and the radius of the molecule r to return an integer number of the maximum number of spherical molecules with radius r that can surround a gridpoint within a grid with edge a. Converting the gridpoint quantity to cell quantity considering the replication, we get $n^{\max}_{\text{mol,gp}} = 8\, n^{\max}_{\text{mol,cell}}$.

The $n^{\max}_{\text{mol,cell}}$ can be calculated with close packing ϕ. Close packing is a fraction of the unit volume that can be filled with the same spheres in a space-efficient way. The closest arrangement is an interesting problem, which is not yet fully explored [227], but for lattice based arranging structures, the value of the closest packing is known [93]. The best lattice packing is attained by face centered cubic (FCC) lattice, which is also used later in this work (for visualization, see fig. 7.7). The coefficient of closest pacing for FCC is given by

$$\phi_{\text{FCC}} = \frac{\sqrt{2}\pi}{6} \approx 0.740. \tag{7.4}$$

[3] imagine cluster with its center at the gridpoint divided into eight cells. Evaluating the gridpoint with the use of the surroundings ensures this situation is detected.

Utilizing the closest packing, here estimate the value for $n^{\max}_{\text{mol,cell}}(a,r)$. Radius of hard spheres is based on the simulated substance as $r = \sigma_{\text{LJ}}/2$. Together with eq. (7.4) an estimate

$$n^{\max}_{\text{mol,cell}}(a,r) = \frac{V_{\text{cell}}\phi_{\max}}{V_{\text{sphere}}} = \frac{a^3\frac{\pi\sqrt{2}}{6}}{\frac{4\pi r^3}{3}} \overset{r=0.5\sigma_{\text{LJ}}}{\leq} \left\lceil \frac{a^3\sqrt{2}}{8\sigma^3_{\text{LJ}}} \right\rceil . \tag{7.5}$$

In this equation, σ_{LJ} represents the underlying substance interaction potential. In this work, the substance is chosen as LJF (see section 3.2.3 for parametrization), and σ_{LJ} for the sphere is then identical to the potential of LJF. For more substances like oxygen or nitrogen, their molecular interaction potential is used instead[4].

We therefore know the limiting values for the bound parameter and can proceed with the description of the comparison. In the core of the grid criterion procedure, the value of the bound parameter b is compared against all grid points. Grid points that do not fulfill the comparison condition are marked (see red grid points in fig. 7.3). The comparison condition is differentiated (reflecting the LDF) for the cases of cluster and void detection as follows:

$$\text{cluster detection} \Rightarrow n_{\text{mol,i}}(a,r) \geq b \tag{7.6}$$

$$\text{void detection} \Rightarrow n_{\text{mol,i}}(a,r) \leq b, \tag{7.7}$$

where $n_{\text{mol,gp,i}}(a,r)$ is number of molecules surrounding gridpoint i.

Critical count Until now, the parameters were directly connected to the atomistic properties. For system-wide awareness, the last value c is constructed to operate on the whole system volume, making a decision based on the number of marked grid points. For simple, robust, and time-independent values, c is constructed as a level-set parameter. The parameter c called the *critical count* and represents the number of density fluctuations permitted to simultaneously exist within the system. Exceeding the value of c the system is classified as no longer metastable. The proper choice for c is crucial for criteria operation, as c can also be described as the strictness of the criteria.

There is an obvious limitation on the value of c for a system with a finite number of molecules. This directly relates to the number of completely packed cells with molecules. In this thought scenario, the value of c related to the number of grid points cannot exceed the number of grid points that the number of molecules N can fill. To again consider the replication in 3D by multiplying the number of observed molecules across all grid points, the upper bound is obtained as

$$c \in \left[0, \frac{8N}{n^{\max}_{\text{mol,gp}}(a,r)}\right] = \left[0, \frac{N}{n^{\max}_{\text{mol,cell}}(a,r)}\right]. \tag{7.8}$$

The condition with c is checked with the prescribed frequency during the simulation. When the number of fluctuations within the system exceeds parameter c, further simulation is terminated and simulation results are collected.

7.2.1.3 Parametrization for detection of critical precursor

The broad scope of the criterion allows for the investigation of multiple phenomena related to the local density[5]. In this section, the parametrization for the metastable region investigation is provided with respect to the detection of the precursor in both metastable vapor and liquid regions. Please note that alternative parametrization are also useful as shown in section 7.4.1.

[4] Molecular interaction potential in this context represents the σ_{LJ} when the molecule is modeled as spherical instead of structure of atoms.

[5] Even spinodal decomposition can be contemplated in terms of LDF.

Because of the general differences in behavior within the metastable region, it is not possible to prescribe one universal set of parameters a, b, c and the effects of temperature need to be accounted for. Therefore, a temperature dependent relationship for the parameters is designed to offer more robust detection.

Observations from chapter 2 are employed as descriptors of the phenomenon, allowing to bypass the need for prior investigation using simulation. In the case where prior simulation results are available, methods such as extrapolation or parameter fitting can be used. Nonetheless, the theoretical investigation provides a reasonable first estimate, which is the desired use-case when the fluid metastable conditions are investigated.

Grid edge In the case of critical precursor detection, the first step is to incorporate the physical dimension of the precursor. In both cases, the geometry of the target is assumed to be spherical in nature, which agrees with our observation of supersonic expansion for large clusters fig. 6.8a.

With the design operating in a neighborhood, the detection or critical precursor can be directly related to setting the grid distance equal to the critical radius:

$$a_{\mathrm{v}} = r^*_{\mathrm{v}} = \frac{2v_1\sigma}{k_{\mathrm{B}}T\ln(p/p_{\mathrm{eq}})} \tag{7.9}$$

$$a_{\mathrm{l}} = r^*_{\mathrm{void}} = \frac{2\sigma}{p^*_{\mathrm{l}} - p}. \tag{7.10}$$

where cluster critical radius r^*_{v} is obtained from eqs. (2.43) and (2.64) and for droplet similar relation is constructed using difference of pressures instead of the chemical potential as shown in work of Kaschiev [116]. In that formulation, the pressure inside the void is $p^*_{\mathrm{void}} = \exp(-v_1(p_{\mathrm{eq}} - p)/k_{\mathrm{B}}T)$. This provides two equations dependent on temperature and pressure with surface tension σ as parameter. All these properties can be calculated from the EoS using temperature and density as inputs[6].

Evaluation bound The bound parameter b is set to reflect the available volume of the cell determined by a. The first estimate is obtained from the number of molecules in the critical precursor, using the eq. (2.63). Following the work of Kashchiev [116], a similar equation for void is obtained

$$n^*_{\mathrm{l}} = \frac{32\pi p_{\mathrm{eq}}\sigma^3}{3k_{\mathrm{B}}T(p^*_{\mathrm{l}} - p)} \tag{7.11}$$

The direct relation of b with the critical cluster number proves to be insufficient because it does not consider the distribution of molecules within the cell. The model case is the spherical precursor, while the rest of the cubic cell is filled with molecules with the density of the system. Adjusting for the surrounding molecules yields:

$$b_v = r^*_{\mathrm{v}} + \Delta V\rho_{\mathrm{l}} \tag{7.12}$$

$$b_l = r^*_{\mathrm{l}} + \Delta V\rho_{\mathrm{v}}, \tag{7.13}$$

Where $\Delta V = (8a)^3 - (4/3)\pi a^3$ is the available volume outside of the precursor (radius of the precursor reflects the choice of $a = r^*$). In the case of metastable vapor, where the reference bulk density is low, this issue is not well pronounced, and the use of correction is optional. Conversely, the metastable liquid criterion's effectiveness is greatly diminished without the correction.

[6] These values are know because they constitute the simulation inputs for the NVE ensemble

Critical count The detection of the critical precursor is primarily captured by the parametrization of a and b. To adhere to our definition of metastable state from definition 7.1, the critical count becomes straightforward.

$$c = 1, \tag{7.14}$$

Note that this is only valid because of the previous construction and the wish to detect a single such entity. For other discretizations, the parameter c is necessary and therefore cannot be removed.

This gives a simple set of equations describing the parametrization of the system under metastable conditions.

It is important to understand that for the described model, the case with a single precursor surrounded by the system of reference density and the case with multiple smaller precursors located within the same neighborhood are equivalent. The criterion in the current implementation makes the decision based on the number of molecules present in the neighborhood. This technical feature offers less computational complexity and is more optimal for the runtime. From thermodynamics and simulations, the argument is that situations with multiple precursors in close proximity quickly combine into one larger precursor, meaning only a minor difference in onset detection is made.

7.3 Solution Method

In this section, the grid criteria are incorporated into the method proposed for the solution of the original task of producing data in the metastable region. The intended thermodynamic property of interest is the pressure of the metastable state obtained for temperature and density sampling. The procedure consists of four steps: region construction, sampling, simulation using the cluster criteria just described, and final analysis. Specific information about the algorithmic side of the method is in appendix B.1, and visualization of the structure is shown in fig. B.1. The solution method is described using LJF as the substance of choice, utilizing the dimensionless scaling of properties outlined in appendix A.1.

7.3.1 Phase region construction

With density and temperature being the input state variables, an EoS is used to calculate pressure for comparison and to support the calculation. This choice is further reflected within the construction of the phase diagram, as illustrated in fig. 7.4. An EoS is used for the evaluation of the individual isotherms as well as phase region boundary curves described in section 2.3.1. From theoretical knowledge, the binodal and spinodal lines need to be constructed first for the estimation of phase regions. During this process, one needs to be aware that errors following from an EoS influence the calculation of spinodal and sometimes even the binodal lines. Binodal curve is calculated with the TREND 4.0 [216] thermophysical properties package. In the case of a spinodal, the construction is performed as described in section 2.3.1. In this manner, a spinodal curve can be constructed even for equations with both single and multiple vdW loops. Full region separation (like the one shown in fig. 2.3) is then completed with the critical isotherm.

It is important to note that this construction is based on EoS predictive capability, and the resulting regions are subject to artifacts of the EoS. But this is required to use the equation to be improved as we can have direct control over the sampling and can saturate areas where the equation is not performing well.

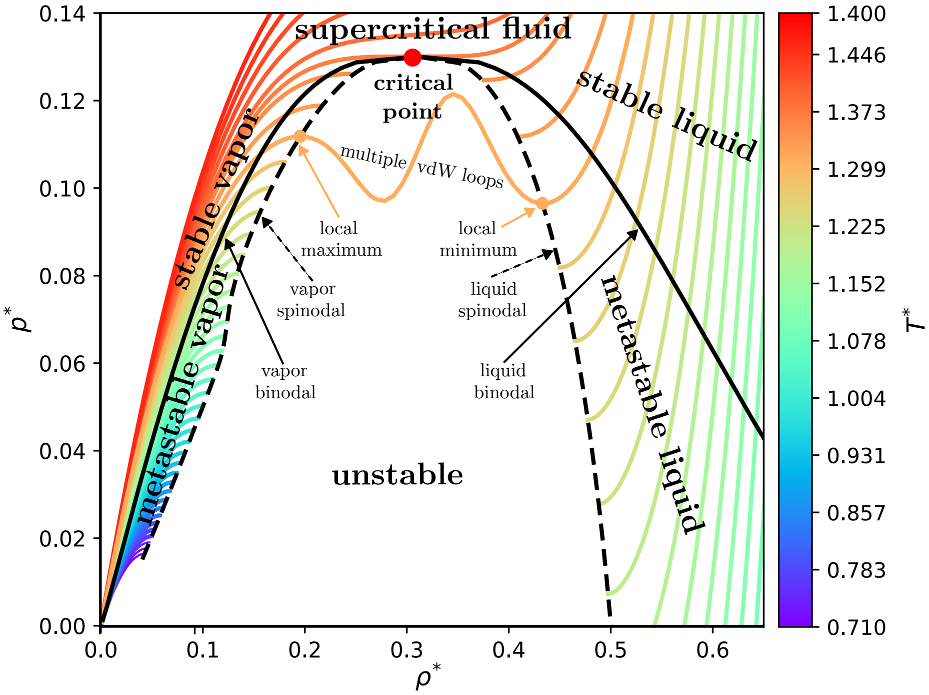


Figure 7.4: Phase diagram of the Lennard-Jones fluid (EoS of Thol *et al.* [229]) in dimensionless variables: pressure over density for varying temperatures. Regions are separated by solid black binodal and dashed black spinodal curves. The red circle shows the critical point. Isotherms are rendered in color corresponding to their temperature while the isotherm $T^* = 1.275$ is extended into unstable region to visualize multiple vdW loops and points used for spinodal calculation.

7.3.2 Sampling of metastable regions

During the construction of the regions, the isotherms were evaluated for a given temperature and density spacing, representing the isotherm as a series of linear segments. In practical terms, the curves are best sampled at the calculated points. But this method of sampling is greatly restricted in cases with coarser refinement, in regions where the isotherm presents fewer points, and in regions where discretization in density does not align well with the intended sampling. More issues appear for the intersections of two curves, i.e., border points of regions. Therefore, all the curves are interpolated and optionally recalculated when higher precision is required.

To obtain samples in the metastable region, further division of the isotherm into stability regions is needed (i.e., the metastable liquid region of the isotherm). The sampling itself is performed on the line segments corresponding to the selected stability region. A simple subdivision of the line segment produces the required number of samples, with points at the boundary optionally excluded.

This is shown with the example sampling of metastable regions in fig. 7.5. For the selected isotherm (thicker lines with temperatures indicated in the color bar), the metastable region is sampled with three points in the metastable vapor region, omitting the point at the binodal. And similarly, with four samples per isotherm, again omitting the point at the binodal.

There are several potential modifications that can be applied to enhance the sampling process, particularly when dealing with narrow stability regions. One useful modification involves sample reduction. Since sampling is conducted with a fixed quantity. This approach helps conserve computational resources by reducing the number of samples generated in close proximity to each other.

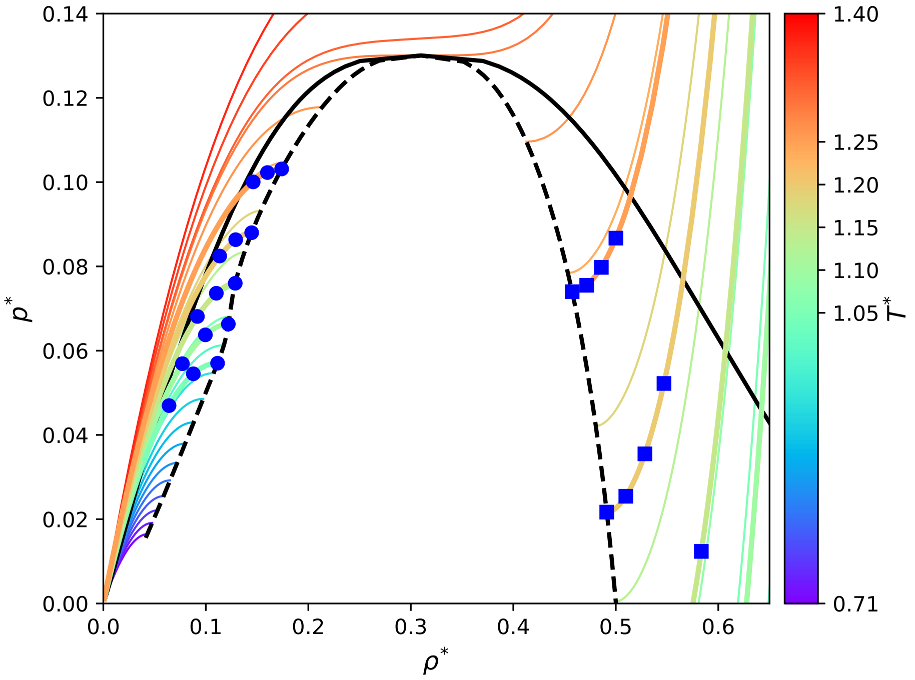


Figure 7.5: Phase diagram of the Lennard-Jones fluid calculated with the EoS of Thol *et al.* [229]. Selected thicker isotherms are separated into metastable regions. Blue symbols depict the sampled points in the metastable regions (circles for vapor and squares for liquid).

7.3.3 Simulation of sampled state points

Given the nature of nucleation, the simulation simulation can be systematically split into two periods separated by the onset of nucleation [116, 118, 239]. Detecting the onset of nucleation allows us to simulate the system only within metastable settings, without the effects introduced by the nucleation taking place.

The simulation is performed in the NVT ensemble, where the input properties are obtained from the samples generated in section 7.3.2. Temperature T^* during the simulation is controlled with the Nose-Hoover thermostat from section 3.4.5.1, and volume is fixed in correspondence to the input density ρ^* and number of simulated molecules N.

The criterion is built into the simulation software as described in section 7.2.1.2 and controls the simulation termination based on the parametrization from section 7.2.1.3. When the simulation is completely terminated, the simulation data are collected and exported into data files used for data analysis and comparison with the literature.

7.4 Properties the of model

In this section, the properties of the designed model and its utilization in the solution method are discussed. The topics of the discussion include: the choice of custom designed parametrization in precursor detection, the effect of equilibration for the simulation of metastable states, and the effect of system size for metastable simulation. These discussions address aspects that are sometimes overlooked in the investigation of metastable states but have substantial implications when disregarded.

7.4.1 Statistical Analysis of Physical Properties for Phase Detection

One of the concerns a user may have is whether parameter c of general criterion is required for nucleation onset detection. In this section, arguments are presented with two different parametrization showing c in comparison with pressure and potential energy that could be considered for the task.

