REPORT SERIES IN AEROSOL SCIENCE
N:o 276 (2024)

MODELING THE ROLE OF IONS IN THE FORMATION AND DETECTION OF ATMOSPHERIC CLUSTERS

IVO NEEFJES

Institute for Atmospheric and Earth System Research / Physics
Faculty of Science
University of Helsinki
Helsinki, Finland

Doctoral dissertation, to be presented for public discussion with the permission of the Faculty of Science of the University of Helsinki, in Auditorium E204, Physicum, Gustaf Hällströmin katu 2, on the 24th of May, 2024 at 12 o'clock.

Aerosolitutkimusseura r.y.
Helsinki 2024

Author's Address: Institute for Atmospheric and Earth System Research / Physics
P.O. Box 64
FI-00014 University of Helsinki
ivo.neefjes@helsinki.fi

Supervisors: Professor Hanna Vehkamäki, Ph.D.
Institute for Atmospheric and Earth System Research / Physics
University of Helsinki

Professor Theo Kurtén, Ph.D.
Institute for Atmospheric and Earth System Research / Chemistry
University of Helsinki

University lecturer Bernhard Reischl, D. Sc. (techn.)
Institute for Atmospheric and Earth System Research / Physics
University of Helsinki

Reviewers: Professor Christopher Hogan, Ph.D.
Department of Mechanical Engineering
University of Minnesota

Professor Karoliina Honkala, Ph.D.
Department of Chemistry
University of Jyväskylä

Opponent: Doctor Mária Lbadaoui-Darvas, Ph.D.
Laboratory of atmospheric processes and their impacts
École Polytechnique Fédérale de Lausanne

ISBN 978-952-7507-29-2 (printed)
ISSN 0784-3496
Helsinki 2024
Unigrafia Oy

ISBN 978-952-7507-30-8 (pdf)
ISSN 2814-4236
Helsinki 2024
http://www.FAAR.fi

Acknowledgements

A thesis is not just the fruit of a single individual's labor. Rather, it is the result of a collaborative effort. I was able to complete this thesis thanks to the wonderful support of my supervisors, colleagues, co-authors, friends, and family. In the space afforded to me here, it is impossible to adequately express my deep gratitude to everyone that has helped me through my scientific journey. Still, I will attempt to highlight some of the individuals that made this thesis possible.

Someone once told me: "Hanna Vehkamäki takes broken scientists and nurses them back to health." For me at least, this is true. After my master thesis, I was burned out, but Hanna's passion for science and her human approach to academic work, reinvigorated me. Hanna, you are a leader that I look up to and a friend that I cherish. Thank you for giving me the opportunity to perform my doctoral research under your supervision.

At the beginning of my doctoral research, I was an unguided projectile, performing quantum chemistry calculations without a clear end goal. It wasn't until Bernhard Reischl took me under his wing, that I really hit my stride. What started as a scientific collaboration, quickly grew into a friendship. Thank you, Bernhard, for supervising me on a near daily basis, and for being a great friend.

My doctoral research was also supervised by Theo Kurtén and Tommaso Zanca, two extremely intelligent scientists, who I could count on to give me a new perspective whenever I got stuck in my work. I furthermore thank my thesis committee members Tuija Jokinen and Juho Helaja for following my progress throughout the past four years, my pre-examiners Chistopher Hogan and Karoliina Honkala for reviewing my thesis, and Mária Lbadaoui-Darvas for being the opponent at the public examination. I am grateful to the Doctoral Programme in Atmospheric Sciences for granting me a salaried position and travel grant. As a computational researcher, I thank the CSC-IT Center for Science for computational resources and technical support.

It was a joy to be a member of the SIMU group. During my doctoral studies, Dina Alfaouri was the linchpin of the group. Dina, I cherish the running, restaurant and cinema visits, and conversations that we shared together. Stephen Ingram arrived in the group on the same day as me and we quickly became friends. Stephen, there is nothing better than going to the Tappen bar at 3 a.m. on a Monday to talk about

fringe political movements and pop culture nonsense. Jakub Kubečka has taught me so much about quantum chemistry. He is the most impressive early-career scientists that I know, with awe-inspiring work ethic and knowledge. On top of that, he is a great travel and drinking partner. Vitus Besel, I greatly admire your passion for everything you set your mind to. I love how you are always up for any activity or weird art project that we conceive of.

I also fondly look back on the collaborations, scientific discussions, conferences, and fun activities that I have had with the other group members, Roope Halonen, Huan Yang, Olli Pakarinen, Golnaz Roudsari, Valtteri Tikkanen, Fatemeh Keshavarz, Monica Passananti, Jing Dou, and Jaakko Kähärä. Beyond the SIMU group, other scientists at INAR, such as Juha Kangasluoma, Angelica Bianco, Stephany Buenrostro Mazon, Magdalena Okuljar, Xinjang Li, and Nanna Myllys have also contributed to making my time at the University of Helsinki both scientifically fruitful and fun. I, furthermore, thank Luis Seijo, who, before Hanna, showed me what a great supervisor looks like.

Besides my coworkers, who have scientifically contributed to my doctoral research, there are also many friends who have supported me by getting me out of my house and keeping me sane. Mats Pylyser, Elisabeth de Smet, Robbe Maes, Karen Verschueren, Timmer Taeymans, Lieke Barthelemy, Joni Schoonjans, Siebe Vanzurpele, Robby Reynaerts, Hans Vandersmissen, Floris Beijersbergen, Tuuli Haappio, Vasilis Symeonidis, and Farshad Ghorbani, I greatly appreciate your friendship.

I am forever grateful to my parents, André and Wilma. They have always been there for me, from helping me with my elementary and high school homework, to supporting me through the twists and turns of my higher education. Despite living in different countries, I have always felt supported by my two brothers, Tim and Marijn. Thank you for being the two friends that I know I can call about anything.

Last, I want to thank my girlfriend, Faustine. Faustine, I know that my move to Helsinki has been very difficult on us. If this was a test of our relationship, I think we passed it. I can think of no one else that I could spend three months in a 20 m^2 apartment with during a lockdown and afterwards still want to spend more time with. You are my best friend. I love you. It's time for us to go home to Belgium and start a family.

Ivo Neefjes
Helsinki, April 2024

Ivo Neefjes
University of Helsinki, 2024

Abstract

The majority of atmospheric aerosol particles are formed through a gas-to-particle conversion process called new particle formation. Upon their formation, aerosol particles impact the climate both directly, by scattering incoming sunlight, and indirectly, by functioning as cloud condensation nuclei. The effect of atmospheric aerosol particles represents one of the largest uncertainties in current climate models. This uncertainty is mostly caused by a lack of knowledge about the initial stages of new particle formation, where gas-phase compounds undergo sticking collisions to form clusters of 2–3 nm in diameter.

This thesis focuses on the role of ions in the formation and detection of atmospheric clusters. Ions typically have strong long-range interactions with other atmospheric compounds and can, therefore, facilitate the formation of atmospheric clusters. The importance of ion-induced new particle formation pathways is, however, still subject of debate. The relative importance of ion-induced pathways can be determined from their cluster formation rates through dynamical models, such as the Atmospheric Cluster Dynamics Code. These models rely on accurate collision and evaporation rate coefficients. We developed a computational approach to calculate accurate estimates for the collision rate coefficient based on the central field approach with interaction parameters fitted to the potential of mean force obtained from free energy calculations. We also developed an analytical interacting hard-sphere model where collision rate coefficient can be calculated for clusters of arbitrary size with low computational cost.

Evaporation rate coefficients are typically derived from binding free energies calculated through quantum chemistry calculations. We developed a machine learning workflow to more efficiently obtain accurate binding free energies for large clusters and used these to investigate the cluster formation rate dependence on the largest cluster size included in the Atmospheric Cluster Dynamics Code.

Ions also play an important role in the detection of atmospheric clusters in chemical ionization and ion mobility mass spectrometry. Ion mobility mass spectrometry is

especially relevant in separating isomers. We performed computational ion mobility calculations to help assign peaks in ion mobility spectra to specific isomers.

The results of these studies can be used to obtain more accurate cluster formation rates and help determine the compounds involved in new particle formation to reduce the uncertainty in the effect of aerosol particles on the climate.

Keywords: new particle formation, nucleation, Atmospheric Cluster Dynamics Code, quantum chemistry, configurational sampling, force fields, molecular dynamics, metadynamics, umbrella sampling, ion mobility, machine learning

Contents

List of symbols **9**

List of abbreviations **16**

List of publications **17**

1 Introduction **19**

1.1 Ions in New Particle Formation 22

1.2 Ions in the detection of atmospheric clusters 24

1.3 Modeling the role of ions in the formation and detection of atmospheric clusters 26

2 Theoretical and computational methods **29**

2.1 Classical nucleation theory 30

2.2 Atmospheric Cluster Dynamics Code 33

2.3 Quantum chemistry 35

2.3.1 The Born–Oppenheimer approximation 37

2.3.2 The Hartree–Fock method 39

2.3.3 Basis sets 42

2.3.4 Post-Hartree–Fock methods 43

2.3.5 Density functional theory 44

2.3.6 Semi-empirical methods 46

2.3.7 Thermochemistry calculations 46

2.4 Configurational sampling 48

2.5 Collision theory 52

2.5.1 The Su & Chesnavich parameterization 56

2.6 Molecular dynamics 57

2.6.1 Statistical mechanics 58

2.6.2 Equations of motion 60

2.6.3 Thermostats 62

2.7 Potential of mean force calculations 65

2.7.1 Umbrella sampling 65
2.7.2 Metadynamics 67
2.8 Force fields 68
2.9 Ion mobility modeling 71
2.10 Machine learning 74

3 Overview of papers and the author's contribution **78**

4 Results and discussion **80**
4.1 Modeling approaches for atmospheric ion–dipole collisions: All-atom trajectory simulations and central field methods (Paper II) 80
4.2 Ion–molecule rate constants for reactions of sulfuric acid with acetate and nitrate ions (Paper III) 81
4.3 Further cautionary tales on thermostatting in molecular dynamics: Energy equipartitioning and non-equilibrium processes in gas-phase simulations (Paper V) 83
4.4 Collision-sticking rates of acid–base clusters in the gas phase determined from atomistic simulation and a novel analytical interacting hard-sphere model (Paper VI) 85
4.5 Atmospheric sulfuric acid–multi-base new particle formation revealed through quantum chemistry enhanced by machine learning (Paper IV) . 86
4.6 Separation of isomers using a differential mobility analyser (DMA): Comparison of experimental vs modelled ion mobility (Paper I) 90

5 Conclusions and future outlook **92**

References **95**

List of symbols

This thesis spans a variety of different topics. These different topics regularly use the same symbols to denote different quantities. An effort has been made to give every quantity a unique symbol. This effort should, however, never stand in the way of the thesis' readability. For example, K conventionally denotes the kinetic energy in thermodynamics, as well as the Hartree–Fock exchange potential in quantum chemistry. Changing the symbol for one of these quantities would only lead to confusion. Thus, duplicates exist in this thesis, but their meaning will be clear from the context.

Vector quantities are denoted by bold lowercase symbols and matrix/tensor quantities by bold uppercase symbols.

Latin alphabet

$\mathcal{A}$	Surface area
A	Interaction coefficient
a	Interaction exponent
a	Multiplication factor
$[\mathbf{B}]$	Drag tensor
B	Drag tensor element
$\mathbf{b}$	Neural network parameters
b	Impact parameter
$\mathcal{C}$	Cluster concentration
$\mathbf{C}$	Coefficient matrix
c	Coefficient
$\mathcal{E}$	Electric field
E	Energy
$\boldsymbol{\mathcal{F}}$	Fock matrix
$\mathcal{F}$	Fock matrix element
F	Free energy shift
$\mathbf{f}$	Force
f	(Distribution) function
G	Gibbs free energy
$\mathbf{h}$	Neural network hidden layer

h	One-electron term of electronic energy
I	Moment of intertia
$\mathcal{J}$	Particle formation rate
J	Hartree–Fock Coulomb potential
$\mathcal{K}$	Kernel function
K	Hartree–Fock exchange potential
K	Kinetic energy
K_{SC}	Su & Chesnavich correction factor
k	Force constant
$\mathcal{L}$	External losses
L	Length
m	Mass
N	Number of entities or elements
O	Example observable
o	Eigenvalue of operator $\hat{O}$
$\mathcal{P}$	Probability
P	Pressure
$\mathbf{p}$	Momentum
p	Momentum coordinate
$\mathcal{Q}$	External sources
q	(Partial) charge
q	Position coordinate
$\mathcal{R}$	Random number
R	Radius
R_{m}	Center-of-mass distance for which ω_v has a minimum
$\mathbf{r}$	Position
r	Distance
$\mathcal{S}$	Entropy
$\mathbf{S}$	Overlap matrix
S	Overlap matrix element
S	Saturation ratio
$\mathbf{s}$	Vector of all collective variables
s	Collective variable
s	Spin coordinate
T	Temperature

ΔT	Bias factor
t	Time
Δt	Time step
U	Interaction potential
$\mathbf{u}_{\text{tot}}$	Sum of velocity changes
$\mathcal{V}$	Voltage
V	Potential energy
V	Fourier coefficient
$\mathbf{v}$	Velocity
v	Velocity magnitude
v_{drift}	Drift velocity
v_{flow}	Velocity of laminar flow
$\mathcal{W}$	Gaussian height
$\mathbf{W}$	Neural network parameters
W	Width
$\mathbf{x}$	Input
$\mathbf{y}$	Output
x, y, z	Cartesian positions
Z	Ion mobility

Greek alphabet

α	Angle-averaged polarizability
β	Collision rate coefficient
Γ	Gamma function
γ	Evaporation rate coefficient
$\boldsymbol{\gamma}$	Friction coefficient
$\boldsymbol{\epsilon}$	Orbital energy matrix
ϵ	Orbital energy
ϵ	Experimental noise
ε	Lennard-Jones energy parameter
ζ	Hyperparameter
ζ	Fictitious degree of freedom
Θ	Vibrational frequency

θ	Angle
$\boldsymbol{\lambda}$	Random vector
μ	Reduced mass
μ_d	Angle-averaged permanent dipole moment
ν	Electronic absorption frequency
ξ	Su & Chesnavich parametrization variable
ρ	Density
Σ	Gaussian width
σ	Lennard-Jones distance parameter
σ	Nonlinear activation function
σ_rot	Symmetry number
ς	Surface tension
$\boldsymbol{\tau}$	Spatial and spin coordinates
τ	Thermostat time constant
τ_G	Metadynamics deposition interval
Φ	Slater determinant
ϕ	Dihedral angle
φ	Basis function
χ	Orbitals
Ψ	Total wavefunction
ψ	Electronic/nuclear wavefunction
Ω	Collision cross section
ω_v	Interacting hard-sphere criterion function

Operators and functionals

$\hat{\mathcal{F}}$	Fock operator
$\hat{g}$	Two-electron operator
$\hat{H}$	Hamiltionian energy operator
$\hat{h}$	One-electron operator
$\hat{J}$	Hartree–Fock Coulomb operator
$\hat{K}$	Hartree–Fock exchange operator
$\hat{O}$	Example operator
$\hat{T}$	Kinetic energy operator

$\hat{V}$	Potential energy operator
$E[\rho_{\mathrm{el}}]$	Electronic energy functional
$T[\rho_{\mathrm{el}}]$	Kinetic energy functional
$V[\rho_{\mathrm{el}}]$	Potential energy functional
Δ	Change, difference
∇	Gradient
$\langle\ \rangle$	Ensemble average
$\{\ \}$	Set
d	Derivative

Sub- and superscripts

0	Equilibrium/ground state/initial/standard value
1	Compound 1
2	Compound 2
12	Between compound 1 and 2
amb	Ambient
angles	Angles
b	Bond
basis	Basis functions
bias	Bias
bins	Umbrella sampling bins
bonds	Bonds
CF	Central field approach
CSVR	Canonical sampling through velocity rescaling
CV	Collective variables
c	Critical
calibrant	Calibrant
center	Center of window
chain	Nosé–Hoover chains
cl	Cluster
com	Between centers of mass
D	Dimensions
d	Droplet

dihedrals	Dihedrals
dof	Degrees of freedom
e, el	Electrons
ee	Between electrons
eff	Effective
eq	Equilibrium
exp	Experimental
ext	External
final	Final value
frb	Freely rotating bonds
G	Metadynamics bias property
gas	Gas property
H	Hartree
HS	Hard-sphere
i	Species i
ij	Between species i and j
initial	Initial value
inter	Intermolecular properties
intra	Intramolecular properties
ion	Ion property
j	Species j
k	Dihedral angle
L	Langevin
L	Neural network layers
LJTM	Lennard-Jones trajectory method
l	Phase angles
MB	Maxwell–Boltzmann
MD	Molecular dynamics
m	Electronic state
micro	Microstates
N	Nuclei
Ne	Between nuclei and electrons
NN	Between nuclei
OPLS	Optimized Potential for Liquid Simulations
occ	Occupied orbitals

PA	Projection approximation
qpol	Quadrupole
ref	Reference
rot	Rotational
SC	Su & Chesnavich parametrization
s	System with no electron–electron interaction
t	Trial
thermal	Internal thermal
tot	Sum of all contributions
trans	Translational
vib	Vibrational
w	Windows
x, y, z	Cartesian positions
xx, yy, zz	Diagonal tensor elements
xc	Exchange correlation
α	Basis function/degree of freedom/matrix row/bin
β	matrix column
θ	Angle
κ	Vibrational mode
$*$	Complex conjugate
$\parallel$	Parallel to trajectory
$\perp$	Perpendicular to trajectory

Constants

e	Elementary charge
h	Planck constant
$\hbar$	Reduced Planck constant
k_B	Boltzmann constant
P_0	Pressure of 1 atm
R	Gas constant
T_0	Temperature of 273.15 K
ε_0	Vacuum permittivity

List of abbreviations

ACDC	Atmospheric Cluster Dynamics Code
AM	Ammonia
APi	Atmospheric pressure interface
CCS	Collision cross section
CF	Central field
CNT	Classical nucleation theory
com	Center of mass
CSVR	Canonical sampling through velocity rescaling
CV	Collective variable
DFT	Density functional theory
DMA	Differential mobility analyzer
DMA	Dimethylamine
dof	Degree of freedom
EDA	Ethylenediamine
HOM	Highly oxygenated organic molecule
HS	Hard-sphere
IMoS	Ion mobility software
IPCC	Intergovernmental Panel on Climate Change
JKCS	Jammy Key for Configurational Sampling
L	Langevin
LJTM	Lennard-Jones trajectory method
MA	Methylamine
MD	Molecular dynamics
NH	Nosé–Hoover
NPF	New particle formation
OPLS	Optimized Potential for Liquid Simulations
PA	Projection approximation
PES	Potential energy surfaces
PMF	Potential of mean force
SC	Su & Chesnavich
SCF	Self-consistent field
TMA	Trimethylamine
VOC	Volatile organic compound
WHO	World Health Organization

List of publications

This thesis consists of an introductory review, followed by 6 research articles. In the introductory part, these papers are cited according to their Roman numerals.

I. Bianco, A., Neefjes, I., Alfaouri, D., Vehkamäki, H., Kurtén, T., Ahonen, L., Passananti, M., & Kangasluoma, J. (2022). Separation of isomers using a differential mobility analyser (DMA): Comparison of experimental vs modelled ion mobility. *Talanta*, *243*(1), 123339. https://doi.org/10.1016/j.talanta.2022.123339

II. Neefjes, I., Halonen, R., Vehkamäki, H., & Reischl, B. (2022). Modeling approaches for atmospheric ion–dipole collisions: all-atom trajectory simulations and central field methods. *Atmospheric Chemistry and Physics*, *22*(17), 11155–11172. https://doi.org/10.5194/acp-22-11155-2022

III. Fomete, S. K. W., Johnson, J. S., Myllys, N., Neefjes, I., Reischl, B., & Jen, C. N. (2022). Ion–Molecule Rate Constants for Reactions of Sulfuric Acid with Acetate and Nitrate Ions. *The Journal of Physical Chemistry A*, *126*(44), 8240–8248. https://doi.org/10.1021/acs.jpca.2c02072

IV. Kubečka, J., Neefjes, I., Besel, V., Qiao, F., Xie, H.-B., & Elm, J. (2023). Atmospheric Sulfuric Acid–Multi-Base New Particle Formation Revealed through Quantum Chemistry Enhanced by Machine Learning. *The Journal of Physical Chemistry A*, *127*(9), 2091–2103. https://doi.org/10.1021/acs.jpca.3c00068

V. Halonen, R., Neefjes, I., & Reischl, B. (2023). Further cautionary tales on thermostatting in molecular dynamics: Energy equipartitioning and non-equilibrium processes in gas-phase simulations. *Journal of Chemical Physics*, *158*(19), 194301. https://doi.org/10.1063/5.0148013

VI. Yang, H., Neefjes, I., Tikkanen, V., Kubečka, J., Kurtén, T., Vehkamäki, H., & Reischl, B. (2023). Collision-sticking rates of acid–base clusters in the gas phase determined from atomistic simulation and a novel analytical interacting hard-sphere model. *Atmospheric Chemistry and Physics*, *23*(10), 5993–6009. https://doi.org/10.5194/acp-23-5993-2023

Paper I is reprinted with permission of Elsevier. **Papers II** and **VI** are reprinted under the Creative Commons Attribution 4.0 License. **Paper III** and **IV** are reprinted

with permission of the American Chemical Society. **Paper V** is reproduced with the permission of AIP Publishing.

The author additionally contributed to the following publications, which are not included in the thesis. These publications are likewise cited according to their Roman numerals.

VII. Neefjes, I., Joos, J. J., Barandiarán, Z., & Seijo, L. (2020). Mixed-Valence Lanthanide-Activated Phosphors: Invariance of the Intervalence Charge Transfer (IVCT) Absorption Onset across the Series. *The Journal of Physical Chemistry C*, *124*(4), 2619–2626. https://doi.org/10.1021/acs.jpcc.9b11084

VIII. Joos, J. J., Neefjes, I., Seijo, L., & Barandiarán, Z. (2021). Charge transfer from Eu^{2+} to trivalent lanthanide co-dopants: Systematic behavior across the series. *The Journal of Chemical Physics*, *154*(6), 064704. https://doi.org/10.1063/5.0037992

IX. Neefjes, I., Laapas, M., Liu, Y., Médus, E., Miettunen, E., Ahonen, L., Quéléver, L., Aalto, J., Bäck, J. K., Kerminen, V.-M., Lamplahti, J., Luoma, K., Mäki, M., Mammarella, I., Petäjä, T., Räty, M., Sarnela, N., Ylivinkka, I., Hakala, S., Kulmala, M., Nieminen, T., & Lintunen, A. (2022). 25 years of atmospheric and ecosystem measurements in a boreal forest – Seasonal variation and responses to warm and dry years. *Boreal Environment Research*, *27*, 1–31

X. Iwahara, N., Huang, Z., Neefjes, I., & Chibotaru, L. F. (2022). Multipolar exchange interaction and complex order in insulating lanthanides. *Physical Review B*, *105*, 144401. https://doi.org/10.1103/PhysRevB.105.144401

XI. Kubečka, J., Besel, V., Neefjes, I., Knattrup, Y., Kurtén, T., Vehkamäki, H., & Elm, J. (2023). Computational Tools for Handling Molecular Clusters: Configurational Sampling, Storage, Analysis, and Machine Learning. *ACS Omega*, *8*(47), 45115–45128. https://doi.org/10.1021/acsomega.3c07412

1 Introduction

The Earth is surrounded by an atmosphere that extends up to 10,000 km from the surface and makes life possible. The atmosphere is mostly composed of gases, but also contains liquid and solid particles. A suspension of particles in a gas is called an aerosol, while the individual particles are referred to as aerosol particles. From the mist that causes morning traffic jams to the sea spray that adds volume to our hair after a day at the beach, aerosol particles are ubiquitous in our daily lives. Aerosol particles, however, also influence the climate in two ways—directly, by scattering incoming sunlight, and indirectly, by serving as nuclei for cloud formation (Haywood & Boucher, 2000). By preventing sunlight from reaching the Earth's surface, aerosol particles have a net cooling effect on the climate. The most recent assessment reports of the Intergovernmental Panel on Climate Change (IPCC) state that clouds and aerosol particles are responsible for the largest uncertainty in estimations and interpretations of the Earth's changing energy budget (Boucher et al., 2013; Chen et al., 2021).

Part of the aerosol particles present in the atmosphere are emitted directly as particles from a variety of sources: sea spray, volcanic ash, dust from deserts or other soils, anthropogenic sources, such as fine particles from fossil fuel combustion, and biogenic sources, such as pollen from trees (Penner et al., 2001). These are called primary aerosol particles. Most atmospheric aerosol particles are, however, formed in the atmosphere by the conversion of gas-phase molecules into solid or liquid particles through a process called new particle formation (NPF; Kulmala et al., 2013). Gas-phase molecules are constantly colliding with each other in the atmosphere when crossing each other's trajectories. In NPF, certain gas-phase molecules can stick together after collision to form atmospheric clusters. These clusters can continue to grow, by addition of further gas-phase molecules and other atmospheric clusters to form aerosol particles (Zhang et al., 2012). Aerosol particles created through NPF are called secondary aerosol particles. The exact border between atmospheric cluster and aerosol particle is ambiguous. Generally, atmospheric clusters contain at the most several tens of molecules, while aerosol particles are generally larger than 2 nm in diameter (Elm et al., 2020). From a physical perspective, aerosol particles can be adequately approximated as a solid or liquid bulk separated from the gas phase by a surface. Their physical properties are, therefore, a monotonic function of the number of molecules in the particle (Smirnov, 2012). For atmospheric clusters, on the other hand, there is no clear solid or liquid bulk yet.

NPF occurs all the time across the entire world (Kerminen et al., 2018). Not every gas-phase molecule can, however, take part in NPF at atmospheric conditions. Water vapor, for instance, does not form liquid droplets by itself through NPF at atmospheric temperature and pressure. The formation of atmospheric clusters requires gas-phase molecules that have a tendency to leave the gas phase, e.g., molecules with a low saturation vapor pressure. In continental regions, it has been found that sulfuric acid plays a crucial role in NPF (Sipilä et al., 2010). Atmospheric sulfuric acid concentrations are generally in the range of 10^6 to 10^8 molecules per cubic centimeter (cm^{-3}) (Beck et al., 2022; Yin et al., 2022), compared to a concentration of around 10^{19} cm^{-3} for the main component of air, nitrogen (N_2). Sulfuric acid has a low saturation vapor pressure. Clusters consisting solely of sulfuric acid and water, however, cannot explain the particle growth rates measured in the atmosphere (Duplissy et al., 2016; Weber et al., 1996). Other atmospheric gas-phase molecules are needed to stabilize clusters, allowing them to grow further to the size of aerosol particles. As the name implies, sulfuric acid is an acid. As such, it can participate in relatively strong interactions with bases. The atmosphere contains several nitrogen-containing bases, such as ammonia, methylamine, dimethylamine, and trimethylamine, at concentrations large enough to participate in NPF (Kurtén et al., 2008). Their ability to enhance NPF depends on several aspects, such as concentration, basicity, stereochemistry, and the capacity to form hydrogen bonds (Harold et al., 2022; Olenius et al., 2017). The stronger the base, the better it stabilizes clusters. The addition of alkyl groups to the nitrogen increases its basicity. Therefore, even though ammonia has the highest atmospheric concentration, it is a weak base, lessening its contribution to NPF. Trimethylamine is the strongest base of the amines, but its methyl groups cause sterical hindrance and thus suppress particle formation. Dimethylamine is the happy medium, being a relatively strong base while not suppressing particle formation through too much sterical hindrance. Dimethylamine likely has a considerable contribution to sulfuric acid-based NPF (**Paper IV**; Almeida et al., 2013; Myllys et al., 2019).

A significant fraction of aerosol particulate mass consists of organic matter (Jimenez et al., 2009). It has, furthermore, been estimated that the majority of these organic aerosol particles are created through NPF (Hallquist et al., 2009). Some of the most abundant gas-phase organic molecules in the atmosphere are volatile organic compounds (VOCs), emitted from both anthropogenic and biogenic sources (Williams & Koppmann, 2007). As the name indicates, VOCs are volatile and, therefore, tend to stay in the gas phase. VOCs can, however, be oxidized to form less-volatile vapors. Recent studies have

found that certain VOCs can under specific circumstances auto-oxidate to form very-low-volatile compounds, called highly oxygenated organic molecules (HOMs; Crounse et al., 2013; Rissanen, 2021). HOMs have also been proposed to play a role in sulfuric acid-based NPF (Bianchi et al., 2019).

While sulfuric acid is a crucial component of continental NPF and most research has been focused on studying sulfuric acid-based NPF, there are other vapors that can form particles under atmospheric conditions, such as methanesulfonic acid (Dawson et al., 2012), nitric acid (Wang et al., 2020), and iodine-containing compounds (Chen et al., 2006). The latter are known to play a role in marine and coastal regions. Iodine (I_2) molecules emitted from the ocean surface and algae can undergo oxidation reactions to form various iodine-containing compounds, which are in turn able to form atmospheric clusters with each other (Finkenzeller et al., 2023; Gómez Martín et al., 2022; O'Dowd et al., 2002). Especially in areas with low pollution, iodine-based NPF can compete with sulfuric acid-based NPF, and therefore have a significant contribution to the particulate mass in the global atmosphere (He et al., 2021).

Besides their climate impact, aerosol particles are also important in the context of air pollution and human health. Especially aerosol particles under 10 μm can penetrate deep in the respiratory system and cause a variety of pulmonary diseases (Gan et al., 2013). Aerosol particles can, furthermore, transport bacteria, viruses, and fungal spores (Falcon-Rodriguez et al., 2016). According to the World Health Organization (WHO), around 8 million premature deaths occur annually worldwide due to air pollution (WHO, 2016).

The effect of aerosol particles on the climate and health is highly size-dependent (Valavanidis et al., 2008). Aerosol particles can be conceptually divided into different modes with distinct sources and properties (AQEG, 2005; Dusek et al., 2006; Jacobson, 2012). Nucleation mode particles are under 10 nm in diameter and mostly originate directly from NPF (Seinfeld & Pandis, 2006). Particles usually rapidly grow out of the nucleation mode by condensation of additional vapor onto the nucleation mode particles. Particles between 10–100 nm belong to the Aitken mode. Aitken mode particles continue to grow through condensation of vapor, as well as coagulation with other particles. Particles between 100 nm to 2.5 μm belong to the accumulation mode. Accumulation mode particles do not easily grow further through either condensation or coagulation due to reduced Brownian motion, diffusion barriers, and relatively low number concentrations of accumulation mode particles (AQEG, 2005; Jacobson, 2012; Kerminen &

Wexler, 1995). At the same time, they are still small enough to not settle due to gravitation. As a result, particles originally formed through NPF tend to accumulate in this mode. Accumulation mode particles have a relatively long atmospheric lifetime, and can function as effective cloud condensation nuclei, allowing cloud droplets to form by condensation of water vapor on the particle surface. Accumulation mode particles are eventually lost through precipitation. Particles with a diameter greater than 2.5 μm belong to the coarse mode. The coarse mode predominantly contains primary aerosol particles. Due to their large size, they are generally lost from the atmosphere through gravitational settling.

Atmospheric aerosol particles play an important role in our daily lives, our health, and the climate. Many aspects of aerosol particle formation are, however, still unknown. In particular, we still do not fully understand which compounds are responsible for the particle concentrations measured in the atmosphere at different places around the world.

