UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2020-71
The Faculty of Science, Technology and Medicine

DISSERTATION

Presented on 11/11/2020 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN PHYSIQUE

by

Martin STÖHR
Born on 22 August 1991 in Garmisch-Partenkirchen (Germany)

VAN DER WAALS DISPERSION INTERACTIONS IN BIOMOLECULAR SYSTEMS: QUANTUM-MECHANICAL INSIGHTS AND METHODOLOGICAL ADVANCES

Dissertation defense committee

Dr. Alexandre Tkatchenko, Dissertation Supervisor
Professor, Université du Luxembourg

Dr. Ludger Wirtz, Chairman
Professor, Université du Luxembourg

Dr. Massimiliano Esposito, Vice Chairman
Professor, Université du Luxembourg

Dr. Marcus Elstner
Professor, Karlsruher Institut für Technologie

Dr. Frank Noé
Professor, Freie Universität Berlin

University of Luxembourg
Campus Limpertsberg
162a, avenue de la Faïencerie
L-1511 Luxembourg
Luxembourg

"I may not have gone where I intended to go, but I think I have ended up where I needed to be."

— Dirk Gently, holistic detective.

From "The Long Dark Tea-Time of the Soul" by Douglas Adams

Abstract

Intermolecular interactions are paramount for the stability, dynamics and response of systems across chemistry, biology and materials science. In biomolecules they govern secondary structure formation, assembly, docking, regulation and functionality. van der Waals (vdW) dispersion contributes a crucial part to those interactions. As part of the long-range electron correlation, vdW interactions arise from Coulomb-coupled quantum-mechanical fluctuations in the instantaneous electronic charge distribution and are thus inherently many-body in nature. Common approaches to describe biomolecular systems (*i.e.*, classical molecular mechanics) fail to capture the full complexity of vdW dispersion by adapting a phenomenological, atom-pairwise formalism. This thesis explores beyond-pairwise vdW forces and the collectivity of intrinsic electronic behaviors in biomolecular systems and discusses their role in the context of biomolecular processes and function. To this end, the many-body dispersion (MBD) formalism parameterized from density-functional tight-binding (DFTB) calculations is used.

The investigation of simple molecular solvents with particular focus on water gives insights into the vdW energetics and electronic response properties in liquids and solvation as well as emergent behavior for coarse-grained models. A detailed study of *intra*-protein and protein–water vdW interactions highlights the role of many-body forces during protein folding and provides a fundamental explanation for the previously observed "unbalanced" description and over-compaction of disordered protein states. Further analysis of the intrinsic electronic behaviors in explicitly solvated proteins indicates a long-range persistence of electron correlation through the aqueous environment, which is discussed in the context of protein–protein interactions, long-range coordination and biomolecular regulation and allostery. Based on the example of a restriction enzyme, the potential role of many-body vdW forces and collective electronic behavior for the long-range coordination of enzymatic activity is discussed. Introducing electrodynamic quantum fluctuations into the classical picture of allostery opens up the path to a more holistic view on biomolecular regulation beyond the traditional focus on merely local structural modifications.

Building on top of the MBD framework, which describes vdW dispersion within the interatomic dipole-limit, a practical extension to higher-order terms is presented. The resulting *Dipole-Correlated Coulomb Singles* account for multipolar as well as dispersion-polarization-like contributions beyond the random phase approximation by means of first-order perturbation theory over the dipole-coupled MBD state. It is shown that Dipole-Correlated Coulomb Singles become particularly relevant for relatively larger systems and can alter qualitative trends in the long-range interaction under (nano-)confinement. Bearing in mind the frequent presence of confinement in biomolecular systems due to cellular crowding, in ion channels or for interfacial water, this so-far neglected contribution is expected to have broad implications for systems of biological relevance.