For a typical molecular dynamics simulation of the system, the physical properties can be understood as random variable observations at each time step (see chapter 3). The average value and running average of the physical properties are used as the results of the simulation. While the quality of the simulation data is determined using the standard deviation and running standard deviation. For the comparison presented in this section, however, the units of the properties are different. Therefore, the absolute value of the coefficient of variation (CV)[7] evaluated for different *property* is defined as

$$CV_{\text{property}} = \left| \frac{\text{stdev}(property)}{\text{mean}(property)} \right| . \tag{7.15}$$

Simulations were performed for Lennard-Jones fluid under metastable vapor ($T^* = 1.0$, $\rho^* = 0.075$) and liquid ($T^* = 1.0$, $\rho^* = 0.06$) conditions with varying system sizes N. Table 7.1 presents the coefficients of variation for the pressure CV_{p^*}, the potential energy $\text{CV}_{\text{E}_{\text{pot}}}$, and the c parameter CV_{c}. Another parametrization utilizing a fine grid $a = 1.0$ with bound related to spinodal density[8] $b = \rho^*_{\text{spinodal}}$ was used to display the fluctuation of parameter c. This parametrization offers a close approximation of the local density of the system but is less capable of detecting the precursors than the one shown insection 7.2.1.3.

From table 7.1, it can be observed that the coefficient of variance for the potential energy is the least fluctuating. Simultaneously, the coefficient of c for the fine grid of $a = 1.0$ shows the highest values. However, selecting the criterion purely based on fluctuation is not preferred. Consider, for example, the potential energy, which is directly influenced by the system size. Moreover, bounds estimation from theory is not readily available, requiring prior simulation data. Even the potential energy per molecule is therefore considered here as unfeasible for efficient execution.

The pressure is a better candidate because it remains similar for varying system sizes [9]. However, an issue arises from the required manipulation of the property by the criterion. This poses a problem for the property that is simultaneously used as a quantity of interest. In this way, the method is reduced to an iterative predictor corrector schema, which is not very suitable as a runtime criterion.

The custom-made parameter c, is therefore constructed on an artificial and non-interacting grid to bypass these issues. It is also independent of the calculated pressure, and the theoretical estimate can be based only on the critical cluster properties. This set of features enables single runtime simulation without the prior need for simulated data.

Advantage of critical precursor parametrization The design of the criteria parametrization described in section 7.2.1.3 results in detection action in the form of the step function (pointed with the arrow in fig. 7.6). This is the reason why a finer grid was used for showing how c and, in consequence, the underlying local density fluctuate.

[7] we are using words for standard deviation and mean to not confuse with already used symbols

[8] The value of ρ^*_{spinodal} is obtained during the phase region construction from fig. 7.4.

[9] This is valid only when a certain size threshold is exceeded, as shown in the following section 7.4.3

system description			coefficient of variation		
	ρ^*	N	$\mathrm{CV}_{\mathrm{p}^*}$	$\mathrm{CV}_{\mathrm{E_{pot}}}$	CV_{c}
metastable liquid	0.06	500	0.3251	0.0067	0.5636
metastable liquid	0.06	864	0.2521	0.0054	0.4304
metastable liquid	0.06	1372	0.2064	0.0049	0.3934
metastable liquid	0.06	2048	0.1755	0.0048	0.3854
metastable liquid	0.06	4000	0.1305	0.0037	0.2896
metastable vapor	0.075	500	0.0874	0.0640	1.2569
metastable vapor	0.075	864	0.0686	0.0540	0.9395
metastable vapor	0.075	1372	0.0547	0.0454	0.8366
metastable vapor	0.075	2048	0.0444	0.0373	0.6708
metastable vapor	0.075	4000	0.0321	0.0299	0.5517

Table 7.1: Comparison of the coefficient of variation for pressure, potential energy, and the parameter c for the LJF system at the metastable vapor ($T^* = 1.0, \rho^* = 0.075$) and liquid ($T^* = 1.0, \rho^* = 0.6$) conditions. N denotes number of molecules used in the simulation.

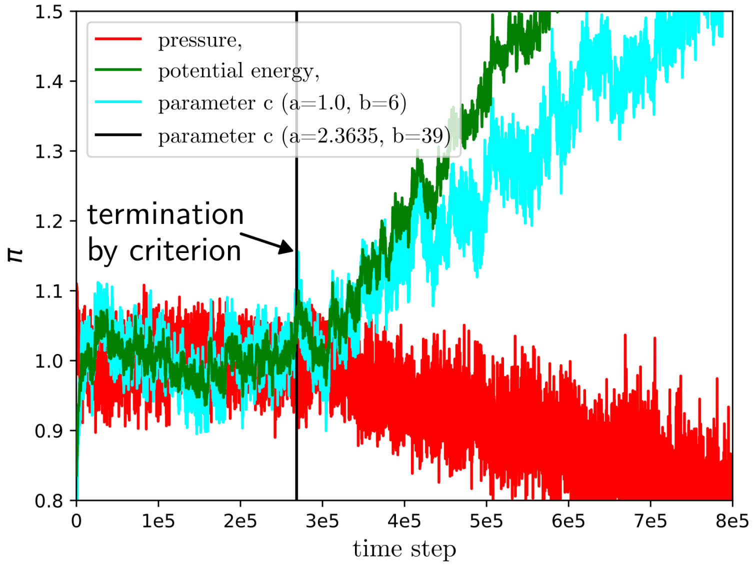


Figure 7.6: Comparison of the scaled pressure $\pi(p^*)$, the potential energy $\pi(E_{\mathrm{pot}})$, and two parameters c with different parameterizations $\pi(c_{1.0,6})$, $\pi(c_{2.3635,39})$ for metastable vapor ($T^* = 1.0, \rho^* = 0.075$) and a system size of $N = 4000$.

Figure 7.6 illustrates the full evolution of the mentioned properties with added termination performed by the metastable vapor criterion (using parametrization from section 7.2.1.3). For the comparison of different properties in the figure, the *property* values were adjusted to their

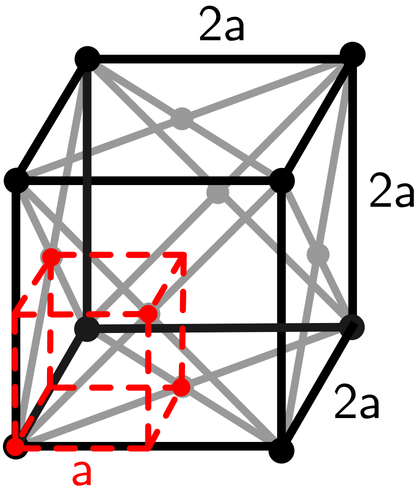


Figure 7.7: Face centered cubic initial placement showing the unit cell of the crystalline structure. Vertices of unit cell are black points and face centers are grey points. Sub cell in red dashed lines showing the reproduction of the patter with the alternating pattern.

respective time step interval: $\Delta t_0 = [0.0,\ 0.26 \cdot 10^6]$ with the equation

$$\pi(property) = \frac{property}{\text{mean}(property(\Delta t_0))}. \tag{7.16}$$

Here *property* represents pressure, potential energy and critical count c.

In fig. 7.6 we see the gradual decline of pressure after the point marked by the metastable vapor criterion, which signifies the nucleation is taking place. The peak in the potential energy and finer parametrization is pointed by the arrow. The termination of the simulation is set at the first peak pointed at by the arrow, as signified by the vertical black line. We can see that the proposed parametrization provides a significant advantage over the other properties in terms of calculation and practical use in the simulation.

7.4.2 Initial simulation stage

All MD simulations performed in this study have two stages: equilibration and production. During the equilibration stage, the system transits from the initially prescribed configuration into a relaxed state, where the desired parameters can be reliably observed without the effects of the initial configuration.

7.4.2.1 Initial configuration generation

In this study, the well-established face centered cubic initial configuration(FCC) is used, which follows the convention of *ms*2 [57, 73, 203]. The placement is performed in an alternating fashion, as illustrated in fig. 7.7. This generated initial configurations of the system aligned to sizes following $N = 4k^3$, where k is the number of red placement cells corresponding to FCC placement shown in fig. 7.7. The initial configuration structure resembles a crystal and needs to be dissolved during the equilibration period.

We further discuss the effect of the initial placement. The method described produces an identical crystal structure, which may be of an issue, when variation in the initial position is desired. To achieve this in our simulation, the generated structure is randomly perturbed up to a maximum distance without molecule contacts. For simple mono-atomic substances within the FCC, this distance is $\sqrt{3}a/2$ where a is the edge of the FCC cell. This removes any doubt that two different runs would produce the same results, diminishing the obtained statistics.

The correlation time and consequently the minimal equilibration time can, therefore, be estimated from this calculated distance with the mean velocity of particles in the system.

7.4.2.2 Equilibration

The equilibration phase of simulation is, in essence, not different from regular simulation, apart from omitting the production outputs. Therefore, it is possible for a nucleation onset to occur during this period or, more likely, for nucleation precursors to be created during this stage. Preexisting of the precursors then influences the simulation, effectively shortening the time for which the metastable system can be sampled (in extreme cases, preventing the sampling as the system nucleates during equilibration).

Performing no equilibration is also not an option, as best shown by the example of a low density system, i.e., $\rho^* = 0.06$. In this case, the initial behavior of the system is that of an ideal gas, where molecules are so far apart ($> 7.5\sigma_{\mathrm{LJ}}$) that no interactions takes place. Therefore, equilibration cannot be omitted, and the free flight period has to be accounted for as $d_{\mathrm{NN}}/v_{\mathrm{mean}}(T)$ with a temperature dependent mean velocity $v_{\mathrm{mean}}(T)$. The nearest neighbor distance d_{NN} in the initial FCC grid with edge a_{FCC} is equal to:

$$d_{\mathrm{NN}} = \frac{\sqrt{3}}{2} a_{\mathrm{FCC}} \tag{7.17}$$

$$a_{\mathrm{FCC}} = \sqrt[3]{\frac{N}{\rho}} \frac{1}{2k-1} \tag{7.18}$$

$$k = \sqrt[3]{\frac{N}{4}}, \tag{7.19}$$

where k is the number of full FCC cells in one dimension.

With this information, equilibration can be performed past the free flight stage, but for further simulation, there is an increasing chance that a nucleation precursor can be created. This technique is known as “quenching”, as explained in the work by Martínez and Müller [158]. The fundamental principle of this technique involves swiftly transitioning the system from one temperature to another, with minimal time for temperature adjustment during the transition.

Quenching in this study is performed from the equilibration run at temperatures above the binodal to the metastable state temperature used for the production run. The temperature drop is performed in a single step, removing the introduction of precursors during temperature adjustment[10]. The system volume is unchanged, and because of the thermostat, the velocities are quickly adjusted to the new temperature. In this way, the initial configuration influence on the production results is significantly diminished.

In situations where temperature quenching is not desired, i.e., in near-critical regions, one has to closely follow the designed minimal equilibration time to prevent free flight from occurring in the production run. In this case, perturbation of the initial grid may be required to introduce a more varied configuration for the production run.

7.4.3 System size

The system size is an important simulation parameter. For metastable simulations, it plays an even more significant role because of its influence on the speed of the nucleation onset (see

[10] Which is, for example, the case for the examined case of supersonic expansion.

section 7.1.1.1. On this note, it can be further reasoned that a system with $N = 500$ molecules cannot exhibit nucleation when $N^* >= 500$. It is less obvious what happens when the N is slightly below N^*.

In both situations, nucleation is restricted by the effect of system size emulated by the increased evaporation/condensation as a mean to equalize precursor with coexisting mother phase. It is questionable that a metastable system, where nucleation is like this, is a good representation. The reason being the difference in kinetic behavior as investigated by Horsch *et al.* [104]. Therefore, the usual reasoning for the system size needs to account for not only the computational power and thermophysical property representation but also whether nucleation is even possible in a given system.

To illustrate the argument about how well the system represents metastable conditions, a set of system sizes $N = 500, 864, 1372, 2048, 4000$ was compared with the EoS prediction. The temperature was set to $T^* = 1.0$ and two densities roughly in the middle of their respective metastable regions were selected (vapor: $\rho^* = 0.075$ and liquid: $\rho^* = 0.66$).

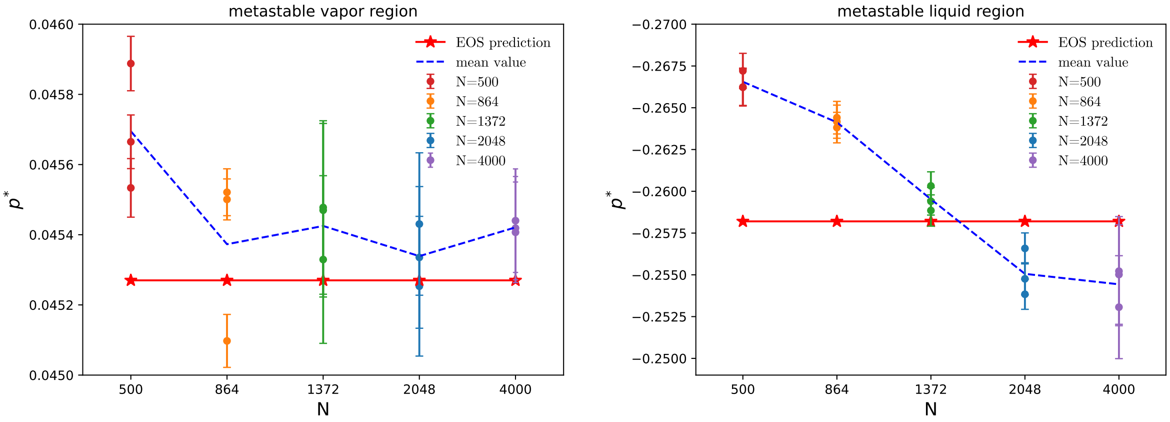


Figure 7.8: Comparison of pressures from simulations of LJF under metastable vapor ($T^* = 1.0, \rho^* = 0.075$) and liquid ($T^* = 1.0, \rho^* = 0.6$) conditions with varying system sizes $N = 500, 864, 1372, 2048, 4000$.

A comparison of the simulation results is shown in fig. 7.8. Using the criteria, the simulations were stopped before the nucleation events, and no phase separation or transition was observed during the analysis of the local density distribution evolution. In consequence, not all simulations have equal runtimes. Associated errors of the simulation were calculated as a standard deviation from pressures collected during the simulation rather than the time averaging natively used by *ms*2.

Within the simulation experiments, it is observed that small systems ($N < 1372$ molecules) seem to be insufficient for predicting the pressure because of issues of self-consistency and increased deviation.

See the simulation results in the metastable vapor region in fig. 7.8 left. For $N = 500, 864$ we notice an increased spread of values exceeding the error estimation. This suggests that the initial configuration either influences the results (which we resolved) or that the system pressure is fluctuating with higher magnitudes, producing inconsistent results.

The second reason is better shown in the metastable liquid case in fig. 7.8 right. Where an obvious convergence trend (dashed line) dependent on the size of the system is present. For smaller systems ($N = 500, 864$), the significant overprediction of the EoS prediction [229] changes into a slight underprediction around $N = 2048$. This illustrated example was observed throughout the metastable liquid region, leading us to the decision to use $N = 2048$ as the primary system size in the metastable liquid region. The current working assumption for this phenomenon is related to the inability of the small system to represent the behavior of the targeted liquid under metastable conditions.

To finalize our statement, we acknowledge that nucleation onset has occurred in any simulation shown, and the simulated time exceeded 10^6 time steps for the simulations with $N = 500, 864, 1372, 2048$ and $0.5 \cdot 10^5$ time steps for $N = 4000$.

In conclusion, the commonly held advice that metastable simulations are better performed with small systems[11] is not entirely beneficial. A smaller system also introduces biases into the predictions. Moreover, the dynamics of the system is changed, which may no longer represent metastable behavior. Small systems can function as some sort of daemon incentivizing the evaporation of larger clusters due to the unavailability of molecules in surrounding vapor [104]. Our approach advocates using a bigger system with criteria and replacing longer runs with multiple runs of varying initial configurations.

7.5 Results and Discussion

In this section, a summary of calculated results is provided as well as a comparison with simulation results from other studies. In the latter parts of this work, our observations from simulations are discussed in relation to the multiparameter equation of state for LJF by Thol *et al.* [229].

7.5.1 Criterion result

With the method from section 7.3 and the tools for sampling metastable regions, the simulation settings are generated and a simulation is run. The visual illustration of the process from the region construction in fig. 7.4 continued by sampling in fig. 7.5 is now finished in the result stage. In fig. 7.9, the sampled points from the EoS are replaced by pressures obtained from our simulations.

All data shown in the following figs. 7.9 and 7.10 are the simulation results of this work using the criterion where only the input points were used from different literature. This is marked in the legend as “from * ref”. Points samples from Baidakov *et al.* are shown as blue x, points samples from Linhart *et al.* are orange circles, and additional points samples from this work are green stars. Our sampling mechanism is described in section 7.3.2. This choice of data points was motivated by deviation graphs shown later in section 7.5.2. The error bars shown for the data points correspond to the duration of the data collection, where a smaller deviation means a longer simulation. The isotherms, binodal and spinodal, are calculated with the EoS from Thol *et al.* [229].

The overall figure presents the data from both metastable vapor and metastable liquid regions. For visualization reasons, the middle portion $p^* \in [0.22, 0.38]$ of the diagram showing primarily the empty unstable region as well as the negative pressure range were removed in favor of metastable regions.