1.1 Ions in New Particle Formation

While the vast majority of compounds in the atmosphere are neutral, there is also a small fraction of charged molecules and atoms, called air ions, as well as charged aerosol particles. Primary air ions, such as N_2^+, O_2^+, NO^+, O^-, and O_2^- are initially formed from a variety of sources, including UV radiation, decay of radioactive elements in the Earth's soil, cosmic rays, interactions between neutral gas molecules and naturally occurring radon gas, corona discharge, waterfalls through the Lenard effect, and thunderstorms (Eisele, 1989; Hirsikko et al., 2011; Israël, 1970; Li, Guo, et al., 2015; Yin et al., 2023). Of these sources, cosmic radiation and radioactive decay are the most prominent in the boundary layer (Harrison & Tammet, 2008), i.e., the lowest part of the atmosphere whose properties are defined by its interactions with the Earth's surface. Cosmic rays consist of charged elementary particles (mostly protons) that impinge on the Earth from extraterrestrial origins (Bazilevskaya et al., 2008). The Earth's soil contains small amounts of radioactive material, such as uranium and radium, that decay over time, emitting radon gas and alpha, beta, and gamma rays able to ionize atmospheric molecules (Chen et al., 2016). Once formed, primary ions can undergo quick ion–molecule reactions and clustering to form small ions (charged clusters with a diameter under 1.6 nm) and eventually charged aerosol particles (over 1.6 nm in

diameter; Luts and Parts, 2002). At any time, there are on average 300–400 small ions per cubic centimeter in the troposphere (Hirsikko et al., 2011; Li, Guo, et al., 2015). The concentration of small ions and charged aerosol particles can, however, increase to several thousands cm^{-3} in the vicinity of sources, such as waterfalls or overhead power lines (Fews et al., 1999; Laakso et al., 2007). Ion pairs are produced at an average rate of 1–10 $cm^{-3}s^{-1}$ at a height of 1 m above the Earth's surface. The ion pair production rate decreases with altitude to a height of 1–2 km, and then increases to a maximum around a height of 15 km (Laakso et al., 2004).

Ions have relatively strong long-range attractive interactions with other molecules. It is, therefore, expected that ions have an increased probability of collisions in the atmosphere. These relatively strong attractive interactions can also assist in keeping the colliding compounds together after collision, and thus in stabilizing clusters. Atmospheric observations, laboratory experiments, and theoretical models have shown that ions have the capability to enhance both cluster formation and their condensational growth (Kirkby et al., 2011; Lehtipalo et al., 2016). There is, however, a lively debate on the extent to which ion-induced NPF plays a role in the global aerosol particle production in the continental boundary layer (Li, Jiang, et al., 2015). Various studies disagree on the fraction of particle formation events that can be attributed to ion-induced pathways.

Several atmospheric observation studies report a contribution of ions to NPF of less than 10% (Gagné et al., 2008; Iida et al., 2006; Kulmala et al., 2010; Manninen et al., 2009). Laboratory experiments further show that a mixture of sulfuric acid and dimethylamine at atmospheric concentrations is sufficient to explain observed particle formation rates (Almeida et al., 2013). However, Yu (2010) claims that these studies make assumptions that result in an underestimation of the importance of ion-induced NPF. Ions have been shown to participate in many boundary layer formation events (Hirsikko et al., 2007). Furthermore, ion-induced NPF models are able to accurately reproduce observed formation events (Yu & Turco, 2008; Yu, 2010). It is thus clear that further studies are necessary to resolve this debate.

Regardless of this debate, other studies have shown that ions likely play a significant role in NPF in pristine environments, such as Antarctica, where particle-forming vapors, like sulfuric acid and amines, have low concentrations compared to sites closer to human habitation and industries (Jokinen et al., 2018; Kirkby et al., 2016). NPF of iodine-containing compounds in marine and coastal regions can also take place through

ion-induced pathways (He et al., 2021). The exact contribution of ions to NPF is, therefore, an important research subject.

1.2 Ions in the detection of atmospheric clusters

To study all aspects of particle formation and their subsequent growth, a daunting range of sizes has to be considered. The participating vapor molecules can have diameters of less than 0.5 nm, while the larger coarse particles go up to sizes of 10 μm. As different measurement techniques have different lower and upper size limits, a variety of instruments have to be used to cover the entire size range. For instance, differential mobility particle sizers are typically used in the size range of 10–800 nm, while particle size magnifiers can be used for particles between 1–3 nm (Aalto et al., 2001; Vanhanen et al., 2011; Wiedensohler et al., 2018).

Mass spectrometry is uniquely suited to measure the earliest stages of NPF, where the first gas-phase molecules collide and stick together to form atmospheric clusters around 1–2 nm. In all mass spectrometers, charged compounds are separated based on their mass-to-charge ratio (Dass, 2007). There are a variety of different mass selection techniques. A straightforward mass selection technique is time-of-flight mass spectrometry. Here, charged compounds are accelerated in vacuum in an electric field to a detector. The time it takes for each compound to reach the detector is directly related to its mass-to-charge ratio. Compounds can thus be identified based on their mass.

As mass spectrometers separate compounds based on their mass-to-charge ratio, they can only be used for ionic compounds. Charged atmospheric clusters can, therefore, be directly measured. Neutral clusters created through electrically neutral NPF pathways, or recombination of positively and negatively charged ionic compounds, however, need to be artificially charged before mass spectrometry can be employed. There are a variety of charging techniques that are separated into hard and soft methods (de Hoffmann & Stroobant, 2007). An example of a hard charging method is electron ionization. In electron ionization, the sample is bombarded by highly energetic electrons. This usually not only charges the neutral compounds in the sample but also results in their fragmentation. When analyzing single molecules, this fragmentation can be desirable. The charged fragments can provide information on the functional groups and connectivity in the studied molecule. However, when we want to determine the identity and amount of each compound in a cluster, fragmentation is not desirable.

Fragmentation leads to loss of neutral molecules from the cluster. The extent of this loss cannot be measured. It is, therefore, important to use a soft charging technique. One of the softest charging techniques is chemical ionization, where a charging reagent is introduced to the sample. The charging reagent either undergoes a charge transfer (e.g., a proton transfer), or clusters with the compounds in the sample. For instance, when analyzing clusters containing sulfuric acid, the nitrate ion is an often-used charging reagent (Eisele & Tanner, 1993). Nitrate attaches itself to the cluster, giving it a negative charge, which allows it to be detected by mass spectrometers. Despite the softness of the chemical ionization technique, the addition of the charge still fundamentally changes the stability of the cluster, and evaporation of molecules from the cluster will likely occur (Ortega et al., 2014).

To accurately separate compounds on their mass-to-charge ratio, the mass selection needs to occur in vacuum. Otherwise, collisions with background gas will alter the detection times. This is problematic when we want to take samples directly from the atmosphere. To go from atmospheric pressure at the inlet to vacuum at the mass selector, an atmospheric pressure interface (APi) can be used (Junninen et al., 2010). An APi consists of several (e.g., three) sections that are maintained at progressively lower pressures by pumps. The charged compounds move through the APi in an electric field. The combination of accelerating charged atmospheric clusters in an electric field and the presence of some remaining background gas can lead to analyte–background gas collisions, resulting again in evaporation of molecules from the clusters (Passananti et al., 2019).

Evaporation of the atmospheric clusters after chemical ionization or by collisions with background gas in the APi results in detection of a different cluster than what was originally in the studied sample. To accurately understand the distribution of atmospheric clusters, freshly formed through NPF, we need to be able to map the cluster that results from the evaporation, incurred through chemical ionization and inside the APi, to the cluster originally present in the atmosphere.

As mass spectrometers measure the mass-to-charge ratio, they are fundamentally incapable of distinguishing between isomers. This is problematic, as there are many isomers among the atmospheric molecules that potentially play a part in NPF, and each isomer usually has different physicochemical properties and distinct biological responses. One option to deal with this issue is to measure a second property, alongside the mass-to-charge ratio, that hopefully differs between isomers. A mass spectrometer can, for

instance, be coupled to an ion mobility spectrometer to additionally measure the ion mobility (Borsdorf & Eiceman, 2006). The ion mobility of a charged compound is a measure of its ability to move through a medium under the influence of an electric field. As the ion mobility is highly dependent on the geometry of the charged compound, it can differ for, especially constitutional, isomers.

Both mass and ion mobility spectrometry rely on ionization methods to measure neutral compounds in a sample. It is important to understand how this ionization process exactly occurs and what its efficiency is. As ionization and collisions in the APi can destabilize clusters, we, furthermore, need to understand how clusters change after charging.

1.3 Modeling the role of ions in the formation and detection of atmospheric clusters

Atmospheric observations of, for instance, meteorological variables and concentrations of atmospheric compounds, allow us to record the specific physical and chemical properties of the atmosphere and how they change over time. Correlations between observations can indicate possible co-dependencies. Experiments in controlled environments help to pinpoint the mechanisms behind atmospheric observations. Theoretical models finally provide a simplified framework to explain the results of atmospheric observations and experiments, and allow us to make predictions for the future. Thus, observations, experiments, and modeling work hand-in-hand.

In atmospheric science, models are utilized over a great range of scales, from climate models predicting global climate changes decades into the future, to models at the atomistic scale that explain how various atmospheric molecules interact with each other to produce the macro-scale observations that we perceive. This thesis is focused on atomistic-scale modeling.

While extensive theoretical research has been done on neutral clusters in the atmosphere, atmospheric ionic compounds have not received similar theoretical attention. The studies in this thesis were, therefore, performed to increase our theoretical understanding of how ionic compounds (1) participate in NPF, (2) are formed from neutral compounds through chemical ionization, and (3) are separated based on ion mobility. The tools developed in these studies can be used by other researchers to create more

accurate models at a larger scale, and to better interpret the results from experimental measurements.

The initial step of NPF is a collision between two gas-phase molecules or ions to form a stable dimer. The kinetics of these collisions can be characterized by a collision rate coefficient. Halonen et al. (2019) showed for neutral molecules (specifically for the collision between two sulfuric acid molecules), that the conventional theoretical models, based on kinetic gas theory, underestimate the collision rate coefficient by a factor of more than two. In **Paper II**, we similarly compared several often-used theoretical models for obtaining the rate coefficient of ion–dipole collisions to atomistic molecular dynamics (MD) collision simulations. We, furthermore, developed a new workflow, using interaction parameters obtained from the potential of mean force (PMF) along the center-of-mass distance between the collision partners within the central field approach to obtain accurate collision rate coefficients at low computational costs.

In **Paper III**, we present an application of our central field method with parameters fitted to the PMF. Experimental measurements of chemical ionization rate coefficients have shown that sulfuric acid has a twice higher rate coefficient with acetate than with nitrate. Although, we do not find a similarly high enhancement through our theoretical calculations, we do find that the sulfuric acid–nitrate rate coefficient is significantly enhanced over what would be expected for pure averaged ion–dipole interactions.

In the central field approach, both collision partners are approximated as point particles. Neglecting the extended size of the collision system can lead to unphysical results in several specific situations. For clusters, it is, furthermore, not straightforward to calculate the interaction parameters needed in the central field approach. In **Paper VI**, we developed an extension to the central field method, named the interacting hard-sphere model, where we introduce an additional collision criterion based on the sum of hard-sphere radii and employ Hamaker's approach (Hamaker, 1937) to calculate interaction parameters for clusters of any size. While the paper focuses on neutral compounds, the interacting hard-sphere model can easily be applied to collisions involving ionic compounds as well.

Special attention needs to be paid when setting up MD simulations in the gas phase. Conventional thermostatting practices for simulations in liquids and solids can produce unphysical distributions of energy between the different internal degrees of freedom of individual polyatomic molecules in the gas phase. These unphysical energy distribu-

tions can be easily overlooked as they are not apparent from the standard MD output, and can lead to systematic errors in the simulation of gas-phase collisions. In **Paper V**, we detail several problems that can occur when thermostats are employed carelessly in gas-phase MD simulations.

Even if the ion mobility of a set of isomers is measured, and they turn out to be significantly different from each other, then we are still not able to assign the ion mobility signals to specific isomers. To do this, we would need literature values of the ion mobility of each of these isomers to compare with the values obtained from our sample. Unfortunately, literature values are not available for many atmospheric compounds. In **Paper I**, we, therefore, tested if computational ion mobility predictions can be used to help distinguish between isomers. Computational ion mobility predictions of three dihydroxybenzene constitutional isomers were calculated and compared to experimental values. While a direct comparison proved impossible, the experimental ion mobility order was preserved in the modeled results for this specific system.

To quantify the role of ions in NPF, we can use dynamical models to calculate the formation rate of clusters with and without ionic pathways. The accuracy of the modeled formation rate, however, depends on the largest clusters included in the model. The main bottleneck for including larger cluster sizes is the high computational cost associated with calculating accurate binding free energies from quantum chemistry. In **Paper IV**, we used machine learning to lower the computational cost of finding the binding free energies of large atmospheric clusters. These binding energies were subsequently used to analyze the dependence of the modeled cluster formation rate on the size of the largest cluster included in the model.

The remainder of this thesis is organized as follows: in section 2, we discuss the theoretical and computational methods employed in this thesis. An overview of the papers in this thesis and the author's contribution to each paper is given in section 3. The results of each study are discussed in section 4. Section 5 concludes the thesis with an outlook toward the future.

2 Theoretical and computational methods

In the following 10 subsections, the various theoretical and computational methods used in this thesis will be discussed.

We can determine the relative importance of ion-induced new particle formation (NPF) pathways to the overall atmospheric aerosol particle population by comparing their particle formation rates $\mathcal{J}$ to those of relevant neutral pathways. **Classical nucleation theory** (CNT; subsection 2.1) provides a theoretical framework to study particle formation kinetics. Starting from birth–death equations, CNT employs various approximations to derive an analytical particle formation rate prediction. In many instances, the approximations made in CNT lead to significant discrepancies with experimental particle formation rates. More accurate particle formation rates can be obtained by numerically solving the birth–death equations as is done in the **Atmospheric Cluster Dynamics Code** (ACDC; subsection 2.2). The accuracy of ACDC's results depends on the accuracy of the collision and evaporation rate coefficients needed for the birth–death equations.

Evaporation rate coefficients can be derived from cluster binding free energies. Binding free energies are typically obtained from **quantum chemistry** (subsection 2.3) calculations. Contrary to classical theory, where the smallest chemical building block is the atom, quantum chemistry applies the rules of quantum mechanics to calculate the interactions between the system's electrons and nuclei. While computationally expensive, quantum chemistry provides very accurate predictions of physical and chemical properties.

As chemical systems tend towards the lowest free energy configuration, we should calculate the binding free energy based on this configuration. Finding the lowest free energy cluster configuration (or preferably all reasonably populated cluster configurations) is, however, challenging. **Configurational sampling** (subsection 2.4) is the process of efficiently finding low free energy configurations.

Collision rate coefficients can be calculated through a variety of different methods. **Collision theory** (subsection 2.5) encompasses several theoretical methods to predict collision rate coefficients. As with CNT, these theoretical methods make approximations that can potentially lead to significant errors in the predicted collision rate coefficients. Collision rate coefficients can also be estimated from **molecular dynamics**

(MD; subsection 2.6) simulations. In MD, the studied system is evolved over discrete time steps. As such, MD provides a method for performing experiments in a computer. We can additionally use **potential of mean force (PMF) calculations** (subsection 2.7) to obtain the energy from MD simulations along a specific system property, e.g., the center-of-mass distance between two collision partners. This energy profile defines the collision trajectory and, therefore, the collision rate coefficient. While PMF and MD simulations can be run using quantum chemistry, the computational cost is typically prohibitive. Thus, the MD and PMF simulations in this thesis have been performed using a classical **force field** (subsection 2.8) description.

To identify peaks in ion mobility spectra, we perform **ion mobility modeling** (subsection 2.9). Ion mobility modeling should be done on the most populated, and therefore lowest free energy, configurations. For neutral molecules charged by clustering to a charging ion in chemical ionization, we employ configurational sampling and quantum chemistry to find the lowest free energy configuration of the analyte–charging ion cluster.

Machine learning (subsection 2.10) is used to find functional relationships between input and output of existing data, so that these functional relationships can be used to predict outputs for new inputs. For computationally demanding methods, such as quantum chemistry, this can result in a significant reduction of the computational cost, allowing for more complex calculations.

2.1 Classical nucleation theory

NPF is an example of a phase transition from the gas phase to either the liquid or solid phase. Phases are physically distinct regions in a chemical system, distinguished by the degree of order between the compounds in these regions (Tester & Modell, 1997). According to thermodynamics, the state with the lowest free energy, given the prevalent ambient conditions (e.g., temperature and pressure) and chemical properties of the compounds in the system, is most stable (Atkins & de Paula, 2014). Free energy is a measure for the ability of a system to cause change. Once in the lowest free energy phase, the system, therefore, has the least ability to undergo a phase transition. However, changes in the conditions (e.g., temperature increase or changes in concentrations of the components) can result in another phase becoming more stable. While thermodynamics indicates that the system will now tend towards the new minimum,

the phase transition doesn't necessarily occur instantaneously. The kinetics of the system dictate how fast a phase transition occurs. Even though the new phase is more stable, it might require a significant reorganisation of the atoms or molecules in the system, requiring free energy. In this case, there is a free energy barrier between the two states. Even if the original phase is now higher in free energy, it can still persist as a so-called metastable state.

In the case of NPF, the phase transition occurs because the solid or liquid phase is more stable for the particle-forming vapors. If the addition of each new compound to the particle has no energy barrier, e.g., growth far outweighs evaporation at every particle size, then the particle formation rate is dictated by the collision rate. Often, however, the smallest clusters are more likely to evaporate again than to grow further. In this case, the particle formation is associated with a formation free energy barrier, and we speak of nucleation. Although the term "nucleation" is often used for phase transitions with and without formation free energy barriers, it, strictly speaking, only applies to the former case.

Classical nucleation theory (CNT) is the most common theoretical framework to study the thermodynamics and kinetics of nucleation (Vehkamäki, 2006). Like any theoretical model, CNT relies on significant approximations to derive analytical expressions for the nucleation rate. To obtain an analytical expression for the free energy barrier of cluster formation, CNT, for instance, assumes that the new phase is an ideal spherical droplet with bulk liquid properties. Using these assumptions, the free energy of droplet formation ΔG in a one-component system as a function of the number of molecules in the droplet N_d is given by

$$\Delta G = -N_\mathrm{d} k_\mathrm{B} T_\mathrm{amb} \ln S + \varsigma \mathcal{A}, \tag{1}$$

where k_B is the Boltzmann constant, T_amb the ambient temperature, S the saturation ratio, ς the surface tension, and $\mathcal{A}$ the surface area. The first and second term in the equation relate to the formation of the bulk volume and surface area of the droplet, respectively (Fig. 1). Surface formation always increases the free energy of the system, while the sign of the bulk volume term depends on the saturation ratio. If the system is supersaturated ($S > 1$) the bulk term is negative. As small clusters have a comparatively large surface area per bulk volume, the initial cluster growth is dominated by the surface area term and, therefore, thermodynamically unfavorable. At larger cluster sizes, the bulk volume term becomes increasingly pronounced, which eventually causes

the free energy to decrease with each increase in cluster size. The transition between these two regimes is the critical cluster size, which corresponds to the maximum of the formation free energy curve plotted as a function of the number of molecules in the cluster. Once the critical cluster size is reached, growth starts to outweigh evaporation. Below the critical cluster size, growth occurs stochastically through a sequence of rare events where compounds are added to the cluster before it evaporates.

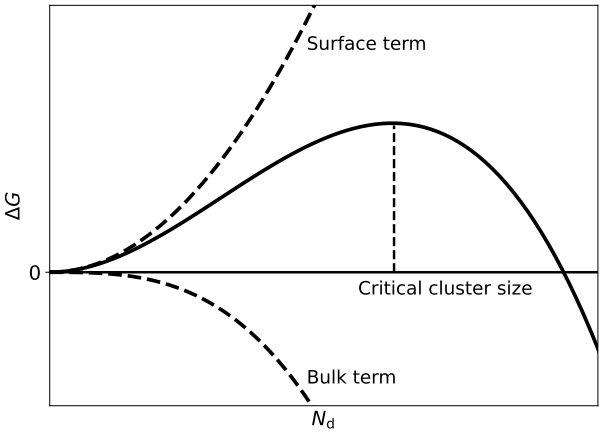


Figure 1: The free energy of droplet formation ΔG in a one-component system as a function of the number of molecules in the droplet N_d (solid line) according to classical nucleation theory. The bulk term ($-N_\mathrm{d} k_\mathrm{B} T_\mathrm{amb} \ln S$) and surface term ($\varsigma \mathcal{A}$) are also indicated separately (dashed lines). The N_d value where ΔG has a maximum corresponds to the critical cluster size.

The kinetics of nucleation can be modeled through birth–death equations. These equations consider the various ways in which the concentration of a cluster with a specific composition can change.

$$\frac{\mathrm{d}\mathcal{C}_i}{\mathrm{d}t} = \sum_{j<i} \beta_{j,(i-j)} \mathcal{C}_j \mathcal{C}_{(i-j)} + \sum_j \gamma_{(i+j)\to i} \mathcal{C}_{i+j} - \sum_j \beta_{i,j} \mathcal{C}_i \mathcal{C}_j - \sum_{j<i} \gamma_{i\to j} \mathcal{C}_i + \mathcal{Q}_i + \mathcal{L}_i, \quad (2)$$

where $\mathcal{C}_i$ is the concentration of cluster i, $\beta_{i,j}$ the collision rate coefficient between cluster/monomer i and j, $\gamma_{i\to j}$ the evaporation rate coefficient of cluster i fragmenting into clusters/monomers i and $i-j$, $\mathcal{Q}_i$ external sources of cluster i, and $\mathcal{L}_i$ all other losses

for cluster i. The collision and evaporation rate coefficients are measures for how often the clusters in question collide and evaporate, respectively. Analytical equations and parameterizations to estimate the collision rate coefficients are discussed in subsection 2.5. An analytical expression for the evaporation rate coefficient can be obtained by assuming that the evaporation rate coefficient does not significantly change compared to equilibrium conditions. At equilibrium, the evaporation of a cluster $i + j$ into fragments i and j is equal to the sticking collisions of i and j to form $i+j$. The evaporation rate coefficient is then obtained from a detailed balance

$$\gamma_{(i+j)\to i} = \beta_{i,j} \frac{\mathcal{C}_i^{\mathrm{eq}} \mathcal{C}_j^{\mathrm{eq}}}{\mathcal{C}_{i+j}^{\mathrm{eq}}} = \beta_{i,j} \mathcal{C}_{\mathrm{ref}} \exp\left\{ \frac{\Delta G_{i+j} - \Delta G_i - \Delta G_j}{k_{\mathrm{B}} T} \right\}, \tag{3}$$

where $\mathcal{C}_i^{\mathrm{eq}}$ is the equilibrium concentration of cluster i and $\mathcal{C}_{\mathrm{ref}}$ is the monomer concentration of a reference vapor. In CNT, it is assumed that the cluster size only changes by the addition or evaporation of monomers (i.e., there are no cluster–cluster collisions or evaporations). External sources $\mathcal{Q}_i$ and losses outside of evaporations $\mathcal{L}_i$ are also neglected. With these assumptions, an analytical equation can be obtained for the steady-state nucleation rate. The analytical CNT nucleation rate equations are able to accurately predict experimental nucleation rates for some specific systems, such as water around ∼260 K (Wölk & Strey, 2001). CNT, however, shows significant deviations from experiments for many systems of interest in atmospheric sciences. This is typically caused by a failure of one or several of the assumptions made in CNT (Merikanto et al., 2007). The liquid droplet model, where clusters are approximated by a liquid bulk separated from the gas phase by a surface with bulk surface tension, breaks down for the smallest clusters. In the atmosphere, external sources and coagulation losses cannot be neglected. In experiments, wall losses constitute a significant sink of clusters. Lastly, compounds such as sulfuric acid tend to form dimers at high concentrations (e.g., in highly polluted cities; Yin et al., 2022). The nucleation of sulfuric acid can, therefore, not be correctly approximated by monomer addition alone.

2.2 Atmospheric Cluster Dynamics Code

The Atmospheric Cluster Dynamics Code (ACDC) is a computer program that numerically solves the birth–death equations given by Eq. (2) to obtain cluster formation rates (McGrath et al., 2012). To run ACDC, we need to define a simulation scheme, i.e., the set of clusters for which the birth–death equations should be solved. For a

nucleating system, there will be a flux from the monomers and smaller clusters to the larger clusters in the simulation scheme. The user has to determine what happens when there is flux beyond the simulation scheme, i.e., the system grows larger than the set of modeled clusters. Clusters outside of the simulation scheme can either be made to instantaneously evaporate back into the simulation scheme through monomer loss, or be removed from the simulation. In the latter case, the cluster is assumed to grow further with negligible evaporation. The cluster formation rate is determined as the sum of the fluxes through the different pathways where the system is allowed to grow out of the simulation scheme. To obtain reasonable simulation results, it is important to choose a suitable simulation scheme and outgrowth criteria. Preferably, the critical cluster size is included in the simulation scheme, as CNT shows that beyond the critical cluster size, the evaporation rate is vastly outweighed by the growth rate, making the outgrowth assumption of no evaporation back into the simulation scheme reasonable.

We can, for example, consider an ACDC calculation of clusters involving sulfuric acid (H_2SO_4) and dimethylamine ($(CH_3)_2NH$). These clusters mostly grow through the successive addition of one acid and one base monomer. If the 4–4 cluster (4 H_2SO_4 and 4 $(CH_3)_2NH$) is the largest cluster in our simulation scheme, then the 5–4 and 4–5 clusters are both reasonable outgrowth criteria (ACDC automatically also considers larger clusters formed through cluster–cluster collisions). We could make a square simulation scheme, including all clusters with 0 to 4 H_2SO_4 and 0 to 4 $(CH_3)_2NH$ (Fig. 2). In this case, fluxes leading to clusters such as 5–1 should evaporate back into the simulation scheme, as such clusters are unstable and hardly contribute to cluster growth. We could likely even leave clusters such as the 4–1 and 1–4 clusters out of our simulation scheme without significantly impacting the resulting cluster formation rate.

ACDC can consider both neutral and ionic cluster formation pathways. The initial concentrations of all monomers and clusters in the simulation scheme need to be set. It is common to assume either a constant concentration of monomers or a source producing monomers at a fixed rate. Additionally, several loss terms can be defined relating to different scenarios (e.g., wall loss in chamber experiments, or loss through coagulation on pre-existing particles).

As input, we need to supply the binding free energies of all the clusters in the simulation scheme. These are used to get the evaporation rate coefficients through Eq. (3). As the binding free energies appear in an exponent, even small errors lead to significant errors in the evaporation rate coefficients. Binding energies are, therefore, typically calculated

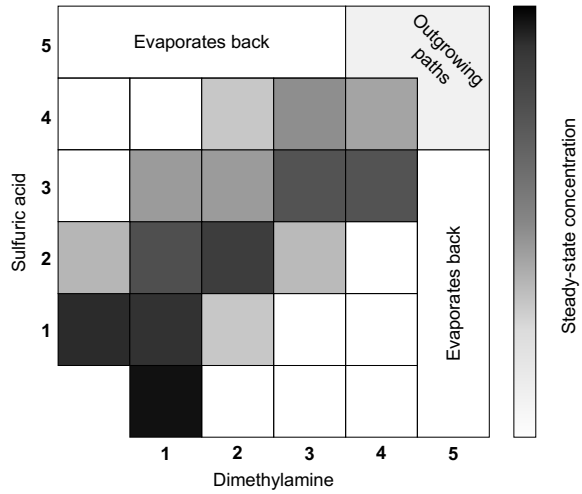


Figure 2: An example of a simulation scheme consisting of 0 to 4 sulfuric acid and 0 to 4 dimethylamine molecules. The grayscale indicates the relative steady-state concentrations of each cluster in the simulation scheme. Only the clusters along the diagonal have significant steady-state concentrations. Figure adapted from McGrath et al. (2012).

through high-quality quantum chemistry calculations. Collision rate coefficients in ACDC are by default calculated using kinetic theory (Eq. (52)) for neutral species and the Su & Chesnavich parameterization (Eq. (53)) for charged species, but can also be defined manually.

2.3 Quantum chemistry

The properties of a chemical compound, such as reactivity, energy, and dipole moment, are determined by the distribution of the electrons around the nuclei. Classical theory fails to correctly describe the behavior of these subatomic particles. The theory of quantum mechanics, on the other hand, has had great success in making accurate predictions at the atomic scale. Quantum mechanics is based on two fundamental concepts, quantization and wave–particle duality. Quantization states that dynamical variables can only take on certain discrete values. According to the wave–particle du-

ality, quantum entities, such as electrons, protons, and neutrons, exhibit both particle and wave properties. These concepts form six postulates (McQuarrie, 2008; Sherrill, 2003):

1. The state of a system composed of quantum entities (e.g., a molecule) is completely described by a function $\Psi(\{\boldsymbol{\tau}\}, t)$, called the wavefunction. Here, $\{\boldsymbol{\tau}\}$ stands for the spatial coordinates of all particles in the system, as well as their spin, an intrinsic form of angular momentum present in quantum entities.

2. For every observable O in classical mechanics (e.g., momentum), there exists a corresponding linear Hermitian operator $\hat{O}$ in quantum mechanics.

3. Upon measurement of the observable associated with the operator $\hat{O}$, the wavefunction 'collapses' into an eigenstate Ψ_i of $\hat{O}$. We will, therefore, only ever observe the eigenvalues o_i that satisfy $\hat{O}\Psi_i = o_i\Psi_i$. Before the measurement, the system does not have to be in this specific eigenstate of $\hat{O}$. Any arbitrary state can, however, be expanded in the complete set of eigenvectors of $\hat{O}$, $\Psi = \sum_i^n c_i\Psi_i$.

4. If a system is in a state described by a normalized wavefunction, then the average value of property O over measurements on an ensemble of identically-prepared systems $\langle O\rangle$ is given by $\langle O\rangle = \int_{-\infty}^{\infty} \Psi^*\hat{O}\Psi \, \mathrm{d}\boldsymbol{\tau}$.

5. The time evolution of the wavefunction is given by the time-dependent Schrödinger equation

$$\hat{H}\Psi = i\hbar\frac{\partial\Psi}{\partial t}, \tag{4}$$

   where $\hat{H}$ is the Hamiltonian operator corresponding to the energy of the system and $\hbar = h/2\pi$ the reduced Planck constant.

6. The wavefunction must be antisymmetric for fermions. Fermions are quantum entities with half-integer spin. Electrons, protons, and neutrons are all fermions. An antisymmetric wavefunction changes sign when the spatial or spin coordinates of two identical particles are interchanged.

Quantization is a result of postulates 3 and 4. The energy E of a system can, for instance, not assume just any value, but only an eigenvalue of the Hamiltonian energy operator $\hat{H}$ applied to the system's wavefunction Ψ. The particle–wave duality

is contained in the fact that a wavefunction is used to represent a system of elementary particles. According to Born's statistical interpretation of quantum mechanics, $|\Psi(\mathbf{r}, s, t)|^2$ represents the probability of finding a particle at position $\mathbf{r}$ with spin s at time t (Born, 1926). Before measuring a quantum system (upon which the wavefunction collapses), its properties are not fixed. Rather, the properties of the system exist as a probability distribution defined by its wavefunction.

In the following, we will use bra-ket notation, defined by

$$|\Psi\rangle \equiv \Psi; \langle\Psi| \equiv \Psi^*, \tag{5}$$

$$\int \Psi^* \Psi \, \mathrm{d}\mathbf{r} = \langle\Psi|\Psi\rangle \,, \tag{6}$$

$$\int \Psi^* \hat{H} \Psi \, \mathrm{d}\mathbf{r} = \left\langle \Psi \,\middle|\, \hat{H} \,\middle|\, \Psi \right\rangle, \tag{7}$$

where Ψ^* is the complex conjugate of Ψ.