Ultimately, this thesis introduces a hybrid approach of DFTB and machine learning for the accurate description of large-scale systems on a robust, albeit approximate, quantum-mechanical level. The developed DFTB-NN_{rep} approach combines the semi-empirical DFTB Hamiltonian with a deep tensor neural network model for localized many-body repulsive potentials. DFTB-NN_{rep} provides an accurate description of energetic, structural and vibrational properties of a wide range of small organic molecules much superior to standard DFTB or machine learning.

Overall, this thesis aims to extend the current view of complex (bio)molecular systems being governed by local, (semi-)classical interactions and develops methodological steps towards an advanced description and understanding including non-local interaction mechanisms enabled by quantum-mechanical phenomena such as long-range correlation forces arising from collective electronic fluctuations.

Preface

Acknowledgments

First of all, I would like to express my gratitude to my supervisor, Prof Alexandre Tkatchenko, for providing the opportunity for my doctoral studies as well as the guidance and numerous discussions. In this context, I would also like to thank my external colleagues and collaborators for all the discussions and our past and ongoing efforts. A special thanks goes to my friends and colleagues at the University of Luxembourg for enriching both my professional and social life and making the past four years such an enjoyable time.

I would also like to acknowledge the financial support from the Fonds National de la Recherche Luxembourg (AFR PhD grant CNDTEC) enabling this work and my attendance at various inspiring international workshops and conferences. The results presented in this thesis heavily relied on the HPC resources of the University of Luxembourg, without which this work would not have been possible. Many thanks to the Institute for Pure and Applied Mathematics at University of California, Los Angeles for hosting me during the early months of my doctorate and providing such a productive, inspiring environment.

Finally, I sincerely thank all the great people that have accompanied and supported me on my academic path in the past and my deepest gratitude to all my friends and family for their untiring support over all these years.

Note on Previous Publication

This thesis is based on and partly published in the following articles:

[1] M. Stöhr, T. Van Voorhis, and A. Tkatchenko "Theory and Practice of Modeling van der Waals Interactions in Electronic-Structure Calculations" *Chem. Soc. Rev.* **48**, 4118 (2019).

[2] M. Stöhr and A. Tkatchenko "Quantum Mechanics of Proteins in Explicit Water: The Role of Plasmon-like Solute–Solvent Interactions" *Sci. Adv.* **5**, eaax0024 (2019).

[3] M. Stöhr, M. Sadhukhan, Y. S. Al-Hamdani, J. Hermann, and A. Tkatchenko "Coulomb Interactions between Dipolar Quantum Fluctuations in van der Waals Bound Molecules and Materials" *Nat. Commun.* accepted (2020). Preprint available at https://arxiv.org/abs/2007.12505.

[4] M. Stöhr, L. Medrano Sandonas, and A. Tkatchenko "Accurate Many-Body Repulsive Potentials for Density-Functional Tight Binding from Deep Tensor Neural Networks" *J. Phys. Chem. Lett.* **11**, 6835 (2020).

Contributions by collaborators are clearly marked as such when used in this thesis. A short note on previously published information is given at the beginning of every chapter or section where applicable.