The simulated sampling of all points in fig. 7.9 shows the Thol *et al.* EoS [229] extrapolates reasonably well into the metastable region.

Conversely, there are areas where the equation could benefit from the simulated data. Notably, the spinodal curve prediction could be adjusted. We can notice a bend in the vapor section of the spinodal around $T^* = 1.1$. This is likely a consequence of the different slope of the isotherm being predicted by the EoS. The simulation results also suggest that the metastable vapor region extends further right. Metastable liquid simulation suggests that the first local minima are located higher than those predicted by the equation. This is illustrated in more detail in fig. 7.10.

[11] The reason is the longer simulation runs, which mean more data can be collected

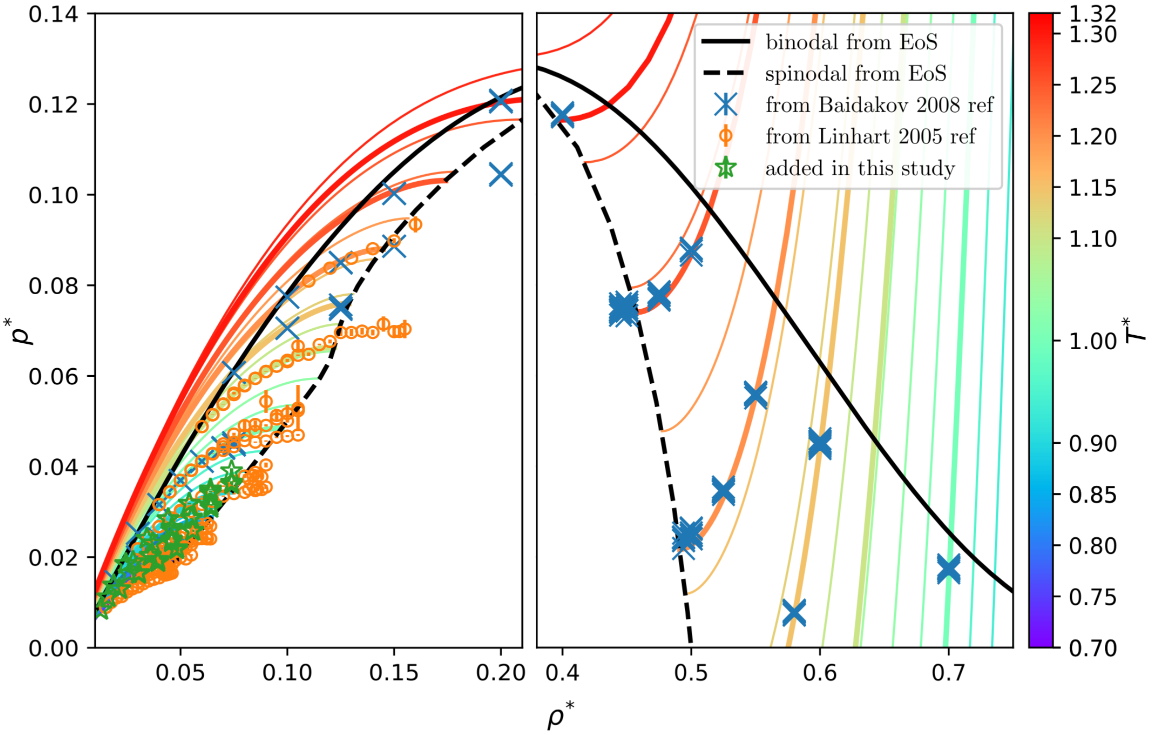


Figure 7.9: p^*, ρ^* diagram of the Lennard-Jones fluid, including different isotherms. Isotherm samplings calculated from the references are denoted on the temperature color bar and accented with a thicker line.

For the metastable liquid region, a similar but less pronounced bend in the spinodal at $T^* = 1.1$ can be observed in fig. 7.10. For lower temperatures, $T^* = [0.7, 1.1)$, this leads to an increased difference between the predicted isotherms and simulated points. With the added data set in a lower temperature range, another effect becomes visible. A slight top-right shift of the first local minimum of metastable liquid isotherms is seen at lower temperatures. The simulation data corresponding to the minima of the isotherms show density and pressure higher than the ones predicted by the EoS. No nucleation event took place to influence the simulation.

In the metastable liquid region, an argument for shifting the spinodal righ can also be made. This is based on the observed decrease in the simulation length for sample points close to the spinodal calculated with the EoS (as could be seen for points $T^* = 0.7, 0.75, \rho^* = [0.65, 0.68)$.

The original equation by Thol *et al.* [229] already provides precise results in the stable region that extrapolate reasonably well into the metastable region for higher temperatures. But because no metastable data was used for the equation development, the predictive capabilities are diminished with the spinodal prediction and the related shape of isotherms. This illustrates author's point about the benefits that metastable data can bring.

7.5.2 Comparison with the equation of state and literature

One of the tasks of this work is to provide data in the metastable region. This has led to the creation of an efficient runtime method that can achieve generality of detection over both metastable regions. Therefore, for the presented comparison with literature data, we will focus on two works that reflect these features.

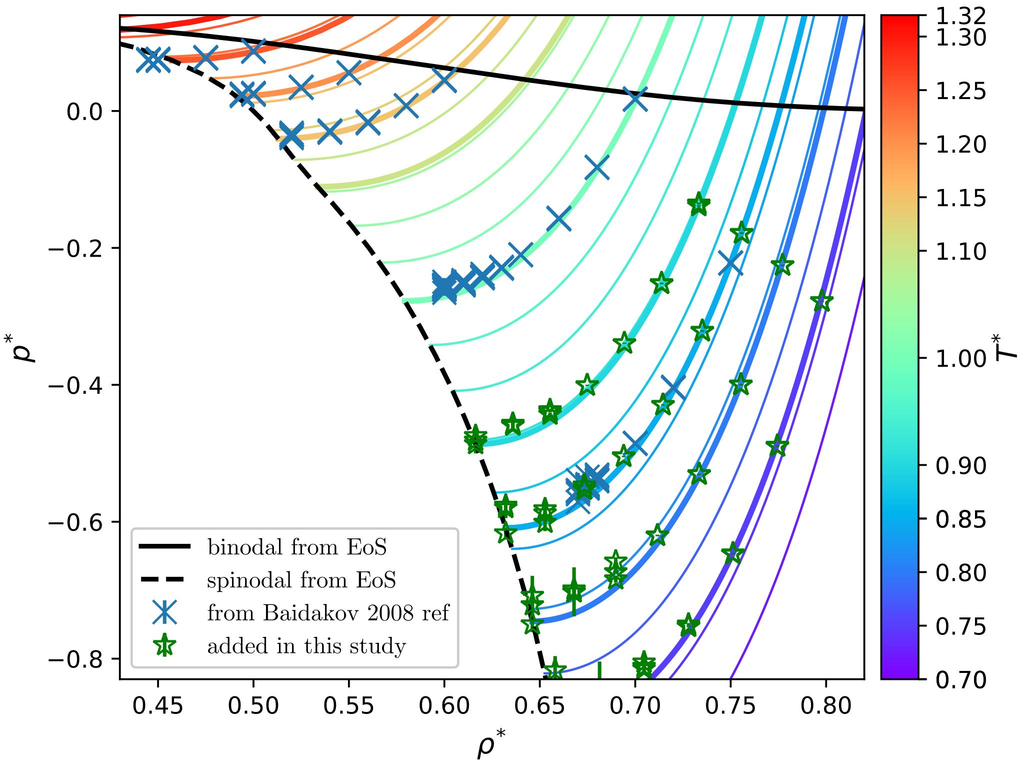


Figure 7.10: Section of the p^*, ρ^* diagram of the Lennard-Jones fluid pressure over density focused on the metastable liquid region.

7.5.2.1 Comparison with work of Linhart

The first is the work of Linhart *et al.* [148] utilizing the older version 1.0 of the same simulation software package *ms*2. They focused on the metastable vapor region, and the role of the criterion was not overly emphasized. We have selected their data here to illustrate the positive effect that more advanced criteria can bring and the effect of the equilibration as discussed in section 7.4.2. In the set of graphs in fig. 7.11, relative pressure deviations from the EoS [229] are compared with simulation data from the work of Linhart *et al.* and this study. In the case of this work, two sets of data were produced, denoted as “support” and “production”.

For the available temperature range, observable differences in trends and significant deviations from the equation predictions can be seen. The support data set does not use quenching equilibration and generally aligns closer to the data of Linhart *et al.* [148] (same system size of $N = 2048$ were used).

In contrast, the production data set from this study utilizes quenching, and a consistent deviation from the EoS prediction is observed when the spinodal is approached. The only alignment of the production data is seen for $T^* = 1.2$.

But this does not mean that it is possible to judge whether the simulation or the equation is right. Instead, we can draw the conclusion that equilibration (use of quenching) has significant consequences for the rest of the simulation. The second observation is that simulations agree with the EoS in the initial near-binodal area of the metastable regions.

For a better understanding of the differences shown in the deviation plots, the simulation results are visualized in coordinates ρ^*, p^* in fig. 7.12, where the trends are observed better. In this figure, we notice the reason for the higher deviation, as the production simulation set shows

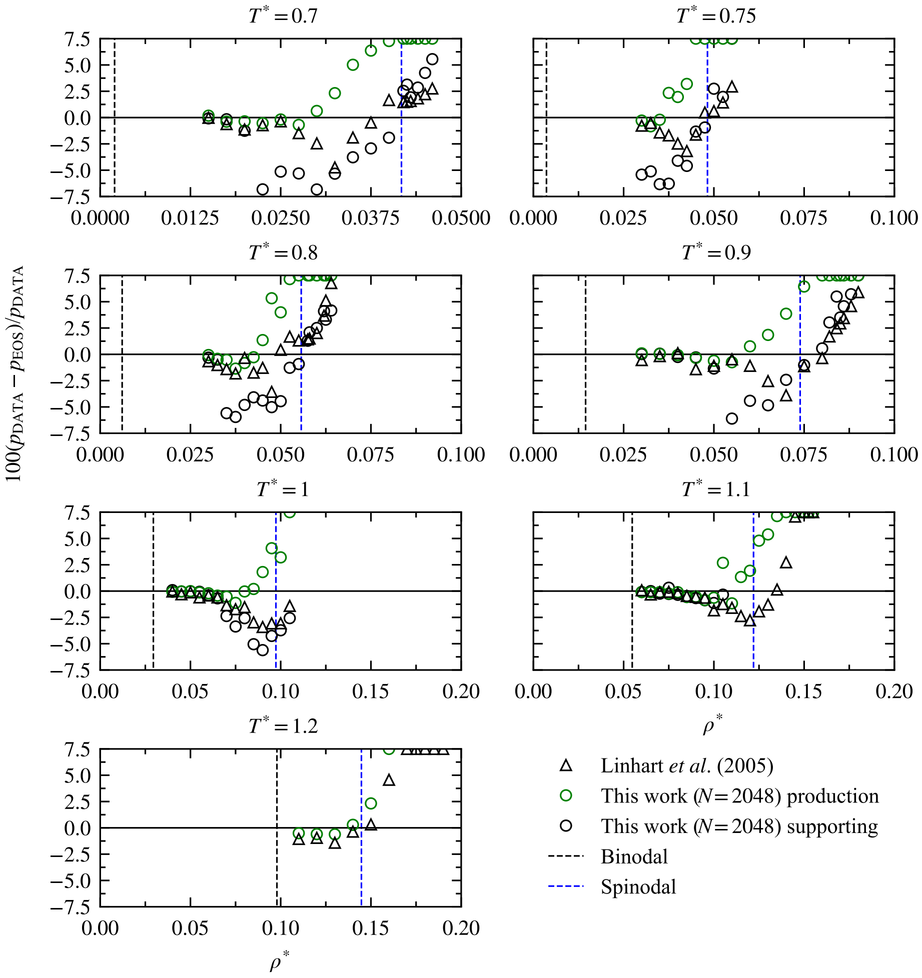


Figure 7.11: Comparison of simulated pressure data of Linhart *et al.* [148] with $N = 2048$ and the two simulation data sets produced in this work with $N = 2048$ using the cluster criterion in the metastable vapor region. Reference pressures, binodals, and spinodals, were calculated from the EoS of Thol *et al.* [229].

a higher peak of the isotherm than predicted by the EoS.

Furthermore, a suspicious flattening is observed for the data of Linhart *et al.* [148], especially for the isotherms $T^* = 1.15$. From the data of Linhart *et al.* [148] we can notice a consistent underprediction of the peak of the isotherm versus the EoS, which in some cases leads to a flattened isotherm. Conversely, the production simulation data set reveals local maxima that exceed the predictions of the EoS. This observation could be interpreted as the simulation indicating that the isotherms are more steeply inclined, suggesting that the spinodal should be situated farther than what is currently predicted by Thol *et al.* EoS.

Another argument for the extension of the metastable region is visible from simulated results crossing the predicted spinodal without significant loss in simulation duration (in the graph shown by the error bars). The smaller the error bar, the longer the production simulation duration.

7.5.2.2 Comparison with work of Baidakov

A second comparison was made with the data of Baidakov *et al.* [13]. In their work, a sequence of simulations with a much more controlled mechanism based on observation of system potential

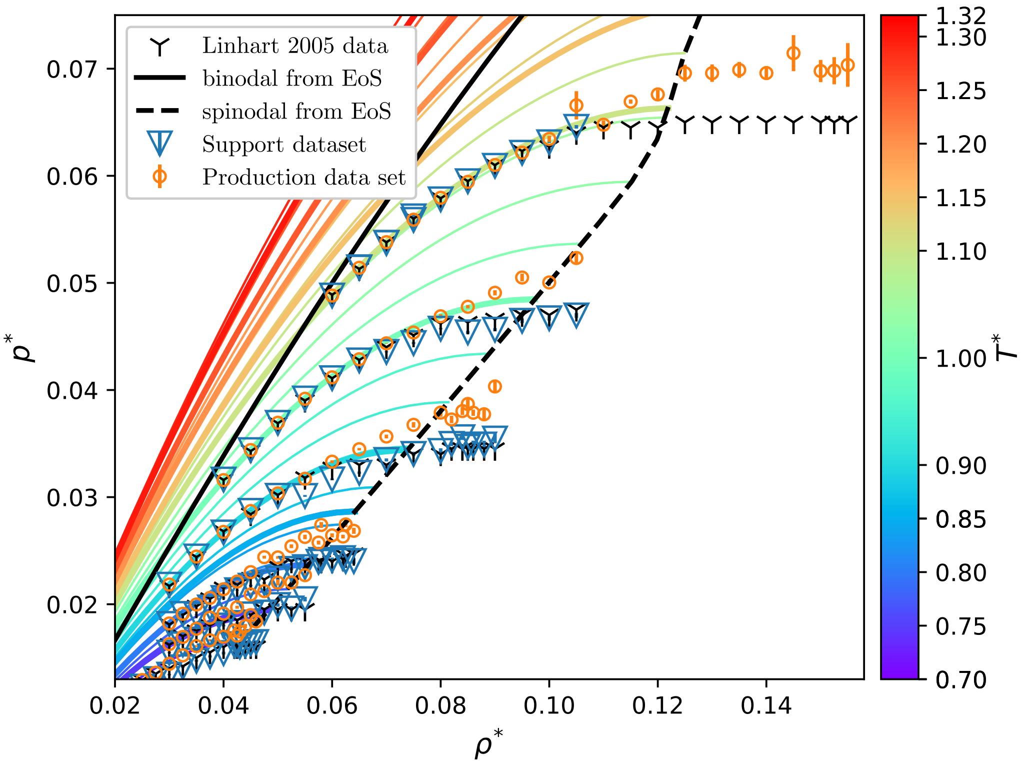


Figure 7.12: Comparison of the simulation pressure prediction in the metastable vapor region. The data of Linhart [148] are shown as black markers, binodal (-) and spinodal (- -) were calculated from the EoS of Thol *et al.* [229]. This work support data set is shown as blue triangles and the production data set as orange circles. The data calculated in this study have error bars corresponding to the length of the simulation.

is utilized to simulate the metastable regions. Badiakov and coauthors perform a sequence of simulations employing a correction scheme to navigate the isotherm from the binodal.

We present first the deviations graph in fig. 7.13 showing the data of Baidakov *et al.* [13] together with the data from this work in relation to the reference provided by EoS from Thol *et al.* [229]. In this work, we included simulations of three different system sizes $N = 1372, 2048, 4000$. The data of Baidakov *et al.* include both metastable regions, thus providing a more general outlook on both metastable regions.

Here we see a good agreement between the simulation datasets of Baidakov *et al.* [13] and this study achieved over both metastable regions. This supports the argument that comparable precision can be achieved with the presented method.

We can further notice the effect of the system size, particularly for systems closer to the spinodal, and the temperature around $T^* = 1.15$. From this comparison, the system size around $N = 2048$ aligns best with the data of Baidakov *et al.* who utilized the systems of same sizes.

Focusing on the comparison of all simulation data to the equation, we notice that the agreement with the EoS is not ideal, which can be attributed to a lack of metastable data during the LJF EoS development. We can identify that the simulation in the vapor region agrees with the equation in the near binodal region. On the side of metastable liquid, this is not the case.