2.3.1 The Born–Oppenheimer approximation

From the previous section, it is clear that if we want to accurately predict the properties of a compound that depend on the interplay between elementary particles, we need to have an accurate representation of the compound's wavefunction. Often in chemistry, we want to know the properties of a compound in a stationary state, i.e., a state where the properties do not change over time. An example is a single molecule in vacuum. However, even in situations where the properties change over time, these changes generally take place in a much slower time frame than the motion of the electrons in the system. Thus, these problems can also be reduced to stationary states. We are, therefore, able to simplify the time-dependent Schrödinger equation of Eq. (4) to a time-independent variation given by (Bransden & Joachain, 2000)

$$\hat{H}\Psi(\{\boldsymbol{\tau}\}) = E\Psi(\{\boldsymbol{\tau}\}). \tag{8}$$

For a system consisting of nuclei and electrons, the Hamiltonian energy operator $\hat{H}$ in Eq. (8) can be written as a sum over different energy contributions (Jensen, 2007; Sherrill, 2005)

$$\hat{H} = \hat{T}_{\mathrm{N}}(\{\mathbf{r}_{\mathrm{N}}\}) + \hat{T}_{\mathrm{e}}(\{\mathbf{r}_{\mathrm{el}}\}) + \hat{V}_{\mathrm{NN}}(\{\mathbf{r}_{\mathrm{N}}\}) + \hat{V}_{\mathrm{ee}}(\{\mathbf{r}_{\mathrm{el}}\}) + \hat{V}_{\mathrm{Ne}}(\{\mathbf{r}_{\mathrm{N}}\}, \{\mathbf{r}_{\mathrm{el}}\}), \tag{9}$$

where $\hat{T}_N$ and $\hat{T}_e$ are the nuclear and electronic kinetic energy operators, respectively. $\hat{V}_{NN}$, $\hat{V}_{ee}$, and $\hat{V}_{Ne}$ are, respectively, the nucleus–nucleus, electron–electron, and nucleus–electron interaction potential operators. $\{\mathbf{r}_\mathrm{N}\}$ and $\{\mathbf{r}_\mathrm{el}\}$ represent the coordinates of all nuclei and electrons, respectively. Note that the energy terms considered here do not depend on the spin coordinates of the nuclei and electrons.

While we can in theory obtain the wavefunction of any system by solving Eq. (8), the calculations are extremely complex in practice. The time-independent Schrödinger equation is a quantum many-body problem. Analytical solutions can only be obtained for a handful of realistic systems, containing two (e.g., the hydrogen atom) or, given certain assumptions, at the most three particles (e.g., the $H_2{}^+$ molecule). Without approximations, even numerical solutions would be too computationally expensive for all but the smallest of systems.

The Born–Oppenheimer approximation is one of the most important approximations in quantum chemistry (Born & Oppenheimer, 1927). As the velocity of electrons is several orders of magnitude larger than that of nuclei, electrons will almost immediately adapt to any motion of the nuclei and find a new equilibrium. In other words, we do not require any information about the past motion of the nuclei to determine the electron distribution around a specific nuclear configuration. We can thus solve an electronic Schrödinger equation for a system where the nuclei are frozen in a specific configuration

$$\begin{aligned}&\hat{H}_\mathrm{el}\psi_\mathrm{el}(\{\mathbf{r}_\mathrm{el}\};\{\mathbf{r}_\mathrm{N}\}) = \\ &\left[\hat{T}_\mathrm{e}(\{\mathbf{r}_\mathrm{el}\}) + \hat{V}_\mathrm{NN}(\{\mathbf{r}_\mathrm{N}\}) + \hat{V}_\mathrm{ee}(\{\mathbf{r}_\mathrm{el}\}) + \hat{V}_\mathrm{Ne}(\{\mathbf{r}_\mathrm{el}\};\{\mathbf{r}_\mathrm{N}\})\right]\psi_\mathrm{el}(\{\mathbf{r}_\mathrm{el}\};\{\mathbf{r}_\mathrm{N}\}) = \\ &E_\mathrm{el}\psi_\mathrm{el}(\{\mathbf{r}_\mathrm{el}\};\{\mathbf{r}_\mathrm{N}\}),\end{aligned} \tag{10}$$

where the semicolon indicates that $\{\mathbf{r}_\mathrm{N}\}$ are parameters and not variables.

The electronic energy E_el is a function of the nuclear degrees of freedom. For N_N nuclei, there are $3N_\mathrm{N} - 6$ nuclear degrees of freedom ($3N_\mathrm{N} - 5$ for linear systems). The E_el thus spans a $(3N_\mathrm{N} - 6)$-dimensional surface that constrains the motion of the nuclei (Fig. 3). This surface is called the potential energy surface (PES). Minima on the PES correspond to stable equilibrium configurations of the system. If we solve the electronic Schrödinger equation for a specific nuclear configuration, we can also derive the forces on the nuclei as the negative gradient of the electronic energy as a function of the nuclear degrees of freedom. Based on the forces on the nuclei, the system can be optimized in steps toward the closest minimum of the PES. The closest minimum

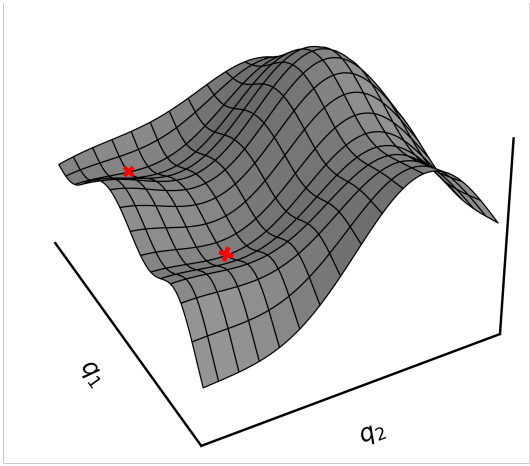


Figure 3: An example of a potential energy surface as a function of two configuration coordinates q_1 and q_2. The minima, indicated by red crosses, correspond to stable equilibrium configurations. Figure adapted from Schlegel (1995).

is not necessarily the global minimum of the PES. To search for the global minimum, we employ configurational sampling (subsection 2.4).

The full wavefunction can be obtained as an expansion of the electronic wavefunction

$$\Psi(\{\mathbf{r}_{\mathrm{el}}\},\{\mathbf{r}_{\mathrm{N}}\}) = \sum_k \psi_{\mathrm{el},m}(\{\mathbf{r}_{\mathrm{el}}\};\{\mathbf{r}_{\mathrm{N}}\})\psi_{\mathrm{N},m}(\{\mathbf{r}_{\mathrm{N}}\}), \tag{11}$$

where the sum runs over electronic states and $\psi_{\mathrm{N}}(\{\mathbf{r}_{\mathrm{N}}\})$ is the nuclear wavefunction. In most quantum chemistry applications, we are only interested in the electronic ground state $\psi_{\mathrm{el},0}$. The Born–Oppenheimer approximation is valid in most situations. It, however, fails in areas where the energies $E_{\mathrm{el},m}$ of two different electronic states are close together.

2.3.2 The Hartree–Fock method

The Born–Oppenheimer approximation reduces the problem of solving the time-independent Schrödinger equation (Eq. (8)) to the problem of solving the electronic Schrödinger equation (Eq. (10)) for specific nuclear configurations. We, however, still

need a method to solve the electronic Schrödinger equation. Hartree–Fock theory provides an approach to find approximate solutions (Szabo & Ostlund, 1996).

To find a reasonable estimation of the ground state wavefunction, Hartree–Fock theory relies on the variational principle. According to the variational principle, the electronic energy obtained from applying the Hamiltonian operator of the studied system on an estimated trial wavefunction $\psi_{\mathrm{el},t}$ will always be higher than the electronic energy of the true electronic ground state $\psi_{\mathrm{el},0}$. In bra-ket notation, this is written as

$$E_{\mathrm{el},0} \leq \left\langle \psi_{\mathrm{el},t} \middle| \hat{H}_{\mathrm{el}} \middle| \psi_{\mathrm{el},t} \right\rangle . \tag{12}$$

In the Hartree–Fock method, the many-body problem of finding the ground state wavefunction is simplified by approximating the ground state wavefunction as a single Slater determinant Φ consisting of one-electron functions $\chi_i(\boldsymbol{\tau}_j)$, also called orbitals:

$$\Phi = \frac{1}{\sqrt{N_{\mathrm{el}}}} \begin{vmatrix} \chi_1(\boldsymbol{\tau}_1) & \chi_2(\boldsymbol{\tau}_1) & \cdots & \chi_n(\boldsymbol{\tau}_1) \\ \chi_1(\boldsymbol{\tau}_2) & \chi_2(\boldsymbol{\tau}_2) & \cdots & \chi_n(\boldsymbol{\tau}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\boldsymbol{\tau}_n) & \chi_2(\boldsymbol{\tau}_n) & \cdots & \chi_n(\boldsymbol{\tau}_n) \end{vmatrix} . \tag{13}$$

The wavefunction is represented by a Slater determinant rather than a simple product of one-electron functions to comply with the requirement of antisymmetry from postulate 6.

We can write the one-electron operators $\hat{T}_{\mathrm{e}}$ and $\hat{V}_{\mathrm{Ne}}$ and the two-electron operator $\hat{V}_{\mathrm{ee}}$ of the electronic Hamiltonian as a sum over the electrons

$$\sum_{i=1}^{N_{\mathrm{el}}} \hat{h}_i = \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{Ne}}, \tag{14}$$

$$\sum_{i=1}^{N_{\mathrm{el}}} \sum_{j>i}^{N_{\mathrm{el}}} \hat{g}_{ij} = \hat{V}_{\mathrm{ee}}. \tag{15}$$

The nuclear–nuclear repulsion is a constant V_{NN} as it does not depend on the electronic coordinates. Given that the one-electron functions χ are orthonormal and the electrons

are indistinguishable, we can express the electronic energy as

$$\begin{aligned}
E_{\mathrm{el}} =& V_{\mathrm{NN}} + \sum_{i=1}^{N_{\mathrm{el}}} \left\langle \chi_i(\boldsymbol{\tau}_1) \,\middle|\, \hat{h}_1 \,\middle|\, \chi_i(\boldsymbol{\tau}_1) \right\rangle + \\
& \sum_{i=1}^{N_{\mathrm{el}}} \sum_{j>i}^{N_{\mathrm{el}}} \left(\langle \chi_i(\boldsymbol{\tau}_1)\chi_j(\boldsymbol{\tau}_2) \,|\, \hat{g}_{12} \,|\, \chi_i(\boldsymbol{\tau}_1)\chi_j(\boldsymbol{\tau}_2) \rangle - \langle \chi_i(\boldsymbol{\tau}_1)\chi_j(\boldsymbol{\tau}_2) \,|\, \hat{g}_{12} \,|\, \chi_j(\boldsymbol{\tau}_1)\chi_i(\boldsymbol{\tau}_2) \rangle \right), \\
=& V_{\mathrm{NN}} + \sum_{i=1}^{N_{\mathrm{el}}} h_i + \sum_{i=1}^{N_{\mathrm{el}}} \sum_{j>i}^{N_{\mathrm{el}}} (J_{ij} - K_{ij}).
\end{aligned} \tag{16}$$

Here, h_i is the one-electron term. J_{ij} is the Coulomb term and represents the repulsion between two electrons. K_{ij} is the exchange term, which has no classical analog, and arises from the exchange of electrons with parallel spin between orbitals. Based on the variational principle, we would like to find the orbitals that minimize this energy. Using Lagrange's method of undetermined multipliers, we can derive that these orbitals adhere to the following equation

$$\hat{\mathcal{F}} \chi_i(\boldsymbol{\tau}_1) = \epsilon_i \chi_i(\boldsymbol{\tau}_1), \tag{17}$$

where $\hat{\mathcal{F}}$ is the Fock operator that works on an orbital as

$$\begin{aligned}
\hat{\mathcal{F}} \left| \chi_i(\boldsymbol{\tau}_1) \right\rangle =& \hat{h}_1 \left| \chi_i(\boldsymbol{\tau}_1) \right\rangle + \\
& \sum_{j}^{N_{\mathrm{el}}} \langle \chi_j(\boldsymbol{\tau}_2) \,|\, \hat{g}_{12} \,|\, \chi_j(\boldsymbol{\tau}_2) \rangle \left| \chi_i(\boldsymbol{\tau}_1) \right\rangle - \sum_{j}^{N_{\mathrm{el}}} \langle \chi_j(\boldsymbol{\tau}_2) \,|\, \hat{g}_{12} \,|\, \chi_i(\boldsymbol{\tau}_2) \rangle \left| \chi_j(\boldsymbol{\tau}_1) \right\rangle, \\
=& \left[\hat{h}_1 + \sum_{j}^{N_{\mathrm{el}}} \left(\hat{J}_j - \hat{K}_j \right) \right] \left| \chi_i(\boldsymbol{\tau}_1) \right\rangle.
\end{aligned} \tag{18}$$

Equation (17) is a pseudo-eigenvalue equation as the Fock operator depends on all the occupied orbitals through the $\hat{J}_j$ and $\hat{K}_j$ operators. We, therefore, need to already know the orbitals in order to calculate the orbitals with Eq. (17). This problem is solved through an iterative process called the self-consistent field (SCF) method. In the SCF method, we start from a set of guess orbitals that are optimized iteratively until a certain convergence criterion is met.

2.3.3 Basis sets

In practice, the orbitals χ_i are represented by a linear combination of known basis functions φ_α,

$$\chi_i = \sum_{\alpha}^{N_{\text{basis}}} c_{\alpha,i}\varphi_\alpha. \tag{19}$$

The most commonly used basis functions are Gaussian functions ($\varphi_\alpha \propto \exp(-ar^2)$). While an individual Gaussian function does not describe the physical situation around the nuclei as good as an exponential function, Gaussian functions are computationally much more efficient, allowing more basis functions to be included (Jensen, 2007).

As a minimum, one basis function should be used for each orbital. The accuracy will, however, increase with an increase in the number of basis functions. There are basis functions to represent each of the electronic shells (i.e., s, p, d, and f). We can, however, add a p-basis function to the representation of an s-orbital to increase flexibility. This is referred to as adding polarization functions (Jensen, 2001). Specialized dispersion basis functions with long tails can also be added to more accurately capture the description of the orbitals further away from the nuclei (Kendall et al., 1992). The total number of basis functions N_{basis} considered in a calculation is called the basis set. The choice of basis set is a compromise between accuracy and computational cost.

We can obtain N_{basis} equations by substituting Eq. 19 into Eq. 17, and each time multiplying from the left with one of the basis functions in the basis set. These equations can then be combined into a concise matrix notation,

$$\begin{aligned} \boldsymbol{\mathcal{F}}\mathbf{C} &= \mathbf{S}\mathbf{C}\boldsymbol{\varepsilon}, \\ \mathcal{F}_{\alpha\beta} &= \left\langle \varphi_\alpha \middle| \hat{F} \middle| \varphi_\beta \right\rangle, \\ S_{\alpha\beta} &= \langle \varphi_\alpha \,|\, \varphi_\beta \rangle, \end{aligned} \tag{20}$$

where $\mathbf{C}$ is a matrix of coefficients $c_{\alpha,i}$ and $\boldsymbol{\epsilon}$ is a diagonal matrix containing the orbital energies ϵ_i. At each SCF iteration, Eq. (20) is a generalized eigenvalue problem. As computers are especially adept at handling matrix manipulations, this matrix notation results in efficient computation of the coefficients that determine the orbitals.

2.3.4 Post-Hartree–Fock methods

The Hartree–Fock method is able to capture ∼99% of the total energy of a system. However, by approximating the wave function as a product of one-electron functions, the Coulombic interaction between the electrons is only accounted for through a mean field. In other words, each electron moves in the mean Coulombic field of the other electrons. Because of this, electrons are able to approach each other closer than in the true system. The Hartree–Fock method thus misses Coulombic electron correlation. This missing ∼1% is especially important to accurately describe chemical phenomena, such as reactions.

The Hartree–Fock energy is the optimal energy for a single Slater determinant given the chosen basis set. If we want to improve on the Hartree–Fock results, one straightforward approach is to include additional Slater determinants. The Hartee–Fock single Slater determinant is created by assuming the electrons are located in the lowest energy orbitals (i.e., the ground state). From a basis set of size N_{basis}, it is possible to construct N_{basis} linearly independent orbitals. If there are N_{occ} occupied lowest energy orbitals in the Hartree–Fock Slater determinant, we can construct additional Slater determinants by moving electrons from the occupied orbitals to the $N_{\text{basis}} - N_{\text{occ}}$ unoccupied orbitals. These additional Slater determinants correspond to exciting electrons to higher energy orbitals. By adding a certain percentage of these excited Slater determinants, we allow the electrons more flexibility in their location, improving the Coulombic correlation.

There are a variety of post-Hartree–Fock methods, that can be distinguished based on the included excitations, as well as other approximations and algorithms used to lower the computational cost. Brillouin's theorem shows that matrix elements between the Hartree–Fock wavefunction and the singly excited wavefunctions are zero (Brillouin, 1934; Surján, 1989). However, because there are relatively few singly excited states, they are usually added, nonetheless. Adding all possible excited states would result in the best possible energy prediction for the given basis set, but most post-Hartree–Fock methods are limited up to doubly or triply excited states as going beyond this would be computationally unfeasible for all but the smallest systems.

2.3.5 Density functional theory

In the Hartree–Fock method, we obtain the ground state energy by first calculating the wavefunction through the electronic Schrödinger equation. We thus calculate a $3N_{\mathrm{el}}$-dimensional function just to get a scalar property. Walter Kohn and Pierre Hohenberg proved that knowledge of the electron density $\rho_{\mathrm{el}}(\mathbf{r})$ is enough to uniquely determine the ground state energy (Hohenberg & Kohn, 1964). Just like the variational principle in the Hartree–Fock method, we can define an energy functional of the electron density, that is minimized by the true electron density

$$E[\rho_{\mathrm{el,t}}] \geq E_0. \tag{21}$$

Here, $\rho_{\mathrm{el,t}}$ is a trial density and $F[\,]$ indicates a functional, which maps a function onto a value, just like a function maps a value onto another value. The energy functional can be expressed in terms of the electronic kinetic energy functional $T_{\mathrm{e}}[\rho_{\mathrm{el}}]$, electron–electron functional $V_{\mathrm{ee}}[\rho_{\mathrm{el}}]$, and external potential functional $V_{\mathrm{ext}}[\rho_{\mathrm{el}}]$

$$E[\rho_{\mathrm{el}}] = T_{\mathrm{e}}[\rho_{\mathrm{el}}] + V_{\mathrm{ee}}[\rho_{\mathrm{el}}] + V_{\mathrm{ext}}[\rho_{\mathrm{el}}]. \tag{22}$$

The external potential functional is the energy contribution of the interaction between the electron density and the nuclei, and is given by

$$V_{\mathrm{ext}}[\rho_{\mathrm{el}}] = \int \hat{V}_{\mathrm{ext}}\rho_{\mathrm{el}}(\mathbf{r})\,\mathrm{d}\mathbf{r} \tag{23}$$

There are, however, no readily available expressions for the $T_{\mathrm{e}}[\rho_{\mathrm{el}}]$ and $V_{\mathrm{ee}}[\rho_{\mathrm{el}}]$ functionals. We can approximate $T_{\mathrm{e}}[\rho_{\mathrm{el}}]$ by the kinetic energy T_{s} of a system with no electron–electron interactions, which is given by a single Slater determinant,

$$T_{\mathrm{s}} = -\frac{1}{2}\sum_{i=1}^{N_{\mathrm{el}}} \left\langle \chi_i \,\middle|\, \nabla^2 \,\middle|\, \chi_i \right\rangle. \tag{24}$$

Similarly, $V_{\mathrm{ee}}[\rho_{\mathrm{el}}]$ can be approximated by a Hartree potential

$$V_{\mathrm{H}} = \frac{1}{2}\int \frac{\rho_{\mathrm{el}}(\mathbf{r}_1)\rho_{\mathrm{el}}(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|}\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 \tag{25}$$

We will have to compensate for these approximations, by adding an additional term $E_{\mathrm{xc}}[\rho_{\mathrm{el}}]$ called the exchange–correlation functional,

$$E[\rho_{\mathrm{el}}] = T_{\mathrm{s}}[\rho_{\mathrm{el}}] + V_{\mathrm{H}}[\rho_{\mathrm{el}}] + V_{\mathrm{ext}}[\rho_{\mathrm{el}}] + E_{\mathrm{xc}}[\rho_{\mathrm{el}}]. \tag{26}$$

where $E_{\mathrm{xc}}[\rho_{\mathrm{el}}]$ is given by

$$E_{\mathrm{xc}}[\rho_{\mathrm{el}}] = (T_{\mathrm{e}}[\rho_{\mathrm{el}}] - T_{\mathrm{s}}[\rho_{\mathrm{el}}]) + (V_{\mathrm{ee}}[\rho_{\mathrm{el}}] - V_{\mathrm{H}}[\rho_{\mathrm{el}}]). \tag{27}$$

By mapping our system onto a fictitious system of non-interacting electrons, we have reintroduced an orbital framework. Expressing the functionals explicitly in terms of the density built from orbitals and applying the variational principle of Eq. (21), we obtain the Kohn–Sham equations (Kohn & Sham, 1965)

$$\left[-\frac{1}{2}\nabla^2 + V_{\mathrm{ext}}(\mathbf{r}) + \int \frac{\rho_{\mathrm{el}}(\mathbf{r}_j)}{|\mathbf{r}_i - \mathbf{r}_j|}\,\mathrm{d}\mathbf{r}_j + V_{\mathrm{xc}}\right]\chi_i = \varepsilon_i \chi_i \tag{28}$$

Like the Hartree–Fock equations (Eq. (17)), the Kohn–Sham equations give the orbitals that minimize the energy. Based on these orbitals, we can calculate the electron density and finally obtain the energy from Eq. (26). While the Hartree–Fock method provides the best solution for a system expressed as a single Slater determinant, density functional theory (DFT) calculates the orbitals of a fictitious non-interacting system where the effective potential is such that the electron density coincides with the true interacting system. Due to this different approach, modern DFT can provide more accurate results at less computational cost than the Hartree–Fock method.

DFT would be an exact theory, if the exchange–correlation functional would be known. The true exchange–correlation functional, however, is likely a pathological functional that is not easily put in an analytical form. Instead, approximate expressions are used, and DFT methods are distinguished based on their specific exchange–correlation expression. Exchange–correlation functionals can be constructed through analytical equations, depending on the density and/or its first- and second-order derivatives, or using parameterizations based on empirical data. The accuracy of a DFT method is highly dependent on its exchange–correlation functional. Some methods are able to quite accurately predict some of the system's properties, while giving large errors for other properties. A specific problem of DFT is that, unlike the Hartree–Fock method, where additional Slater determinants can be introduced, there is no systematic way of improving on DFT.

While Hartree–Fock explicitly contains exchange interaction through the exchange term K_{ij}, DFT has to include exchange interactions through the exchange–correlation functional. Hybrid DFT methods account for exchange interaction by incorporating a certain amount of exact Hartee–Fock exchange into the exchange–correlation functional.

2.3.6 Semi-empirical methods

Despite increasing capabilities of computers and advances in algorithms, quantum chemistry calculations are still a compromise between accuracy and computational cost. A post-Hartree–Fock method like coupled clusters with single and double exited states included, is very accurate, but too costly for systems with more than a few atoms (Bartlett & Musiał, 2007). DFT with the local density approximation functional is cheap to use, but might not accurately predict the system's properties (Kohn & Sham, 1965; Perdew, 1985).

Semi-empirical quantum chemistry methods attempt to decrease the computational cost while maintaining (qualitative) accuracy. Most semi-empirical methods are based on the zero differential overlap approximation (Jensen, 2007). In this approximation, any product of two basis functions that are centered on different atoms is set to zero. As a result, the overlap matrix $\mathbf{S}$ in Eq. (20) becomes a unit matrix and all two-electron integrals with basis functions centered on more than two atoms disappear. This greatly speeds up the calculation, but neglects a significant amount of information. To compensate for this, the results are parameterized using empirical data.

Semi-empirical methods can also be derived from DFT. For example, the density functional tight-binding method is based on a Taylor expansion of the DFT energy expression around a reference density (Elstner & Seifert, 2014).

In certain situations, semi-empirical methods can give qualitatively correct results at low computational cost. Semi-empirical methods are, however, very dependent on the empirical data used in their parameterizations. If the behavior of the studied system is not captured by the empirical data, the semi-empirical methods will give incorrect results.

2.3.7 Thermochemistry calculations

The discussed quantum chemistry methods (Hartree–Fock, Post-Hartree–Fock, DFT, semi-empirical) allow us to calculate the total ground state electronic energy. At non-zero temperatures, the system, however, also possesses thermal energy and entropy, which together determine the system's (Gibbs) free energy. The free energy is the energy actually available to the system to do work (e.g., change configuration or undergo

a chemical reaction) and, therefore, the main quantity of interest when assessing the stability of a certain nuclear configuration. In quantum chemistry, calculations are typically not performed directly on the free energy surface due to the computational complexity. Rather, the free energy is calculated as a correction to the total electronic energy. The correction is divided in contributions from translation, electronic motion, vibration, and rotation to the internal thermal energy and the entropy (McQuarrie & Simon, 1999; Ochterski, 2000).

To calculate the translational internal thermal energy $E_{\text{thermal,trans}}$ and entropy $\mathcal{S}_{\text{trans}}$ contributions, the system is approximated as an ideal gas, resulting in

$$\mathcal{S}_{\text{trans}} = R\left(\ln\left[\left(\frac{2\pi m k_{\text{B}} T}{h^2}\right)^{3/2}\frac{k_{\text{B}}T}{P}\right] + \frac{5}{2}\right), \tag{29}$$

$$E_{\text{thermal,trans}} = \frac{3}{2}RT, \tag{30}$$

where R is the universal gas constant, m the system's mass, and P the pressure. For the electronic internal thermal energy $E_{\text{thermal,e}}$ and entropy $\mathcal{S}_{\text{e}}$ contributions, it is assumed that the first and higher excited states are entirely inaccessible, resulting in

$$\mathcal{S}_{\text{e}} = R\ln(2s_{\text{tot}} + 1), \tag{31}$$

$$E_{\text{thermal,e}} = 0, \tag{32}$$

where s_{tot} is the total spin of the system. The vibrational internal thermal energy $E_{\text{thermal,vib}}$ and entropy $\mathcal{S}_{\text{vib}}$ contributions are calculated by approximating each vibrational mode of the system by a quantum harmonic oscillator. This is a reasonable approximation close to the minimum, but shows increasing discrepancies with increasing amplitude (Fig. 4).

$$\mathcal{S}_{\text{vib}} = R\sum_{\kappa=1}\Theta_{\text{vib},\kappa}\left(\frac{\Theta_{\text{vib},\kappa}/T}{e^{-\Theta_{\text{vib},\kappa}/T}-1} - \ln\left(1-e^{-\Theta_{\text{vib},\kappa}/T}\right)\right), \tag{33}$$

$$E_{\text{thermal,vib}} = R\sum_{\kappa=1}\Theta_{\text{vib},\kappa}\left(\frac{1}{2} + \frac{1}{e^{-\Theta_{\text{vib},\kappa}/T}-1}\right), \tag{34}$$

where $\Theta_{\text{vib},\kappa}$ is the vibrational frequency of vibrational mode κ and the sums run over all vibrational modes. To find the vibrational frequencies, a frequency calculation has to be performed, where the second derivatives of the energy with respect to the nuclear coordinates are calculated. Note that the vibrational thermal energy contribution $E_{\text{thermal,vib}}$ is nonzero even at $T = 0$ K for a quantum harmonic oscillator. This is

referred to as the zero-point energy (Levine, 2014). For the rotational internal thermal energy $E_{\text{thermal,rot}}$ and entropy $\mathcal{S}_{\text{rot}}$ contributions, it is assumed that the system behaves like a rigid rotor, i.e., that the distances between the atoms do not change. This assumption is at odds with the vibrational motion of the molecule, which involves bond stretching. The amplitude of bond stretching is, however, typically quite small so that the rigid rotor approximation is reasonable. For a linear molecule, we have

$$\mathcal{S}_{\text{rot}} = R\left(\ln\frac{8\pi^2 I k_{\text{B}} T}{\sigma_{\text{rot}} h^2} + 1\right), \tag{35}$$

$$E_{\text{thermal,rot}} = RT, \tag{36}$$

and for a nonlinear polyatomic molecule

$$\mathcal{S}_{\text{rot}} = R\left(\ln\left(\frac{8\pi^2 I k_{\text{B}} T}{\sigma_{\text{rot}} h^2}\right) + \frac{3}{2}\right), \tag{37}$$

$$E_{\text{thermal,rot}} = \frac{3}{2}RT. \tag{38}$$

Here, $I = \sum_{i=1}^{N}(r_i - r_{\text{com}})^2$ is the moment of inertia of the N-atom system and σ_{rot} the symmetry number, which represents the number of indistinguishable orientations of the system. The Gibbs free energy is finally given by

$$G = E_{\text{el}} + E_{\text{thermal,tot}} + k_{\text{B}}T - T\mathcal{S}_{\text{tot}}, \tag{39}$$

with

$$E_{\text{thermal,tot}} = E_{\text{thermal,trans}} + E_{\text{thermal,e}} + E_{\text{thermal,vib}} + E_{\text{thermal,rot}} \tag{40}$$

$$\mathcal{S}_{\text{tot}} = \mathcal{S}_{\text{trans}} + \mathcal{S}_{\text{e}} + \mathcal{S}_{\text{vib}} + \mathcal{S}_{\text{rot}} \tag{41}$$

2.4 Configurational sampling

In the Born–Oppenheimer approximation, the nuclei of a system move along a $3N_{\text{N}}-6$-dimensional PES ($3N_{\text{N}} - 5$ for linear systems). The minima on this PES correspond to stable equilibrium geometries of the system, also called configurations. We are specifically interested in the lowest energy minima, as these are the most populated at thermal equilibrium according to the Maxwell–Boltzmann population distribution

$$\frac{N_i}{N_{\text{tot}}} = \frac{\exp\left(-\frac{E_i}{k_{\text{B}}T}\right)}{\sum_{j=1}\exp\left(-\frac{E_j}{k_{\text{B}}T}\right)} \tag{42}$$

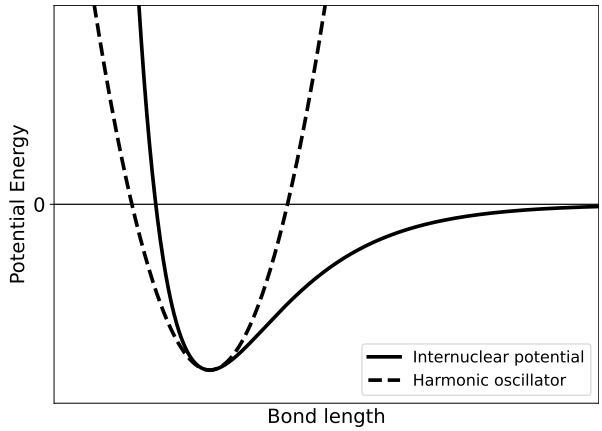


Figure 4: Comparison between a realistic internuclear potential (solid line) and the corresponding harmonic oscillator (dashed line) for a bond stretching. The harmonic oscillator is only a good approximation for the real internuclear potential around the minimum bond length.

where N_i/N_{tot} is the fraction of the total number of systems in configuration i and E_i is the energy associated with configuration i. The configuration with the lowest energy has the highest population and, therefore, the largest contribution to the system's properties. If we want to determine the system's equilibrium properties, we need to find the lowest energy configurations.

There are a variety of quite sophisticated geometry optimization algorithms to optimize an initial geometry to a minimum (Schlegel, 2011). These algorithms, however, typically optimize the system towards the minimum nearest to the initial geometry, which doesn't necessarily correspond to one of the lowest energy configurations. To find the lowest energy configurations, we need to employ configurational sampling.