Table of Contents

List of Abbreviations

ACFD	adiabatic-connection fluctuation-dissipation
AFM	atomic force microscopy
ATM	Axilrod-Teller-Muto three-body van der Waals dispersion term
CCSD(T)	coupled-cluster theory with single, double and perturbative triple excitations
CNT	carbon nanotube
DC	dipole-coupled
DCS	Dipole-Correlated Coulomb Singles
DFA	density-functional approximation
DFT	density-functional theory
DFTB	density-functional tight-binding
DOS	density of states
DQMC	diffusion quantum Monte-Carlo
DTNN	deep tensor neural network
Fip35-WW	Fip35 Hpin1 WW domain
GGA	generalized gradient approximation
HF	Hartree-Fock
HP35-NleNle	35-residue villin headpiece (double norleucine mutant)
kcal	kilocalories ($\hat{=}$ 4187 kg·m^2/s^2 = 4187 *Joule*)
KS	Kohn-Sham
LCAO	linear combination of atomic orbitals
LDA	local density approximation
MAE	mean absolute error
MARE	mean absolute relative error
MBD	many-body dispersion (formalism)
MD	molecular dynamics
ML	machine learning
MM	molecular mechanics
MP2	second-order Møller-Plesset perturbation theory
MSE	mean signed error
NN	neural network
PT	perturbation theory
QDO	quantum Drude oscillator
QHO	quantum harmonic oscillator
QM	quantum-mechanical
(D)QMC	(diffusion) quantum Monte-Carlo
RMSD	root-mean-square deviation
RPA	random phase approximation
(rs)SCS	(range-separated) self-consistent screening
SE	Schrödinger equation
vdW	van der Waals
vdW(TS)	van der Waals method by Tkatchenko and Scheffler, also referred to as **TS**
xc	exchange-correlation

List of Symbols and Notation

$\oplus$	direct sum (also generalized to vectors)
$\otimes$	outer (vector) product, $\mathbf{u}\otimes\mathbf{v} = \vert\mathbf{u}\rangle\langle\mathbf{v}\vert = \mathbf{u}\mathbf{v}^{\mathrm{T}}$ and $(\mathbf{u}\otimes\mathbf{v})_{ij} = u_i v_j$
$\dagger$	complex conjugate: $\mathbf{A}^{\dagger} = (\mathbf{A}^*)^{\mathrm{T}}$
$\partial/\partial x$	partial derivative with respect to x
$\nabla_{\mathbf{x}}$	Nabla operator with respect to $\mathbf{x}$: $\left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, ..., \frac{\partial}{\partial x_n}\right)^{\mathrm{T}}$
$\int d\mathbf{x}$	(multidimensional) integral over $\mathbf{x}$: $\int\cdots\int dx_1 dx_2 ... dx_n$
$\mathbb{1}_n$	n-dimensional identity matrix
α	(isotropic) scalar dipole polarizability
$\boldsymbol{\alpha}$	dipole polarizability tensor
δ_{ij}	Kronecker delta
$\delta^n(x)$	(n-dimensional) Dirac delta function
$\delta/\delta X$	functional derivative with respect to X
ζ	mass-weighted displacement of quantum harmonic oscillator
η	(effective) excitation frequency of uncoupled oscillator
ν	(spectroscopic) frequency
ξ	eigenmode coordinate
ρ	electronic charge density
χ	density-density response function
Ψ	(total) wave function
ω	(effective) excitation frequency of (dipole-coupled) oscillator
C_6	two-body dipole–dipole van der Waals interaction coefficient
C_8	two-body dipole–quadrupole van der Waals interaction coefficient
C_9	three-body dipole–dipole van der Waals interaction coefficient
E	(total, relative, or interaction) energy
$\mathbf{F}$	(atomic) force
f_{damp}	damping function
g_{rs}	range-separating function
$\mathcal{H}$	(quantum-mechanical) Hamiltonian
m	(effective) mass
q	(effective) charge
$\mathbf{r}$	electronic or pseudo-particle position
$\mathbf{R}$	(atomic) position
R_{vdW}	van der Waals radius
$\mathbb{R}^n$	n-dimensional (coordinate) space of real numbers
$\mathcal{T}$	kinetic energy
$\mathbf{T}$	dipole–dipole coupling tensor
$\mathbf{T}_{\mathrm{sr/lr}}$	short-range/long-range dipole–dipole coupling tensor
$\mathbf{T}_{\mathrm{xc}}$	dipolar exchange-correlation kernel
u	(angular) frequency
$\mathcal{V}$	potential
V	volume
Z	atomic number

Bold characters denote vectors (lower case) or matrices/tensors (upper case), while normal-weight italic characters denote scalars.