In fig. 7.13 the right set of graphs shows a shift in deviation between temperatures $T^* = 1.0$ and $T^* = 1.15$. While the data are grouped together, there is a five percent relative difference

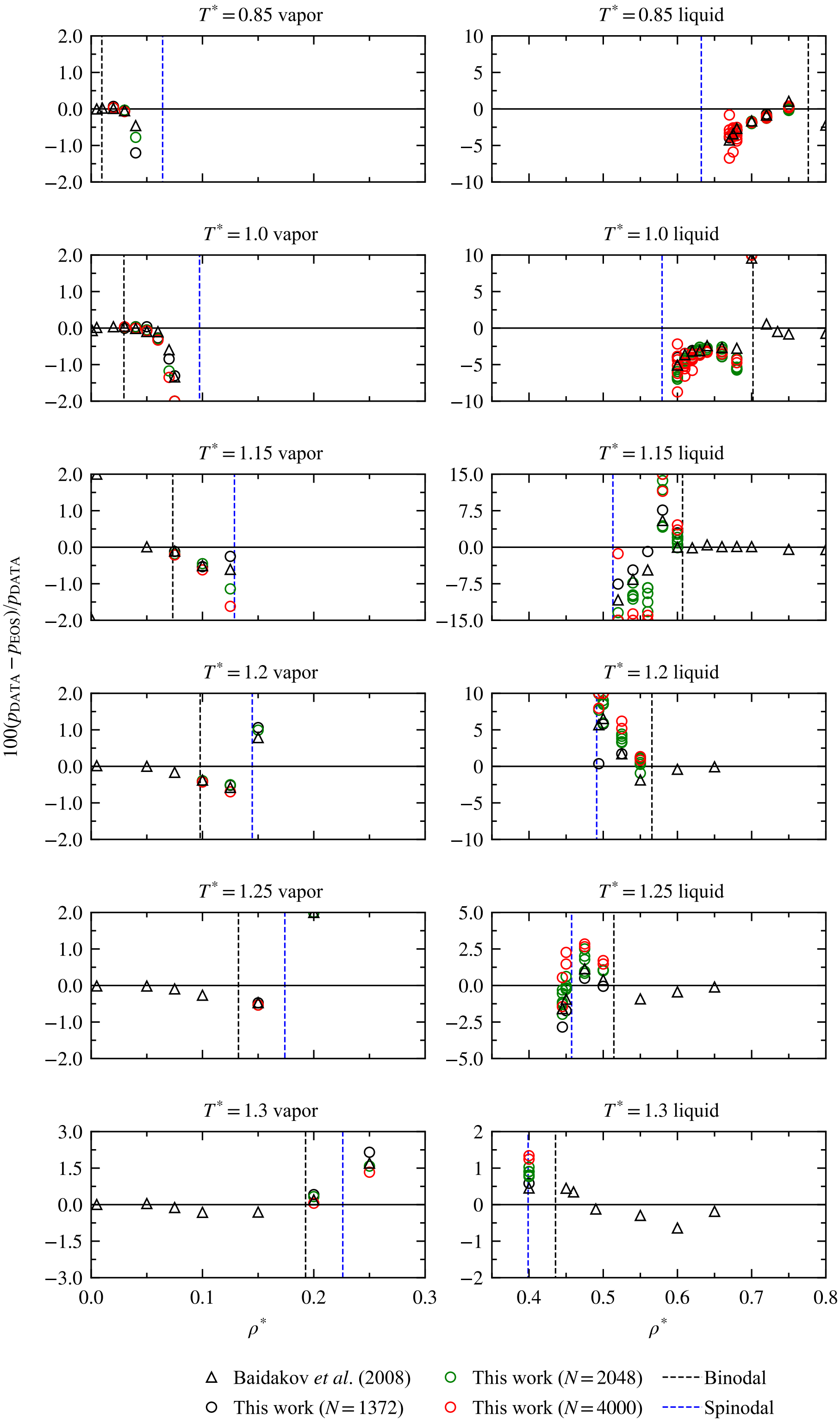


Figure 7.13: Comparison of Baidakov *et al.* [13] with $N = 2048$ and simulations in this study with $N = 1372, 2048, 4000$ using the cluster criterion in metastable vapor and liquid regions for selected isotherms. The LJF Eos of Thol *et al.* [229] have been used to calculate the reference pressures, binodals, and spinodals.

in pressures collectively predicted by simulation from the pressures predicted by the EoS for $T^* = 1.0$. Moreover, a suspicious artifact is identified at the binodal $T^* = 1.0, \rho^* = 0.7$ where an unexpectedly high deviation from the EoS for all simulated and reference points is observed.

Further investigation will be required to find the cause of this artifact. But our assumption is related to the multiple vdW loops that manifest already in the metatable area. If this is true, then this particular issue may be corrected with the data provided by this method. With the comparison with the data of Baidakov *et al.* [13], we have shown that the method can reach similar accuracy over both metastable regions. But as an added benefit, the method does not rely on any iterative procedure with post-processing of the system potential but executes along the simulation. This creates a powerful tool for exploring metastable regions. In consequence, metastable region data can be supplied to new generations of equations of state and research on nucleation.

7.6 Conclusions

This part of the dissertation has focused on the metastable state condition. The main aim was to provide a fast runtime method for system properties of metastable vapor and, more importantly, liquid conditions. The goal has been realized with the development of a new universal grid based cluster and void detection criterion. In this way, it is possible to detect the onset of nucleation during the running molecular dynamics simulation and sample only the metastable conditions of the system.

The presented method comprises metastable region identification, sampling, and the criteria implemented into the molecular dynamics package *ms*2 [73]. The method was also connected with the thermodynamic package TREND [216] providing access to the EoS for thermophysical property calculation. With the theoretical properties of critical size and critical number of molecules, the method operates in both metastable regions without the need for prior simulation data to classify the clusters and voids.

The proposed method was tested for internal consistency and compared with literature data and the EoS in the case of the Lennard Jones fluid [229]. The presented method performs better than alternative simple criteria approach and attains a similar level of accuracy as the more complicated and time-consuming successive iteration simulation scheme. Furthermore, with its simple, generalized design, there is a great potential for further extension into the class of simple fluids. Investigation of other phenomena within the metastable region, like spinodal decomposition, may also be possible in the future.

This makes the method a strong contender for fast exploration of vapor and liquid metastable regions using molecular dynamics simulations. The results of this research are of interest for research of nucleation and the development of new equations of state. A few suggestions have already been proposed for the tested equation by Thol *et al.* [229].

Concluding remarks 8

The theoretical foundations of nucleation, molecular simulation, and clustering motivated us to study three related topics: phase interface research, supersonic expansion simulation, and the investigation of properties of metastable vapor and liquid system. We have developed or extended the models necessary for the problem solution. Procedures and utilities facilitating the solution were implemented as part of the solution. The focus was given to the underlying challenges related to the development of optimal parallel algorithms for GPUs as well as the development of highly efficient method that can be integrated into molecular simulation to detect phase transitions. The proposed methods have been verified and their efficiency measured in comparison to existing software. The collected results were compared with available literature data including more general models, molecular simulations, and experiments. The conclusions pertaining to the individual topics are provided at the end of the respective chapters section 5.5, section 6.7, and section 7.6. The general conclusion highlighting the key areas of the research performed in the field of nucleation is presented next.

Nucleation is an important phenomenon present in everyday life, and as such, nucleation which plays an important role in many natural processes a is utilized in a broad spectrum of industrial applications. It is a rich field that a single thesis cannot hope to map completely; we have therefore focused on three interconnected topics related to nucleation. Of particular interest was the use of molecular simulation to uncover the characteristics of nucleation. During the simulation of expansion, we have observed the evolution of clusters and characterized their shape. We have provided insight into the experimentally observed phenomenon of highly nonspherical clusters. The time period before and during the initial phases of the nucleation were investigated, utilizing highly efficient methods specifically designed for the purpose. We have identified what constitutes the onset of nucleation and designed a criterion to consistently detect it. We have also provided suggestions for improvement of the Lennard-Jones fluid equation of state based on the data we simulated in the metastable region. We have consistently fulfilled the tasks that were given.

During the research, cooperation between research groups was fostered by connecting teams from the Institute of Thermomechanics, the University of Chemistry and Technology, Ruhr Universität Bochum, Technische Universität Berlin, and Technische Universität Dresden. Author's work and contributions has been published, generating ongoing impact in the scientific community. With the work summarized in this thesis, innovation was brought into the fields of molecular simulation, equations of state, nucleation and phase interfaces.

Appendices

Nucleation from the equilibrium view A

A.1 Dimensionless scaling used for Lennard-Jones EoS

In the investigated case of the Lennard-Jones fluid, it is common to use dimensionless scaling for calculations and visualizations. The scaling is based on interaction potential parameters σ_{LJ} and ϵ_{LJ} as follows:

$$T^* = \frac{k_{\mathrm{B}}T}{\epsilon_{\mathrm{LJ}}}, \tag{A.1}$$

$$\rho^* = \rho\sigma_{\mathrm{LJ}}^3, \tag{A.2}$$

$$p^* = \frac{p\sigma_{\mathrm{LJ}}^3}{\epsilon_{\mathrm{LJ}}} \tag{A.3}$$

A.2 Dimensionless scaling used for phase interface

For purposes of the artificial variable X a different scaling is required to transform the property into dimensionless form. The derivation of the scaling is based on eq. (5.14). The dimensionless variables are also denoted with an asterisk. There is no need for confusion, as only the artificial variable X is present in the main body and the rest is utilized internally for the calculation.

In this case, temperature is not scaled, i.e. $T^* = T$.The partial molar densities are scaled with the critical densities of individual components.

$$\rho_i^* = \frac{\rho_i}{\rho_{i,\mathrm{crit}}}. \tag{A.4}$$

Since the temperature is not scaled, the work of formation can be scaled as follows

$$\Delta\Omega^* = \frac{\Delta\Omega}{\mathrm{k_B}T}. \tag{A.5}$$

A general scaling is used for lengths, where a general coordinate s represents either an axial or a radial coordinate depending on the computed interface geometry

$$s^* = \frac{s}{\mathrm{L_s}}. \tag{A.6}$$

In eq. (A.6), $\mathrm{L_s}$ marks a typical magnitude of the investigated phase interface. In our calculations, it is set to a constant arbitrary value of $\mathrm{L_s} = 1\mathrm{nm}$.

In order to re-scale the algebraic eq. (5.14), the scaling for the *artificial* variable X and the influence parameter $c_{i,j}$ is needed. The definition of the work of formation given by eqs. (2.88) and (2.89) can be used in this case. From these equations, one can see that scaling down the work of formation $\Delta\Omega$ leads to the scaling of the influence parameter. The density derivatives in eqs. (2.88) and (2.89) can be scaled in the following manner

$$\left(\frac{\partial \rho_i}{\partial s}\right)^* = \frac{\mathrm{L_s}}{\rho_{i,\mathrm{crit}}}\left(\frac{\partial \rho_i}{\partial s}\right). \tag{A.7}$$

Considering eqs. (A.4), (A.6) and (A.7) and applying them on the original dimensional equations, one gets the following scaling laws for the influence parameter

$$c^*_{i,j} = \frac{\rho_{i,\mathrm{crit}}\rho_{j,\mathrm{crit}}\mathrm{L_s}}{\mathrm{k_B}T} c_{i,j}. \tag{A.8}$$

Scaling of the modified density $\tilde{\rho}$ is performed according to defintion in eq. (5.11), employing formula for $c^*_{i,j}$ from eq. (A.8) and ρ^*_i from eq. (A.4).

The grand potential density can be scaled in similar style using previously derived scaling laws as

$$\Delta\omega^* = \frac{\mathrm{L_s}^3}{\mathrm{k_B}T}\Delta\Omega. \tag{A.9}$$

Using the definition for the grand potential density given in eq. (2.87), the scaling formula for the chemical potential is obtained as

$$\left(\mu_i^V\right)^* = \frac{\rho_{i,\mathrm{crit}}\mathrm{L_s}^3}{\mathrm{k_B}T}\mu_i^V. \tag{A.10}$$

With all equations given above, the scaling formula for the *artificial* variable X can be derived as follows

$$X^* = \frac{\mathrm{L_s}^{\frac{5}{2}}}{\sqrt{\mathrm{k_B}T}} X. \tag{A.11}$$

Formulas given in eqs. (A.4), (A.6) to (A.8) and (A.11) represent equations required for the dimensionless scaling of the algebraic solution. The inverted formulas are also employed in our calculations in the subsequent backward scaling to dimensional quantities.

A.3 PC-SAFT parameters of the investigated fluids

Table A.1 summarizes parameters of the PC-SAFT equation of state in section 2.1.2.3. For CO_2, two sets of parameters are provided: one for PC-SAFT, and one for PCP-SAFT accounting for quadrupolarity of CO_2, which is denoted as CO_2 (Q). Parameters for SF_6 were correlated to the saturated liquid density and the vapor pressure evaluated from the multiparameter EoS by Guder and Wagner [91] in the temperature range from 223.6 to 317.6 K. For simplicity of initial investigation mixtures calculation were performed with a binary interaction parameter $k_{ij} = 0$. A non-zero k_{ij} could improve the agreement with the experimental data presented in section 6.6.5, however more independent data would be needed for fiting the k_{ij}.

The influence parameter $c_{i,i}$ was correlated to the experimental data for the surface tension of pure substances together with the critical point properties and the molar weight are provided in table A.2.

Substance	Reference	m[-]	σ_{LJ}[Å]	$\frac{\epsilon_{\mathrm{LJ}}}{\mathrm{k_B}}$[K]	q[D]	n_q[-]
n-butane	[88]	2.34212	3.70240	222.38	0	0
n-nonane	[88]	4.20737	3.84480	244.51	0	0
n-decane	[88]	4.66325	3.83840	243.87	0	0
CO_2	[88]	2.07274	2.78520	169.21	0	0
CO_2 (Q)	[86]	1.51310	3.18690	163.33	4.4	1
SF_6	this work	2.51811	3.31188	160.54	0	0

Table A.1: PC-SAFT and PCP-SAFT parameters for the selected substances.

Substance	$M[\frac{\mathrm{kg}}{\mathrm{mol}}]$	T_{c}[K]	p_{c}[Pa]	$\rho_{\mathrm{c}}[\frac{\mathrm{mol}}{\mathrm{m^3}}]$	$c_{i,i}[\frac{\mathrm{J \cdot m^5}}{\mathrm{mol^2}}]$
n-butane	0.0581222	425.1250	37.960	3922.8	1.8274123E-19
n-nonane	0.1282551	594.5500	22.810	1810.0	7.7169704E-19
n-decane	0.1422817	617.7000	21.030	1640.0	9.6732246E-19
CO_2	0.0440095	304.1282	73.773	10624.9	1.8972774E-20
CO_2 (Q)	0.0440095	304.1282	73.773	10624.9	2.1338599E-20
SF_6	0.14605542	318.7232	37.550	5082.3	7.4922648E-20

Table A.2: Molar weight, critical point properties, and influence parameter for the selected substances.

Investigation of metastable system properties B

B.1 Program design structure around the cluster and void criteria

The developed software design encompasses an overarching control structure capable of managing, calculating, and analyzing numerous datasets with varying thermodynamic state variables, simulation properties, and criteria parameters. Configuration files serve as a centralized means of transferring information between software layers, while a console application facilitates user input processing.

The method faces two main implementation challenges: fast runtime requirements and a reliance on various computational packages. To address these challenges, a unifying interface was developed, allowing the method to visualize data on personal computers and utilize calculation machines or supercomputers. Python was chosen as the interface language, integrating APIs written in FORTRAN 95 for performance-critical code, specifically the *ms*2 v. 4.0 [73] (molecular simulation) and TREND [216] (used for EoS evaluation). Additional Python libraries, including Click [186] for command line interface and Signac [3, 58] for data flow organization, were utilized. In consequence, Linux is the targeted platform for compatibility, other platforms are not actively supported but should still work when incorporated with the libraries of the packages provided for Windows.

The method organization structure is shown in fig. B.1. Encapsulated parts are bundled together into working modules. Interaction between modules is mediated by files in csv, json, or custom formatting. The data flow from sampled points into the simulation and analysis modules are managed by Signac, while the simulation module is designed to run on data center. Data analysis can also be done data center, whereas graphic related outputs need to be generated locally. For the purpose of transferring images, a separate custom library enabling Matplotlib figure saving was developed.