For single molecules, sampling different conformers is straightforward. Various algorithms can systematically manipulate the torsion angles in a molecule to generate unique conformers. For ringed structures, various ring configurations can also be sampled. As the number of freely rotating bonds (N_{frb}) increases, the number of possible conformers increases rapidly ($\sim 3^{N_{\text{frb}}}$). For large molecules, it is, therefore, necessary to forego a systematic search and perform a randomized search instead.

The configurational space of a molecular cluster is vastly greater than that of single

molecules. Not only do the different conformers of each molecule have to be considered, also the orientation of the molecules relative to each other is important. For small clusters, the orientation might be found through chemical intuition (although chemical intuition can also be misleading). For larger clusters, it becomes impossible to create correct configurations by hand. Even if we know the lowest energy conformers of the molecules that make up the cluster, it is not necessarily the case that these conformers also appear in the cluster: interactions in the cluster environment can make a different conformer more stable. Furthermore, molecules can undergo reactions inside clusters. Acid–base clusters typically involve proton transfer (Myllys et al., 2019), while two iodic acid molecules can react to form water and iodine pentoxide in iodine-containing clusters (He et al., 2021). All relevant molecule conformers, orientations, and reactions have to be considered when searching for the lowest energy cluster configurations.

Various methods have been developed to perform configurational sampling. Kildgaard et al. (2018a, 2018b) developed a method to obtain the structure of hydrated clusters based on the structure of the dry clusters. In this method, each water molecule is placed consecutively, by sampling different orientations and specific positions around the existing cluster and in between existing hydrogen bonds. By only sampling specific positions for the introduction of the next water molecule, this method severely reduces the sampled configurational space. The problem is that the structure of the N-cluster (where N is the number of molecules in the cluster) needs to be known to find the structure of the $(N + 1)$-cluster.

In the Jammy Key for Configurational Sampling (JKCS) workflow (**Paper XI**), a large number of potential cluster configurations is generated at a low level of theory and then optimized at progressively higher levels of theory. Between each optimization step, a filter and selection procedure is performed to limit the number of cluster configurations taken to the next step. This ensures that the optimization steps remain feasible even at computationally demanding levels of theory. As the energy order of cluster configurations typically does not match between different levels of theory, we cannot optimize just the lowest energy structure found at the low level of theory.

The JKCS workflow employs various methods to generate the initial guess configurations. The first is to use an artificial bee colony algorithm with rigid monomers interacting through classic force fields, as implemented in the ABCluster program (Zhang & Dolg, 2015, 2016). In the artificial bee colony algorithm, a randomly distributed initial population of guess configurations is created. The algorithm then looks for new

cluster geometries around each of the initial configurations and constructs new geometries as a linear combination of properties of the initial geometries. If a new geometry is found to have a lower energy, it replaces the initial geometry. This process is repeated for a user-set number of generations. If the algorithm fails to find any lower energy geometries for a certain number of generations, a new randomly generated guess structure is generated. This method is based on rigid monomers. For each molecule in the cluster, we, therefore, have to include all relevant conformers. To consider acid–base reactions, all relevant conformers of conjugated acids and bases should also be included. As this exploration relies on classical force fields, it is extremely computationally efficient. When dealing with flexible monomers with many unique conformers, it, however, quickly becomes cumbersome to add all relevant monomer conformers.

The ABCluster program also includes a second approach to explore cluster configurations. In this approach, new cluster configurations are generated from a single initial structure by simultaneously altering the conformation of the molecules in the cluster and their relative orientation. Here, the semi-empirical GFN-xTB tight-binding model is used to calculate cluster interactions and energies (Bannwarth et al., 2019). The major advantage of this method is that only one conformer for each molecule in the cluster has to be supplied as input. The computational cost of the semi-empirical calculations is, however, larger than for a classical force field approach, limiting the amount of cluster configurations that can be generated.

A typical JKCS workflow generates $\sim 10^6$ cluster configurations using ABCluster with rigid monomers, optimizes a subset of $\sim 10^4$ of the configurations using GFN-xTB, and optimizes a subset of $\sim 10^2$ of those configurations using high-quality DFT. Between each of these steps, we have to filter and select the configurations to keep a feasible number for the next level of theory. In JKCS, we can use uniqueness filtering to remove clusters with similar energy, gyration radius, and dipole moment. We can also filter out clusters with high energies and large gyration radii, which will likely not correspond to a low energy minimum at higher levels of theory. Sometimes, unwanted reactions occur inside a cluster. These clusters can be filtered out through an algorithm that compares the connections between non-hydrogen atoms of each configuration. The configurations with connections that deviate from the majority are filtered out. Note that this algorithm fails when a majority of the clusters have undergone an unwanted reaction.

After the final optimization step, we can still apply post-processing on a few of the

lowest energy configurations. Post-processing includes:

- Calculating the Gibbs free energy of the configuration
- Replacing low vibrational contributions to the entropy with internal rotational modes, as their harmonic treatment can result in unphysical low energies
- Correcting for vibrational anharmonicities with a scaling factor
- Correcting the internal energy at an even higher level of theory, such as coupled cluster

The assumption that the lowest energy cluster configuration defines the cluster properties relies on the idea that clusters are both liquid-like in the sense that they immediately reorganize to their global minimum structure, and crystal-like in the sense that the clusters subsequently remain in this global minimum. In reality, complex clusters likely possess many low energy minima that easily interconvert (Elm et al., 2020). Still, the properties of low energy cluster configurations can provide valuable insight into the cluster formation process.

2.5 Collision theory

Compounds in the atmosphere are in constant motion along trajectories in three-dimensional space with velocities that depend on the temperature. While on their trajectories, these compounds interact with each other through long-range attractive forces. Long-range attractive forces are the result of electrostatic interactions between the (dynamic) charge distributions of the compounds (Israelachvili, 2011). The charge distribution of a compound in turn depends on how the negative electrons are spread around the positive nuclei. Ions have either a lack or surplus of electrons, resulting in strong electrostatic interactions with other compounds. Dipolar molecules have the same number of negative electrons around the nuclei as positive protons inside the nuclei. The electrons are, however, distributed unequally around the nuclei, because some nuclei have a larger affinity for electrons. Ions and dipolar molecules can also affect the charge distribution of molecules without a permanent dipole moment, creating an induced dipole moment. Even between two nonpolar molecules, there are still some

weak attractive electrostatic forces, due to synchronization of the normal fluctuations in their charge distributions.

When atmospheric compounds cross each other's trajectory or sufficiently attract each other with their intermolecular forces, a collision occurs. The collision rate coefficient is a measure for how many collisions occur in a given time for a specific collision system. The collision rate coefficient generally depends on the relative velocity and the fluid density regime. For atmospheric compounds, we can assume that collisions occur in the free-molecular regime with the velocities of the collision partners following the Maxwell–Boltzmann speed distribution $f_{\mathrm{MB}}(v)$

$$f_{\mathrm{MB}}(v) = \left(\frac{m}{2\pi k_{\mathrm{B}} T}\right)^{3/2} 4\pi v^2 \exp\left(-\frac{m v^2}{2\ k_{\mathrm{B}} T}\right) \tag{43}$$

To model the collision between two atmospheric compounds, we can use the central field approach (Halonen, 2021; Landau & Lifshitz, 1976). In the central field approach, both collision partners are represented as point particles, and the three-dimensional two-body motion is reduced to a one-dimensional single-body motion by choosing a frame of reference that moves with one of the collision partners (Fig. 5). This collision partner, therefore, remains stationary at the origin, while the other collision partner is assumed to approach from far away, so that its initial trajectory is unaffected by the intermolecular forces between the two collision partners. As the collision partners form an isolated system, both energy and angular momentum are conserved, i.e.,

$$\frac{1}{2}\mu v_0^2 = U + \frac{1}{2}\mu v^2 = U_{\mathrm{eff}} + \frac{1}{2}\mu v_{\parallel}^2, \tag{44}$$

where v_0 is the initial relative velocity, U the interaction potential between the point particles, v the instantaneous relative velocity, and $v_{\parallel}$ the instantaneous relative velocity component parallel to the center-to-center direction. The reduced mass between the two collision partners with masses m_1 and m_2 is given by $\mu = m_1 m_2/(m_1 + m_2)$. The effective interaction potential U_{eff} is given by

$$U_{\mathrm{eff}} = U + \frac{1}{2}\mu v_{\perp}^2 = U + \frac{\mu v_0^2 b^2}{2r^2}, \tag{45}$$

where r is the distance between the centers of mass of the collision partners, $v_{\perp}$ the instantaneous relative velocity component perpendicular to the center-to-center direction, and b the impact parameter, which is the distance perpendicular to the center-to-center direction from the origin to the direction of the initial trajectory, i.e., the minimum distance the collision partners will reach in absence of long-range attractive forces. The

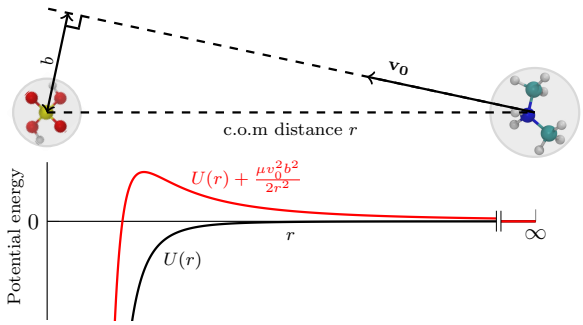


Figure 5: Schematic diagram of the central field approach with the corresponding potential energy profile. The molecule on the left is at rest and its center of mass (c.o.m.; center of the gray circle) designates the center of the field, while the molecule on the right initially moves along a trajectory set by velocity vector $\mathbf{v}_0$. At the start, the colliding molecules are infinitely far away from each other and the intermolecular potential energy $U(r)$ equals zero. The perpendicular distance between the trajectory and the center of the field is the impact parameter b. If $b > 0$, the orbital angular momentum gives rise to a centrifugal barrier shown by the red curve.

effective interaction is a result of reducing the orbital motion to the one-dimensional motion along the center-to-center direction.

A collision occurs when $r = 0$. As the centrifugal barrier is the only obstacle in achieving this criterion, $v_\parallel$ has to be positive at the peak of the centrifugal barrier, i.e.,

$$\frac{1}{2}\mu v_\parallel^2 = \frac{1}{2}\mu v_0^2 - U_{\text{eff}} = \frac{1}{2}\mu v_0^2 - U - \frac{\mu v_0^2 b^2}{2r^2} > 0. \tag{46}$$

Solving this inequality for the impact parameter b, we obtain

$$b^2 < r^2 \left[1 - \frac{2U}{\mu v_0^2}\right]. \tag{47}$$

The square of the impact parameter has to be lower than the right-hand side of Eq. 47 for a collision to occur.

At this point, we need a theoretical expression for the interaction potential U between the collision partners. The interaction potential can be represented by a variety of functional forms. Pair potentials, which are solely a function of the center-to-center distance (i.e., $U(r)$), are often used. We can, for instance, model the intermolecular interactions between the two collision partners through a general attractive pair potential

of the form

$$U(r) = -A\left(\frac{r}{r_0}\right)^a, \tag{48}$$

where A is an interaction coefficient, r_0 a distance parameter, and $a < 0$ the characteristic interaction exponent. Inserting the interaction potential of Eq. (48) into Eq. (47), we can obtain the maximum (or critical) impact parameter $b_\mathbf{c}$ for which a collision is still possible, which can then be used to express the collision cross section (CCS) Ω_{CF} in an ideal, isotropic, central field, as follows:

$$\Omega_{\mathrm{CF}}(v_0) = \pi b_\mathrm{c}^2 = \frac{\pi a}{a+2}\left(-\frac{\mu v_0^2}{A(a+2)}\right)^{2/a} r_0^2. \tag{49}$$

In thermal equilibrium, the initial velocity v_0 follows the Maxwell–Boltzmann relative speed distribution, $f_{\mathrm{MB}}(v_0)$ (Eq. (43) with μ instead of m). For interaction exponent $a < -2$, the collision rate coefficient β_{CF} in a central field can be calculated as

$$\begin{aligned}\beta_{\mathrm{CF}}(T) &= \int_0^\infty \mathrm{d}v_0 \;\; v_0 f_{\mathrm{MB}}(v_0)\Omega_{\mathrm{CF}}(v_0), \\ &= \pi r_0^2 \sqrt{\frac{8k_\mathrm{B}T}{\pi\mu}}\Gamma\left(\frac{2+a}{a}\right)\left(-\frac{2k_\mathrm{B}T}{A(a+2)}\right)^{2/a},\end{aligned} \tag{50}$$

where k_B is the Boltzmann constant, T is the temperature, and $\Gamma(x)$ denotes the Gamma function of x.

The expressions for both the collision cross section and rate coefficient are derived for the general attractive pair potential given by Eq. (48). Analytical angle-averaged expressions have, however, been derived for pair potentials of the specific molecular interactions (Israelachvili, 2011):

- Ion–ion (Coulomb): $\frac{q_1 q_2}{4\pi\varepsilon_0} r^{-1}$
- Ion–dipole: $-\frac{q^2\mu_\mathrm{d}^2}{6(4\pi\varepsilon_0)^2 k_\mathrm{B}T} r^{-4}$
- Ion–induced dipole: $-\frac{q^2\alpha}{2(4\pi\varepsilon_0)^2 k_\mathrm{B}T} r^{-4}$
- Dipole–dipole (Keesom): $-\frac{\mu_{\mathrm{d},1}^2\mu_{\mathrm{d},2}^2}{3(4\pi\varepsilon_0)^2 k_\mathrm{B}T} r^{-6}$
- Dipole–induced dipole (Debye): $-\frac{\mu_\mathrm{d}^2\alpha^2}{(4\pi\varepsilon_0)^2} r^{-6}$

- Induced dipole–induced dipole (London dispersion): $-\frac{3}{2}\frac{h\nu_1\nu_2}{\nu_1+\nu_2}\frac{\alpha_1\alpha_2}{(4\pi\varepsilon_0)^2}r^{-6}$

Here, q is the charge, ε_0 the vacuum permittivity, μ_d the angle-averaged permanent dipole moment, k_B the Boltzmann constant, α the angle-averaged polarizability, h the Planck constant, and ν the electronic absorption frequency. It is important to note that in reality, molecules and ions do not interact solely through one specific type of interaction, rather their interactions consist of a complex interplay of different contributions.

The ion–induced dipole pair potential is typically able to represent the main interaction between an ion and neutral molecule. In Gaussian cgs (centimeter–gram–second) units, the ion–induced dipole pair potential can be compactly written as $U(r) = -q\alpha/(2r^4)$. Substituting $A = q\alpha/2$, $a = -4$, and $r_0 = 1$ into Eq. (49), the collision rate coefficient within the central field model becomes

$$\beta_\mathrm{L} = 2\pi q\sqrt{\frac{\alpha}{\mu}}. \tag{51}$$

This is known as the Langevin–Gioumousis–Stevenson expression (Gioumousis & Stevenson, 1958; Langevin, 1905). Due to the specific interaction exponent $a = -4$, the temperature dependence of Eq. (49) is no longer present in Eq. (51).

For the collision between two neutral gas-phase molecules, collision rate coefficients have been conventionally calculated from kinetic gas theory. Here, the long-range interactions between the collision partners are assumed to be negligible (i.e., the interaction exponent a approaches $-\infty$), while the molecules are represented by hard spheres of specific radii R_1 and R_2, instead of point particles. In this non-interacting hard-sphere model, the CCS $\Omega_\mathrm{HS} = \pi\left(R_1 + R_2\right)^2$ and has a clear geometrical interpretation. It is the area perpendicular to the trajectory of one of the collision partners that the other collision partner needs to cross for a collision to occur. The non-interacting hard-sphere CCS is given by

$$\beta_\mathrm{HS} = \pi\left(R_1 + R_2\right)^2\sqrt{\frac{8k_\mathrm{B}T}{\pi\mu}}. \tag{52}$$

2.5.1 The Su & Chesnavich parameterization

While the Langevin–Gioumousis–Stevenson expression provides a straightforward method to calculate the rate coefficients of collisions between ions and neutral com-

pounds, it does have various shortcomings when applied to ion–dipole collisions. The Langevin–Gioumousis–Stevenson expression has no dependence on either the temperature or the permanent dipole moment μ_{D} of the dipolar compound. To address these problems, Su and Chesnavich developed a parameterized correction term K_{SC} to the Langevin–Gioumousis–Stevenson expression based on trajectory simulations between a point charge and a two-dimensional polar rigid rod (Su & Chesnavich, 1982; Su et al., 1978). The Su & Chesnavich collision rate coefficient β_{SC} is, therefore, given by

$$\beta_{\mathrm{SC}} = K_{\mathrm{SC}}\beta_{\mathrm{L}}, \tag{53}$$

where K_{SC} depends on a parameter $\xi = \mu_{\mathrm{D}}/(2\alpha k_{\mathrm{B}}T)^{1/2}$ as

$$K_{\mathrm{SC}} = \begin{cases} 0.4767\xi + 0.6200; & \xi \geq 2, \\ \frac{(\xi+0.5090)^2}{10.526} + 0.9754; & \xi \leq 2 \end{cases}, \tag{54}$$

For almost all realistic systems, $\xi > 2$. Through ξ, the Su & Chesnavich parameterization reintroduces a dependency on the temperature and the permanent dipole moment of the dipolar compound. Several studies have shown that the Su & Chesnavich parameterization predicts collision rate coefficients in good agreement with experiments for collisions between small molecules and ions (Amelynck et al., 2005; Midey et al., 2001; Strekowski et al., 2019; Williams et al., 2004; Woon & Herbst, 2009). As a default option, ion–dipole collision rate coefficients are calculated using the Su & Chesnavich parameterization in ACDC.

2.6 Molecular dynamics

Molecular dynamics (MD) is a computational method that allows us to follow the time evolution of a system on an atomistic level. We can generally assume that the nuclei in a system behave classically. While there are situations where nuclear quantum effects need to be considered (e.g., when significant nuclear delocalization occurs), even these systems can typically be mapped onto a classical system with an extended phase space (Unke et al., 2021). The dynamics of the nuclei are determined by the forces acting on them, which can be calculated at any time through either quantum chemistry or a force field description. Given the positions $\{\mathbf{r}\}$ and momenta $\{\mathbf{p}\}$ (where $\mathbf{p}_i = m_i\mathbf{v}_i$ with $\mathbf{v}_i$ the instantaneous velocity of nuclei i) of the nuclei at some initial time t_0, the positions and momenta of the nuclei at any other time t are completely determined

by Newton's equations of motion. In MD simulations, we thus integrate Newton's equations of motion to follow the atomistic trajectories of the nuclei over time.

Where quantum chemistry calculations generally focus on stationary points of the PES, MD can be used to study dynamical and non-equilibrium properties. The process of setting up and running an MD simulation closely mirrors real-life experiments and is, therefore, sometimes referred to as a computer experiment (Frenkel & Smit, 2002). MD is able to visualize atomistic trajectories of a system, providing insight into the atomistic dynamics as well as macroscopic properties through statistical mechanics.

2.6.1 Statistical mechanics

Statistical mechanics is a formalism that connects the atomistic dynamics of a system (which are simulated in MD) to its macroscopic properties (Tuckerman, 2010). All possible position coordinates $\{q\}$ (where q denotes a position component) and momentum coordinates $\{p\}$ of the nuclei in the system span a $6N_\mathrm{N}$-dimensional space, called the phase space. Each point of this phase space corresponds to a unique set of position and momentum coordinates of the nuclei in the system. Such a point is referred to as a microstate. Usually, not all microstates are accessible to the system. A system can, for instance, be constrained to certain constant macroscopic properties, such as total energy, number of compounds, temperature, pressure, or volume. These constraints are only fulfilled by a subset of the total number of microstates. Still, there will typically be many microstates that correspond to a specific set of constant macroscopic properties. If we, for example, consider a system with a constant total energy E, there are generally many ways to distribute the energy over the N_N nuclei to obtain the same total system energy. Each of these energy distributions corresponds to a unique microstate.

One of the most important concepts in statistical mechanics is the 'ensemble'. An ensemble is a large number of virtual copies of the system, possessing the same set of constant macroscopic properties, but existing in different microstates. For each macroscopic observable O, there exists a function $o(\{q\},\{p\})$ to calculate its value from the microstates. Given the probability $\mathcal{P}$ of the system being found in each of the N_micro microstates, we can calculate the expectation value of O as the ensemble

average $\langle o \rangle$ through

$$\langle o \rangle = \sum_{i=1}^{N_{\text{micro}}} o(\{q\}, \{p\}) \mathcal{P}_i \tag{55}$$

The probability of the system being found in each of the microstates depends on the specific ensemble. To approximate the ensemble average, we could run a large number of MD simulations starting from different initial positions and momenta. At any time t, the simulations will be in different microstates according to the probability $\mathcal{P}$. An estimate of a macroscopic property's ensemble average is then given by an average over the different simulations at time t.

In many cases, it is cumbersome to set up a large number of distinct MD simulations. It is, however, typically assumed that given enough time, a system will visit all relevant accessible microstates. This assumption is called ergodicity (Frenkel & Smit, 2002). As a result, the time average over different microstates along a single simulation trajectory will for sufficiently long simulation times equal the ensemble average. The ergodicity assumption fails when two accessible regions of the phase space are separated by an insurmountable energy barrier. When a simulation is started in one of the regions, it will never visit microstates in the other region no matter the simulation time.

In **Papers II**, **III**, **V** and **VI**, we run a large number of MD collision simulations with different initial coordinates to obtain an estimate of the collision probability for a specific combination of initial relative velocity and impact parameter. MD simulations were performed in the microcanonical and canonical ensemble.

Microcanonical ensemble

In the microcanonical or NVE ensemble, the total energy E, volume V, and number of compounds N are constant. This corresponds to a system that is completely isolated from its surroundings. As all microstates have the same energy, statistical mechanics postulates that the system is equally likely to exist in any of the accessible microstates. The probability of a specific microstate is then equal to $1/N_{\text{micro}}$.

Canonical ensemble

In the canonical or NVT ensemble, the temperature T, volume V, and number of compounds N are constant. This corresponds to a system in thermal equilibrium with a heat reservoir. The probability $\mathcal{P}_i$ of the system being in microstate i is given by

$$\mathcal{P}_i = \frac{\exp(-E_i/k_{\text{B}}T)}{\sum_{j=1}^{N_{\text{micro}}} \exp(-E_j/k_{\text{B}}T)}. \tag{56}$$

For classical systems, where the energy is a continuum, the sum in the denominator of Eq. (56) becomes an integral.

Most energy terms of interest have a quadratic form az^2 (e.g., kinetic energy $p^2/2$ or the harmonic potential $kq^2/2$). For a quadratic energy contribution, the corresponding average energy is given by (Longair, 2005)

$$\frac{\int az^2 \exp(-az^2/k_\mathrm{B}T)\,\mathrm{d}z}{\int \exp(-az^2/k_\mathrm{B}T)\,\mathrm{d}z} = \frac{1}{2}k_\mathrm{B}T. \tag{57}$$

As this was derived for a random quadratic energy term, it shows that at thermal equilibrium, all degrees of freedom that appear only quadratically in the energy contribute on average $k_\mathrm{B}T/2$ to the total energy. This is known as the equipartition theorem. To run physically sound MD simulations at thermal equilibrium, the system should obey the equipartition theorem. Typically, this is not the case for the initial coordinates of the nuclei as the momenta are randomly chosen from the Maxwell–Boltzmann speed distribution and the relative positions do not exactly correspond to equilibrium. It is, therefore, necessary to equilibrate the system over a certain time to obtain an equilibrium where the total energy is properly partitioned over all degrees of freedom.

For quantum systems, the equipartition theorem is not always valid. The theorem only applies when the quadratic degrees of freedom can be freely excited (Pathria & Beale, 2011). Quantum systems can, for instance, have high-frequency vibrational modes that are not freely accessible and, therefore, possess no or at least significantly less energy than $k_\mathrm{B}T/2$.

2.6.2 Equations of motion

To simulate the time evolution of a system, we have to solve Newton's equation of motion for each nucleus $i \in N_\mathrm{N}$

$$m_i \frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} = m_i \frac{\mathrm{d}^2\mathbf{r}_i}{\mathrm{d}t^2} = \mathbf{f}_i, \tag{58}$$

where m_i, $\mathbf{v}_i$, and $\mathbf{r}_i$ are the mass, velocity, and position of nucleus i, and $\mathbf{f}_i$ is the force exerted on nucleus i. Integrating Eq. (58) once or twice gives us the time evolution of the velocity and position, respectively. The combined equations of motion for a system containing more than two nuclei are, however, impossible to solve analytically. Instead, the continuous time evolution is discretized into time steps Δt. At a specific time step,

all forces on a nucleus are calculated. The nucleus is then propagated according to the net force exerted on it until the next time step, where the forces are recalculated. In this approach, the net force can change from one time step to the next, but not in between time steps. The time step Δt should be small enough to resolve the fastest dynamics of the system, such as bond stretching involving hydrogen. Typical Δt values are between 0.1 to 2 fs. Larger time steps will not properly capture the dynamics of most systems.

Regardless of the size of the time step, the error resulting from discretizing the integration will lead the simulated trajectory to exponentially diverge from the true trajectory (without discretization) for the specific initial coordinates. It is, however, believed that the discretized trajectory closely follows at least some true trajectory in phase space (Frenkel & Smit, 2002). In the end, we are not interested in a specific unique trajectory, but rather that the simulated trajectory corresponds to at least one of the possible trajectories of the system. This belief in a true 'shadow' orbital that closely follows our simulated orbital is not proven. Good agreement with experiments has, however, been observed across many MD studies.

The forces can be calculated at each time step as the negative gradient of the energy obtained from either a quantum chemistry calculation or by describing the interactions between the nuclei with a force field. Once we have obtained the forces exerted on a nucleus at a certain time step t, the position and velocity at the next time step $t + \Delta t$ can be calculated through various algorithms. For the MD simulations in this thesis, we used the velocity Verlet algorithm (Swope et al., 1982), where the position and velocity of a nucleus i at time step $t + \Delta t$ is obtained by first calculating the half-step velocity $\mathbf{v}_i(t + \Delta t/2)$,

$$\mathbf{v}_i\left(t + \frac{\Delta t}{2}\right) = \mathbf{v}_i(t) + \frac{\mathbf{f}_i(t)\Delta t}{2m_i}, \tag{59}$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t + \frac{1}{2}\Delta t)\Delta t, \tag{60}$$

$$\mathbf{f}_i(t + \Delta t) = \mathbf{f}_i(\mathbf{r}_i(t + \Delta t)), \tag{61}$$

$$\mathbf{v}_i(t + \Delta t) = \mathbf{v}_i\left(t + \frac{\Delta t}{2}\right) + \frac{\mathbf{f}_i(t + \Delta t)\Delta t}{2m_i}. \tag{62}$$

Here, $\mathbf{f}_i(t + \Delta t)$ can be calculated after $\mathbf{r}_i(t + \Delta t)$ is known. We can follow the time evolution of the system by solving these equations for all nuclei at successive time steps.

Cluster formation in the atmosphere occurs at rates of around 10^2 cm^{-3}s^{-1} for sulfuric

acid concentrations of 10^7 cm^{-3}. To capture this in an MD simulation would require computationally prohibitive system sizes and simulation times. It is clear that MD cannot simulate the actual atmospheric conditions. Rather, MD simulations need to be designed in such a way that information can be obtained about (part of) the particle formation process, even when realistic conditions are not reproduced. In **Papers II**, **III**, **V** and **VI**, we follow the collision trajectories of an isolated system containing two collision partners to obtain a prediction of their collision rate coefficient.

2.6.3 Thermostats

The canonical ensemble is conceptualized as a system in thermal equilibrium with a heat reservoir. Simulating an actual heat reservoir (i.e., a large number of reservoir molecules) in MD would, however, waste a lot of computational resources on a system whose specifics do not interest us. Instead, MD simulations in constant temperature ensembles rely on thermostat algorithms that artificially maintain the temperature around a constant value. This is typically done by altering the velocity of all or several nuclei, as equipartition shows that the instantaneous kinetic temperature at time t is given by

$$T(t) = \frac{2K}{N_{\mathrm{dof}} k_{\mathrm{B}}} = \sum_{i=1}^{N} \frac{m_i}{N_{\mathrm{dof}} k_{\mathrm{B}}} v_i^2, \tag{63}$$

where K is the kinetic energy. There are a variety of different thermostat algorithms that are distinguished based on the method used to maintain the system at a constant temperature. Thermostat algorithms can be stochastic or deterministic. They can, furthermore, affect specific nuclei (local) or the system as a whole (global). In the thesis, we used the canonical sampling through velocity rescaling (Bussi et al., 2007), Nosé–Hoover (Hoover, 1985; Nosé, 1984), and Langevin thermostats (Bussi & Parrinello, 2007).

Canonical sampling through velocity rescaling

Velocity rescaling is a straightforward method to achieve constant temperature. In its simplest form, the velocities of all nuclei are multiplied every time step by the same factor a:

$$a = \sqrt{\frac{\langle K \rangle}{K}}, \tag{64}$$

where $\langle K \rangle$ is the average kinetic energy at the desired constant temperature. As a result, the temperature is exactly equal to the desired constant temperature at each time step. The instantaneous kinetic temperature of a canonical ensemble, however, fluctuates with deviations in the order of $1/\sqrt{N_{\text{dof}}}$, where N_{dof} is the number of degrees of freedom. Simply rescaling the velocities by a every time step thus does not reproduce a canonical ensemble. The canonical sampling through velocity rescaling (CSVR) thermostat instead uses a rescaling factor a_{CSVR} (Bussi et al., 2007)

$$a_{\text{CSVR}}^2 = e^{-\Delta t/\tau} + \frac{\langle K \rangle}{N_{\text{dof}} K}(1 - e^{-\Delta t/\tau}) \left(\mathcal{R}_1^2 + \sum_{i=2}^{N_{\text{dof}}} \mathcal{R}_i^2 \right) + 2e^{-\Delta t/\tau} \sqrt{\frac{\langle K \rangle}{N_{\text{dof}} K}(1 - e^{-\Delta t/\tau})} \mathcal{R}_1, \tag{65}$$

where $\mathcal{R}_i$ denotes independent random numbers from a Gaussian distribution with unitary variance. τ is a time constant that determines the strength of the temperature/velocity corrections provided by the thermostat, e.g., how fast the temperature is adjusted. $\tau \geq 20\Delta t$ is generally considered a reasonable choice. a_{CSVR} both maintains a canonical ensemble and ensures a smooth evolution towards the desired constant temperature. The rescaling is done for all nuclei and is dependent on the random numbers $\mathcal{R}_i$. The CSVR thermostat is therefore global and stochastic.