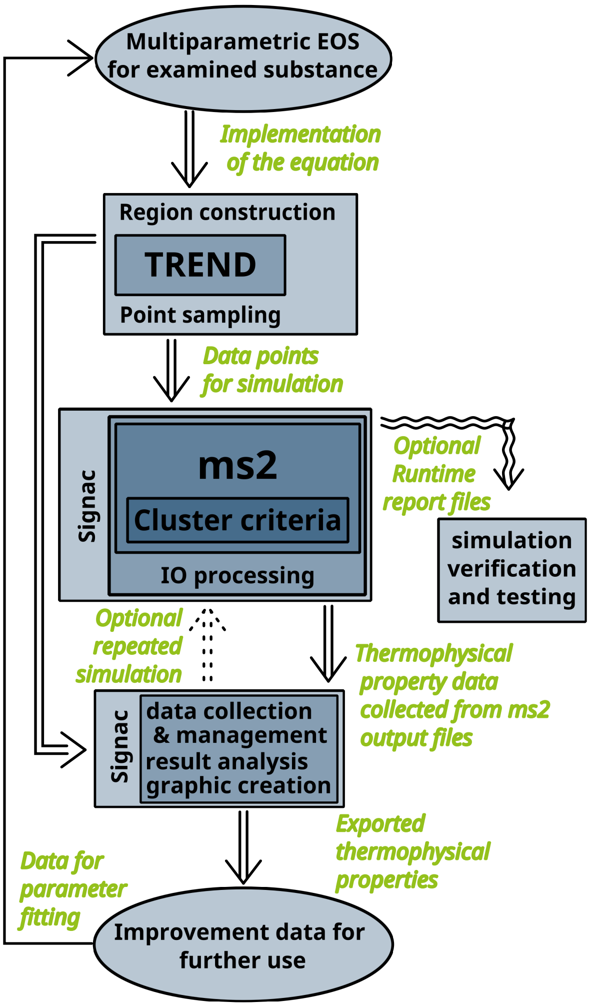

Multiparametric EOS for examined substance
Implementation of the equation
Region construction
TREND
Point sampling
Data points for simulation
Signac
ms2
Cluster criteria
IO processing
Optional Runtime report files
simulation verification and testing
Optional repeated simulation
Thermophysical property data collected from ms2 output files
Signac
data collection & management
result analysis
graphic creation
Exported thermophysical properties
Data for parameter fitting
Improvement data for further use

Figure B.1: Execution diagram for the example of the grid cluster criterion. Two line arrows denote the main operation direction; a dashed arrow shows optional user action; and a wavy arrow points at optional verification capabilities. Rectangular areas show modules and code sections, while ellipses are code-unrelated steps in the process.

Bibliography

A

[1] A. Aasen, E. M. Blokhuis, and Ø. Wilhelmsen, *Tolman lengths and rigidity constants of multicomponent fluids: Fundamental theory and numerical examples*, The Journal of chemical physics, 148 (2018), p. 204702.

[2] M. I. L. Abutaqiya, C. J. Sisco, Y. Khemka, M. A. Safa, E. F. Ghloum, A. M. Rashed, R. Gharbi, S. Santhanagopalan, M. Al-Qahtani, E. Al-Kandari, and F. M. Vargas, *Accurate Modeling of Asphaltene Onset Pressure in Crude Oils Under Gas Injection Using Peng–Robinson Equation of State*, Energy & Fuels, 34 (2020), pp. 4055–4070. https://pubs.acs.org/doi/10.1021/acs.energyfuels.9b04030.

[3] C. S. Adorf, P. M. Dodd, V. Ramasubramani, and S. C. Glotzer, *Simple data and workflow management with the signac framework*, Computational Materials Science, 146 (2018), pp. 220–229. https://linkinghub.elsevier.com/retrieve/pii/S0927025618300429.

[4] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, *Automatic subspace clustering of high dimensional data for data mining applications*, in Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, 1998, pp. 94–105.

[5] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Clarendon Press, Oxford, 1987.

[6] A. Aminian, D. Celný, E. Mickoleit, A. Jäger, and V. Vinš, *Ideal Gas Heat Capacity and Critical Properties of HFE-Type Engineering Fluids: Ab Initio Predictions of Cpig, Modeling of Phase Behavior and Thermodynamic Properties Using Peng–Robinson and Volume-Translated Peng–Robinson Equations of State*, International Journal of Thermophysics, 43 (2022), p. 87. https://link.springer.com/10.1007/s10765-022-03006-z.

[7] H. C. Andersen, *Molecular dynamics simulations at constant pressure and/or temperature*, The Journal of Chemical Physics, 72 (1980), pp. 2384–2393. https://pubs.aip.org/jcp/article/72/4/2384/218722/Molecular-dynamics-simulations-at-constant.

[8] R. Angélil, J. Diemand, K. K. Tanaka, and H. Tanaka, *Homogeneous SPC/e water nucleation in large molecular dynamics simulations*, J. Chem. Phys., 143 (2015), p. 064507.

[9] J. M. Anglada, G. J. Hoffman, L. V. Slipchenko, M. M.Costa, M. F. Ruiz-López, and J. S. Francisco, *Atmospheric significance of water clusters and ozone–water complexes*, J. Phys. Chem. A, 117 (2013), pp. 10381–10396.

[10] M. E. Araújo and M. A. Meireles, *Improving phase equilibrium calculation with the Peng–Robinson EOS for fats and oils related compounds/-supercritical CO2 systems*, Fluid Phase Equilibria, 169 (2000), pp. 49–64. https://www.sciencedirect.com/science/article/pii/S0378381200003071.

[11] P. Atkins, P. W. Atkins, and J. de Paula, *Atkins' Physical Chemistry*, Oxford university press, 2014.

[12] S. Ayuba, D. Suh, K. Nomura, T. Ebisuzaki, and K. Yasuoka, *Kinetic analysis of homogeneous droplet nucleation using large-scale molecular dynamics simulations*, The Journal of Chemical Physics, 149 (2018), p. 044504. http://aip.scitation.org/doi/10.1063/1.5037647.

B

[13] V. Baidakov, S. Protsenko, and Z. Kozlova, *Thermal and caloric equations of state for stable and metastable Lennard-Jones fluids: I. Molecular-dynamics simulations*, Fluid Phase Equilibria, 263 (2008), pp. 55–63. https://linkinghub.elsevier.com/retrieve/pii/S0378381207006085.

[14] V. G. Baidakov, *Spontaneous cavitation in a Lennard-Jones liquid: Molecular dynamics simulation and the van der Waals-Cahn-Hilliard gradient theory*, The Journal of Chemical Physics, 144 (2016), p. 074502. http://aip.scitation.org/doi/10.1063/1.4941689.

[15] V. G. Baidakov and S. P. Protsenko, *Metastable Lennard-Jones fluids. II. Thermal conductivity*, The Journal of Chemical Physics, 140 (2014), p. 214506. http://aip.scitation.org/doi/10.1063/1.4880958.

[16] ——, *Metastable Lennard-Jones fluids. III. Bulk viscosity*, The Journal of Chemical Physics, 141 (2014), p. 114503. http://aip.scitation.org/doi/10.1063/1.4895624.

[17] V. G. Baidakov, S. P. Protsenko, and Z. R. Kozlova, *Metastable Lennard-Jones fluids. I. Shear viscosity*, The Journal of Chemical Physics, 137 (2012), p. 164507. http://aip.scitation.org/doi/10.1063/1.4758806.

[18] J. A. Barker and D. Henderson, *Perturbation Theory and Equation of State for Fluids: The Square-Well Potential*, The Journal of Chemical Physics, 47 (1967), pp. 2856–2861. https://pubs.aip.org/jcp/article/47/8/2856/85369/Perturbation-Theory-and-Equation-of-State-for.

[19] L. S. Bartell and P. J. Lennon, *Generation of protosnowflakes in supersonic flow*, J. Chem. Phys., 130 (2009), p. 084303.

[20] D. Becker, C. W. Dierking, J. Suchan, F. Zurheide, J. Lengyel, M. Fárník, P. Slavíček, U. Buck, and T. Zeuch, *Temperature evolution in IR action spectroscopy experiments with sodium doped water clusters*, Physical Chemistry Chemical Physics, 23 (2021), pp. 7682–7695. http://xlink.rsc.org/?DOI=D0CP05390B.

[21] R. Becker and W. Döring, *Kinetische Behandlung der Keimbildung in übersättigten Dämpfen*, Annalen der Physik, 416 (1935), pp. 719–752. https://onlinelibrary.wiley.com/doi/10.1002/andp.19354160806.

[22] I. H. Bell, J. Wronski, S. Quoilin, and V. Lemort, *Pure and Pseudo-pure Fluid Thermophysical Property Evaluation and the Open-Source Thermophysical Property Library CoolProp*, Industrial & Engineering Chemistry Research, 53 (2014), pp. 2498–2508. https://pubs.acs.org/doi/10.1021/ie4033999.

[23] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, *The missing term in effective pair potentials*, The Journal of Physical Chemistry, 91 (1987), pp. 6269–6271. https://pubs.acs.org/doi/abs/10.1021/j100308a038.

[24] 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 Intermolecular Forces, B. Pullman, ed., vol. 14, Springer Netherlands, Dordrecht, 1981, pp. 331–342. http://link.springer.com/10.1007/978-94-015-7658-1_21.

[25] D. E. Bohn, N. Sürken, and F. Kreitmeier, *Nucleation phenomena in a multi-stage low pressure steam turbine*, Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy, 217 (2003), pp. 453–460. http://journals.sagepub.com/doi/10.1243/095765003322315513.

[26] S. Bonella, A. Coretti, R. Vuilleumier, and G. Ciccotti, *Adiabatic motion and statistical mechanics via mass-zero constrained dynamics*, Physical Chemistry Chemical Physics, 22 (2020), pp. 10775–10785.

[27] A. Borner, Z. Li, and D. A. Levin, *Development of a molecular-dynamics-based cluster-heat-capacity model for study of homogeneous condensation in supersonic water-vapor expansions*, J. Chem. Phys., 138 (2013), p. 064302.

[28] T. Boublík, *Hard-Sphere Equation of State*, The Journal of Chemical Physics, 53 (1970), pp. 471–472. https://pubs.aip.org/jcp/article/53/1/471/82532/Hard-Sphere-Equation-of-State.

[29] E. Brauer and E. Hough, *Interfacial tension of the normal butane-carbon dioxide system*, Producers Monthly, 29 (1965), pp. 13–...

[30] P. Brault, S. Chuon, and J.-M. Bauchire, *Molecular Dynamics Simulations of Platinum Plasma Sputtering: A Comparative Case Study*, Frontiers in Physics, 4 (2016). http://journal.frontiersin.org/Article/10.3389/fphy.2016.00020/abstract.

[31] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of Markov Chain Monte Carlo*, CRC press, 2011.

[32] R. A. Buckingham, *The classical equation of state of gaseous helium, neon and argon*, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 168 (1938), pp. 264–283. https://royalsocietypublishing.org/doi/10.1098/rspa.1938.0173.

[33] G. Bussi, D. Donadio, and M. Parrinello, *Canonical sampling through velocity rescaling*, The Journal of Chemical Physics, 126 (2007), p. 014101. https://pubs.aip.org/jcp/article/126/1/014101/186581/Canonical-sampling-through-velocity-rescaling.

C

[34] J. Cahn, *Free energy of a nonuniform system .2. thermodynamic basis*, J. Chem. Phys., 30 (1959), pp. 1121–1124.

[35] J. Cahn and J. Hilliard, *Free energy of a nonuniform system .1. interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.

[36] ——, *Free energy of a nonuniform system .3. nucleation in a 2-component incompressible fluid*, J. Chem. Phys., 31 (1959), pp. 688–699.

[37] H. B. Callen, *Thermodynamics and an Introduction to Thermostatistics*, Wiley, 2nd ed., 1985.

[38] K. Calvin, D. Dasgupta, G. Krinner, A. Mukherji, P. W. Thorne, C. Trisos, J. Romero, P. Aldunce, K. Barrett, G. Blanco, W. W. Cheung, S. Connors, F. Denton, A. Diongue-Niang, D. Dodman, M. Garschagen, O. Geden, B. Hayward, C. Jones, F. Jotzo, T. Krug, R. Lasco, Y.-Y. Lee, V. Masson-Delmotte, M. Meinshausen, K. Mintenbeck, A. Mokssit, F. E. Otto, M. Pathak, A. Pirani, E. Poloczanska, H.-O. Pörtner, A. Revi, D. C. Roberts, J. Roy, A. C. Ruane, J. Skea, P. R. Shukla, R. Slade, A. Slangen, Y. Sokona, A. A. Sörensson, M. Tignor, D. van Vuuren, Y.-M. Wei, H. Winkler, P. Zhai, Z. Zommers, J.-C. Hourcade, F. X. Johnson, S. Pachauri, N. P. Simpson, C. Singh, A. Thomas, E. Totin, P. Arias, M. Bustamante, I. Elgizouli, G. Flato, M. Howden, C. Méndez-Vallejo, J. J. Pereira, R. Pichs-Madruga, S. K. Rose, Y. Saheb, R. Sánchez Rodríguez, D. Ürge-Vorsatz, C. Xiao, N. Yassaa, A. Alegría, K. Armour, B. Bednar-Friedl, K. Blok, G. Cissé, F. Dentener, S. Eriksen, E. Fischer, G. Garner, C. Guivarch, M. Haasnoot, G. Hansen, M. Hauser, E. Hawkins, T. Hermans, R. Kopp, N. Leprince-Ringuet, J. Lewis, D. Ley, C. Ludden, L. Niamir, Z. Nicholls, S. Some, S. Szopa, B. Trewin, K.-I. van der Wijst, G. Winter, M. Witting, A. Birt, M. Ha, J. Romero, J. Kim, E. F. Haites, Y. Jung, R. Stavins, A. Birt, M. Ha, D. J. A. Orendain, L. Ignon, S. Park, Y. Park, A. Reisinger, D. Cammaramo, A. Fischlin, J. S. Fuglestvedt, G. Hansen, C. Ludden, V. Masson-Delmotte, J. R. Matthews, K. Mintenbeck, A. Pirani, E. Poloczanska, N. Leprince-Ringuet, and C. Péan, *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.*, tech. rep., Intergovernmental Panel on Climate Change (IPCC), July 2023. https://www.ipcc.ch/report/ar6/syr/.

[39] Gilles. Celeux, Didier. Chauveau, and Jean. Diebolt, *Stochastic versions of the em algorithm: An experimental study in the mixture case*, Journal of Statistical Computation and Simulation, 55 (1996), pp. 287–314. http://www.tandfonline.com/doi/abs/10.1080/00949659608811772.

[40] D. Celny, *Mathematical modeling of planar and spherical phase interfaces for multicomponent fluids*, Master's thesis, Czech Technical University in Prague, Prague, Jan. 2016.

[41] D. Celný, M. Klíma, and J. Kolafa, *Molecular Dynamics of Heterogeneous Systems on GPUs and Their Application to Nucleation in Gas Expanding to a Vac-*

uum, Journal of Chemical Theory and Computation, 17 (2021), pp. 7397–7405. https://pubs.acs.org/doi/10.1021/acs.jctc.1c00736.

[42] D. Celný, S. Pohl, M. Thol, V. Vinš, R. Span, and J. Vrabec, *Thermodynamic properties of metastable liquid and vapor phases by molecular dynamics with grid cluster criteria*, (submitted for publication).

[43] D. Celný, V. Vinš, and J. Hrubý, *Modelling of planar and spherical phase interfaces for multicomponent systems using density gradient theory*, Fluid Phase Equilibria, 483 (2019), pp. 70–83. https://linkinghub.elsevier.com/retrieve/pii/S0378381218304369.

[44] D. Celný, V. Vinš, B. Planková, and J. Hrubý, *Mathematical modeling of planar and spherical vapor–liquid phase interfaces for multicomponent fluids*, EPJ Web of Conferences, 114 (2016), p. 02011. http://www.epj-conferences.org/10.1051/epjconf/201611402011.

[45] D. Celný, V. Vinš, B. Planková, and J. Hrubý, *Mathematical modeling of planar and spherical vapor-liquid phase interfaces for multicomponent fluids*, in EFM15 - EXPERIMENTAL FLUID MECHANICS 2015, P. Dancova and M. Vesely, eds., vol. 114 of EPJ Web of Conferences, DANTEC Dynam GmbH; LAVISION; LENAM; MECAS ESI s r o; MIT s r o; TSI GmbH, 2016. 10th Anniversary International Conference on Experimental Fluid Mechanics, Prague, CZECH REPUBLIC, NOV 17-20, 2015.

[46] M. Čenský, J. Hrubý, V. Vinš, J. Hykl, and B. Šmíd, *Investigation of droplet nucleation in CCS relevant systems – design and testing of the expansion chamber*, EPJ Web of Conferences, 180 (2018), p. 02015. https://www.epj-conferences.org/10.1051/epjconf/201818002015.

[47] W. Chapman, K. Gubbins, G. Jackson, and M. Radosz, *SAFT: Equation-of-state solution model for associating fluids*, Fluid Phase Equilibria, 52 (1989), pp. 31–38. https://linkinghub.elsevier.com/retrieve/pii/0378381289803085.

[48] W. G. Chapman, K. E. Gubbins, G. Jackson, and M. Radosz, *New reference equation of state for associating liquids*, Industrial & Engineering Chemistry Research, 29 (1990), pp. 1709–1721. https://pubs.acs.org/doi/abs/10.1021/ie00104a021.

[49] G. Ciccotti and M. Ferrario, *Holonomic Constraints: A Case for Statistical Mechanics of Non-Hamiltonian Systems*, Computation, 6 (2018), p. 11. http://www.mdpi.com/2079-3197/6/1/11.

[50] Content at Scale, *Content at Scale's Advanced AI Detector*, (2021). https://contentatscale.ai/.

[51] P. Cornelisse, C. Peters, and J. de Swaan Arons, *Application of the peng-robinson equation of state to calculate interfacial tensions and profiles at vapour-liquid interfaces*, Fluid Phase Equilib, 82 (1993), pp. 119 – 129.

[52] C. F. Curtiss and J. O. Hirschfelder, *Integration of Stiff Equations*, Proceedings of the National Academy of Sciences, 38 (1952), pp. 235–243. https://pnas.org/doi/full/10.1073/pnas.38.3.235.

D

[53] H. Davis and L. Scriven, *Stress and structure in fluid interfaces*, Adv. Chem. Phys., 49 (1982), pp. 357–454.

[54] H. T. Davis, *Statistical mechanics of phases. interfaces, and thin films*, Wiley-VCH, Inc., 1996.

D

[55] S. W. de Leeuw, J. W. Perram, and H. G. Petersen, *Hamilton's equations for constrained dynamical systems*, Journal of Statistical Physics, 61 (1990), pp. 1203–1222. http://link.springer.com/10.1007/BF01014372.

D

[56] A. P. Dempster, N. M. Laird, and D. B. R. R. work(s):, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society. Series B (Methodological), 39 (1977), pp. 1–38. http://www.jstor.org/stable/2984875.

[57] S. Deublein, B. Eckl, J. Stoll, S. V. Lishchuk, G. Guevara-Carrion, C. W. Glass, T. Merker, M. Bernreuther, H. Hasse, and J. Vrabec, *{ms2}: A molecular simulation tool for thermodynamic properties*, Computer Physics Communications, 182 (2011), pp. 2350–2367. https://linkinghub.elsevier.com/retrieve/pii/S0010465511002025.

[58] B. Dice, B. Butler, V. Ramasubramani, A. Travitz, M. Henry, H. Ojha, K. Wang, C. Adorf, E. Jankowski, and S. Glotzer, *Signac: Data Management and Workflows for Computational Researchers*, in Python in Science Conference, Austin, Texas, 2021, pp. 23–32. https://conference.scipy.org/proceedings/scipy2021/bradley_dice.html.

[59] J. Diemand, R. Angélil, K. K. Tanaka, and H. Tanaka, *Large scale molecular dynamics simulations of homogeneous nucleation*, J. Chem. Phys., 139 (2013), p. 074309.

[60] K. K. Dingilian, R. Halonen, V. Tikkanen, B. Reischl, H. Vehkamäki, and B. E. Wyslouzil, *Homogeneous nucleation of carbon dioxide in supersonic nozzles I: Experiments and classical theories*, Physical Chemistry Chemical Physics, 22 (2020), pp. 19282–19298. http://xlink.rsc.org/?DOI=D0CP02279A.

[61] K. K. Dingilian, R. Halonen, V. Tikkanen, B. Reischl, H. Vehkamaki, and B. E. Wyslouzil, *Homogeneous nucleation of carbon dioxide in supersonic nozzles I: experiments and classical theories*, Phys. Chem. Chem. Phys., 22 (2020), pp. 19282–19298.

[62] V. T. Do and J. Straub, *Surface tension, coexistence curve, and vapor pressure of binary liquid-gas mixtures*, Int. J. Thermophys., 7 (1986), pp. 41–51.

[63] J. Dormand and P. Prince, *A family of embedded runge-kutta formulae*, Journal of Computational and Applied Mathematics, 6 (1980).

E

[64] C. Ebner, W. Saam, and D. Stroud, *Density-functional theory of simple classical fluids. i. surfaces*, Physical Review A, 14 (1976), p. 2264.

[65] M. Einsiedler and T. Ward, *Ergodic Theory: With a View towards Number Theory*, Springer London, London, 2011. https://link.springer.com/10.1007/978-0-85729-021-2.

[66] H. M. Ellerby, C. L. Weakliem, and H. Reiss, *Toward a molecular theory of vapor-phase nucleation. I. Identification of the average embryo*, The Journal of Chemical Physics, 95 (1991), pp. 9209–9218. https://pubs.aip.org/jcp/article/95/12/9209/97318/Toward-a-molecular-theory-of-vapor-phase.

[67] M. Ester, H.-P. Kriegel, and X. Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, 96 (1996), pp. 226–231.

[68] R. Evans, *The nature of the liquid-vapour interface and other topics in the statistical mechanics of non-uniform, classical fluids*, Adv. Phys., 28 (1979), pp. 143–200.

F

[69] D. G. Fahrenheit, *VIII. Experimenta & observationes de congelatione aquæ in vacuo factæ a DG Fahrenheit, RS S*, Philosophical Transactions of the Royal Society of London, 33 (1724), pp. 78–84.

[70] M. Fárník, J. Fedor, J. Kočišek, J. Lengyel, E. Pluhařová, V. Poterya, and A. Pysanenko, *Pickup and reactions of molecules on clusters relevant for atmospheric and interstellar processes*, Physical Chemistry Chemical Physics, 23 (2021), pp. 3195–3213. http://xlink.rsc.org/?DOI=D0CP06127A.

[71] M. Farnik, J. Fedor, J. Kocisek, J. Lengyel, E. Pluharova, V. Poterya, and A. Pysanenko, *Pickup and reactions of molecules on clusters relevant for atmospheric and interstellar processes*, Phys. Chem. Chem. Phys., 23 (2021), pp. 3195–3213.

[72] M. Field, *A Practical Introduction to the Simulation of Molecular Systems*, Cambridge University Press, Cambridge ; New York, 2nd ed ed., 2007.

[73] R. Fingerhut, G. Guevara-Carrion, I. Nitzke, D. Saric, J. Marx, K. Langenbach, S. Prokopev, D. Celný, M. Bernreuther, S. Stephan, et al., *{ms2}: A molecular simulation tool for thermodynamic properties, release 4.0*, Computer Physics Communications, 262 (2021), p. 107860.

[74] J. Fischer, R. Lustig, H. Breitenfelder-Manske, and W. Lemming, *Influence of intermolecular potential parameters on orthobaric properties of fluids consisting of spherical and linear molecules*, Molecular Physics, 52 (1984), pp. 485–497. http://www.tandfonline.com/doi/abs/10.1080/00268978400101351.

[75] C. Fraley and A. E. Raftery, *Model-Based Clustering, Discriminant Analysis, and Density Estimation*, Journal of the American Statistical Association, 97 (2002), pp. 611–631. http://www.tandfonline.com/doi/abs/10.1198/016214502760047131.

[76] JP. Franck and HG. Hertz, *Messung der kritischen Übersättigung von Dämpfen mit der Diffusionsnebelkammer*, Zeitschrift für Physik, 143 (1956), pp. 559–590.

[77] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, vol. 1, Elsevier, 2001.

[78] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications, 2nd edition*, Academic Press, San Diego, 2002.

[79] K.-S. Fu and S.-Y. Lu, *A clustering procedure for syntactic patterns*, IEEE Transactions on Systems, Man, and Cybernetics, 7 (1977), pp. 734–742.

G

[80] C. W. Gear, *The numerical integration of ordinary differential equations of various orders*, tech. rep., Argonne National Lab., Ill., 1966.

[81] D. Ghosh, A. Manka, R. Strey, S. Seifert, R. E. Winans, and B. E. Wyslouzil, *Using small angle x-ray scattering to measure the homogeneous nucleation rates of n-propanol, n-butanol, and n-pentanol in supersonic nozzle expansions*, J. Chem. Phys., 129 (2008), p. 124302.

[82] J. W. Gibbs, *On the equilibrium of heterogeneous substances*, American Journal of Science, 3 (1878), pp. 441–458.

[83] H. Goldstein, C. Poole, and J. Safko, *Classical Mechanics*, Addison Wesley, 3 ed., 2001.

[84] P. Gonnet, *Pairwise verlet lists: Combining cell lists and verlet lists to improve memory locality and parallelism*, J. Comput. Chem., 33 (2012), pp. 76–81.

[85] ——, *Pseudo-verlet lists: a new, compact neighbour list representation*, Molecular Simulation, 39 (2013), pp. 721–727.

[86] J. Gross, *An equation-of-state contribution for polar components: Quadrupolar molecules*, AIChE J., 51 (2005), pp. 2556–2568.

[87] J. Gross and G. Sadowski, *Perturbed-Chain SAFT: An Equation of State Based on a Perturbation Theory for Chain Molecules*, Industrial & Engineering Chemistry Research, 40 (2001), pp. 1244–1260. https://pubs.acs.org/doi/10.1021/ie0003887.

[88] ——, *Perturbed-chain saft: An equation of state based on a perturbation theory for chain molecules*, Ind. Eng. Chem. Res., 40 (2001), pp. 1244–1260.

[89] ——, *Application of the Perturbed-Chain SAFT Equation of State to Associating Systems*, Industrial & Engineering Chemistry Research, 41 (2002), pp. 5510–5515. https://pubs.acs.org/doi/10.1021/ie010954d.

[90] J. Gross and J. Vrabec, *An equation-of-state contribution for polar components: Dipolar molecules*, AIChE Journal, 52 (2006), pp. 1194–1204. https://onlinelibrary.wiley.com/doi/10.1002/aic.10683.

[91] C. Guder and W. Wagner, *A reference equation of state for the thermosynamic properties of sulfur hexafluoride (sf6) for temperatures from the melting line to 625 k and pressures up to 150 mpa*, J. Phys. Chem. Ref. Data, 38 (2009), pp. 33–94.

[92] S. Guha, R. Rastogi, and K. Shim, *CURE: An efficient clustering algorithm for large databases*, ACM Sigmod record, 27 (1998), pp. 73–84.

H

[93] J. D. Haley and C. McCabe, *Predicting the phase behavior of fluorinated organic molecules using the GC-SAFT-VR equation of state*, Fluid Phase Equilibria, 440 (2017), pp. 111–121. https://linkinghub.elsevier.com/retrieve/pii/S0378381217300225.

[94] J. W. Halley, *Statistical Mechanics: From First Principles to Macroscopic Phenomena*, Cambridge University Press, 2006.

[95] R. Halonen, E. Zapadinsky, and H. Vehkamaki, *Deviation from equilibrium conditions in molecular dynamic simulations of homogeneous nucleation*, J. Chem. Phys., 148 (2018).

[96] D. A. Hegg and M. B. Baker, *Nucleation in the atmosphere*, Reports on Progress in Physics, 72 (2009), p. 056801. https://iopscience.iop.org/article/10.1088/0034-4885/72/5/056801.

[97] R. H. Heist and H. He, *Review of Vapor to Liquid Homogeneous Nucleation Experiments from 1968 to 1992*, Journal of Physical and Chemical Reference Data, 23 (1994), pp. 781–805. http://aip.scitation.org/doi/10.1063/1.555951.

[98] E. A. Hemmingsen, *Cavitation in gas-supersaturated solutions*, Journal of Applied Physics, 46 (1975), pp. 213–218. https://pubs.aip.org/jap/article/46/1/213/1031644/Cavitation-in-gas-supersaturated-solutions.

[99] P. G. Hill, *Condensation of water vapour during supersonic expansion in nozzles*, Journal of Fluid Mechanics, (1966).

[100] T. L. Hill, *Molecular Clusters in Imperfect Gases*, The Journal of Chemical Physics, 23 (1955), pp. 617–622. http://aip.scitation.org/doi/10.1063/1.1742067.

[101] P. Hirunsit, Z. Huang, T. Srinophakun, M. Charoenchaitrakool, and S. Kawi, *Particle formation of ibuprofen–supercritical CO2 system from rapid expansion of supercritical solutions (RESS): A mathematical model*, Powder Technol., 154 (2005), pp. 83–94.

[102] W. G. Hoover, *Canonical dynamics: Equilibrium phase-space distributions*, Physical Review A, 31 (1985), pp. 1695–1697. https://link.aps.org/doi/10.1103/PhysRevA.31.1695.

[103] M. Horsch and H. Hasse, *Reprint of: Molecular simulation of nano-dispersed fluid phases*, Chemical Engineering Science, 115 (2014), pp. 195–204. https://linkinghub.elsevier.com/retrieve/pii/S0009250914002437.

[104] M. Horsch, S. Miroshnichenko, and J. Vrabec, *Steady-state molecular dynamics simulation of vapour to liquid nucleation with McDonald's daemon.* http://arxiv.org/abs/0911.5485, Nov. 2009.

[105] M. Horsch and J. Vrabec, *Grand canonical steady-state simulation of nucleation*, The Journal of Chemical Physics, 131 (2009), p. 184104. http://aip.scitation.org/doi/10.1063/1.3259696.

[106] J. Hrubý, D. G. Labetski, and M. E. H. van Dongen, *Gradient theory computation of the radius-dependent surface tension and nucleation rate for n-nonane clusters*, The Journal of Chemical Physics, 127 (2007), p. 164720.

[107] J. Hsu, N. Nagarajan, and R. Robinson, *Equilibrium phase compositions, phase densities, and interfacial-tensions for co2 + hydrocarbon systems .1. co2 + normal-butane*, J. Chem. Eng. Data, 30 (1985), pp. 485–491.

J

[108] R. T. Jacobsen and R. B. Stewart, *Thermodynamic Properties of Nitrogen Including Liquid and Vapor Phases from 63 K to 2000 K with Pressures to 10,000 Bar*, Journal of Physical and Chemical Reference Data, 2 (1973), pp. 757–922. https://pubs.aip.org/jpr/article/2/4/757/241463/Thermodynamic-Properties-of-Nitrogen-Including.

[109] A. K. Jain, M. N. Murty, and P. J. Flynn, *Data clustering: A review*, ACM Computing Surveys, 31 (1999), pp. 264–323. https://dl.acm.org/doi/10.1145/331499.331504.

[110] J. Janek and J. Kolafa, *Novel gear-like predictor–corrector integration methods for molecular dynamics*, Mol. Phys., 118 (2019), p. e1674937.

[111] J. Janek and J. Kolafa, *Novel Gear-like predictor–corrector integration methods for molecular dynamics*, Molecular Physics, 118 (2020), p. e1674937. https://www.tandfonline.com/doi/full/10.1080/00268976.2019.1674937.

K

[112] H. K. and S. E., *Calculation of surface properties of pure fluids using density gradient theory and saft-eos*, Fluid Phase Equilibria, 172 (2000), pp. 27 – 42.

[113] V. Kalikmanov, *Nucleation Theory*, vol. 860 of Lecture Notes in Physics, Springer Netherlands, Dordrecht, 2013. https://link.springer.com/10.1007/978-90-481-3643-8.

[114] A. Kantrowitz, *Nucleation in very rapid vapor expansions*, (1951).

[115] G. Karypis, E. Han, and V. Kumar, *A hierarchical clustering algorithm using dynamic modeling*, (1999).

[116] D. Kashchiev, *Nucleation: Basic Theory with Applications*, Butterworth Heinemann, Oxford ; Boston, 2000.

[117] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley Series in Probability and Mathematical Statistics, Wiley, Hoboken, N.J, 2005.

[118] K. F. Kelton and A. L. Greer, *Nucleation in Condensed Matter - Applications in Materials and Biology*, vol. 15 of Pergamon Materials Series, Elsevier, 2010. https://linkinghub.elsevier.com/retrieve/pii/C20090045000.

[119] V.-M. Kerminen, T. Petäjä, H. E. Manninen, P. Paasonen, T. Nieminen, M. Sipilä, H. Junninen, M. Ehn, S. Gagné, L. Laakso, I. Riipinen, H. Vehkamäki, T. Kurten, I. K. Ortega, M. Dal Maso, D. Brus, A. Hyvärinen, H. Lihavainen, J. Leppä, K. E. J. Lehtinen, A. Mirme, S. Mirme, U. Hõrrak, T. Berndt,

F. Stratmann, W. Birmili, A. Wiedensohler, A. Metzger, J. Dommen, U. Baltensperger, A. Kiendler-Scharr, T. F. Mentel, J. Wildt, P. M. Winkler, P. E. Wagner, A. Petzold, A. Minikin, C. Plass-Dülmer, U. Pöschl, A. Laaksonen, and M. Kulmala, *Atmospheric nucleation: Highlights of the EUCAARI project and future directions*, Atmospheric Chemistry and Physics, 10 (2010), pp. 10829–10848. https://acp.copernicus.org/articles/10/10829/2010/.

[120] A. Khan, C. H. Heath, U. M. Dieregsweiler, B. E. Wyslouzil, and R. Strey, *Homogeneous nucleation rates for d2o in a supersonic laval nozzle*, J. Chem. Phys., 119 (2003), pp. 3138–3147.

[121] S. Khosharay, *Linear gradient theory for modeling investigation on the surface tension of (ch4+h2o), (n-2+h2o) and (ch4+n-2)+h2o systems*, Journal of natural gas science and engineering, 23 (2015), pp. 474–480.

[122] T. Kinjo, K. Ohguchi, K. Yasuoka, and M. Matsumoto, *Computer simulation of ¯uid phase change: Vapor nucleation and bubble formation dynamics*, Computational Materials Science, (1999).

[123] M. Klíma, D. Celný, J. Janek, and J. Kolafa, *Properties of water and argon clusters developed in supersonic expansions*, (submitted for publication).

[124] M. Klíma and J. Kolafa, *Direct molecular dynamics simulation of nucleation during supersonic expansion of gas to a vacuum*, J. Chem. Theory Comput., 14 (2018), pp. 2332–2340.

[125] M. Klíma and J. Kolafa, *Direct Molecular Dynamics Simulation of Nucleation during Supersonic Expansion of Gas to a Vacuum*, Journal of Chemical Theory and Computation, 14 (2018), pp. 2332–2340. https://pubs.acs.org/doi/10.1021/acs.jctc.8b00066.

[126] J. Kolafa, *MACSIMUS*. https://old.vscht.cz/fch/software/macsimus/.

[127] J. Kolafa and M. Lísal, *Time-Reversible Velocity Predictors for Verlet Integration with Velocity-Dependent Right-Hand Side*, Journal of Chemical Theory and Computation, 7 (2011), pp. 3596–3607. https://pubs.acs.org/doi/10.1021/ct200108g.

[128] J. Kolafa and M. Lísal, *Time-reversible velocity predictors for verlet integration with velocity-dependent right-hand side*, J. Chem. Theory Comput., 7 (2011), pp. 3596–3607.