Nosé–Hoover

The Nosé–Hoover (NH) thermostat introduces a fictitious degree of freedom ζ to the system, which acts as a friction, slowing down or accelerating the nuclei to achieve the correct kinetic energy distribution. As a result of this additional degree of freedom the velocity Verlet algorithm of Eq. (59–62) is modified to (Marenduzzo, 2015)

$$\mathbf{v}_i \left(t + \frac{\Delta t}{2} \right) = \mathbf{v}_i(t) + \frac{\Delta t}{2} \left(\frac{\mathbf{f}_i(t)}{m_i} - \zeta(t)\mathbf{v}_i(t) \right), \tag{66}$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i \Delta t + \left(\frac{\mathbf{f}_i(t)}{m_i} - \zeta(t)\mathbf{v}_i(t) \right) \frac{\Delta t^2}{2}, \tag{67}$$

$$\mathbf{f}_i(t + \Delta t) = \mathbf{f}_i(\mathbf{r}_i(t + \Delta t)), \tag{68}$$

$$\zeta \left(t + \frac{\Delta t}{2} \right) = \zeta(t) + \frac{\Delta t}{2\tau} \left[\sum_{i=1}^{N_{\text{N}}} m_i \frac{\mathbf{v}_i(t)^2}{2} - \frac{3N_{\text{N}} + 1}{2} k_{\text{B}} T \right], \tag{69}$$

$$\zeta(t + \Delta t) = \zeta(t + \frac{\Delta t}{2}) + \frac{\Delta t}{2\tau} \left[\sum_{i=1}^{N_{\text{N}}} m_i \frac{\mathbf{v}_i \left(t + \frac{\Delta t}{2} \right)^2}{2} - \frac{3N_{\text{N}} + 1}{2} k_{\text{B}} T \right], \tag{70}$$

$$\mathbf{v}_i(t + \Delta t) = \frac{\mathbf{v}_i \left(t + \frac{\Delta t}{2} \right) + \frac{\Delta t}{2} \frac{\mathbf{f}_i(t + \Delta t)}{m_i}}{1 + \frac{\Delta t}{2} \zeta(t + \Delta t)}. \tag{71}$$

Here, τ is again the time constant. A single NH thermostat can produce dynamics that are non-ergodic, e.g., it does not properly sample the accessible phase space. To minimize this issue, we employ a NH chain thermostat of size N_{chain} (Martyna et al., 1992). The temperature of each NH thermostat $i < N_{\text{chain}}$ in the chain is controlled by thermostat $i + 1$ in the chain. While this improves the ergodicity of the dynamics, complete ergodicity is not achieved. The NH thermostat is global and deterministic.

Langevin

The Langevin thermostat introduces a virtual solvent around the nuclei. The correct kinetic energy distribution is achieved through random collisions between the virtual solvent and nuclei. Including the virtual solvent, the velocity Verlet algorithm becomes (Marenduzzo, 2015)

$$\mathbf{v}_i\left(t+\frac{\Delta t}{2}\right) = \mathbf{v}_i(t) + \frac{\Delta t}{2}\left(\frac{\mathbf{f}_i(t)\Delta t}{m_i} - \gamma \mathbf{v}_i(t)\right) + \sqrt{\frac{\Delta t k_{\text{B}} T \gamma}{m_i}}\boldsymbol{\lambda}_i, \tag{72}$$

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i\left(t+\frac{\Delta t}{2}\right)\Delta t, \tag{73}$$

$$\mathbf{f}_i(t+\Delta t) = \mathbf{f}_i(\mathbf{r}_i(t+\Delta t)), \tag{74}$$

$$\mathbf{v}_i(t+\Delta t) = \mathbf{v}_i\left(t+\frac{\Delta t}{2}\right) + \frac{\Delta t}{2}\left(\frac{\mathbf{f}_i(t)\Delta t}{m_i} - \gamma \mathbf{v}_i\left(t+\frac{\Delta t}{2}\right)\right) + \sqrt{\frac{\Delta t k_{\text{B}} T \gamma}{m_i}}\boldsymbol{\lambda}_i, \tag{75}$$

where $\boldsymbol{\lambda}_i$ is a random vector which components are chosen from a Gaussian distribution with mean 0 and unit variance. $\gamma \propto \tau^{-1}$ is the friction coefficient which determines the strength of the friction induced by the virtual solvent. The system undergoes semi-Brownian motion to a degree set by γ. This shift toward Brownian dynamics by the Langevin thermostat results in notable dampening of dynamical processes, which may not reflect the desired environment and transport properties of the studied system. The systems using global NH and CSVR thermostats produce more realistic gas-phase dynamics regardless of applied τ. The Langevin thermostat is local and stochastic.

In this thesis, thermostats are used during equilibration to achieve equipartition over the degrees of freedom. In **Paper V**, we looked at the thermostatting efficiency of the Langevin, NH, and CSVR thermostats for gas-phase systems in the free-molecular regime. All three thermostats are found to maintain equipartition once it is achieved, but only the Langevin thermostat is able to efficiently bring the system to equipartition.

2.7 Potential of mean force calculations

MD simulations sample a complex high-dimensional phase space. When modeling the sticking collisions that constitute the initial stages of NPF, we are, however, primarily interested in the free energy profile along the center-of-mass distance between the collision partners. This free energy profile determines the interactions between the collision partners, and therefore the rate coefficient. In conventional MD simulations, the forces drive the system towards a free energy minimum. As a result, the simulation trajectory predominantly stays around this minimum and only sporadically visits other regions of the free energy profile, leading to inadequate sampling. To properly sample the entire free energy profile along the center-of-mass distance, we have to employ enhanced sampling methods, such as umbrella sampling (Torrie & Valleau, 1977) or metadynamics (Laio & Parrinello, 2002). These methods introduce a bias potential that forces the system to sample all regions along one or several selected degrees of freedom.

The degrees of freedom along which metadynamics and umbrella sampling perform the enhanced sampling are referred to as collective variables (CVs). There is no universal recipe for selecting appropriate CVs (Barducci et al., 2011; Chen, 2021). Examples of frequently used CVs are interatomic distances, angles, dihedrals, radii of gyration, and dipole moments. The selected CVs should describe all stages of the event of interest. It is assumed that all other variables can rapidly adjust to any changes in the CVs. If this is not the case, e.g., if another variable cannot be adequately sampled in the simulation time, then it needs to be included as a CV. Otherwise, the enhanced sampling can lead to erroneous free energy profiles.

To obtain the potential energy, also called the potential of mean force (PMF), from the free energy profile, we may need to remove an entropic contribution. For example, the center-of-mass distance r_{com} is a radial CV which contains an entropic contribution given by $-2k_{\mathrm{B}}T\ln(r_{\mathrm{com}})$ (Fig. 6; McQuarrie, 1976; Piskulich and Thompson, 2020).

2.7.1 Umbrella sampling

In umbrella sampling, the free energy profile is calculated for a specified range of the chosen CVs by dividing this range into intervals, or 'windows'. For each window, an MD simulation is run, where the system is kept close to the center of the window

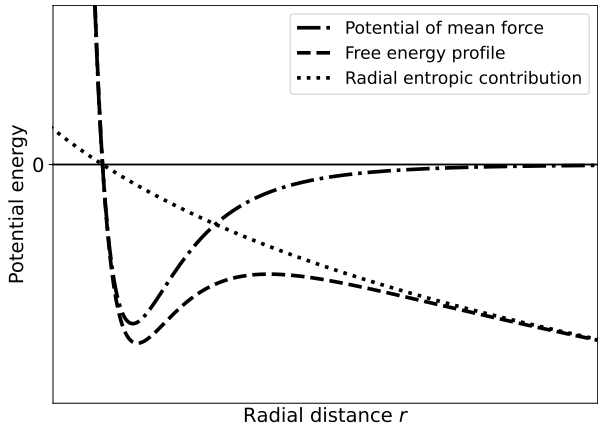


Figure 6: An example of a free energy profile (dashed line), the corresponding potential of mean force (dashed–dotted line), and the radial entropic contribution $-2k_{\mathrm{B}}T\ln(r)$ (dotted line).

through a harmonic potential (Fig. 7)

$$V_{\mathrm{bias}}(\mathbf{s}) = -\frac{1}{2}k(\mathbf{s}-\mathbf{s}_{\mathrm{center}})^2, \tag{76}$$

where $\mathbf{s}$ is the set of chosen CVs. Given a long enough simulation time, the probability of finding the system at a certain value of the CVs will closely follow a Gaussian distribution around the center of the window. In post-processing, we need to retrieve the unbiased free energy profile from the biased potentials in the N_{w} windows. This can be done through the weighted histogram analysis method (Kumar et al., 1992). In the weighted histogram analysis method, each collective variable s is divided into N_{bins}. The unbiased probability distribution for bin α, $\mathcal{P}(s^\alpha)$, is given by

$$\mathcal{P}(s^\alpha) = \frac{\sum_{i=1}^{N_{\mathrm{w}}} N_i^\alpha}{\sum_{i=1}^{N_{\mathrm{w}}} N_i \exp\left[F_i - V_{\mathrm{bias},i}(s^\alpha)/k_{\mathrm{B}}T\right]}, \tag{77}$$

where N_i is the total number of datapoints (e.g., time steps) from the simulation in window i, and N_i^α is the number of datapoints from the simulation in window i where the system was in bin α. F_i is the free energy shift of window i, calculated through

$$F_i = -k_{\mathrm{B}}T\ln\left(\sum_{\alpha=1}^{N_{\mathrm{bins}}} \mathcal{P}(s^\alpha)\exp\left[-V_{\mathrm{bias},i}(s^\alpha)/k_{\mathrm{B}}T\right]\right) \tag{78}$$

As F_i depends on $\mathcal{P}(s^\alpha)$ and visa versa, Eq. (77) and (78) have to be solved iteratively.

Umbrella sampling provides extensive control over the evolution of the system. The accuracy of calculations can also be systematically improved by increasing the number of windows and the simulation time. To start an umbrella sampling calculation, we, however, need to supply the initial configuration for each of the windows. This can be problematic when the configurations along the CVs are not known or difficult to construct. The number of simulations to run also increases rapidly with the number of CVs, making calculations with more than one CV cumbersome.

2.7.2 Metadynamics

In metadynamics, an MD simulation is run for a specified range of the CVs. To prevent the system from being stuck in the free energy minima along the collective variable, a Gaussian bias potential is dropped at certain time intervals τ_{G}. This raises the energy around the particular value of the CV, discouraging the system from revisiting configurations that have already been sampled. For N_{CV} collective variables $s_i(q)$, the overall bias potential at a time t is given by

$$V_{\mathrm{G}}(\mathbf{s}, t) = \sum_{t'=0,\tau_{\mathrm{G}},2\tau_{\mathrm{G}},\ldots}^{t'<t} \mathcal{W} \exp\left(-\sum_{i=1}^{N_{\mathrm{CV}}} \frac{(s_i(q)) - s_i(q(t')))^2}{2\Sigma_i^2} \right), \tag{79}$$

where $\mathcal{W}$ and Σ are the Gaussian height and width. Given a long enough simulation time, all the features (i.e., minima) of the free energy profile will be filled in by the consecutive drops of Gaussian potentials (Fig. 7). The sum over all Gaussian potentials will then be equal to the negative of the free energy profile. Initially, it is advantageous to have large Gaussian potentials to quickly fill up the free energy minima. Once the deepest minima are filled, it becomes more advantageous to fill the spaces between previous drops with smaller Gaussian potentials. This is accounted for in well-tempered metadynamics calculations, by changing the height of the Gaussian potentials from $\mathcal{W}$ to $\mathcal{W}e^{-V(\mathbf{s}(q(t')))/\Delta T}$, where ΔT is a user-defined bias factor. The well-tempered metadynamics Gaussian height decreases over time to create a smooth convergence towards the negative of the free energy profile (Barducci et al., 2008).

As the bias potential is a simple sum of Gaussian potentials, we can run multiple simulations at the same time, and sum up the Gaussian deposits of each simulation. A big advantage of metadynamics over umbrella sampling is that each simulations can start from the same initial configuration, although with staggered start times. Furthermore,

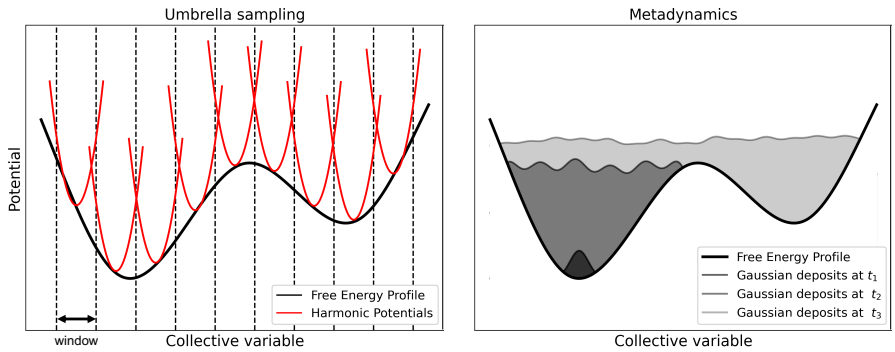


Figure 7: Visual examples of umbrella sampling (left) and metadynamics (right). In umbrella sampling, the collective variable is divided into windows. In each window, the system is constrained by the addition of harmonic bias potentials (red lines). In metadynamics, the system is evolved from an initial configuration. At time $t_1 = \tau_G$, a first Gaussian potential is deposited. At some time $t_2 > t_1$, enough Gaussian potentials have been deposited to fill the left minimum. At some time $t_3 > t_2$, both minima are filled and the sum of Gaussian potentials is an estimate for the negative of the free energy profile.

no *a priori* information is necessary about any configuration along the CVs beyond the initial configuration. The number of simulations to run is also independent of the number of CVs.

On the other hand, the quality of the free energy profile obtained from metadynamics is highly dependent on the calculation parameters (e.g., W, σ, and τ_G). These parameters are system specific and *a priori* difficult to determine. Non-ideal parameters can result in significant noise and convergence issues in the estimated free energy profile.

2.8 Force fields

In the Born–Oppenheimer approximation, the nuclei move on the PES created by the electron distribution. To study the dynamics of the system, we, therefore, have to map the PES by calculating the electronic energy for given positions of the nuclei. The complexity and dimensionality of the PES, however, increases with the number of atoms in the system. Mapping the PES from first principles thus quickly becomes

computationally extremely expensive.

In force field methods, first principle calculations are bypassed by describing the potential energy between the atoms as a parametric function of the nuclear coordinates (Jensen, 2007). A force field is thus determined by the functional form of the potential energy description and the specific set of parameters used. These parameters can be fitted either to quantum chemical or experimental data. Force fields are a model and their usefulness is determined by how well they approximate the quantum chemical or experimental behavior of the system. Contrary to quantum chemistry, atoms are not separated in nuclei and electrons, but considered as a single entity with a certain partial charge. Force fields are typically composed of intramolecular terms for interactions between atoms separated by one or more covalent bonds and intermolecular terms for nonbonded interactions. Generally, force field methods are many orders of magnitude faster than quantum chemistry methods. While MD simulations can be done by calculating the forces through quantum chemistry calculations at each time step (i.e., *ab initio* MD), the computational cost is prohibitive for most systems, and we have to rely on a force field description instead.

The OPLS (Optimized Potential for Liquid Simulations) all-atom force field is a good example of the functional form of a classical force field (Jorgensen et al., 1996). The intramolecular interactions consist of harmonic bond potentials between covalently bonded atoms, harmonic angle potentials between atoms separated by two covalent bonds, and dihedral angle potentials between atoms separated by three covalent bonds,

$$U_{\text{intra}}^{\text{OPLS}} = \sum_{i=1}^{N_{\text{bonds}}} \frac{k_i^{\text{b}}}{2} \left(r_i - r_i^0\right)^2 + \sum_{j=1}^{N_{\text{angles}}} \frac{k_j^{\theta}}{2} \left(\theta_j - \theta_j^0\right)^2 + \sum_{k=1}^{N_{\text{dihedrals}}} \sum_{l=1}^{4} \frac{V_l}{2} \left[1 + \cos\left(l\phi^k - \phi_l^k\right)\right], \tag{80}$$

where k_i^{b}, r_i, and r_i^0 are the force constant, instantaneous, and equilibrium length of bond i, k_j^{θ}, θ_j, and θ_j^0 are the force constant, instantaneous, and equilibrium value of angle j, and V_n, ϕ_n^k, and ϕ^k are the Fourier coefficients, phase angles, and instantaneous value of the dihedral angle k.

The intermolecular interactions, as well as intramolecular interactions between atoms separated by more than three covalent bonds, are described by Lennard-Jones potentials between atoms i and j separated by a distance r_{ij}, with distance and energy parameters σ_{ij} and ϵ_{ij}, and Coulomb interactions between the atoms' partial charges

q_i and q_j,

$$U_{\text{inter}} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}}, \tag{81}$$

where ϵ_0 is the vacuum permittivity. As bonds are described through a harmonic potential, the OPLS force field (and classical force fields in general) cannot simulate bond breaking. Chemical reactions are, therefore, not accounted for by the force field. While molecules as a whole can be polarized by changing their geometry in an electric field, the atoms themselves have a fixed partial charge q.

It is usually only possible to fit the parameters of simple force fields, like the OPLS force field, to provide good agreement for some of the system's properties, but not all. For example, it is often not possible to accurately reproduce the binding energies of multiple distinct cluster configurations with the OPLS force field. In this case, we can choose to focus on the most important configuration (usually the lowest energy configuration) and accept that other configurations deviate from the empirical data. Alternatively, a larger number of properties can be accurately reproduced by including additional terms to the force field description or changing to a more complex functional form. Terms can, for instance, be added to explicitly model hydrogen bonds. Some force fields can also model atomic polarizability. In core-shell models, for example, atoms are separated into a positive core surrounded by a negative shell, connected through a harmonic oscillator (Dick & Overhauser, 1958). Other functional forms can model reactivity. The ReaxFF force field uses a bond order parameter that changes according to the environment surrounding the atoms participating in the bond (van Duin et al., 2001). If properly fitted, the bond order parameter evolves in such a way that it accurately reproduces bond breaking and formation. Essentially, any number of properties can be accurately modeled by increasing the complexity of the force field.

Adding degrees of freedom, however, makes it more difficult and time-consuming to fit the parameters. There is also a risk of overfitting, where the force field perfectly matches the data used to fit it, but fails to extrapolate outside that data. Complex force fields are also generally more computationally expensive. One of the strengths of a classical force field is the clear distinction between the various contributions to the energy. Complex functional forms do not allow for a straightforward interpretation of the energy contributions.

2.9 Ion mobility modeling

Chemical-ionization atmospheric-pressure-interface mass spectrometry is uniquely suited to measure freshly formed atmospheric clusters, clustering vapors, and their precursors at low concentrations and ambient pressure. Among the possible clustering vapors and precursors, there are, however, many isomers which cannot be separated using mass spectrometry alone. A possible solution is to measure a second property. The ion mobility Z is an obvious choice, as mass spectrometers already require the sample to be charged. It is a measure for the ability of ions to move through a fluid under the influence of an electric field $\mathcal{E}$,

$$v_{\text{drift}} = Z\mathcal{E}, \tag{82}$$

where v_{drift} is the drift velocity, i.e., the terminal velocity reached by the ion as it moves through the fluid under the influence of the electric field. Ion mobility is a property of both the ion and the fluid it travels through. There are various instruments to measure the ion mobility, which are typically fitted in between the charging unit and the mass spectrometer. The differential mobility analyzer (DMA) is one such instrument, which has an instructive ion mobility measurement approach (Fig. 8; Knutson and Whitby, 1975; Xu et al., 2023). When ions enter the DMA, they are subjected to an electric field parallel to the inlet as well as a perpendicular laminar flow of carrier gas. Opposite the inlet, there is an exit hole at a distance L in the direction of the laminar flow. Given a fixed laminar flow velocity v_{flow} and electric field $\mathcal{E}$, only ions with a specific ion mobility will traverse the DMA width W in the same time it takes to travel the length L in the direction of the gas flow, and thus reach the exit hole. This leads to a straightforward equation for deriving the ion mobility

$$Z = \frac{v_{\text{flow}} W^2}{L\mathcal{V}}, \tag{83}$$

where $\mathcal{V} = W\mathcal{E}$ is the voltage. During a measurement, the voltage is altered until all ions in the sample have reached the exit hole. In practice, the flow is not exactly laminar but can contain localized eddy currents. Furthermore, there are often slight deviations of W and L from the values indicated by the manufacturer. Thus, the ion mobility is usually measured using a calibration ion for which the ion mobility $Z_{\text{calibrant}}$ is known.

$$Z = \frac{\mathcal{V}_{\text{calibrant}}}{\mathcal{V}} Z_{\text{calibrant}}. \tag{84}$$

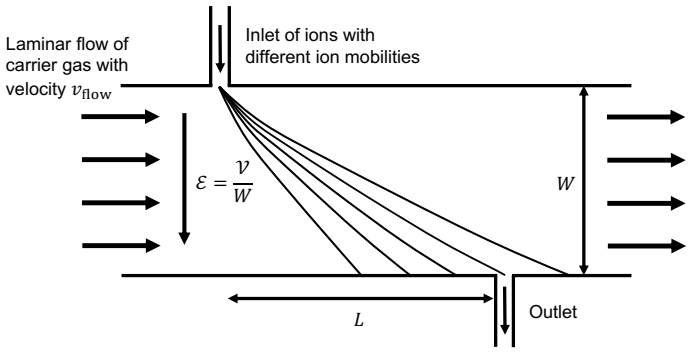


Figure 8: Sideview of a differential mobility analyzer. Figure adapted from Yale DMA-MS Facility (2015).

To promote unity across different experiments, the ion mobility is typically reported as a reduced ion mobility Z_0 given by

$$Z_0 = \frac{P}{P_0}\frac{T_0}{T}Z, \tag{85}$$

where $T_0 = 273.15$ K and $P_0 = 1$ atm. Unlike the mass-to-charge ratio, the ion mobility cannot be easily calculated from a compound's chemical formula. We, therefore, have to rely on literature values to determine the peaks in an ion mobility spectrum. Unfortunately, there are not many ion mobility values available for the compounds of potential interest in NPF. This is where computational ion mobility predictions can assist.

There are several methods to obtain computational ion mobility predictions. The projection approximation method is simple, although not very accurate (Larriba-Andaluz, 2023). In this method, the CCS Ω is equated to the area of interaction between a carrier gas molecule and the ion. This area of interaction is the projection of the carrier gas–ion complex, at the moment of collision, onto a plane. By averaging over many random orientations of the carrier gas–ion collision complex, we can get a rough estimate of the CCS. Using the Mason–Schamp equation, we can then calculate the ion mobility from the CCS (Viehland & Mason, 1975)

$$Z = \frac{3}{16}\frac{qe}{\rho_{\text{gas}}}\left(\frac{2\pi}{\mu k_{\text{B}} T}\right)^{1/2}\frac{1}{\Omega}, \tag{86}$$

where q is the formal charge of the ion, e the elementary charge, μ the reduced mass, and ρ_{gas} the gas density.

Currently, the most sophisticated ion mobility calculation method is the Lennard-Jones trajectory method. In this method, it is assumed that the ion is moving through the carrier gas fluid with a constant velocity v_{drift}, while continuously re-orientating itself, making all orientations equally possible (Larriba & Hogan, 2013). We choose a frame of reference moving along with the ion, so that the ion is stationary in this frame. A control volume is placed around the ion and the trajectories of N_{gas} carrier gas molecules moving through the control volume are simulated. When a gas molecule collides with the stationary ion, a momentum transfer occurs. The sum over the velocity changes of all collided gas molecules is recorded as $\mathbf{u}_{\text{tot}}$

$$\mathbf{u}_{\text{tot}} = \sum_{i=1}^{N_{\text{gas}}} \mathbf{v}_{\text{final},i} - \mathbf{v}_{\text{initial},i}. \tag{87}$$

The total change in velocity $\mathbf{u}_{\text{tot}}$ is proportional to the drag force on the ion. As we assume all ion orientations to be equally possible, we can orientate the ion in three perpendicular directions and calculate the drag force components for each of these orientations to form a drag tensor $[\mathbf{B}]$. The components B_{ij} of this tensor are obtained from Eq. (87), where j is the component (x, y, or z) of the carrier gas velocity given that v_{drift} is aligned in direction i. The drag tensor $[\mathbf{B}]$ is symmetric and positive-definite and can, therefore be diagonalized. The diagonal elements B_{xx}, B_{yy}, and B_{zz} are then used to determine the ion mobility through (Larriba-Andaluz et al., 2017)

$$Z = \left(1 + \frac{m_{\text{gas}}}{m_{\text{ion}}}\right)^{1/2} \frac{\sqrt{\pi} q e t_{\text{tot}}}{8 m_{\text{gas}}} \left(\frac{3}{B_{xx} + B_{yy} + B_{zz}}\right), \tag{88}$$

where t_{tot} is the total simulation time. Long-range interactions between the carrier gas molecules and the ion are described by a 4-6-12 potential of the form (Mesleh et al., 1996)

$$U(x,y,z) = 4\varepsilon \sum_{i=1}^{N} \left[\left(\frac{\sigma}{r_i}\right)^{12} - \left(\frac{\sigma}{r_i}\right)^{6}\right] - \frac{\alpha}{2}\left(\frac{qe}{N}\right)^2 \left[\left(\sum_{i=1}^{N} \frac{x_i}{r_i^3}\right) + \left(\sum_{i=1}^{N} \frac{y_i}{r_i^3}\right) + \left(\sum_{i=1}^{N} \frac{z_i}{r_i^3}\right)\right], \tag{89}$$

where $r_i = (x_i, y_i, z_i)$ is the relative distance between the gas molecules and atom i of the N atoms of the ion, α is the polarizability of the gas molecules due to the ion–induced dipole potential. ε and σ are the energy and distance Lennard-Jones parameters between the gas molecules and atom i.

Pure air or N_2 are typically used as carrier gas in experimental ion mobility measurements. As air mostly consists of N_2, and O_2 is very similar to N_2, the ion mobility in

air can be approximated by assuming pure N_2. For the N_2 molecule, we can include quadrupole contributions for improved accuracy. The quadrupole moment is obtained by placing a negative charge of $-0.4825e$ on each nitrogen atom and a positive charge of $0.9650e$ in the middle of the N_2 molecule. The quadrupole potential U_{qpol} between the N_2 molecule and the ion is then given by

$$U_{\text{qpol}}(x, y, z) = \sum_{j=1}^{3} \sum_{i=1}^{N} q_i q_j \frac{e^2}{r_{ij}^3}, \tag{90}$$

where the j index runs over the three different charges on the N_2 molecule, and the i index over the N partial charges on the ion. It is assumed that the N_2 carrier gas molecules rotate freely, and the quadrupole potential can be averaged over the rotational degrees of freedom of the molecules.

The computational ion mobility predictions in this thesis are obtained using the Ion Mobility Software (IMoS) program (Coots et al., 2020). As input, IMoS requires the Cartesian coordinates and partial charges of the atoms of the studied system. These values are obtained from quantum chemistry, with the most probable system configurations found through configurational sampling. IMoS and similar programs, such as MobCal (Haack et al., 2023; Mesleh et al., 1996), have been used extensively for computational ion mobility calculations. The discrepancies between the Lennard-Jones trajectory method and experimental data are typically less than 5%. This discrepancy is in part due to several approximations that are made in the Lennard-Jones trajectory method, such as the assumption that the quadrupole potential can be averaged over the N_2 molecule rotation, the absence of higher order poles, as well as various rotational and vibrational effects and alignments between the carrier gas molecules and the ion. Despite these approximations, the Lennard-Jones trajectory method represents the most accurate method currently available.

2.10 Machine learning

Machine learning algorithms aim to find structural information in large sets of data (Unke et al., 2021). As there is an enormous number of molecules that potentially play a part in atmospheric NPF, each with its own set of relevant configurations and properties, machine learning is an essential tool to deal with all these data points. While machine learning can be employed for a variety of different use cases, such

as data classification, one of the most common applications is to predict output $\mathbf{y}$ associated with input $\mathbf{x} \in \mathbb{R}^{\mathrm{D}}$ based on existing input and output data.

$$\mathbf{y} = f(\mathbf{x}) + \epsilon, \tag{91}$$

where ϵ denotes measurement noise. There are many different machine learning algorithms to establish these functional relationships, of which kernel-based methods and neural networks are some of the most common.

In kernel-based methods, the functional relationship is given by the sum of a kernel function working on the N input values $\mathbf{x}'_i \in \mathbb{R}^{\mathrm{D}}$ for which the output $\mathbf{y}'_i$ is known.

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i \mathcal{K}(\mathbf{x}, \mathbf{x}'_i), \tag{92}$$

where c_i are coefficients that need to be determined. The kernel function $\mathcal{K}(\mathbf{x}, \mathbf{x}'_i)$ is a function that typically evaluates the similarity between $\mathbf{x}$ and $\mathbf{x}'_i$ in a higher dimensional Hilbert space without actually having to map $\mathbf{x}$ and $\mathbf{x}'_i$ onto that space. A variety of different functional forms can perform this task. Gaussian kernels are widely used, as they evaluate the similarity between $\mathbf{x}$ and $\mathbf{x}'_i$ in a ∞-dimensional Hilbert space,

$$\mathcal{K}(\mathbf{x}, \mathbf{x}'_i) = \exp\big(\zeta ||\mathbf{x} - \mathbf{x}'_i||^2\big), \tag{93}$$

where ζ is a user-defined hyperparameter and $||\cdot||$ is the L^2 norm. Using matrix algebra, we can calculate the coefficients c_i that fully determine the machine learning model, given a dataset of inputs $[\mathbf{x}'_1 \cdots \mathbf{x}'_N]$ and outputs $[\mathbf{y}'_1 \cdots \mathbf{y}'_N]$. With the trained machine learning model, we can predict output $\mathbf{y}$ for any unseen input $\mathbf{x}$.

Neural networks contain a series of hidden layers $\mathbf{h}$ that predict output $\mathbf{y}$ based on input $\mathbf{x}$. Each hidden layer consists of a linear transformation. The hidden layers are combined through a nonlinear activation function σ. A neural network with N_{L} layers would look like

$$\begin{aligned}
\mathbf{h}_1 &= \sigma\left(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1\right) \\
\mathbf{h}_2 &= \sigma\left(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2\right) \\
&\vdots \\
\mathbf{h}_{N_{\mathrm{L}}} &= \sigma\left(\mathbf{W}_{N_{\mathrm{L}}} \mathbf{h}_{N_{\mathrm{L}}-1} + \mathbf{b}_{N_{\mathrm{L}}}\right) \\
\mathbf{y} &= \mathbf{W}_{N_{\mathrm{L}}+1} \mathbf{h}_{N_{\mathrm{L}}} + \mathbf{b}_{N_{\mathrm{L}}+1},
\end{aligned} \tag{94}$$

where $\{\mathbf{W}_i, \mathbf{b}_i\}_{i=1}^{N_\mathrm{L}+1}$ are parameters that are randomly initialized and optimized to minimize a loss function that measures the difference between output $\mathbf{y}$ and a dataset of inputs for which the outputs are known. The sizes of the $\mathbf{W}$ matrix and $\mathbf{b}$ vector determine the size of each hidden layer and its corresponding $\mathbf{h}$ vector. Neural networks typically need more data to achieve the same accuracy as kernel-based methods. However, evaluating a neural network trained model is significantly faster. For large datasets, neural networks are, therefore, the preferred choice.

Machine learning algorithms can be used to predict any output as long as there is a reasonable dependency on the input. In machine learning applications of chemical systems on an atomistic level, the input is often given by the geometry of the system. Cartesian coordinates are, however, not a sensible descriptor to define a chemical system. A molecule can, for instance, slightly rotate around its center of mass. This will not change the molecule, nor even its configuration, but it will change the Cartesian coordinates. We, therefore, need a descriptor of the chemical system that encodes structural information about the relative positions of the atoms in space. Common descriptors for machine learning applications include vectors of internal coordinates, Coulomb matrices, or representations of atomic environments (e.g., the SOAP and FCHL descriptors; Musil et al., 2021). The choice of descriptor depends on the machine learning application. Some descriptors, for instance, can distinguish between different configurations of the same system. Depending on the application, this can be necessary or unwanted. A specific class of neural networks, end-to-end neural networks, create a suitable descriptor alongside the prediction of the output property. End-to-end neural networks start from a random initial descriptor, which is optimized at each hidden layer. Thus, the atomic positions can be given as input.

In **Paper IV**, we used a kernel-based machine learning method with the FCHL descriptor (Christensen et al., 2020) to predict the electronic energy of acid–base clusters at a high-quality DFT level of theory based on electronic energies and cluster geometries obtained with the semi-empirical GFN-xTB method (Bannwarth et al., 2019). Machine learning can additionally be used to fit machine learning force fields for MD simulations, or to predict ion mobility based on the ion's geometry.

While machine learning algorithms can efficiently provide highly accurate predictions, it is important to be aware of their limitations. The functional relationship that machine learning algorithms find between input and output is typically extremely complex and, therefore, difficult or impossible to interpret. As a result, it is not clear how the

algorithm determines its predictions. Usually, the performance is evaluated by calculating the mean absolute error or root mean square error when the machine learning model is applied to a dataset not included in the training data. There could, however, be biases shared between the training and test datasets, which are not revealed by these simple scalar values. In short, machine learning algorithms can be extremely helpful tools, that need to be tested rigorously to achieve a correct mapping of input properties to output properties.

3 Overview of papers and the author's contribution

Paper I. Separation of isomers using a differential mobility analyser (DMA): Comparison of experimental vs modelled ion mobility (2022)
Bianco, A., Neefjes, I., Alfaouri, D., Vehkamäki, H., Kurtén, T., Ahonen, L., Passananti, M., & Kangasluoma, J.
Author contribution: *The author was responsible for all computational modeling, including configurational sampling, quantum chemistry calculations, and ion mobility modeling. The author, furthermore, wrote all computational modeling sections of the manuscript.*

Paper II. Modeling approaches for atmospheric ion–dipole collisions: All-atom trajectory simulations and central field methods (2022)
Neefjes, I., Halonen, R., Vehkamäki, H., & Reischl, B.
Author contribution: *The author planned the study with H. Vehkamäki and B. Reischl, designed the simulation workflows with B. Reischl and R. Halonen, and carried out the force-field-based molecular dynamics collision and well-tempered metadynamics simulations with B. Reischl. The first draft of the manuscript was written by the author, B. Reischl, and R. Halonen.*

Paper III. Ion–molecule rate constants for reactions of sulfuric acid with acetate and nitrate ions (2022) Fomete, S. K. W., Johnson, J. S., Myllys, N., Neefjes, I., Reischl, B., & Jen, C. N.
Author contribution: *The author carried out well-tempered metadynamics simulations with B. Reischl in support of the experimental results, and contributed to the manuscript.*

Paper IV. Atmospheric sulfuric acid–multi-base new particle formation revealed through quantum chemistry enhanced by machine learning (2023)
Kubečka, J., Neefjes, I., Besel, V., Qiao, F., Xie, H.-B., & Elm, J.
Author contribution: *The author carried out cluster dynamics calculations using the ACDC code and contributed to the first draft of the manuscript.*

Paper V. Further cautionary tales on thermostatting in molecular dynamics: Energy equipartitioning and non-equilibrium processes in gas-phase simulations (2023)
Halonen, R., Neefjes, I., & Reischl, B.
Author contribution: *The author helped conceptualize the study, performed pre-*

liminary molecular dynamics simulations, and contributed to the first draft of the manuscript.

Paper VI. Collision-sticking rates of acid–base clusters in the gas phase determined from atomistic simulation and a novel analytical interacting hard-sphere model (2023) Yang, H., Neefjes, I., Tikkanen, V., Kubečka, J., Kurtén, T., Vehkamäki, H., & Reischl, B.

Author contribution: *The author assisted in developing the theoretical framework of the interacting hard-sphere model and carried out the molecular dynamics collision simulations together with H. Yang. The author, H. Yang, and B. Reischl analyzed the simulation data and wrote the manuscript with comments from the other authors.*

4 Results and discussion

4.1 Modeling approaches for atmospheric ion–dipole collisions: All-atom trajectory simulations and central field methods (Paper II)

To model the relative importance of ion-induced NPF pathways, we can calculate the cluster formation rates of the different pathways with the Atmospheric Cluster Dynamics Code (ACDC). As stated in subsection 2.5, ACDC calculates the collision rate coefficients of ion-induced pathways by default using the Su & Chesnavich parameterization. In **Paper II**, we compared ion–dipole collision rate coefficients obtained from the Su & Chesnavich parameterization to those obtained from force-field-based atomistic MD collision simulations for several atmospherically relevant collision systems. The studied systems included both molecular and dimeric dipoles and ions. We developed a MD workflow where the two collision partners are first equilibrated separately beyond the force field cutoff using Langevin thermostats. The collision partners are then moved within the force field cutoff and both given a velocity towards each other. These simulations are repeated a large number of times for different values of the initial relative velocity v_0 and impact parameter b to obtain the collision probability $P(v_0, b)$ as a function of v_0 and b. We can then obtain the collision cross section $\Omega_{\mathrm{MD}}(v_0)$ and rate coefficient β_{MD} through

$$\Omega_{\mathrm{MD}}(v_0) = \pi \int_0^\infty \mathrm{d}b^2 \, P(v_0, b), \tag{95}$$

$$\beta_{\mathrm{MD}} = \int_0^\infty \mathrm{d}v_0 \, v_0 f_{\mathrm{MB}}(v_0) \Omega_{\mathrm{MD}}(v_0), \tag{96}$$

where $f_{\mathrm{MB}}(v_0)$ is the Maxwell–Boltzmann speed distribution. Additionally, we developed a complementary approach where the collision rate coefficient is calculated from the central field approach (Eq. (49)) with interaction parameters a, A, and r_0 obtained from fitting a general attractive potential to the potential of mean force (PMF). The PMF was calculated from well-tempered metadynamics simulations. In this approach, we use the framework of the central field approach, but with interaction parameters obtained using the same force field description as the MD simulations.

Table 6 shows the ratio of the collision rate coefficient obtained from atomistic MD simulations over the collision rate coefficient obtained from the Langevin–Gioumousis–

Stevenson model, Su & Chesnavich parameterization, and central field model with interaction parameters fitted to the PMF. We use the atomistic force-field-based MD simulations as a reference, as the outcome of a collision trajectory is determined at a relatively large center-of-mass distance (i.e., where the collision barrier is located). Therefore, any short-range quantum effects, reactivity, and atom polarizability should be negligible in determining the collision rate coefficient and the classical force field description should be adequate. It can be seen from Table 6 that the Langevin–Gioumousis–Stevenson model performs poorly for the studied systems, with a discrepancy of a factor 3 to 13 compared to the MD collision rate coefficients. The Su & Chesnavich model performs much better, specifically for collision systems with sulfuric acid as the dipole, where the discrepancy is 10–20%. However, for the more complex dipole [$HSO_4^- \cdot (CH_3)_2NH_2^+$], the discrepancy increases to 30–40%. The central field approach with interaction parameters fitted to the PMF, on the other hand, performs well for all studied collision systems, with a discrepancy from the MD collision rate coefficient of at most 10%. The central field approach with interaction parameters fitted to the PMF is, therefore, an elegant and computationally efficient alternative to brute force sampling of trajectories over a range of impact parameters and relative velocities. The good agreement between atomistic MD and the central field approach also indicates that the assumptions made in the central field approach are largely valid for the studied systems. The ion–dipole collision process can be accurately described by isotropic and adiabatic intermolecular dynamics.

4.2 Ion–molecule rate constants for reactions of sulfuric acid with acetate and nitrate ions (Paper III)

Paper III is a collaboration with the Jen research lab at the Carnegie Mellon University. They measured chemical ionization rate coefficient of sulfuric acid with both nitrate and acetate using chemical ionization mass spectrometry. It was found that the chemical ionization rate coefficient with acetate ($3.6–4.6 \cdot 10^{-15}$ m^3s^{-1}) is a factor 1.9–2.4 higher than with nitrate ($1.9 \cdot 10^{-15}$ m^3s^{-1}). Given the good agreement with atomistic MD simulations in **Paper II**, we used the central field approach with interaction parameters fitted to the PMF to calculate computational collision rate coefficient predictions for the same collision systems. As the experiments were run at a relative humidity of around 20%, the most populated hydration state of acetate is monohydrated. Furthermore, nitrate mostly appears as the nitric acid–nitrate dimer (Kurtén

Table 6: Collision rate coefficients calculated from molecular dynamics collision simulations β_{MD}, along with their ratio to collision rate coefficients calculated from the Langevin–Gioumousis–Stevenson model β_L, Su & Chesnavich parameterization β_{SC}, and central field model with interaction parameters fitted to the potential of mean force β_{CF} in 10^{-15} m^3s^{-1}.

System	β_{MD}	β_{MD}/β_L	β_{MD}/β_{SC}	β_{MD}/β_{CF}
H_2SO_4 – HSO_4^-	2.62	3.21	1.20	1.04
H_2SO_4 – $(CH_3)_2NH_2^+$	3.01	2.96	1.11	0.92
H_2SO_4 – NH_4^+	4.43	3.04	1.14	1.05
H_2SO_4 – NO_3^-	2.85	3.09	1.16	1.04
H_2SO_4 – $[H_2SO_4 \cdot HSO_4^-]$	2.18	3.10	1.16	0.98
$[HSO_4^- \cdot (CH_3)_2NH_2^+]$ – HSO_4^-	9.50	13.15	1.39	0.93
$[HSO_4^- \cdot (CH_3)_2NH_2^+]$ – $(CH_3)_2NH_2^+$	11.62	12.48	1.32	0.90
$[HSO_4^- \cdot (CH_3)_2NH_2^+]$ – $[H_2SO_4 \cdot HSO_4^-]$	7.50	12.40	1.31	1.01

et al., 2011). Figure 9 shows the collision rate coefficients from the Su & Chesnavich parameterization and the central field model with interaction parameters fitted to the PMF, as a function of $1/\sqrt{\mu}$, where μ is the reduced mass. All collision systems where sulfuric acid is the dipole from **Paper II** are also included. The Su & Chesnavich parameterization is linear with respect to $1/\sqrt{\mu}$. The collision systems of **Paper II** largely follow a similar linear trend. Acetate, both dry and monohydrated, however, has significantly higher collision rate coefficients than the linear trend would indicate. The linear trend suggests that the ionic collision partner only influences the collision rate coefficient through its mass. The deviation of acetate from this trend thus implies that acetate influences the collision rate coefficient through other properties than just its mass. Further research is necessary to determine exactly what properties these would be. The central field collision rate coefficient prediction of the sulfuric acid–acetate system is in good agreement with the lower end of the experimental results, while the prediction of the sulfuric acid–nitrate system is significantly higher than the experimental results. This indicates that the nitric acid–nitrate dimer might not be correctly described by the force field.

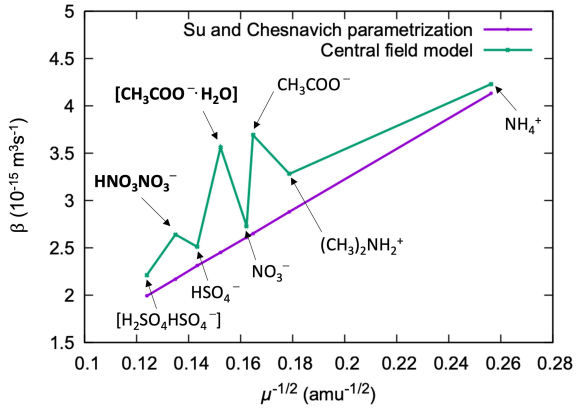


Figure 9: Collision rate coefficients β plotted as a function of $1/\sqrt{\mu}$ for collisions involving H_2SO_4 as a dipole, from the Su & Chesnavich parameterization and the central field model with interaction parameters fitted to the potential of mean force in 10^{-15} m^3s^{-1}.

4.3 Further cautionary tales on thermostatting in molecular dynamics: Energy equipartitioning and non-equilibrium processes in gas-phase simulations (Paper V)

In the collision simulation setup of Halonen et al. (2019), the collision partners were equilibrated using a Nosé–Hoover chain thermostat. During the development of the collision simulation workflow of **Paper II**, we, however, discovered that this thermostat does not lead to equipartition in a reasonable timescale. This issue, along with other pitfalls of incorrect thermostatting, were collected in **Paper V**. Figure 10 shows the time evolution of the translational, rotational, and vibrational temperature partitions for a simulation box containing 30 $(CH_3)_2NH_2^+$ ions during equilibration using the Langevin, Canonical Sampling through Velocity Rescaling (CSVR) and Nosé–Hoover thermostats with time constants $\tau = 0.1$, 1, and 10 ps. The Langevin thermostat takes less than 0.1 ns to reach equipartition among the different energy partitions. CSVR and Nosé–Hoover, however, take more than 100 ns to reach equipartition.

The slow equilibration of CSVR and NH is a result of their global thermostatting procedure. After each time step, they change the momenta of all atoms in the system without any energy exchange between the translational/rotational and vibrational de-

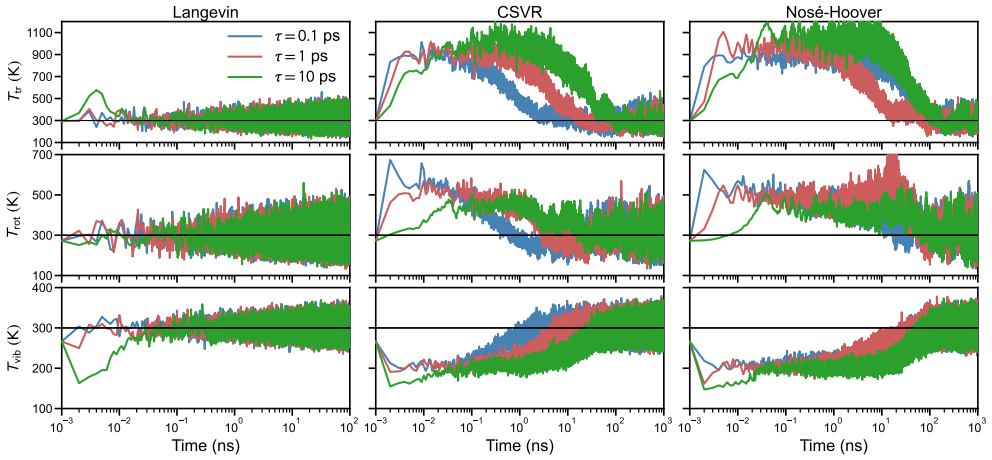


Figure 10: Time evolution of different temperature partitions of 30 $(CH_3)_2NH_2^+$ ions during equilibration using different thermostats set to a target temperature $T_0 = 300$ K. From left to right: Langevin, Canonical Sampling through Velocity Rescaling (CSVR), and Nosé–Hoover thermostats. From top to bottom: translational, rotational, and vibrational temperatures. Thermostat time constants $\tau = 0.1$, 1, and 10 ps are represented by blue, red, and green lines, respectively.

grees of freedom. This exchange of energy requires physical interactions, i.e., collisions, between the compounds or with an inert carrier gas. In condensed matter systems, these interactions occur frequently, leading to efficient equilibration with all three thermostats. However, when studying gas-phase systems near the free-molecular regime, equilibration through a large number of consecutive collisions is very inefficient. The Langevin thermostat does not suffer from these problems, as the momenta are not conserved and the energy exchange is facilitated by the fictitious solvent.

In **Paper V**, we also discussed the issue of latent heat when performing MD simulations of dimer/cluster formation in an ensemble with constant temperature. Cluster formation lowers the system's energy. Due to energy conservation, this energy is released from the system in the form of latent heat, affecting the momenta, and thus the temperature, of the atoms participating in the cluster formation. Global thermostats, such as CSVR and Nosé–Hoover, maintain a constant temperature by adjusting the momenta of all atoms, not distinguishing the atoms involved in cluster formation. As a result, the kinetic energy of the atoms not involved in the cluster formation is

artificially scaled down, affecting the dynamics of the system. While the Langevin thermostat works locally, its working principle of introducing a fictitious solvent is incompatible with cluster formation simulations in the free-molecular regime. Only the NVE ensemble strictly represents physically well-defined Hamiltonian dynamics.

4.4 Collision-sticking rates of acid–base clusters in the gas phase determined from atomistic simulation and a novel analytical interacting hard-sphere model (Paper VI)

While **Paper II** showed that the central field approach with interaction parameters fitted to the PMF is in good agreement with atomistic MD simulations, this method has several limitations. First, as the central field approach represents the collision partners by point particles, a collision only occurs when the two particles coincide in the center of the field. Real collision partners can, however, already form short-range 'sticking' interactions, such as hydrogen bonds, at a center-of-mass distance of several Ångströms. At large relative velocities, the central field CCS can, furthermore, become vanishingly small due to the point particle approximation. Secondly, we rely on the PMF to obtain the interaction parameters needed for the central field approach. While it is relatively straightforward to calculate the PMF for molecular compounds and dimers, it becomes cumbersome for larger clusters.

Paper VI addresses these issues by developing an interacting hard-sphere model. In the interacting hard-sphere model, the critical impact parameter, that determines if a collision will take place, relies on two cases,

$$b_{\mathrm{c}}^2 = \begin{cases} \omega_v(R_{\mathrm{m}}), & \text{if } R_{\mathrm{m}} > R_i + R_j \\ \omega_v(R_i + R_j), & \text{if } R_{\mathrm{m}} \leq R_i + R_j \end{cases}, \tag{97}$$

where

$$\omega_v(r) = r^2 \left[1 - \frac{2U(r)}{\mu v_0^2} \right], \tag{98}$$

and R_{m} is the center-of-mass distance r for which $\omega_v(r)$ has a minimum, and R_i and R_j are the radii of the collision partners. The first case is the same as the conventional central field approach. The second case, however, considers that the collision partners already collide at some center-of-mass distance, given here by the sum of hard-sphere

radii. In other words, we follow the normal collision criterion of the central field approach, unless the collision partners are able to come within a distance $R_i + R_j$ of each other, which is now considered to always lead to a collision. The difference between the central field approach and the interacting hard-sphere model is, therefore, only evident for trajectories that come within a distance $R_i + R_j$ of each other but do not collide according to the central field approach (Fig. 11). For most systems, this only occurs at high relative velocities, which only a few systems possess according to the Maxwell–Boltzmann speed distribution. There are, however, certain systems, like CO_2, where there is a significant difference between the central field approach and the interacting hard-sphere model, and the latter should be considered.

The interacting hard-sphere model, furthermore, allows predictions of rate coefficients for collisions between a molecule and a cluster of any size by employing Hamaker's approach (Hamaker, 1937). Here, we assume that the free molecule and the monomers in the cluster interact through a Lennard-Jones potential (although the model can readily be expanded to include other interactions, such as those involving ions). The potential U_{mc} between a molecule and multicomponent cluster is then given by

$$U_{\mathrm{mc}}(r) = -\frac{4}{\left(r^2 - R_{\mathrm{cl}}^2\right)^3} \sum_{i=1}^{\text{monomer type}} N_{\mathrm{cl},i}\varepsilon_i\sigma_i^6, \tag{99}$$

where R_{cl} is the cluster radius, $N_{\mathrm{cl},i}$ the number of molecules of type i in the cluster, and ε and σ the Lennard-Jones energy and distance parameters. Based on this potential, the distance R_{m} and, therefore, the collision rate coefficient can be calculated for a molecule–cluster collision through the equations derived in **Paper VI**.

4.5 Atmospheric sulfuric acid–multi-base new particle formation revealed through quantum chemistry enhanced by machine learning (Paper IV)

Advances in quantum chemistry and computing have led to increasingly accurate predictions of the cluster binding free energies that are used to evaluate the evaporation rate coefficients in ACDC. The number of cluster configurations, however, increases dramatically with the size of the cluster. As a result, more cluster configurations need to be considered at a high level of theory during configurational sampling and the computational cost quickly becomes prohibitively high. In **Paper IV**, we performed

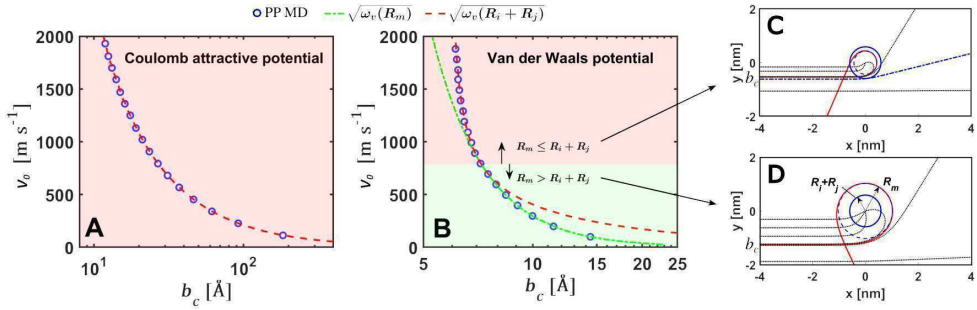


Figure 11: Critical impact parameters b_c from the interacting hard-sphere model and from point particle molecular dynamics simulations (PP MD) for (A) Coulomb attractive potential for the $(CH_3)_2NH_2^+$ and HSO_4^- ion pair, and (B) Van der Waals attractive potential for the H_2SO_4 and $[HSO_4^- \cdot (CH_3)_2NH_2^+]_1$ neutral pair. As the relative initial velocity v_0 increases, the system in panel B transitions from a regime where the critical impact parameter $b_c = \sqrt{\omega_v(r_m)}$ to $b_c = \sqrt{\omega_v(R_i + R_j)}$. Example orbits for $v_0 = 2000$ ms^{-1} and $v_0 = 150$ ms^{-1} are shown in panels (C) and (D), respectively.

configurational sampling of sulfuric acid–base clusters, with ammonia (AM), methylamine (MA), dimethylamine (DMA), trimethylamine (TMA), and ethylenediamine (EDA) as the bases. Initial guess configurations were obtained with the rigid monomer optimizer in ABCluster. These configurations were subsequently optimized using first the semi-empirical GFN1-XTB method and then high-quality DFT calculations at the ωB97X-D/6-31++G(d,p) level of theory (Bannwarth et al., 2019; Chai & Head-Gordon, 2008; Hehre et al., 1972). To improve the correlation between the GFN1-XTB and DFT steps, we trained a kernel ridge regression machine learning model on the difference in electronic binding energy obtained through both methods. The training set consisted of clusters containing up to 3 sulfuric acid molecules and up to 3 bases, where the cluster configurations were represented by the FCHL descriptor (Christensen et al., 2020).

The trained model was used to enhance the configurational sampling of clusters containing up to 4 sulfuric acid molecules and up to 4 bases, where all combinations of the studied bases were considered. We find that the lowest binding energy configurations after the machine learning step are significantly more likely to contain the lowest binding energy configuration according to the DFT step than without the machine learning

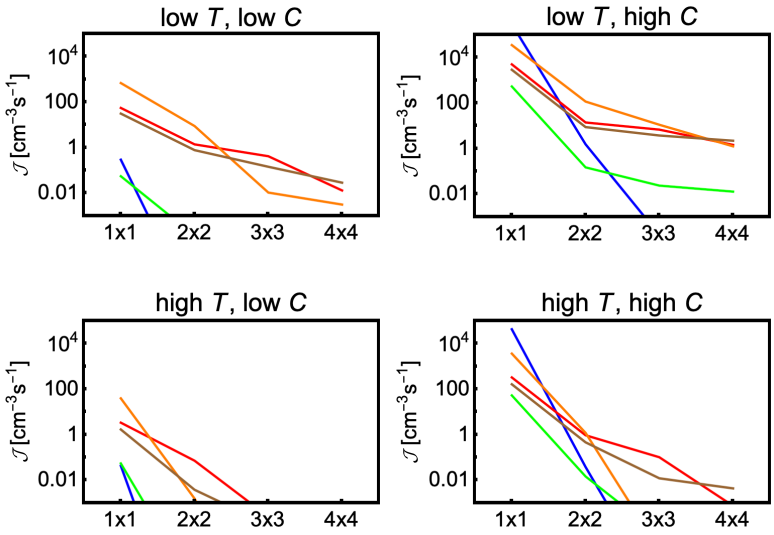


Figure 12: Cluster formation rates $\mathcal{J}$ obtained at different simulation scheme sizes. The lines correspond to the binary sulfuric acid–base systems with bases: ammonia (AM; blue), methylamine (MA; green), dimethylamine (DMA; red), trimethylamine (TMA; orange), and ethylenediamine (EDA; brown). Concentrations: [AM] = 10–10000 ppt, [MA] = 1–100 ppt, [DMA/TMA/EDA] = 1–10 ppt.

step. As a result, less configurations have to be passed from the machine learning step to the DFT step, leading to more efficient configurational sampling.

The binding free energies of the lowest free energy configuration according to the configurational sampling were used to estimate the cluster formation rate in ACDC. Figure 12 shows the cluster formation rate at different simulation scheme sizes (up to 4×4), temperatures, and monomer concentrations. For all temperatures and concentrations, the cluster formation rates decrease with increasing simulation scheme size. This is reasonable, as the smaller simulation schemes probably do not include the critical cluster size. The outgrowing clusters, therefore, still have a significant evaporation rate which is not considered, leading to an artificially high cluster formation rate. Surprisingly, even in the case of low temperatures and high monomer concentrations, where the critical cluster size is included in the larger simulation scheme sizes, the cluster formation rate still doesn't seem to converge. Apparently, clusters well beyond the critical cluster size need to be included to obtain the true cluster formation rate.

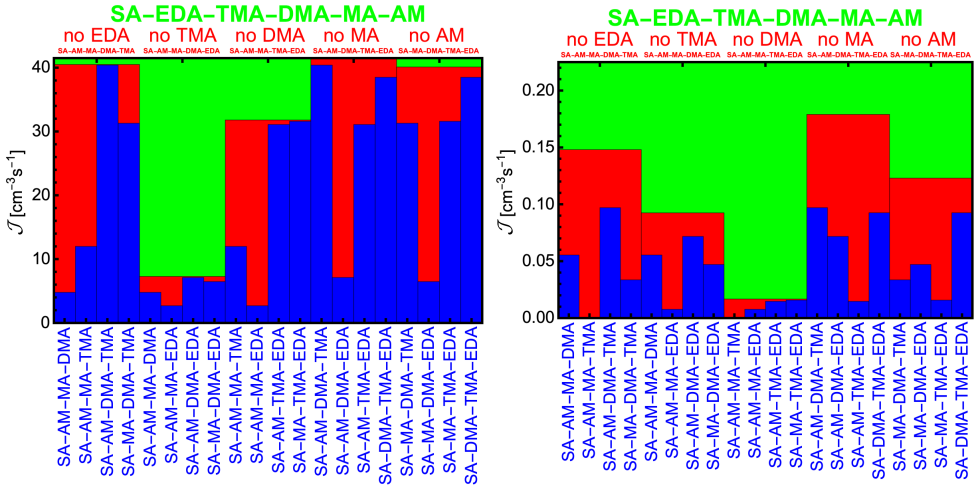


Figure 13: Synergy in cluster formation at high monomer concentrations at 278.15 K (left) and 298.15 K (right) for the binary sulfuric acid–base systems with bases: ammonia (AM), methylamine (MA), dimethylamine (DMA), trimethylamine (TMA), and ethylenediamine (EDA). Green represents cluster formation rates $\mathcal{J}$ including all 5 bases, red columns represent $\mathcal{J}$ including 4 bases, and blue columns (twice repeated within each graph) represent $\mathcal{J}$ including only 3 bases. Concentrations: [AM] = 10000 ppt, [MA] = 100 ppt, [DMA/TMA/EDA] = 10 ppt.

Figure 13 shows the cluster formation rates of systems containing different combinations of bases at two different temperatures. At the lower temperature, sulfuric acid is saturated by the bases and only three bases (e.g., AM–DMA–TMA or DMA–TMA–EDA) are needed to capture most of the cluster formation rate. At the higher temperature, all bases contribute significantly to the cluster formation rate. These results show that depending on the concentrations and temperature, there is a significant amount of synergy between the bases in sulfuric acid–base systems. All relevant bases need to be included to model realistic atmospheric nucleation rates.

4.6 Separation of isomers using a differential mobility analyser (DMA): Comparison of experimental vs modelled ion mobility (Paper I)

In **Paper I**, we examined if computational ion mobility predictions can be used to assign peaks in an ion mobility spectrum to specific isomers. This is of particular importance to NPF studies, as there are many isomers among the compounds that (potentially) contribute to atmospheric cluster formation. As a proof of concept, we focused on the three structural isomers of dihydroxybenzene: catechol, resorcinol, and hydroquinone. The ion mobility of these compounds was measured using chemical ionization mass spectrometry with both nitrate and acetate as the charging ion. The measured ion mobility is, therefore, that of the dihydroxybenzene–charging ion dimer. We performed configurational sampling according to the JKCS procedure (**Paper XI**) to determine the geometry and partial charges of the lowest free energy dihydroxybenzene–charging ion configuration for each isomer. Computational ion mobility predictions were subsequently calculated using the Lennard-Jones trajectory method with quadrupole corrections in the IMoS program.

Table 7 compares the experimental and modeled ion mobilities. Overall, the computational ion mobility predictions are in excellent agreement with the experimental values, especially for the dihydroxybenzene–nitrate dimers, where the discrepancy is below 2%. However, the experimental ion mobilities for the different isomers are very close in value. Thus, it is not possible to directly assign peaks based on modeled ion mobilities. For example, resorcinol–nitrate's modeled ion mobility is closer to the experimental ion mobility of hydroquinone–nitrate than its own experimental value.

On the other hand, both with nitrate and acetate as charging ion, the experimental ion mobility order between the isomers is conserved in the computational predictions. This preservation of the experimental order also turns out to be quite robust. For instance, while the projection approximation method predicts far too large ion mobility values, it still maintains the correct order. This is mainly due to significant differences in geometry between the isomers. The ortho-placement of the hydroxyl groups allows catechol to form two hydrogen bonds with both nitrate and acetate, creating a relatively compact dimer with larger ion mobility. Resorcinol and hydroquinone, on the other hand, can only form one hydrogen bond. For other isomers with significant structural differences, it is similarly expected that peaks in the ion mobility spectrum can be

Table 7: The experimental reduced ion mobility (Z_0^{exp}), Lennard-Jones trajectory method reduced ion mobility (Z_0^{LJTM}), percentual difference between experimental and modeled ion mobility ($\Delta Z_0/Z_0^{\mathrm{exp}}$, where $\Delta Z_0 = Z_0^{\mathrm{LJTM}} - Z_0^{\mathrm{exp}}$), and projected area ion mobility (Z^{PA}) for dimers of catechol, resorcinol, and hydroquinone with either nitrate or acetate. Dimer geometries were obtained with the ωB97X-D functional and 6–31++G(d,p) basis set. Partial charges were obtained with the CHelp method.

System	Z_0^{exp} ($\mathrm{cm^2V^{-1}s^{-1}}$)	Z_0^{LJTM} ($\mathrm{cm^2V^{-1}s^{-1}}$)	$\Delta Z_0/Z_0^{\mathrm{exp}}$ (%)	Z^{PA} ($\mathrm{cm^2V^{-1}s^{-1}}$)
catechol–nitrate	1.76	1.74	-1.4	2.78
resorcinol–nitrate	1.71	1.68	-1.7	2.76
hydroquinone–nitrate	1.69	1.66	-1.6	2.72
catechol–acetate	1.65	1.73	4.6	2.64
resorcinol–acetate	1.60	1.70	6.3	2.62
hydroquinone–acetate	1.58	1.67	5.6	2.58

identified by comparing their order to that of computational ion mobility predictions.

5 Conclusions and future outlook

Aerosol particles represent one of the largest uncertainties in current climate models. This uncertainty is mainly due to a lack of knowledge about the initial particle formation steps, where clusters of ∼2 nm are formed from vapors. The exact pathways that lead to stable clusters able to function as cloud condensation nuclei are still unknown. In particular, the extent to which ion-induced pathways play a role in new particle formation in different environments is still subject of lively debate.

Where the majority of current research is focused on neutral pathways, especially those involving acid–base clusters, this thesis has a particular focus on ions in cluster formation. The competitiveness of ion-induced cluster formation pathways depends on the cluster formation rate, which can be modeled through dynamical models. The accuracy of these models, however, relies on the accuracy of the collision and evaporation rate coefficients used in the model.