[129] J. Kolafa, F. Moucka, and I. Nezbeda, *Handling electrostatic interactions in molecular simulations: A systematic study*, Coll. Czech. Chem. Comm., 73 (2008), pp. 481–506.

[130] M. Kolovratník, J. Hrubỳ, V. Ždímal, O. Bartoš, I. Jiříček, P. Moravec, and N. Zíková, *Nanoparticles found in superheated steam: A quantitative analysis of possible heterogeneous condensation nuclei*, Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy, 228 (2014), pp. 186–193.

[131] D. Kondepudi and I. Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures*, John Wiley & Sons Inc, Chichester, West Sussex Hoboken, NJ, second edition ed., 2015.

[132] G. Kontogeorgis and I. Economou, *Equations of state: From the ideas of van der waals to association theories*, J. Supercrit. Fluids, 55 (2010), pp. 421–437.

[133] S. B. Kotsiantis and P. E. Pintelas, *Recent Advances in Clustering: A Brief Survey*, 1 (2004), pp. 73–81.

[134] A. Krishnamoorthy, K.-i. Nomura, N. Baradwaj, K. Shimamura, P. Rajak, A. Mishra, S. Fukushima, F. Shimojo, R. Kalia, A. Nakano, and P. Vashishta, *Dielectric Constant of Liquid Water Determined with Neural Network Quantum Molecular Dynamics*, Physical Review Letters, 126 (2021), p. 216403. https://link.aps.org/doi/10.1103/PhysRevLett.126.216403.

[135] W. D. Kristensen, E. J. Jensen, and R. M. J. Cotterill, *Thermodynamics of small clusters of atoms: A molecular dynamics simulation*, The Journal of Chemical Physics, 60 (1974), pp. 4161–4169. https://pubs.aip.org/jcp/article/60/11/4161/775061/Thermodynamics-of-small-clusters-of-atoms-A.

L

[136] T. Lafitte, B. Mendiboure, M. Pineiro, D. Bessieres, and C. Miqueu, *Interfacial properties of water/co2: A comprehensive description through a gradient theory-saft-vr mie approach*, J. Phys. Chem. B, 114 (2010), pp. 11110–11116.

[137] L. D. Landau and E. M. Lifshitz, *Statistical Physics: Volume 5*, vol. 5, 1980.

[138] L. D. Landau and E. M. Lifshitz, *Course of theoretical physics. 6. Fluid mechanics, 2nd ed.*, Pergamon Press, Oxford, 1987.

[139] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, no. v. 6 in Course of Theoretical Physics, Pergamon Press, Oxford, England ; New York, 2nd ed., 2nd english ed., rev ed., 1987.

[140] J. K. Lee, J. A. Barker, and F. F. Abraham, *Theory and Monte Carlo simulation of physical clusters in the imperfect vapor*, The Journal of Chemical Physics, 58 (1973), pp. 3166–3180. https://pubs.aip.org/jcp/article/58/8/3166/462788/Theory-and-Monte-Carlo-simulation-of-physical.

[141] J. Lengyel, J. Kočišek, V. Poterya, A. Pysanenko, P. Svrčková, M. Fárník, D. K. Zaouris, and J. Fedor, *Uptake of atmospheric molecules by ice nanoparticles: Pickup cross sections*, J. Chem. Phys., 137 (2012), p. 034304.

[142] J. Lengyel, A. Pysanenko, V. Poterya, P. Slavíček, M. Fárník, J. Kočišek, and J. Fedor, *Irregular Shapes of Water Clusters Generated in Supersonic Expansions*, Physical Review Letters, 112 (2014), p. 113401. https://link.aps.org/doi/10.1103/PhysRevLett.112.113401.

[143] Z. Li, A. Borner, and D. A. Levin, *Multi-scale study of condensation in water jets using ellipsoidal-statistical Bhatnagar-Gross-Krook and molecular dynamics modeling*, The Journal of Chemical Physics, 140 (2014), p. 224501. https://pubs.aip.org/jcp/article/140/22/224501/352487/Multi-scale-study-of-condensation-in-water-jets.

[144] Z. Li, A. Borner, and D. A. Levin, *Multi-scale study of condensation in water jets using ellipsoidal-statistical bhatnagar-gross-krook and molecular dynamics modeling*, J. Chem. Phys., 140 (2014), p. 224501.

[145] X. Liang and M. L. Michelsen, *General approach for solving the density gradient theory in the interfacial tension calculations*, Fluid Phase Equilibria, 451 (2017), pp. 79–90.

[146] X. Liang, M. L. Michelsen, and G. M. Kontogeorgis, *Pitfalls of using the geometric-mean combining rule in the density gradient theory*, Fluid Phase Equilib., 415 (2016), pp. 75–83.

[147] H. Lin, Y.-Y. Duan, and Q. Min, *Gradient theory modeling of surface tension for pure fluids and binary mixtures*, Fluid Phase Equilib., 254 (2007), pp. 75 – 90.

[148] A. Linhart, C.-C. Chen, J. Vrabec, and H. Hasse, *Thermal properties of the metastable supersaturated vapor of the Lennard-Jones fluid*, The Journal of Chemical Physics, 122 (2005), p. 144506. http://aip.scitation.org/doi/10.1063/1.1872774.

[149] D. Lovelock and H. Rund, *Tensors, Differential Forms, and Variational Principles*, Courier Corporation, 1989.

[150] S. Lubetkin and M. Blackwell, *The nucleation of bubbles in supersaturated solutions*, Journal of Colloid and Interface Science, 126 (1988), pp. 610–615. https://linkinghub.elsevier.com/retrieve/pii/0021979788901610.

[151] R. Lustig, *Statistical thermodynamics in the classical molecular dynamics ensemble. I. Fundamentals*, The Journal of Chemical Physics, 100 (1994), pp. 3048–3059. https://pubs.aip.org/jcp/article/100/4/3048/112740/Statistical-thermodynamics-in-the-classical.

[152] N. Lümmen and T. Kraska, *Investigation of the formation of iron nanoparticles from the gas phase by molecular dynamics simulation*, Nanotech., 15 (2004), pp. 525–533.

M

[153] K. Magoulas and D. Tassios, *Thermophysical properties of n-Alkanes from C1 to C20 and their prediction for higher ones*, (1990).

[154] J. Mairhofer and J. Gross, *Modeling of interfacial properties of multicomponent systems using density gradient theory and pcp-saft*, Fluid Phase Equilib., 439 (2017), pp. 31–42.

[155] ——, *Modeling properties of the one-dimensional vapor-liquid interface: Application of classical density functional and density gradient theory*, Fluid Phase Equilib., 458 (2018), pp. 243–252.

[156] A. Manka, H. Pathak, S. Tanimura, J. Woelk, R. Strey, and B. E. Wyslouzil, *Freezing water in no-man's land*, Phys. Chem. Chem. Phys., 14 (2012), pp. 4505–4516.

[157] G. A. Mansoori, N. F. Carnahan, K. E. Starling, and T. W. Leland, *Equilibrium Thermodynamic Properties of the Mixture of Hard Spheres*, The Journal of Chemical Physics, 54 (1971), pp. 1523–1525. https://pubs.aip.org/jcp/article/54/4/1523/445610/Equilibrium-Thermodynamic-Properties-of-the.

[158] F. Martínez-Veracoechea * and E. Müller, *Temperature-quench Molecular Dynamics Simulations for Fluid Phase Equilibria*, Molecular Simulation, 31 (2005), pp. 33–43. http://www.tandfonline.com/doi/abs/10.1080/08927020412331298991.

[159] G. Martyna, M. E. Tuckerman, D. J. Tobias, and M. L. Klein, *Explicit reversible integrators for extended systems dynamics*, Mol. Phys., 87 (1996), pp. 1117–1157.

[160] G. J. Martyna, M. E. Tuckerman, D. J. Tobias, and M. L. Klein, *Explicit reversible integrators for extended systems dynamics*, Molecular Physics, 87 (1996), pp. 1117–1157. http://www.tandfonline.com/doi/abs/10.1080/00268979600100761.

[161] D. J. McGinty, *Molecular dynamics studies of the properties of small clusters of argon atoms*, The Journal of Chemical Physics, 58 (1973), pp. 4733–4742. https://pubs.aip.org/jcp/article/58/11/4733/87684/Molecular-dynamics-studies-of-the-properties-of.

[162] A. D. McLaren, *Optimal Numerical Integration on a Sphere*, vol. 17, 1963.

[163] J. B. McQueen, *Some methods of classification and analysis of multivariate observations*, in Proc. of 5th Berkeley Symposium on Math. Stat. and Prob., 1967, pp. 281–297.

[164] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, The journal of chemical physics, 21 (1953), pp. 1087–1092.

[165] B. Metz, O. Davidson, H. Coninck, M. Loos, and L. Meyer, *IPCC special report carbon dioxide capture and storage summary for policymakers*, in United Nations Framework Convention on Climate Change, 2005.

[166] J. M. Miguez, M. G. J., F. J. Blas, H. Segura, A. Mejia, and M. M. Pineiro, *Comprehensive characterization of interfacial behavior for the mixture co2 + h2o + ch4: Comparison between atomistic and coarse grained molecular simulation models and density gradient theory*, J. Phys. Chem. C, 118 (2014), pp. 24504–24519.

[167] C. Miqueu, B. Mendiboure, A. Graciaa, and J. Lachaise, *Modelling of the surface tension of pure components with the gradient theory of fluid interfaces: a simple and accurate expression for the influence parameters*, Fluid Phase Equilib., 207 (2003), pp. 225–246.

[168] ——, *Modeling of the surface tension of multicomponent mixtures with the gradient theory of fluid interfaces*, Industrial & engineering chemistry research, 44 (2005), pp. 3321–3329.

[169] C. Miqueu, B. Mendiboure, A. Graciaa, and J. Lachaise, *Petroleum mixtures: An efficient predictive method for surface tension estimations at reservoir conditions*, Fuel, 87 (2008), pp. 612–621. https://linkinghub.elsevier.com/retrieve/pii/S0016236107002839.

[170] C. Miqueu, B. Mendiboure, C. Graciaa, and J. Lachaise, *Modelling of the surface tension of binary and ternary mixtures with the gradient theory of fluid interfaces*, Fluid Phase Equilibria, 218 (2004), pp. 189 – 203.

[171] E. A. Müller and A. Mejía, *Interfacial properties of selected binary mixtures containing n-alkanes*, Fluid Phase Equilibria, 282 (2009), pp. 68 – 81.

N

[172] N. Nagarajan and R. Robinson Jr, *Equilibrium phase compositions, phase densities, and interfacial tensions for carbon dioxide+ hydrocarbon systems. 2. carbon dioxide+ n-decane*, Journal of Chemical and Engineering Data, 31 (1986), pp. 168–171.

[173] I. Napari, J. Julin, and H. Vehkamäki, *Cluster sizes in direct and indirect molecular dynamics simulations of nucleation*, The Journal of Chemical Physics, 131 (2009), p. 244511. https://pubs.aip.org/aip/jcp/article/190041.

[174] A. V. Neimark and A. Vishnyakov, *The birth of a bubble: A molecular simulation study*, The Journal of Chemical Physics, 122 (2005), p. 054707. https://pubs.aip.org/jcp/article/122/5/054707/187061/The-birth-of-a-bubble-A-molecular-simulation-study.

[175] I. Nezbeda, M. Kotrla, and J. Kolafa, *Úvod Do Počítačovỳch Simulací: Metody Monte Carlo a Molekulární Dynamiky*, Karolinum, 2003.

[176] E. Noether, *Invariante variationsprobleme. Nachr. vd ges. d. Wiss. zu göttingen (1918) 235; E. Noether e MA tavel*, Transport Theor. Stat. Phys, 1 (1971), p. 183.

[177] S. Nosé, *A molecular dynamics method for simulations in the canonical ensemble*, Molecular Physics, 52 (1984), pp. 255–268. http://www.tandfonline.com/doi/abs/10.1080/00268978400101201.

[178] S. Nosé, *Constant-temperature molecular dynamics*, Journal of Physics: Condensed Matter, 2 (1990), pp. SA115–SA119. https://iopscience.iop.org/article/10.1088/0953-8984/2/S/013.

[179] NVIDIA, *CUDA C++ programming guide, release: 9.2.* https://developer.nvidia.com/cuda-toolkit, Aug. 2018.

[180] ——, *CUDA C++ programming guide, release: 12.2.* https://developer.nvidia.com/cuda-toolkit, July 2023.

O

[181] A. Obeidat, M. Gharaibeh, H. Ghanem, F. Hrahsheh, N. Al-Zoubi, and G. Wilemski, *Nucleation rates of methanol using the saft-0 equation of state*, ChemPhysChem, 11 (2010), pp. 3987–3995.

[182] M. G. Omran, A. P. Engelbrecht, and A. Salman, *An overview of clustering methods*, Intelligent Data Analysis, 11 (2007), pp. 583–605. https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/IDA-2007-11602.

[183] OpenAI, *ChatGPT3: A large-scale language model for conversational AI*, (2021). https://openai.com.

[184] W. Ostwald, *Studien über die Bildung und Umwandlung fester Körper: 1. Abhandlung: Übersättigung und Überkaltung*, Zeitschrift für Physikalische Chemie, 22U (1897), pp. 289–330. https://www.degruyter.com/document/doi/10.1515/zpch-1897-2233/html.

[185] Kl. Oswatitsch, *Kondensationserscheinungen in überschalldüsen .*, Zamm-zeitschrift Fur Angewandte Mathematik Und Mechanik, (1942).

P

[186] pallets, *Click.* https://click.palletsprojects.com/en/8.1.x/#miscellaneous-pages, Mar. 2023.

P

[187] A. Péneloux, E. Rauzy, and R. Fréze, *A consistent correction for Redlich-Kwong-Soave volumes*, Fluid Phase Equilibria, 8 (1982), pp. 7–23. https://linkinghub.elsevier.com/retrieve/pii/0378381282800022.

[188] D.-Y. Peng and D. B. Robinson, *A New Two-Constant Equation of State*, Industrial & Engineering Chemistry Fundamentals, 15 (1976), pp. 59–64. https://pubs.acs.org/doi/abs/10.1021/i160057a011.

[189] L. M. Pereira, A. Chapoy, R. Burgass, M. B. Oliveira, J. A. Coutinho, and B. Tohidi, *Study of the impact of high temperatures and pressures on the equilibrium densities and interfacial tension of the carbon dioxide/water system*, The Journal of Chemical Thermodynamics, 93 (2016), pp. 404–415.

[190] A. Pérez and A. Rubio, *A molecular dynamics study of water nucleation using the TIP4p/2005 model*, J. Chem. Phys., 135 (2011), p. 244505.

[191] Perplexity.ai, *Perplexity AI-powered search engine to provide accurate and comprehensive answers to user queries.*, (2021). https://www.perplexity.ai/.

[192] B. Planková, V. Vinš, J. Hrubý, M. Duška, T. Němec, and D. Celný, *Molecular simulation of water vapor–liquid phase interfaces using TIP4P/2005 model*, EPJ Web of Conferences, 92 (2015), p. 02071. http://www.epj-conferences.org/10.1051/epjconf/20159202071.

[193] B. Planková, V. Vinš, and J. Hrubý, *Predictions of homogeneous nucleation rates for n-alkanes accounting for the diffuse phase interface and capillary waves*, J. Chem. Phys., 147 (2017).

[194] S. Pohl, R. Fingerhut, M. Thol, J. Vrabec, and R. Span, *Equation of state for the Mie (λ_r ,6) fluid with a repulsive exponent from 11 to 13*, The Journal of Chemical Physics, 158 (2023), p. 084506. https://aip.scitation.org/doi/10.1063/5.0133412.

[195] A. J. Proctor, T. J. Lipscomb, A. Zou, J. A. Anderson, and S. S. Cho, *Performance analyses of a parallel verlet neighbor list algorithm for gpu-optimized md simulations*, in 2012 ASE/IEEE International Conference on BioMedical Computing (BioMedCom), IEEE, 2012, pp. 14–19.

Q

[196] QuillBot (Course Hero), *QuillBot's suite of tools employs cutting-edge AI technology in order to make writing painless*, (2021). https://quillbot.com/.

R

[197] D. C. Rapaport, *The Art of Molecular Dynamics Simulation*, Cambridge University Press, Cambridge, UK ; New York, NY, 2. ed., 2004.

[198] P. Rehner and J. Gross, *Surface tension of droplets and tolman lengths of real substances and mixtures from density functional theory*, The Journal of chemical physics, 148 (2018), p. 164703.

[199] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer Texts in Statistics, Springer New York, New York, NY, 2004. http://link.springer.com/10.1007/978-1-4757-4145-2.

[200] L. Rovigatti, P. Šulc, I. Z. Reguly, and F. Romano, *A comparison between parallelization approaches in molecular dynamics simulations on gpus*, Journal of computational chemistry, 36 (2015), pp. 1–8.

[201] D. Rupp, M. Adolph, T. Gorkhover, S. Schorb, D. Wolter, R. Hartmann, N. Kimmel, C. Reich, T. Feigl, A. R. B. de Castro, R. Treusch, L. Strueder, T. Moeller, and C. Bostedt, *Identification of twinned gas phase clusters by single-shot scattering with intense soft x-ray pulses*, New J. Phys., 14 (2012), p. 055016.

[202] A. Rusanov and E. Brodskaya, *The molecular dynamics simulation of a small drop*, Journal of Colloid and Interface Science, 62 (1977), pp. 542–555. https://linkinghub.elsevier.com/retrieve/pii/0021979777901059.