We have shown that the Su & Chesnavich parameterization, commonly used to predict ion–dipole collision rate coefficients, performs well for many small and relatively simple collision systems, but shows larger discrepancies for systems containing larger and more complex clusters. It also fails to reproduce the large collision rate coefficient for the sulfuric acid–acetate collision system, as it neglects key properties, such as the dipole moment of the ion. To accurately model collision rate coefficients of systems involving ions, we thus need better theoretical or computational models. An example of an improvement is our method to compute collision rate coefficients using the central field approach with interaction parameters calculated from a fit to the potential of mean force. The interaction parameters are calculated using a force field description, but the method is far more efficient than brute force sampling over many force-field-based molecular dynamics trajectory simulations. Another improvement is the interacting hard-sphere model, which addresses two issues of the central field approach: (1) the introduction of a collision criterion based on the hard-sphere radii, while unnecessary for some systems, is important for accurate results in collision systems such as CO_2 cluster formation, and (2) the application of Hamaker's approach to calculate collision rate coefficients for clusters of any size solves the problem of obtaining interaction parameters (e.g., Lennard-Jones parameters) for larger cluster. Any theoretical or computational model has to be carefully scrutinized to determine if it actually produces physically sound behavior. We showed that global thermostats like Nosé–Hoover and

Canonical Sampling through Velocity Rescaling take an excessive amount of time to achieve equipartition for molecular dynamics simulations in the free-molecular regime. Additionally, these thermostats unphysically alter the temperature of the vapor in cluster formation simulations by globally adjusting the temperature after latent heat release.

Ions also play a role in the detection of species of interest in new particle formation, both in chemical ionization and in ion mobility mass spectrometry. We developed a proof of concept that computational ion mobility predictions can be used to assign peaks in ion mobility spectra to specific isomers by comparing the experimental and modeled ion mobility order.

Based on the above results, there are several avenues to further advance these research topics, such as

- Using collision rate coefficients from the central field approach with interaction parameters fitted to the potential of mean force and evaporation rate coefficients from quantum chemistry based on the lowest free energy structures found through configurational sampling, we can compare neutral and ion-induced pathways of specific nucleating systems. A good candidate system is clusters of iodine oxoacids, where the neutral pathway occurs through sequential addition of iodic acid and iodous acid, and the ion-induced pathway through addition of iodic acid onto an iodate nucleus (He et al., 2021).

- It is not clear why the Su & Chesnavich parameterization fails to provide collision rate coefficients in good agreement with experiments for the sulfuric acid–acetate collision system. *Ab initio* molecular dynamics collision simulations could be performed to exactly elucidate which properties of the system are missing or misrepresented in the Su & Chesnavich model.

- We should establish a clear rule grounded in readily accessible parameters of the collision system to assess when a collision criterion based on the hard-sphere radii of the collision partners is needed.

- We should compare experimental and modeled ion mobilities for more complex isomer systems to ascertain if the experimental ion mobility order is generally preserved in computational predictions.

In a broader context, future research should increasingly use the capabilities of machine learning. New particle formation is characterized by numerous compounds, a wide size range, and extended timescales from a molecular modeling perspective. Even with considerable computational advances, brute force computation will be too time and resource intensive. The ability of machine learning to find functional relations between data has the potential to circumvent expensive computations and can lead to novel insights into the physics governing the behavior of a system. Machine learning algorithms can, for instance, be trained to provide machine learned force fields for molecular dynamics simulations, providing *ab initio* accuracy with the efficiency of classical force field methods. The use of machine learning algorithms should, however, be accompanied by robust tests, as the derived functional relationship can typically not be mapped to physically intuitive contributions.

Accurate, scalable, and transferable machine learning models generally need a large amount of training data. Besides machine learning, the field of atmospheric cluster studies, therefore, also needs smart theoretical approaches to overcome computational obstacles. For example, Zanca et al. (2020) treated the binding energy as a parameter to determine the survival probability of various clusters containing highly oxygenated organic molecules in the atmospheric pressure interface (APi). In this work, the significant computational cost and complexity to calculate accurate binding energies of clusters containing flexible highly oxygenated organic molecules is circumvented by treating the binding energies as a parameter. Similar smart approaches are needed in the future to deal with the complexity of atmospheric new particle formation. Given the escalating impacts of climate change and air pollution globally, a better understanding of new particle formation is vital.

References

Aalto, P., Hämeri, K., Becker, E., Weber, R., Salm, J., Mäkelä, J. M., Hoell, C., O'dowd, C. D., Karlsson, H., Hansson, H.-C., Väkevä, M., Koponen, I. K., Buzorius, G., & Kulmala, M. (2001). Physical characterization of aerosol particles during nucleation events. *Tellus B: Chemical and Physical Meteorology*, *53*, 344–358. https://doi.org/10.3402/tellusb.v53i4.17127

Almeida, J., Schobesberger, S., Kürten, A., Ortega, I. K., Kupiainen-Määttä, O., Praplan, A. P., Adamov, A., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Dommen, J., Donahue, N. M., Downard, A., Dunne, E., Duplissy, J., Ehrhart, S., Flagan, R. C., Franchin, A., Guida, R., Hakala, J., Hansel, A., Heinritzi, M., Henschel, H., Jokinen, T., Junninen, H., Kajos, M., Kangasluoma, J., Keskinen, H., Kupc, A., Kurtén, T., Kvashin, A. N., Laaksonen, A., Lehtipalo, K., Leiminger, M., Leppä, J., Loukonen, V., Makhmutov, V., Mathot, S., McGrath, M. J., Nieminen, T., Olenius, T., Onnela, A., Petäjä, T., Riccobono, F., Riipinen, I., Rissanen, M., Rondo, L., Ruuskanen, T., Santos, F. D., Sarnela, N., Schallhart, S., Schnitzhofer, R., Seinfeld, J. H., Simon, M., Sipilä, M., Stozhkov, Y., Stratmann, F., Tomé, A., Tröstl, J., Tsagkogeorgas, G., Vaattovaara, P., Viisanen, Y., Virtanen, A., Vrtala, A., Wagner, P. E., Weingartner, E., Wex, H., Williamson, C., Wimmer, D., Ye, P., Yli-Juuti, T., Carslaw, K. S., Kulmala, M., Curtius, J., Baltensperger, U., Worsnop, D. R., Vehkamäki, H., & Kirkby, J. (2013). Molecular understanding of sulphuric acid–amine particle nucleation in the atmosphere. *Nature*, *502*(7471), 359–363. https://doi.org/10.1038/nature12663

Amelynck, C., Schoon, N., Kuppens, T., Bultinck, P., & Arijs, E. (2005). A selected ion flow tube study of the reactions of H_3O^+, NO^+ and $O_2{}^+$ with some oxygenated biogenic volatile organic compounds. *International Journal of Mass Spectrometry*, *247*(1), 1–9. https://doi.org/10.1016/j.ijms.2005.08.010

AQEG. (2005). *Particulate Matter in the UK*. Defra Publications.

Atkins, P., & de Paula, J. (2014). *Physical Chemistry: Thermodynamics, Structure, and Change* (10th ed.). W. H. Freeman; Company.

Bannwarth, C., Ehlert, S., & Grimme, S. (2019). GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions.

Journal of Chemical Theory and Computation, *15*(3), 1652–1671. https://doi.org/10.1021/acs.jctc.8b01176

Barducci, A., Bonomi, M., & Parrinello, M. (2011). Metadynamics. *WIREs Computational Molecular Science*, *1*(5), 826–843. https://doi.org/10.1002/wcms.31

Barducci, A., Bussi, G., & Parrinello, M. (2008). Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters*, *100*(2), 020603. https://doi.org/10.1103/PhysRevLett.100.020603

Bartlett, R. J., & Musiał, M. (2007). Coupled-cluster theory in quantum chemistry. *Reviews of Modern Physics*, *79*, 291–352. https://doi.org/10.1103/RevModPhys.79.291

Bazilevskaya, G. A., Usoskin, I. G., Flückiger, E. O., Harrison, R. G., Desorgher, L., Bütikofer, R., Krainev, M. B., Makhmutov, V. S., Stozhkov, Y. I., Svirzhevskaya, A. K., Svirzhevsky, N. S., & Kovaltsov, G. A. (2008). Cosmic Ray Induced Ion Production in the Atmosphere. *Space Science Reviews*, *137*(1-4), 149–173. https://doi.org/10.1007/s11214-008-9339-y

Beck, L. J., Schobesberger, S., Sipilä, M., Kerminen, V.-M., & Kulmala, M. (2022). Estimation of sulfuric acid concentration using ambient ion composition and concentration data obtained with atmospheric pressure interface time-of-flight ion mass spectrometer. *Atmospheric Measurement Techniques*, *15*(6), 1957–1965. https://doi.org/10.5194/amt-15-1957-2022

Bianchi, F., Kurtén, T., Riva, M., Mohr, C., Rissanen, M. P., Roldin, P., Berndt, T., Crounse, J. D., Wennberg, P. O., Mentel, T. F., Wildt, J., Junninen, H., Jokinen, T., Kulmala, M., Worsnop, D. R., Thornton, J. A., Donahue, N., Kjaergaard, H. G., & Ehn, M. (2019). Highly Oxygenated Organic Molecules (HOM) from Gas-Phase Autoxidation Involving Peroxy Radicals: A Key Contributor to Atmospheric Aerosol. *Chemical Reviews*, *119*(6), 3472–3509. https://doi.org/10.1021/acs.chemrev.8b00395

Bianco, A., Neefjes, I., Alfaouri, D., Vehkamäki, H., Kurtén, T., Ahonen, L., Passananti, M., & Kangasluoma, J. (2022). Separation of isomers using a differential mobility analyser (DMA): Comparison of experimental vs modelled ion mobility. *Talanta*, *243*(1), 123339. https://doi.org/10.1016/j.talanta.2022.123339

Born, M., & Oppenheimer, R. (1927). Zur Quantentheorie der Molekeln. *Annalen der Physik*, *389*(20), 457–484. https://doi.org/10.1002/andp.19273892002

Born, M. (1926). Zur Quantenmechanik der Stoßvorgänge. *Zeitschrift für Physik*, *37*(12), 863–867. https://doi.org/10.1007/bf01397477

Borsdorf, H., & Eiceman, G. A. (2006). Ion mobility spectrometry: principles and applications. *Applied Spectroscopy Reviews*, *41*(4), 323–375. https://doi.org/10.1080/05704920600663469

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M., Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S., Sherwood, S., Stevens, B., & Zhang, X. (2013). Clouds and Aerosols. In T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, & P. Midgley (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 571–657). Cambridge University Press.

Bransden, B. H., & Joachain, C. J. (2000). *Quantum Mechanics*. Pearson Education Limited.

Brillouin, L. (1934). Les Champs 'Self-consistent' de Hartree et de Fock. *Actualités scientifiques et industrielles*, (159), 37.

Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, *126*(1), 014101. https://doi.org/10.1063/1.2408420

Bussi, G., & Parrinello, M. (2007). Accurate sampling using Langevin dynamics. *Physical Review E*, *75*(5), 056707. https://doi.org/10.1103/PhysRevE.75.056707

Chai, J.-D., & Head-Gordon, M. (2008). Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.*, *10*, 6615–6620. https://doi.org/10.1039/B810189B

Chen, D., Rojas, M., Samset, B. H., Cobb, K., Niang, A. D., Edwards, P., Emori, S., Faria, S., Hawkins, E., Hope, P., Huybrechts, P., Meinshausen, M., Mustafa, S., Plattner, G.-K., Edwards, P., Emori, S., Faria, S., Hawkins, E., Hope, P., Huybrechts, P., Meinshausen, M., & Tréguier, A.-M. (2021). Framing, Context, and Methods. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (Eds.), *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 147–286). Cambridge University Press. https://doi.org/10.1017/9781009157896.003

Chen, H., Brandt, R., Bandur, R., & Hoffmann, T. (2006). Characterization of iodine species in the marine aerosol: To understand their roles in particle formation processes. *Frontiers of Chemistry in China*, *1*(2), 1673–3495. https://doi.org/10.1007/s11458-006-0001-5

Chen, M. (2021). Collective variable-based enhanced sampling and machine learning. *The European Physical Journal B*, *94*(10). https://doi.org/10.1140/epjb/s10051-021-00220-w

Chen, X., Kerminen, V.-M., Paatero, J., Paasonen, P., Manninen, H. E., Nieminen, T., Petäjä, T., & Kulmala, M. (2016). How do air ions reflect variations in ionising radiation in the lower atmosphere in a boreal forest? *Atmospheric Chemistry and Physics*, *16*(22), 14297–14315. https://doi.org/10.5194/acp-16-14297-2016

Christensen, A. S., Bratholm, L. A., Faber, F. A., & Anatole von Lilienfeld, O. (2020). FCHL revisited: Faster and more accurate quantum machine learning. *The Journal of Chemical Physics*, *152*(4), 044107. https://doi.org/10.1063/1.5126701

Coots, J., Gandhi, V., Onakoya, T., Chen, X., & Larriba-Andaluz, C. (2020). A parallelized tool to calculate the electrical mobility of charged aerosol nanoparticles and ions in the gas phase. *Journal of Aerosol Science*, *147*, 105570. https://doi.org/10.1016/j.jaerosci.2020.105570

Crounse, J. D., Nielsen, L. B., Jørgensen, S., Kjaergaard, H. G., & Wennberg, P. O. (2013). Autoxidation of Organic Compounds in the Atmosphere. *The Journal of Physical Chemistry Letters*, *4*(20), 3513–3520. https://doi.org/10.1021/jz4019207

Dass, C. (2007). Basics of Mass Spectrometry. In *Fundamentals of Contemporary Mass Spectrometry* (pp. 1–14). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470118498.ch1

Dawson, M. L., Varner, M. E., Perraud, V., Ezell, M. J., Gerber, R. B., & Finlayson-Pitts, B. J. (2012). Simplified mechanism for new particle formation from methanesulfonic acid, amines, and water via experiments and ab initio calculations. *Proceedings of the National Academy of Sciences*, *109*(46), 18719–18724. https://doi.org/10.1073/pnas.1211878109

de Hoffmann, E., & Stroobant, V. (2007). Ion Sources. In *Mass Spectrometry: Principles and Applications* (pp. 15–84). John Wiley & Sons, Ltd.

Dick, B. G., & Overhauser, A. W. (1958). Theory of the Dielectric Constants of Alkali Halide Crystals. *Physical Review*, *112*, 90–103. https://doi.org/10.1103/PhysRev.112.90

Duplissy, J., Merikanto, J., Franchin, A., Tsagkogeorgas, G., Kangasluoma, J., Wimmer, D., Vuollekoski, H., Schobesberger, S., Lehtipalo, K., Flagan, R. C., Brus, D., Donahue, N. M., Vehkamäki, H., Almeida, J., Amorim, A., Barmet, P., Bianchi, F., Breitenlechner, M., Dunne, E. M., Guida, R., Henschel, H., Junninen, H., Kirkby, J., Kürten, A., Kupc, A., Määttänen, A., Makhmutov, V., Mathot, S., Nieminen, T., Onnela, A., Praplan, A. P., Riccobono, F., Rondo, L., Steiner, G., Tome, A., Walther, H., Baltensperger, U., Carslaw, K. S., Dommen, J., Hansel, A., Petäjä, T., Sipilä, M., Stratmann, F., Vrtala, A., Wagner, P. E., Worsnop, D. R., Curtius, J., & Kulmala, M. (2016). Effect of ions on sulfuric acid-water binary particle formation: 2. Experimental data and comparison with QC-normalized classical nucleation theory. *Journal of Geophysical Research: Atmospheres*, *121*(4), 1752–1775. https://doi.org/10.1002/2015JD023539

Dusek, U., Frank, G. P., Hildebrandt, L., Curtius, J., Schneider, J., Walter, S., Chand, D., Drewnick, F., Hings, S., Jung, D., Borrmann, S., & Andreae, M. O. (2006). Size Matters More Than Chemistry for Cloud-Nucleating Ability of Aerosol Particles. *Science*, *312*(5778), 1375–1378. https://doi.org/10.1126/science.1125261

Eisele, F. L. (1989). Natural and anthropogenic negative ions in the troposphere. *Journal of Geophysical Research: Atmospheres*, *94*(D2), 2183–2196. https://doi.org/10.1029/JD094iD02p02183

Eisele, F. L., & Tanner, D. J. (1993). Measurement of the gas phase concentration of H_2SO_4 and methane sulfonic acid and estimates of H_2SO_4 production and loss in the atmosphere. *Journal of Geophysical Research: Atmospheres*, *98*(D5), 9001–9010. https://doi.org/10.1029/93JD00031

Elm, J., Kubečka, J., Besel, V., Jääskeläinen, M. J., Halonen, R., Kurtén, T., & Vehkamäki, H. (2020). Modeling the formation and growth of atmospheric molecular clusters: A review. *Journal of Aerosol Science*, *149*, 105621. https://doi.org/10.1016/j.jaerosci.2020.105621

Elstner, M., & Seifert, G. (2014). Density functional tight binding. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *372*(2011), 20120483. https://doi.org/10.1098/rsta.2012.0483

Falcon-Rodriguez, C. I., Osornio-Vargas, A. R., Sada-Ovalle, I., & Segura-Medina, P. (2016). Aeroparticles, Composition, and Lung Diseases. *Frontiers in Immunology*, *7*, 3. https://doi.org/10.3389/fimmu.2016.00003

Fews, A., Henshaw, D. L., Wilding, R. J., & Keitch, P. A. (1999). Corona ions from powerlines and increased exposure to pollutant aerosols. *International Journal of Radiation Biology*, *75*(12), 1523–1531. https://doi.org/10.1080/095530099139124

Finkenzeller, H., Iyer, S., He, X.-C., Simon, M., Koenig, T. K., Lee, C. F., Valiev, R., Hofbauer, V., Amorim, A., Baalbaki, R., Baccarini, A., Beck, L., Bell, D. M., Caudillo, L., Chen, D., Chiu, R., Chu, B., Dada, L., Duplissy, J., Heinritzi, M., Kemppainen, D., Kim, C., Krechmer, J., Kürten, A., Kvashnin, A., Lamkaddam, H., Lee, C. P., Lehtipalo, K., Li, Z., Makhmutov, V., Manninen, H. E., Marie, G., Marten, R., Mauldin, R. L., Mentler, B., Müller, T., Petäjä, T., Philippov, M., Ranjithkumar, A., Rörup, B., Shen, J., Stolzenburg, D., Tauber, C., Tham, Y. J., Tomé, A., Vazquez-Pufleau, M., Wagner, A. C., Wang, D. S., Wang, M., Wang, Y., Weber, S. K., Nie, W., Wu, Y., Xiao, M., Ye, Q., Zauner-Wieczorek, M., Hansel, A., Baltensperger, U., Brioude, J., Curtius, J., Donahue, N. M., Haddad, I. E., Flagan, R. C., Kulmala, M., Kirkby, J., Sipilä, M., Worsnop, D. R., Kurten, T., Rissanen, M., & Volkamer, R. (2023). The gas-phase formation mechanism of iodic acid as an atmospheric aerosol source. *Nature Chemistry*, *15*(1), 129–135. https://doi.org/10.1038/s41557-022-01067-z

Fomete, S. K. W., Johnson, J. S., Myllys, N., Neefjes, I., Reischl, B., & Jen, C. N. (2022). Ion–Molecule Rate Constants for Reactions of Sulfuric Acid with Acetate and Nitrate Ions. *The Journal of Physical Chemistry A*, *126*(44), 8240–8248. https://doi.org/10.1021/acs.jpca.2c02072

Frenkel, D., & Smit, B. (2002). *Understanding Molecular Simulation: From Algorithms to Applications* (2nd ed.). Academic Press. https://doi.org/10.1016/B978-0-12-267351-1.X5000-7

Gagné, S., Laakso, L., Petäjä, T., Kerminen, V.-M., & Kulmala, M. (2008). Analysis of one year of Ion-DMPS data from the SMEAR II station, Finland. *Tellus B*, *60*(3), 318–329. https://doi.org/10.1111/j.1600-0889.2008.00347.x

Gan, W. Q., FitzGerald, J. M., Carlsten, C., Sadatsafavi, M., & Brauer, M. (2013). Associations of Ambient Air Pollution with Chronic Obstructive Pulmonary Disease Hospitalization and Mortality. *American Journal of Respiratory and Critical Care Medicine*, *187*(7), 721–727. https://doi.org/10.1164/rccm.201211-2004oc

Gioumousis, G., & Stevenson, D. P. (1958). Reactions of Gaseous Molecule Ions with Gaseous Molecules. V. Theory. *Journal of Chemical Physics*, *29*(2), 294–299. https://doi.org/10.1063/1.1744477

Gómez Martín, J. C., Lewis, T. R., James, A. D., Saiz-Lopez, A., & Plane, J. M. C. (2022). Insights into the Chemistry of Iodine New Particle Formation: The Role of Iodine Oxides and the Source of Iodic Acid. *Journal of the American Chemical Society*, *144*(21), 9240–9253. https://doi.org/10.1021/jacs.1c12957

Haack, A., Ieritano, C., & Hopkins, W. S. (2023). MobCal-MPI 2.0: an accurate and parallelized package for calculating field-dependent collision cross sections and ion mobilities. *Analyst*, *148*, 3257–3273. https://doi.org/10.1039/D3AN00545C

Hallquist, M., Wenger, J. C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M., Dommen, J., Donahue, N. M., George, C., Goldstein, A. H., Hamilton, J. F., Herrmann, H., Hoffmann, T., Iinuma, Y., Jang, M., Jenkin, M. E., Jimenez, J. L., Kiendler-Scharr, A., Maenhaut, W., McFiggans, G., Mentel, T. F., Monod, A., Prévôt, A. S. H., Seinfeld, J. H., Surratt, J. D., Szmigielski, R., & Wildt, J. (2009). The formation, properties and impact of secondary organic aerosol: current and emerging issues. *Atmospheric Chemistry and Physics*, *9*(14), 5155–5236. https://doi.org/10.5194/acp-9-5155-2009

Halonen, R. (2021). *Mechanisms of cluster formation in the gas phase* (PhD thesis). University of Helsinki.

Halonen, R., Neefjes, I., & Reischl, B. (2023). Further cautionary tales on thermostatting in molecular dynamics: Energy equipartitioning and non-equilibrium processes in gas-phase simulations. *Journal of Chemical Physics*, *158*(19), 194301. https://doi.org/10.1063/5.0148013

Halonen, R., Zapadinsky, E., Kurtén, T., Vehkamäki, H., & Reischl, B. (2019). Rate enhancement in collisions of sulfuric acid molecules due to long-range intermolecular forces. *Atmospheric Chemistry and Physics Discussions*, *19*(21), 13355–13366. https://doi.org/10.5194/acp-2019-400

Hamaker, H. (1937). The London–van der Waals attraction between spherical particles. *Physica*, *4*(10), 1058–1072. https://doi.org/10.1016/S0031-8914(37)80203-7

Harold, S. E., Bready, C. J., Juechter, L. A., Kurfman, L. A., Vanovac, S., Fowler, V. R., Mazaleski, G. E., Odbadrakh, T. T., & Shields, G. C. (2022). Hydrogen-Bond Topology Is More Important Than Acid/Base Strength in Atmospheric Prenucleation Clusters. *The Journal of Physical Chemistry A*, *126*(10), 1718–1728. https://doi.org/10.1021/acs.jpca.1c10754

Harrison, R. G., & Tammet, H. (2008). Ions in the Terrestrial Atmosphere and Other Solar System Atmospheres. *Space Science Reviews*, *137*(1–4), 107–118. https://doi.org/10.1007/s11214-008-9356-x

Haywood, J., & Boucher, O. (2000). Estimates of the direct and indirect radiative forcing due to tropospheric aerosols: A review. *Reviews of Geophysics*, *38*(4), 513–543. https://doi.org/10.1029/1999RG000078

He, X.-C., Tham, Y. J., Dada, L., Wang, M., Finkenzeller, H., Stolzenburg, D., Iyer, S., Simon, M., Kürten, A., Shen, J., Rörup, B., Rissanen, M., Schobesberger, S., Baalbaki, R., Wang, D. S., Koenig, T. K., Jokinen, T., Sarnela, N., Beck, L. J., Almeida, J., Amanatidis, S., Amorim, A., Ataei, F., Baccarini, A., Bertozzi, B., Bianchi, F., Brilke, S., Caudillo, L., Chen, D., Chiu, R., Chu, B., Dias, A., Ding, A., Dommen, J., Duplissy, J., Haddad, I. E., Carracedo, L. G., Granzin, M., Hansel, A., Heinritzi, M., Hofbauer, V., Junninen, H., Kangasluoma, J., Kemppainen, D., Kim, C., Kong, W., Krechmer, J. E., Kvashin, A., Laitinen, T., Lamkaddam, H., Lee, C. P., Lehtipalo, K., Leiminger, M., Li, Z., Makhmutov, V., Manninen, H. E., Marie, G., Marten, R., Mathot, S., Mauldin, R. L., Mentler, B., Möhler, O., Müller, T., Nie, W., Onnela, A., Petäjä, T., Pfeifer, J., Philippov, M., Ranjithkumar, A., Saiz-Lopez, A., Salma, I., Scholz, W., Schuchmann, S., Schulze, B., Steiner, G., Stozhkov, Y., Tauber, C., Tomé, A., Thakur, R. C., Väisänen, O., Vazquez-Pufleau, M., Wagner, A. C., Wang, Y., Weber, S. K., Winkler, P. M., Wu, Y., Xiao, M., Yan, C., Ye, Q., Ylisirniö, A., Zauner-Wieczorek, M., Zha, Q., Zhou, P., Flagan, R. C., Curtius, J., Baltensperger, U., Kulmala, M., Kerminen, V.-M., Kurtén, T., Donahue, N. M., Volkamer, R., Kirkby, J., Worsnop, D. R., & Sipilä, M. (2021). Role of iodine oxoacids in atmospheric aerosol nucleation. *Science*, *371*(6529), 589–595. https://doi.org/10.1126/science.abe0298

Hehre, W. J., Ditchfield, R., & Pople, J. A. (1972). Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *The Journal of Chemical Physics*, *56*(5), 2257–2261. https://doi.org/10.1063/1.1677527

Hirsikko, A., Bergman, T., Laakso, L., Maso, M. D., Riipinen, I., Hõrrak, U., & Kulmala, M. (2007). Identification and classification of the formation of intermediate ions measured in boreal forest. *Atmospheric Chemistry and Physics*, *7*(1), 201–210. https://doi.org/10.5194/acp-7-201-2007

Hirsikko, A., Nieminen, T., Gagné, S., Lehtipalo, K., Manninen, H. E., Ehn, M., Hõrrak, U., Kerminen, V.-M., Laakso, L., McMurry, P. H., Mirme, A., Mirme, S., Petäjä, T., Tammet, H., Vakkari, V., Vana, M., & Kulmala, M. (2011). Atmospheric ions and nucleation: a review of observations. *Atmospheric Chemistry and Physics*, *11*(2), 767–798. https://doi.org/10.5194/acp-11-767-2011

Hohenberg, P., & Kohn, W. (1964). Inhomogeneous Electron Gas. *Physical Review*, *136*, B864–B871. https://doi.org/10.1103/PhysRev.136.B864

Hoover, W. G. (1985). Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, *31*(3), 1695. https://doi.org/10.1103/PhysRevA.31.1695

Iida, K., Stolzenburg, M., McMurry, P., Dunn, M. J., Smith, J. N., Eisele, F., & Keady, P. (2006). Contribution of ion-induced nucleation to new particle formation: Methodology and its application to atmospheric observations in Boulder, Colorado. *Journal of Geophysical Research: Atmospheres*, *111*(D23). https://doi.org/10.1029/2006jd007167

Israël, H. (1970). *Atmospheric electricity*. Israel Program for Scientific Translations.

Israelachvili, J. N. (2011). *Intermolecular and Surface Forces*. Academic Press, London.

Iwahara, N., Huang, Z., Neefjes, I., & Chibotaru, L. F. (2022). Multipolar exchange interaction and complex order in insulating lanthanides. *Physical Review B*, *105*, 144401. https://doi.org/10.1103/PhysRevB.105.144401

Jacobson, M. (2012). *Air Pollution and Global Warming: History, Science, and Solutions* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139109444

Jensen, F. (2001). Polarization consistent basis sets: Principles. *The Journal of Chemical Physics*, *115*(20), 9113–9125. https://doi.org/10.1063/1.1413524

Jensen, F. (2007). *Introduction to Computational Chemistry*. John Wiley & Sons Ltd.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., DeCarlo, P. F., Allan, J. D., Coe, H., Ng, N. L., Aiken, A. C., Docherty, K. S., Ulbrich, I. M., Grieshop, A. P., Robinson, A. L., Duplissy, J., Smith, J. D., Wilson, K. R., Lanz, V. A., Hueglin, C., Sun, Y. L., Tian, J., Laaksonen, A., Raatikainen, T., Rautiainen, J., Vaattovaara, P., Ehn, M., Kulmala, M., Tomlinson, J. M., Collins, D. R., Cubison, M. J., E., Dunlea, J., Huffman, J. A., Onasch, T. B., Alfarra, M. R., Williams, P. I., Bower, K., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Salcedo, D., Cottrell, L., Griffin, R., Takami, A., Miyoshi, T., Hatakeyama, S., Shimono, A., Sun, J. Y., Zhang, Y. M., Dzepina, K., Kimmel, J. R., Sueper,

D., Jayne, J. T., Herndon, S. C., Trimborn, A. M., Williams, L. R., Wood, E. C., Middlebrook, A. M., Kolb, C. E., Baltensperger, U., & Worsnop, D. R. (2009). Evolution of Organic Aerosols in the Atmosphere. *Science*, *326*(5959), 1525–1529. https://doi.org/10.1126/science.1180353

Jokinen, T., Sipilä, M., Kontkanen, J., Vakkari, V., Tisler, P., Duplissy, E.-M., Junninen, H., Kangasluoma, J., Manninen, H. E., Petäjä, T., Kulmala, M., Worsnop, D. R., Kirkby, J., Virkkula, A., & Kerminen, V.-M. (2018). Ion-induced sulfuric acid–ammonia nucleation drives particle formation in coastal Antarctica. *Science Advances*, *4*(11). https://doi.org/10.1126/sciadv.aat9744

Joos, J. J., Neefjes, I., Seijo, L., & Barandiarán, Z. (2021). Charge transfer from Eu^{2+} to trivalent lanthanide co-dopants: Systematic behavior across the series. *The Journal of Chemical Physics*, *154*(6), 064704. https://doi.org/10.1063/5.0037992

Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, *118*(45), 11225–11236. https://doi.org/10.1021/ja9621760

Junninen, H., Ehn, M., Petäjä, T., Luosujärvi, L., Kotiaho, T., Kostiainen, R., Rohner, U., Gonin, M., Fuhrer, K., Kulmala, M., & Worsnop, D. R. (2010). A high-resolution mass spectrometer to measure atmospheric ion composition. *Atmospheric Measurement Techniques*, *3*(4), 1039–1053. https://doi.org/10.5194/amt-3-1039-2010

Kendall, R. A., Dunning, J., Thom H., & Harrison, R. J. (1992). Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *The Journal of Chemical Physics*, *96*(9), 6796–6806. https://doi.org/10.1063/1.462569

Kerminen, V.-M., Chen, X., Vakkari, V., Petäjä, T., Kulmala, M., & Bianchi, F. (2018). Atmospheric new particle formation and growth: review of field observations. *Environmental Research Letters*, *13*(10), 103003. https://doi.org/10.1088/1748-9326/aadf3c

Kerminen, V.-M., & Wexler, A. S. (1995). Growth laws for atmospheric aerosol particles: An examination of the bimodality of the accumulation mode. *Atmospheric Environment*, *29*(22), 3263–3275. https://doi.org/10.1016/1352-2310(95)00249-X

Kildgaard, J. V., Mikkelsen, K. V., Bilde, M., & Elm, J. (2018a). Hydration of atmospheric molecular clusters: A new method for systematic configurational sampling. *The Journal of Physical Chemistry A*, *122*, 5026–5036. https://doi.org/10.1021/acs.jpca.8b02758

Kildgaard, J. V., Mikkelsen, K. V., Bilde, M., & Elm, J. (2018b). Hydration of atmospheric molecular clusters: Organic Acid–Water Clusters. *The Journal of Physical Chemistry A*, *122*, 8549–8556. https://doi.org/10.1021/acs.jpca.8b07713

Kirkby, J., Curtius, J., Almeida, J., Dunne, E., Duplissy, J., Ehrhart, S., Franchin, A., Gagné, S., Ickes, L., Kürten, A., Kupc, A., Metzger, A., Riccobono, F., Rondo, L., Schobesberger, S., Tsagkogeorgas, G., Wimmer, D., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Dommen, J., Downard, A., Ehn, M., Flagan, R. C., Haider, S., Hansel, A., Hauser, D., Jud, W., Junninen, H., Kreissl, F., Kvashin, A., Laaksonen, A., Lehtipalo, K., Lima, J., Lovejoy, E. R., Makhmutov, V., Mathot, S., Mikkilä, J., Minginette, P., Mogo, S., Nieminen, T., Onnela, A., Pereira, P., Petäjä, T., Schnitzhofer, R., Seinfeld, J. H., Sipilä, M., Stozhkov, Y., Stratmann, F., Tomé, A., Vanhanen, J., Viisanen, Y., Vrtala, A., Wagner, P. E., Walther, H., Weingartner, E., Wex, H., Winkler, P. M., Carslaw, K. S., Worsnop, D. R., Baltensperger, U., & Kulmala, M. (2011). Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation. *Nature*, *476*(7361), 429–433. https://doi.org/10.1038/nature10343

Kirkby, J., Duplissy, J., Sengupta, K., Frege, C., Gordon, H., Williamson, C., Heinritzi, M., Simon, M., Yan, C., Almeida, J., Tröstl, J., Nieminen, T., Ortega, I. K., Wagner, R., Adamov, A., Amorim, A., Bernhammer, A.-K., Bianchi, F., Breitenlechner, M., Brilke, S., Chen, X., Craven, J., Dias, A., Ehrhart, S., Flagan, R. C., Franchin, A., Fuchs, C., Guida, R., Hakala, J., Hoyle, C. R., Jokinen, T., Junninen, H., Kangasluoma, J., Kim, J., Krapf, M., Kürten, A., Laaksonen, A., Lehtipalo, K., Makhmutov, V., Mathot, S., Molteni, U., Onnela, A., Peräkylä, O., Piel, F., Petäjä, T., Praplan, A. P., Pringle, K., Rap, A., Richards, N. A. D., Riipinen, I., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger, S., Scott, C. E., Seinfeld, J. H., Sipilä, M., Steiner, G., Stozhkov, Y., Stratmann, F., Tomé, A., Virtanen, A., Vogel, A. L., Wagner, A. C., Wagner, P. E., Weingartner, E., Wimmer, D., Winkler, P. M., Ye, P., Zhang, X., Hansel, A., Dommen, J., Donahue, N. M., Worsnop, D. R., Baltensperger, U., Kulmala, M., Carslaw, K. S., & Curtius, J. (2016). Ion-induced nucleation of pure biogenic particles. *Nature*, *533*(7604), 521–526. https://doi.org/10.1038/nature17953

Knutson, E., & Whitby, K. (1975). Aerosol classification by electric mobility: apparatus, theory, and applications. *Journal of Aerosol Science*, *6*(6), 443–451. https://doi.org/10.1016/0021-8502(75)90060-9

Kohn, W., & Sham, L. J. (1965). Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, *140*, A1133–A1138. https://doi.org/10.1103/PhysRev.140.A1133

Kubečka, J., Besel, V., Neefjes, I., Knattrup, Y., Kurtén, T., Vehkamäki, H., & Elm, J. (2023). Computational Tools for Handling Molecular Clusters: Configurational Sampling, Storage, Analysis, and Machine Learning. *ACS Omega*, *8*(47), 45115–45128. https://doi.org/10.1021/acsomega.3c07412

Kubečka, J., Neefjes, I., Besel, V., Qiao, F., Xie, H.-B., & Elm, J. (2023). Atmospheric Sulfuric Acid–Multi-Base New Particle Formation Revealed through Quantum Chemistry Enhanced by Machine Learning. *The Journal of Physical Chemistry A*, *127*(9), 2091–2103. https://doi.org/10.1021/acs.jpca.3c00068

Kulmala, M., Riipinen, I., Nieminen, T., Hulkkonen, M., Sogacheva, L., Manninen, H. E., Paasonen, P., Petäjä, T., Dal Maso, M., Aalto, P. P., Viljanen, A., Usoskin, I., Vainio, R., Mirme, S., Mirme, A., Minikin, A., Petzold, A., Hõrrak, U., Plaß-Dülmer, C., Birmili, W., & Kerminen, V.-M. (2010). Atmospheric data over a solar cycle: no connection between galactic cosmic rays and new particle formation. *Atmospheric Chemistry and Physics*, *10*(4), 1885–1898. https://doi.org/10.5194/acp-10-1885-2010

Kulmala, M., Kontkanen, J., Junninen, H., Lehtipalo, K., Manninen, H. E., Nieminen, T., Petäjä, T., Sipilä, M., Schobesberger, S., Rantala, P., Franchin, A., Jokinen, T., Järvinen, E., Äijälä, M., Kangasluoma, J., Hakala, J., Aalto, P. P., Paasonen, P., Mikkilä, J., Vanhanen, J., Aalto, J., Hakola, H., Makkonen, U., Ruuskanen, T., Mauldin, R. L., Duplissy, J., Vehkamäki, H., Bäck, J., Kortelainen, A., Riipinen, I., Kurtén, T., Johnston, M. V., Smith, J. N., Ehn, M., Mentel, T. F., Lehtinen, K. E. J., Laaksonen, A., Kerminen, V.-M., & Worsnop, D. R. (2013). Direct Observations of Atmospheric Aerosol Nucleation. *Science*, *339*(6122), 943–946. https://doi.org/10.1126/science.1227385

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., & Kollman, P. A. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, *13*(8), 1011–1021. https://doi.org/10.1002/jcc.540130812

Kurtén, T., Loukonen, V., Vehkamäki, H., & Kulmala, M. (2008). Amines are likely to enhance neutral and ion-induced sulfuric acid-water nucleation in the atmosphere more effectively than ammonia. *Atmospheric Chemistry and Physics*, *8*(14), 4095–4103. https://doi.org/10.5194/acp-8-4095-2008

Kurtén, T., Petäjä, T., Smith, J., Ortega, I. K., Sipilä, M., Junninen, H., Ehn, M., Vehkamäki, H., Mauldin, L., Worsnop, D. R., & Kulmala, M. (2011). The effect of H_2SO_4–amine clustering on chemical ionization mass spectrometry (CIMS) measurements of gas-phase sulfuric acid. *Atmospheric Chemistry and Physics*, *11*(6), 3007–3019. https://doi.org/10.5194/acp-11-3007-2011

Laakso, L., Hirsikko, A., Grönholm, T., Kulmala, M., Luts, A., & Parts, T.-E. (2007). Waterfalls as sources of small charged aerosol particles. *Atmospheric Chemistry and Physics*, *7*(9), 2271–2275. https://doi.org/10.5194/acp-7-2271-2007

Laakso, L., Petäjä, T., Lehtinen, K. E. J., Kulmala, M., Paatero, J., Hõrrak, U., Tammet, H., & Joutsensaari, J. (2004). Ion production rate in a boreal forest based on ion, particle and radiation measurements. *Atmospheric Chemistry and Physics*, *4*(7), 1933–1943. https://doi.org/10.5194/acp-4-1933-2004

Laio, A., & Parrinello, M. (2002). Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, *99*(20), 12562–12566. https://doi.org/10.1073/pnas.202427399

Landau, L. D., & Lifshitz, E. M. (1976). *Mechanics* (Vol. 1). Butterworth-Heinemann.

Langevin, P. (1905). A fundamental formula of kinetic theory. *Annales de chimie et de physique*, *5*(0365-1444), 245–288.

Larriba, C., & Hogan, C. J. J. (2013). Ion Mobilities in Diatomic Gases: Measurement versus Prediction with Non-Specular Scattering Models. *The Journal of Physical Chemistry A*, *117*(19), 3887–3901. https://doi.org/10.1021/jp312432z

Larriba-Andaluz, C. (2023). *IMoS (Ion Mobility Software) User Guide Version 1.13* [Available at https://www.imospedia.com/imos/explanation-of-the-code/].

Larriba-Andaluz, C., Nahin, M., & Shrivastav, V. (2017). A contribution to the amaranthine quarrel between true and average electrical mobility in the free molecular regime. *Aerosol Science and Technology*, *51*(7), 887–895. https://doi.org/10.1080/02786826.2017.1316829

Lehtipalo, K., Rondo, L., Kontkanen, J., Schobesberger, S., Jokinen, T., Sarnela, N., Kürten, A., Ehrhart, S., Franchin, A., Nieminen, T., Riccobono, F., Sipilä, M., Yli-Juuti, T., Duplissy, J., Adamov, A., Ahlm, L., Almeida, J., Amorim, A., Bianchi, F., Breitenlechner, M., Dommen, J., Downard, A. J., Dunne, E. M.,

Flagan, R. C., Guida, R., Hakala, J., Hansel, A., Jud, W., Kangasluoma, J., Kerminen, V.-M., Keskinen, H., Kim, J., Kirkby, J., Kupc, A., Kupiainen-Määttä, O., Laaksonen, A., Lawler, M. J., Leiminger, M., Mathot, S., Olenius, T., Ortega, I. K., Onnela, A., Petäjä, T., Praplan, A., Rissanen, M. P., Ruuskanen, T., Santos, F. D., Schallhart, S., Schnitzhofer, R., Simon, M., Smith, J. N., Tröstl, J., Tsagkogeorgas, G., Tomé, A., Vaattovaara, P., Vehkamäki, H., Vrtala, A. E., Wagner, P. E., Williamson, C., Wimmer, D., Winkler, P. M., Virtanen, A., Donahue, N. M., Carslaw, K. S., Baltensperger, U., Riipinen, I., Curtius, J., Worsnop, D. R., & Kulmala, M. (2016). The effect of acid–base clustering and ions on the growth of atmospheric nano-particles. *Nature Communications*, *7*(1), 11594. https://doi.org/10.1038/ncomms11594

Levine, I. (2014). *Quantum Chemistry* (7th ed.). Pearson.

Li, Q., Jiang, J., & Hao, J. (2015). A Review of Aerosol Nanoparticle Formation from Ions. *KONA Powder and Particle Journal*, *32*(0), 57–74. https://doi.org/10.14356/kona.2015013

Li, Y., Guo, X., Wang, T., Zhao, Y., Zhang, H., & Wang, W. (2015). Characteristics of Atmospheric Small Ions and Their Application to Assessment of Air Quality in a Typical Semi-Arid City of Northwest China. *Aerosol and Air Quality Research*, *15*(3), 865–874. https://doi.org/10.4209/aaqr.2014.06.0123

Longair, M. (2005). Statistical ands Quantum Physics [accessed: 29-01-2024]. https://edisciplinas.usp.br/mod/folder/view.php?id=70077

Luts, A., & Parts, T. (2002). Evolution of negative small air ions at two different temperatures. *Journal of Atmospheric and Solar-Terrestrial Physics*, *64*(7), 763–774. https://doi.org/https://doi.org/10.1016/S1364-6826(02)00076-7

Manninen, H. E., Nieminen, T., Riipinen, I., Yli-Juuti, T., Gagné, S., Asmi, E., Aalto, P. P., Petäjä, T., Kerminen, V.-M., & Kulmala, M. (2009). Charged and total particle formation and growth rates during EUCAARI 2007 campaign in Hyytiälä. *Atmospheric Chemistry and Physics*, *9*(12), 4077–4089. https://doi.org/10.5194/acp-9-4077-2009

Marenduzzo, D. (2015). Thermostats in Molecular dynamics [accessed: 06-02-2024]. www2.ph.ed.ac.uk/~dmarendu/MVP/MVP03.pdf

Martyna, G. J., Klein, M. L., & Tuckerman, M. (1992). Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *The Journal of Chemical Physics*, *97*(4), 2635–2643. https://doi.org/10.1063/1.463940

McGrath, M. J., Olenius, T., Ortega, I. K., Loukonen, V., Paasonen, P., Kurtén, T., Kulmala, M., & Vehkamäki, H. (2012). Atmospheric Cluster Dynamics Code: a flexible method for solution of the birth-death equations. *Atmospheric Chemistry and Physics*, *12*, 2345–2355. https://doi.org/10.5194/acp-12-2345-2012

McQuarrie, D. A. (1976). *Statistical Mechanics*. Harper & Row.

McQuarrie, D. A. (2008). *Quantum Chemistry*. University Science Books.

McQuarrie, D. A., & Simon, J. D. (1999). *Molecular Thermodynamics*. University Science Books.

Merikanto, J., Zapadinsky, E., Lauri, A., & Vehkamäki, H. (2007). Origin of the Failure of Classical Nucleation Theory: Incorrect Description of the Smallest Clusters. *Physical Review Letters*, *98*, 145702. https://doi.org/10.1103/PhysRevLett.98.145702

Mesleh, M. F., Hunter, J. M., Shvartsburg, A. A., Schatz, G. C., & Jarrold, M. F. (1996). Structural Information from Ion Mobility Measurements: Effects of the Long-Range Potential. *The Journal of Physical Chemistry*, *100*(40), 16082–16086. https://doi.org/10.1021/jp961623v

Midey, A. J., Williams, S., & Viggiano, A. A. (2001). Reactions of NO^+ with Isomeric Butenes from 225 to 500 K. *J. Phys. Chem. A*, *105*(9), 1574–1578. https://doi.org/10.1021/jp0019005

Musil, F., Grisafi, A., Bartók, A. P., Ortner, C., Csányi, G., & Ceriotti, M. (2021). Physics-Inspired Structural Representations for Molecules and Materials. *Chemical Reviews*, *121*(16), 9759–9815. https://doi.org/10.1021/acs.chemrev.1c00021

Myllys, N., Kubečka, J., Besel, V., Alfaouri, D., Olenius, T., Smith, J. N., & Passananti, M. (2019). Role of base strength, cluster structure and charge in sulfuric-acid-driven particle formation. *Atmospheric Chemistry and Physics*, *19*(15), 9753–9768. https://doi.org/10.5194/acp-19-9753-2019

Neefjes, I., Halonen, R., Vehkamäki, H., & Reischl, B. (2022). Modeling approaches for atmospheric ion–dipole collisions: all-atom trajectory simulations and central field methods. *Atmospheric Chemistry and Physics*, *22*(17), 11155–11172. https://doi.org/10.5194/acp-22-11155-2022

Neefjes, I., Joos, J. J., Barandiarán, Z., & Seijo, L. (2020). Mixed-Valence Lanthanide-Activated Phosphors: Invariance of the Intervalence Charge Transfer (IVCT) Absorption Onset across the Series. *The Journal of Physical Chemistry C*, *124*(4), 2619–2626. https://doi.org/10.1021/acs.jpcc.9b11084

Neefjes, I., Laapas, M., Liu, Y., Médus, E., Miettunen, E., Ahonen, L., Quéléver, L., Aalto, J., Bäck, J. K., Kerminen, V.-M., Lamplahti, J., Luoma, K., Mäki, M., Mammarella, I., Petäjä, T., Räty, M., Sarnela, N., Ylivinkka, I., Hakala, S., Kulmala, M., Nieminen, T., & Lintunen, A. (2022). 25 years of atmospheric and ecosystem measurements in a boreal forest – Seasonal variation and responses to warm and dry years. *Boreal Environment Research*, *27*, 1–31.

Nosé, S. (1984). A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, *81*(1), 511–519. https://doi.org/10.1063/1.447334

Ochterski, J. W. (2000). Thermochemistry in Gaussian [accessed: 01-01-2024]. https://gaussian.com/thermo/

O'Dowd, C. D., Jimenez, J. L., Bahreini, R., Flagan, R. C., Seinfeld, J. H., Hämeri, K., Pirjola, L., Kulmala, M., Jennings, S. G., & Hoffmann, T. (2002). Marine aerosol formation from biogenic iodine emissions. *Nature*, *417*(6889), 632–636. https://doi.org/10.1038/nature00775

Olenius, T., Halonen, R., Kurtén, T., Henschel, H., Kupiainen-Määttä, O., Ortega, I. K., Jen, C. N., Vehkamäki, H., & Riipinen, I. (2017). New particle formation from sulfuric acid and amines: Comparison of monomethylamine, dimethylamine, and trimethylamine. *Journal of Geophysical Research: Atmospheres*, *122*(13), 7103–7118. https://doi.org/10.1002/2017JD026501

Ortega, I. K., Olenius, T., Kupiainen-Määttä, O., Loukonen, V., Kurtén, T., & Vehkamäki, H. (2014). Electrical charging changes the composition of sulfuric acid–ammonia/dimethylamine clusters. *Atmospheric Chemistry and Physics*, *14*(15), 7995–8007. https://doi.org/10.5194/acp-14-7995-2014

Passananti, M., Zapadinsky, E., Zanca, T., Kangasluoma, J., Myllys, N., Rissanen, M. P., Kurtén, T., Ehn, M., Attoui, M., & Vehkamäki, H. (2019). How well can we predict cluster fragmentation inside a mass spectrometer? *Chemical Communications*, *55*(42), 5946–5949. https://doi.org/10.1039/c9cc02896j

Pathria, R. K., & Beale, P. D. (2011). *Statistical Mechanics* (3rd ed.). Elsevier. https://doi.org/10.1016/C2009-0-62310-2

Penner, J., Andreae, M., Annegarn, H., Barrie, L., Feichter, J., Hegg, D., Jayaraman, A., Leaitch, R., Murphy, D., Nganga, J., Pitari, G., Ackerman, A., Adams, P., Austin, P., Boers, R., Boucher, O., Chin, M., Chuang, C., Collins, B., Cooke, W., DeMott, P., Feng, Y., Fischer, H., Fung, I., Ghan, S., Ginoux, P., Gong, S.-L., Guenther, A., Herzog, M., Higurashi, A., Kaufman, Y., Kettle, A., Kiehl,

J., Koch, D., Lammel, G., Land, C., Lohmann, U., Madronich, S., Mancini, E., Mishchenko, M., Nakajima, T., Quinn, P., Rasch, P., Roberts, D., Savoie, D., Schwartz, S., Seinfeld, J., Soden, B., Tanré, D., Taylor, K., Tegen, I., Tie, X., Vali, G., Dingenen, R. V., van Weele, M., & Zhang, Y. (2001). Aerosols, their Direct and Indirect Effects. In J. Houghton, Y. Ding, D. Griggs, M. Noguer, P. van der Linden, X. Dai, K. Maskell, & C. Johnson (Eds.), *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 291–336). Cambridge University Press.

Perdew, J. P. (1985). Density functional theory and the band gap problem. *International Journal of Quantum Chemistry*, *28*(S19), 497–523. https://doi.org/10.1002/qua.560280846

Piskulich, Z. A., & Thompson, W. H. (2020). On the temperature dependence of liquid structure. *The Journal of Chemical Physics*, *152*(1), 011102. https://doi.org/10.1063/1.5135932

Rissanen, M. (2021). Anthropogenic Volatile Organic Compound (AVOC) Autoxidation as a Source of Highly Oxygenated Organic Molecules (HOM). *The Journal of Physical Chemistry A*, *125*(41), 9027–9039. https://doi.org/10.1021/acs.jpca.1c06465

Schlegel, H. B. (1995). Geometry Optimization on Potential Energy Surfaces. In *Modern Electronic Structure Theory*. World Scientific Publishing Company.

Schlegel, H. B. (2011). Geometry optimization. *WIREs Computational Molecular Science*, *1*(5), 790–809. https://doi.org/10.1002/wcms.34

Seinfeld, J., & Pandis, S. (2006). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change* (2nd ed.). John Wiley & Sons, Inc.

Sherrill, C. D. (2005). The Born-Oppenheimer Approximation [accessed: 27-12-2023]. http://vergil.chemistry.gatech.edu/notes/bo/bo.pdf

Sherrill, D. (2003). Postulates of Quantum Mechanics [accessed: 05-02-2024]. http://vergil.chemistry.gatech.edu/notes/intro_estruc/node4.html

Sipilä, M., Berndt, T., Petäjä, T., Brus, D., Vanhanen, J., Stratmann, F., Patokoski, J., III, R. L. M., Hyvärinen, A.-P., Lihavainen, H., & Kulmala, M. (2010). The Role of Sulfuric Acid in Atmospheric Nucleation. *Science*, *327*(5970), 1243–1246. https://doi.org/10.1126/science.1180315

Smirnov, B. M. (2012). *Clusters and small particles: In Gases and Plasmas* (R. Berry Stephen, L. Birman Joseph, W. Lynn Jeffrey, P. Silverman Mark, E. Stanley H.,

& M. Voloshin, Eds.). Springer New York, NY. https://doi.org/10.1007/978-1-4612-1294-2

Strekowski, R. S., Alvarez, C., Petrov-Stojanović, J., Hagebaum-Reignier, D., & Wortham, H. (2019). Theoretical chemical ionization rate constants of the concurrent reactions of hydronium ions (H_3O^+) and oxygen ions ($O_2{}^+$) with selected organic iodides. *J. Mass Spectrom.*, *54*(5), 422–428. https://doi.org/10.1002/jms.4349

Su, T., & Chesnavich, W. J. (1982). Parametrization of the ion–polar molecule collision rate constant by trajectory calculations. *Journal of Chemical Physics*, *76*(10), 5183–5185. https://doi.org/10.1063/1.442828

Su, T., Su, E. C., & Bowers, M. T. (1978). Ion–polar molecule collisions. Conservation of angular momentum in the average dipole orientation theory. The AADO theory. *Journal of Chemical Physics*, *69*(5), 2243–2250. https://doi.org/10.1063/1.436783

Surján, P. R. (1989). The Brillouin Theorem. In *Second Quantized Approach to Quantum Chemistry: An Elementary Introduction* (pp. 87–92). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-74755-7_11

Swope, W. C., Andersen, H. C., Berens, P. H., & Wilson, K. R. (1982). A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, *76*(1), 637–649. https://doi.org/10.1063/1.442716

Szabo, A., & Ostlund, N. S. (1996). *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory* (1st ed.). Dover Publications, Inc.

Tester, J. W., & Modell, M. (1997). *Thermodynamics and Its Applications* (N. R. Amundson, Ed.). Prentice Hall, Upper Saddle River, NJ.

Torrie, G., & Valleau, J. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, *23*(2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8

Tuckerman, M. E. (2010). *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press.

Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., Tkatchenko, A., & Müller, K.-R. (2021). Machine Learning Force Fields. *Chemical Reviews*, *121*(16), 10142–10186. https://doi.org/10.1021/acs.chemrev.0c01111

Valavanidis, A., Fiotakis, K., & Vlachogianni, T. (2008). Airborne Particulate Matter and Human Health: Toxicological Assessment and Importance of Size and Composition of Particles for Oxidative Damage and Carcinogenic Mechanisms. *Journal of Environmental Science and Health, Part C*, *26*(4), 339–362. https://doi.org/10.1080/10590500802494538

van Duin, A. C. T., Dasgupta, S., Lorant, F., & Goddard, W. A. (2001). ReaxFF: A Reactive Force Field for Hydrocarbons. *The Journal of Physical Chemistry A*, *105*(41), 9396–9409. https://doi.org/10.1021/jp004368u

Vanhanen, J., Mikkilä, J., Lehtipalo, K., Sipilä, M., Manninen, H. E., Siivola, E., Petäjä, T., & Kulmala, M. (2011). Particle Size Magnifier for Nano-CN Detection. *Aerosol Science and Technology*, *45*(4), 533–542. https://doi.org/10.1080/02786826.2010.547889

Vehkamäki, H. (2006). *Classical Nucleation Theory in Multicomponent Systems*. Springer Berlin, Heidelberg. https://doi.org/10.1007/3-540-31218-8

Viehland, L. A., & Mason, E. A. (1975). Gaseous lon mobility in electric fields of arbitrary strength. *Annals of Physics*, *91*(2), 499–533. https://doi.org/10.1016/0003-4916(75)90233-X

Wang, M., Kong, W., Marten, R., He, X.-C., Chen, D., Pfeifer, J., Heitto, A., Kontkanen, J., Dada, L., Kürten, A., Yli-Juuti, T., Manninen, H. E., Amanatidis, S., Amorim, A., Baalbaki, R., Baccarini, A., Bell, D. M., Bertozzi, B., Bräkling, S., Brilke, S., Murillo, L. C., Chiu, R., Chu, B., Menezes, L.-P. D., Duplissy, J., Finkenzeller, H., Carracedo, L. G., Granzin, M., Guida, R., Hansel, A., Hofbauer, V., Krechmer, J., Lehtipalo, K., Lamkaddam, H., Lampimäki, M., Lee, C. P., Makhmutov, V., Marie, G., Mathot, S., Mauldin, R. L., Mentler, B., Müller, T., Onnela, A., Partoll, E., Petäjä, T., Philippov, M., Pospisilova, V., Ranjithkumar, A., Rissanen, M., Rörup, B., Scholz, W., Shen, J., Simon, M., Sipilä, M., Steiner, G., Stolzenburg, D., Tham, Y. J., Tomé, A., Wagner, A. C., Wang, D. S., Wang, Y., Weber, S. K., Winkler, P. M., Wlasits, P. J., Wu, Y., Xiao, M., Ye, Q., Zauner-Wieczorek, M., Zhou, X., Volkamer, R., Riipinen, I., Dommen, J., Curtius, J., Baltensperger, U., Kulmala, M., Worsnop, D. R., Kirkby, J., Seinfeld, J. H., El-Haddad, I., Flagan, R. C., & Donahue, N. M. (2020). Rapid growth of new atmospheric particles by nitric acid and ammonia condensation. *Nature*, *581*(7807), 184–189. https://doi.org/10.1038/s41586-020-2270-4

Weber, R. J., Marti, J. J., McMurry, P. H., Eisele, F. L., Tanner, D. J., & Jefferson, A. (1996). Measured Atmospheric New Particle Formation Rates: Implications for Nucleation Mechanisms. *Chemical Engineering Communications*, *151*(1), 53–64. https://doi.org/10.1080/00986449608936541

WHO. (2016). An estimated 12.6 million deaths each year are attributable to unhealthy environments [accessed: 14-01-2024]. https://www.who.int/news/item/15-03-2016-an-estimated-12-6-million-deaths-each-year-are-attributable-to-unhealthy-environments

Wiedensohler, A., Wiesner, A., Weinhold, K., Birmili, W., Hermann, M., Merkel, M., Müller, T., Pfeifer, S., Schmidt, A., Tuch, T., Velarde, F., Quincey, P., Seeger, S., & Nowak, A. (2018). Mobility particle size spectrometers: Calibration procedures and measurement uncertainties. *Aerosol Science and Technology*, *52*(2), 146–164. https://doi.org/10.1080/02786826.2017.1387229

Williams, J., & Koppmann, R. (2007). Volatile Organic Compounds in the Atmosphere: An Overview. In *Volatile Organic Compounds in the Atmosphere* (pp. 1–32). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470988657.ch1

Williams, S., Knighton, W. B., Midey, A. J., Viggiano, A. A., Irle, S., Wang, Q., & Morokuma, K. (2004). Oxidation of Alkyl Ions in Reactions with O_2 and O_3 in the Gas Phase. *J. Phys. Chem. A*, *108*(11), 1980–1989. https://doi.org/10.1021/jp031099+

Wölk, J., & Strey, R. (2001). Homogeneous Nucleation of H_2O and D_2O in Comparison: The Isotope Effect. *The Journal of Physical Chemistry B*, *105*(47), 11683–11701. https://doi.org/10.1021/jp0115805

Woon, D., & Herbst, E. (2009). Quantum chemical predictions of the properties of known and postulated neutral interstellar molecules. *Astrophys. J., Suppl. Ser.*, *185*(2), 273–288. https://doi.org/10.1088/0067-0049/185/2/273

Xu, Z., Gao, J., Xu, Z., Attoui, M., Pei, X., Amo-González, M., Zhang, K., & Wang, Z. (2023). Characterization of the planar differential mobility analyzer (DMA P5): resolving power, transmission efficiency and its application to atmospheric relevant cluster measurements. *Atmospheric Measurement Techniques*, *16*(24), 5995–6006. https://doi.org/10.5194/amt-16-5995-2023

Yale DMA-MS Facility. (2015). Yale's open facility on ion mobility spectrometry/mass spectrometry (IMS-MS) [accessed: 21-01-2024]. https://www.eng.yale.edu/DMAMSfacility/

Yang, H., Neefjes, I., Tikkanen, V., Kubečka, J., Kurtén, T., Vehkamäki, H., & Reischl, B. (2023). Collision-sticking rates of acid–base clusters in the gas phase determined from atomistic simulation and a novel analytical interacting hard-sphere model. *Atmospheric Chemistry and Physics*, *23*(10), 5993–6009. https://doi.org/10.5194/acp-23-5993-2023

Yin, K., Mai, S., & Zhao, J. (2022). Atmospheric Sulfuric Acid Dimer Formation in a Polluted Environment. *International Journal of Environmental Research and Public Health*, *19*(11), 6848. https://doi.org/10.3390/ijerph19116848

Yin, R., Li, X., Yan, C., Cai, R., Zhou, Y., Kangasluoma, J., Sarnela, N., Lampilahti, J., Petäjä, T., Kerminen, V.-M., Bianchi, F., Kulmala, M., & Jiang, J. (2023). Revealing the sources and sinks of negative cluster ions in an urban environment through quantitative analysis. *Atmospheric Chemistry and Physics*, *23*(9), 5279–5296. https://doi.org/10.5194/acp-23-5279-2023

Yu, F., & Turco, R. (2008). Case studies of particle formation events observed in boreal forests: implications for nucleation mechanisms. *Atmospheric Chemistry and Physics*, *8*(20), 6085–6102. https://doi.org/10.5194/acp-8-6085-2008

Yu, F. (2010). Ion-mediated nucleation in the atmosphere: Key controlling parameters, implications, and look-up table. *Journal of Geophysical Research: Atmospheres*, *115*(D3). https://doi.org/10.1029/2009jd012630

Zanca, T., Kubečka, J., Zapadinsky, E., Passananti, M., Kurtén, T., & Vehkamäki, H. (2020). Highly oxygenated organic molecule cluster decomposition in atmospheric pressure interface time-of-flight mass spectrometers. *Atmospheric Measurement Techniques*, *13*(7), 3581–3593. https://doi.org/10.5194/amt-13-3581-2020

Zhang, J., & Dolg, M. (2015). ABCluster: the artificial bee colony algorithm for cluster global optimization. *Physical Chemistry Chemical Physics*, *17*, 24173–24181. https://doi.org/10.1039/C5CP04060D

Zhang, J., & Dolg, M. (2016). Global optimization of clusters of rigid molecules using the artificial bee colony algorithm. *Physical Chemistry Chemical Physics*, *18*, 3003–3010. https://doi.org/10.1039/C5CP06313B

Zhang, R., Khalizov, A., Wang, L., Hu, M., & Xu, W. (2012). Nucleation and Growth of Nanoparticles in the Atmosphere. *Chemical Reviews*, *112*(3), 957–2011. https://doi.org/10.1021/cr2001756