[203] G. Rutkai, A. Köster, G. Guevara-Carrion, T. Janzen, M. Schappals, C. W. Glass, M. Bernreuther, A. Wafai, S. Stephan, M. Kohns, S. Reiser, S. Deublein, M. Horsch, H. Hasse, and J. Vrabec, *{ms2} : A molecular simulation tool for thermodynamic properties, release 3.0*, Computer Physics Communications, 221 (2017), pp. 343–351. https://linkinghub.elsevier.com/retrieve/pii/S0010465517302527.

[204] 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*, Journal of Computational Physics, 23 (1977), pp. 327–341. https://linkinghub.elsevier.com/retrieve/pii/0021999177900985.

[205] F. Römer, B. Fischer, and T. Kraska, *Investigation of the nucleation and growth of methanol clusters from supersaturated vapor by molecular dynamics simulations*, Soft Materials, 10 (2012), pp. 130–152.

S

[206] R. J. Sadus, *Molecular Simulation of Fluids - Theory, Algorithms and Object-Orientation*, Elsevier, Amsterdam, 1. ed., 1999.

[207] A. V. Samodurov, S. V. Vosel', A. M. Baklanov, A. A. Onishchuk, and V. V. Karasev, *A study of homogeneous nucleation of ibuprofen in a flow chamber. Determination of the surface tension of critical nuclei*, Colloid Journal, 75 (2013), pp. 397–408. http://link.springer.com/10.1134/S1061933X13040078.

[208] W. Sarlet and F. Cantrijn, *Generalizations of Noether's Theorem in Classical Mechanics*, SIAM Review, 23 (1981), pp. 467–494. http://epubs.siam.org/doi/10.1137/1023098.

[209] P. Schaaf, B. Senger, and H. Reiss, *Defining Physical Clusters in Nucleation Theory from the* N *-Particle Distribution Function*, The Journal of Physical Chemistry B, 101 (1997), pp. 8740–8747. https://pubs.acs.org/doi/10.1021/jp970428t.

[210] K. A. G. Schmidt, G. K. Folas, and B. Kvamme, *Calculation of the interfacial tension of the methane-water system with the linear gradient theory*, Fluid Phase Equilib., 261 (2007), pp. 230–237. 11th International Conference on Propeties and Phase Equilibria for Product and Process Design, Crete, GREECE, MAY 20-25, 2007.

[211] B. Senger, P. Schaaf, D. S. Corti, R. Bowles, J.-C. Voegel, and H. Reiss, *A molecular theory of the homogeneous nucleation rate. I. Formulation and fundamental issues*, The Journal of Chemical Physics, 110 (1999), pp. 6421–6437. https://pubs.aip.org/aip/jcp/article/110/13/6421-6437/476049.

[212] M. J. Sewell, *Maximum and Minimum Principles: A Unified Approach with Applications*, vol. 1, CUP Archive, 1987.

[213] N. Shardt, Y. Wang, Z. Jin, and J. A. Elliott, *Surface tension as a function of temperature and composition for a broad range of mixtures*, Chemical Engineering Science, 230 (2021), p. 116095. https://linkinghub.elsevier.com/retrieve/pii/S0009250920306278.

[214] P. H. A. Sneath, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, A Series of Books in Biology, W H Freeman & Co (Sd), 1973.

[215] R. Span, *Multiparameter Equations of State*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000. http://link.springer.com/10.1007/978-3-662-04092-8.

[216] R. Span, R. Beckmüller, T. Eckermann, S. Herrig, S. Hielscher, A. Jäger, E. Mickoleit, T. Neumann, S. Pohl, B. Semrau, and M. Thol, *TREND. Thermodynamic Reference and Engineering Data 4.0.*, Lehrstuhl für Thermodynamik, Ruhr-Universität Bochum., (2019).

[217] R. Span, E. W. Lemmon, R. T. Jacobsen, W. Wagner, and A. Yokozeki, *A Reference Equation of State for the Thermodynamic Properties of Nitrogen for Temperatures from 63.151 to 1000 K and Pressures to 2200 MPa*, Journal of Physical and Chemical Reference Data, 29 (2000), pp. 1361–1433. https://pubs.aip.org/jpr/article/29/6/1361/241876/A-Reference-Equation-of-State-for-the.

[218] G. Spilková, *Optimalizace Parametrů Stavové Rovnice PC-SAFT v Modelování Termofyzikálních Vlastností Tekutin*, BSc thesis, České vysoké učení technické v Praze. Vypočetní a informační centrum., 2022.

[219] G. D. Stein and P. P. Wegener, *Experiments on the Number of Particles Formed by Homogeneous Nucleation in the Vapor Phase*, The Journal of Chemical Physics, 46 (1967), pp. 3685–3686. https://pubs.aip.org/jcp/article/46/9/3685/82708/Experiments-on-the-Number-of-Particles-Formed-by.

[220] S. Stephan, *Vapor-Liquid Interfaces: Molecular Simulation, Density Gradient Theory, and Experiments*, no. volume 33 in Scientific Report Series, Laboratory of Engineering Thermodynamics (LTD), TU Kaiserslautern, Kaiserslautern, 2020.

[221] S. Stephan, M. Thol, J. Vrabec, and H. Hasse, *Thermophysical Properties of the Lennard-Jones Fluid: Database and Data Assessment*, Journal of Chemical Information and Modeling, 59 (2019), pp. 4248–4265. https://pubs.acs.org/doi/10.1021/acs.jcim.9b00620.

[222] J. A. Stevens, F. Grünewald, P. A. Marco van Tilburg, M. König, B. R. Gilbert, T. A. Brier, Z. R. Thornburg, Z. Luthey-Schulten, and S. J. Marrink, *Molecular dynamics simulation of an entire cell*, Frontiers in Chemistry, 11 (2023), p. 1106495. https://www.frontiersin.org/articles/10.3389/fchem.2023.1106495/full.

[223] F. H. Stillinger, *Rigorous Basis of the Frenkel-Band Theory of Association Equilibrium*, The Journal of Chemical Physics, 38 (1963), pp. 1486–1494. http://aip.scitation.org/doi/10.1063/1.1776907.

[224] R. Strey, P. E. Wagner, and Y. Viisanen, *The Problem of Measuring Homogeneous Nucleation Rates and the Molecular Contents of Nuclei: Progress in the Form of Nucleation Pulse Measurements*, The Journal of Physical Chemistry, 98 (1994), pp. 7748–7758. https://pubs.acs.org/doi/abs/10.1021/j100083a003.

T

[225] K. K. Tanaka, J. Diemand, R. Angélil, and H. Tanaka, *Free energy of cluster formation and a new scaling relation for the nucleation rate*, The Journal of Chemical Physics, 140 (2014), p. 194310. https://pubs.aip.org/jcp/article/140/19/194310/351064/Free-energy-of-cluster-formation-and-a-new-scaling.

[226] S. Tanimura, H. Pathak, and B. E. Wyslouzil, *Binary nucleation rates for ethanol/water mixtures in supersonic laval nozzles: Analyses by the first and second nucleation theorems*, J. Chem. Phys., 139 (2013), p. 174311.

[227] M. Tatarevic, *On Limits of Dense Packing of Equal Spheres in a Cube*, The Electronic Journal of Combinatorics, 22 (2015), p. P1.35. https://www.combinatorics.org/ojs/index.php/eljc/article/view/v22i1p35.

[228] P. R. ten Wolde and D. Frenkel, *Computer simulation study of gas–liquid nucleation in a Lennard-Jones system*, The Journal of Chemical Physics, 109 (1998), pp. 9901–9918. https://pubs.aip.org/jcp/article/109/22/9901/476853/Computer-simulation-study-of-gas-liquid-nucleation.

[229] M. Thol, G. Rutkai, A. Köster, R. Lustig, R. Span, and J. Vrabec, *Equation of state for the Lennard-Jones fluid*, Journal of Physical and Chemical Reference Data, (2016).

[230] S. M. Thompson, K. E. Gubbins, J. P. R. B. Walton, R. A. R. Chantry, and J. S. Rowlinson, *A molecular dynamics study of liquid drops*, Journal of Chemical Physics, (1984).

[231] S. Toxvaerd, *Algorithms for canonical molecular dynamics simulations*, Mol. Phys., 72 (1991), pp. 159–168.

[232] A. S. Tucker and C. A. Ward, *Critical state of bubbles in liquid-gas solutions*, Journal of Applied Physics, 46 (1975), pp. 4801–4808. https://pubs.aip.org/jap/article/46/11/4801/7111/Critical-state-of-bubbles-in-liquid-gas-solutions.

[233] M. E. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation*, Oxford University Press, Oxford ; New York, 2010.

V

[234] J. van der Waals, *Thermodynamische theorie der kapillarität unter voraussetzung stetiger dichteänderung*, J. Phys. Chem., 13 (1894), pp. 657–725.

[235] J. D. Van Der Waals and J. S. Rowlinson, *On the Continuity of the Gaseous and Liquid States*, Dover publications, 2004.

[236] JD. Van der Waals, *Continuity of the gaseous and liquid state of matter*, University of Leiden, (1873).

[237] C. Vega, E. Sanz, and J. L. F. Abascal, *The melting temperature of the most common models of water*, The Journal of Chemical Physics, 122 (2005), p. 114507. https://pubs.aip.org/jcp/article/122/11/114507/929655/The-melting-temperature-of-the-most-common-models.

[238] L. F. Vega, O. Vilaseca, F. Llovell, and J. S. Andreu, *Modeling ionic liquids and the solubility of gases in them: Recent advances and perspectives*, Fluid Phase Equilibria, 294 (2010), pp. 15 – 30. Ionic Liquids Special Issue.

[239] H. Vehkamäki, *Classical Nucleation Theory in Multicomponent Systems*, Springer, Berlin ; New York, 2006.

[240] JA. Venables, GDT. Spiller, and M. Hanbucken, *Nucleation and growth of thin films*, Reports on progress in physics, 47 (1984), p. 399.

[241] L. Verlet, *Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules*, Physical Review, 159 (1967), pp. 98–103. https://link.aps.org/doi/10.1103/PhysRev.159.98.

[242] Y. Viisanen, R. Strey, and H. Reiss, *Homogeneous nucleation rates for water*, 99 (1993).

[243] V. Vinš, A. Aminian, D. Celný, M. Součková, J. Klomfar, M. Čenský, and O. Prokopová, *Surface tension and density of dielectric heat transfer fluids of HFE type-experimental data at 0.1 MPa and modeling with PC-SAFT equation of state and density gradient theory*, International Journal of Refrigeration, 131 (2021), pp. 956–969. https://linkinghub.elsevier.com/retrieve/pii/S0140700721002668.

[244] V. Vinš, D. Celný, B. Planková, T. Němec, M. Duška, and J. Hrubý, *Molecular Simulations of the Vapor–Liquid Phase Interfaces of Pure Water Modeled with the SPC/E and the TIP4P/2005 Molecular Models*, EPJ Web of Conferences, 114 (2016), p. 02136. http://www.epj-conferences.org/10.1051/epjconf/201611402136.

[245] V. Vinš, J. Hykl, J. Hrubỳ, A. Blahut, D. Celnỳ, M. Čenskỳ, and O. Prokopová, *Possible anomaly in the surface tension of supercooled water: New experiments at extreme supercooling down to -31.4 C*, 11 (2020), pp. 4443–4447.

[246] V. Vinš, J. Hrubý, and B. Planková, *Surface tension of binary mixtures including polar components modeled by the density gradient theory combined with the pc-saft equation of state*, International Journal of Thermophysics, 34 (2013), pp. 792–812.

[247] V. Vinš, B. Planková, J. Hrubý, and D. Celný, *Density gradient theory combined with the pc-saft equation of state used for modeling the surface tension of associating systems*, EPJ Web Conferences, 67 (2014).

[248] V. Vinš, M. Čenský, J. Hykl, and J. Hrubý, *Investigation of droplet nucleation in ccs relevant systems – design and testing of a co2 branch of the mixture preparation device*, EPJ Web Conf., 143 (2017), p. 02140.

[249] M. Volmer and A. Weber, *Keimbildung in übersättigten gebilden*, Zeitschrift für physikalische Chemie, 119 (1926), pp. 277–301.

[250] J. Vrabec and J. Gross, *Vapor-Liquid Equilibria Simulation and an Equation of State Contribution for Dipole-Quadrupole Interactions*, The Journal of Physical Chemistry B, 112 (2008), pp. 51–60. https://pubs.acs.org/doi/10.1021/jp072619u.

W

[251] H. Wadell, *Volume, shape, and roundness of quartz particles*, J. Geology, 43 (1935), pp. 250–280.

[252] W. Wang, J. Yang, and R. Muntz, *STING : A Statistical Information Grid Approach to Spatial Data Mining*, 97 (1997), pp. 186–195.

[253] J. Wedekind, R. Strey, and D. Reguera, *New method to analyze simulations of activated processes*, The Journal of Chemical Physics, 126 (2007), p. 134103. https://pubs.aip.org/jcp/article/126/13/134103/188198/New-method-to-analyze-simulations-of-activated.

[254] P. Wegener and L. Mack, *Condensation in supersonic and hypersonic wind tunnels*, Advances in Applied Mechanics, (1958).

[255] P. Wegener and A. Pouring, *Experiments on condensation of water vapor by homogeneous nucleation in nozzles*, (1964).

[256] K. Wegner, P. Piseri, H. V. Tafreshi, and P. Milani, *Cluster beam deposition: A tool for nanoscale science and technology*, Journal of Physics D: Applied Physics, 39 (2006), pp. R439–R459. https://iopscience.iop.org/article/10.1088/0022-3727/39/22/R02.

[257] M. S. Wertheim, *Fluids with highly directional attractive forces. II. Thermodynamic perturbation theory and integral equations*, Journal of Statistical Physics, 35 (1984), pp. 35–47. http://link.springer.com/10.1007/BF01017363.

[258] Ø. Wilhelmsen, A. Aasen, G. Skaugen, P. Aursand, A. Austegard, E. Aursand, M. A. Gjennestad, H. Lund, G. Linga, and M. Hammer, *Thermodynamic Modeling with Equations of State: Present Challenges with Established Methods*, Industrial & Engineering Chemistry Research, 56 (2017), pp. 3503–3515. https://pubs.acs.org/doi/10.1021/acs.iecr.7b00317.

[259] O. Wilhelmsen, D. Bedeaux, and D. Reguera, *Communication: Tolman length and rigidity constants of water and their role in nucleation*, J. Chem. Phys., 142 (2015).

[260] C. T. R. Wilson, *XI. Condensation of water vapour in the presence of dust-free air and other gases*, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, (1897), pp. 265–307.

[261] G. Winkler and G. H. Schnerr, *Nucleating unsteady flows in low-pressure steam turbine stages*, Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy, 215 (2001), pp. 773–781. http://journals.sagepub.com/doi/10.1243/0957650011538901.

[262] Wu, *A MOLECULAR DYNAMICS SIMULATION OF BUBBLE NUCLEATION IN HOMOGENEOUS LIQUID UNDER HEATING WITH CONSTANT MEAN NEGATIVE PRESSURE*, Microscale Thermophysical Engineering, 7 (2003), pp. 137–151. http://www.tandfonline.com/doi/abs/10.1080/10893950390203323.

X

[263] D. Xu, S. Luo, J. Song, J. Liu, and W. Cao, *Direct numerical simulations of supersonic compression-expansion slope with a multi-GPU parallel algorithm*, Acta Astronautica, 179 (2021), pp. 20–32. https://linkinghub.elsevier.com/retrieve/pii/S0094576520306421.

[264] D. Xu and Y. Tian, *A Comprehensive Survey of Clustering Algorithms*, Annals of Data Science, 2 (2015), pp. 165–193. http://link.springer.com/10.1007/s40745-015-0040-1.

Y

[265] K. Yasuoka and M. Matsumoto, *Molecular dynamics of homogeneous nucleation in the vapor phase. I. Lennard-Jones fluid*, The Journal of Chemical Physics, 109 (1998), pp. 8451–8462. http://aip.scitation.org/doi/10.1063/1.477509.

[266] ——, *Molecular dynamics of homogeneous nucleation in the vapor phase. II. Water*, The Journal of Chemical Physics, 109 (1998), pp. 8463–8470. http://aip.scitation.org/doi/10.1063/1.477510.

Z

[267] Y. B. Zeldovich, *On the theory of new phase formation: Cavitation*, Acta Physicochem., USSR, 18 (1943), p. 1.

[268] R. Zhang, A. Khalizov, L. Wang, M. Hu, and W. Xu, *Nucleation and Growth of Nanoparticles in the Atmosphere*, Chemical Reviews, 112 (2012), pp. 1957–2011. https://pubs.acs.org/doi/10.1021/cr2001756.

[269] T. Zhang, R. Ramakrishnan, and M. Livny, *BIRCH: An efficient data clustering method for very large databases*, ACM sigmod record, 25 (1996), pp. 103–114.

[270] J. Zhong, M. I. Zeifman, and D. A. Levin, *Direct simulation of condensation in a one-dimensional unsteady expansion: Microscopic mechanisms*, Physics of Fluids, 17 (2005), p. 128102. https://pubs.aip.org/pof/article/17/12/128102/932718/Direct-simulation-of-condensation-in-a-one.

[271] Y. Zuo and E. Stenby, *Calculation of interfacial tensions with gradient theory*, Fluid Phase Equilib., 132 (1997), pp. 139–158.