List of Figures

List of Tables

Chapter 1

Introduction

Computational modeling and simulation has by now become an integral part of understanding systems across chemistry, biology and materials science. It plays a key-role for pharmacological and technological developments from drug therapy to energy materials. On a meso- to macroscopic scale, systems can be described by the rules of classical mechanics, but below that one enters the peculiar realm of quantum phenomena. The behaviors and properties of matter thereby follow the laws of quantum mechanics as established in the early 20$^{\text{th}}$ century. Quoting Paul A.M. Dirac, *"The underlying physical laws necessary for the mathematical theory (...) are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble."*[5] Overcoming this seemingly minor difficulty, however, turned out to become a major subject of research, where various developments approached higher and higher levels of accuracy and sophistication over the past century. Among others, it gave rise to the fields of theoretical chemistry and chemical physics, which even today are as much concerned with the identification and improvement of the deficiencies of such approximate theories as with their application. As a general rule of thumb, the more sophisticated and reliable a given approach is, the more computationally demanding it is and the more steeply its computational workload scales with the size of the system. Despite the tremendous growth of the availability of computational resources and high-performance implementations, this limits the most accurate theories to rather small systems and studying phenomena at increasing length and time scales requires to resort to increasingly approximate approaches. This bears an inherent complication: While applicable methods become less and less complete, many-body (quantum) effects make the systems increasingly more complex before transitioning to classical behaviors. This *quantum-to-classical transition* can be qualitatively explained in various ways, *e.g.*, based on Heisenberg's uncertainty relation, Bohr's correspondence principle, decoherence theory, the emergence of statistical determinism when averaging over a sufficiently large number of (independent) quantum phenomena or simply the limited measurement accuracy on the macroscale. While a quantitative or formal description of the classical limit of quantum mechanics is still heavily debated, the (broad) transition region can generally be expected around the nano- to microscale, where systems are small enough to render classical treatments insufficient, but large enough for complex many-body effects to cause decisive deviations from the well-studied behaviors of few-particle quantum systems. Identifying such deviations, potential shortcomings of

conventional approximations and their implications for systems in this intermediate regime is paramount for a reliable description of the emergent phenomena and effective interaction laws as well as for a better understanding of the quantum-to-classical transition in general. Gaining a more fundamental understanding and description of systems at the nano- or microscale is further motivated by the ongoing miniaturization of technological devices and especially the ever-growing interest in an atomistic and molecular understanding of biomolecular systems and processes.

Molecular Mechanics as Workhorse Method for Biomolecular Simulations

Since the first successful simulations of protein structure and dynamics in atomistic detail,[6,7] computational and theoretical modeling of biomolecular systems has boosted our understanding of the building blocks of living organisms and how their dynamics and response provide the function needed. While the tremendous progress in the experimental study of biomolecules cannot be discarded, this seldom allows to directly investigate full dynamical pathways. Based on physical models for the interactions and forces acting on the individual atoms of the system, computer simulations provide access to such pathways by consecutively solving Newton's equations of motion for the individual atoms. Thanks to manifold methodological advances and the increasing availability of (special-purpose) computational resources, this approach known as *molecular dynamics simulation* by now provides access to the mapping of folding processes of proteins with a few tens of residues and folding times of a few milliseconds or less at full atomic resolution.[8] This allows to not only predict protein behavior and likely folding path ensembles, but together with the advances of experimental techniques also enables the atomistic interpretation of experimental observations.[9,10] Computational modeling and simulation have since provided essential insights to, *e.g.*, the structure of membranes[11,12] and molecular mechanisms of protein folding,[13–15] signal transduction,[16,17] enzymatic activity[18,19] or ion channels.[20–22]

The physical models that these studies are based on are predominantly molecular mechanics (MM) "force fields", which provide a physically-motivated, yet empirical, description of the free energy landscape of the system. As a first classification, MM approaches distinguish *bonded* and *non-bonded* (or *non-covalent*) contributions. Spatial proximity of atoms can lead to a localization of electrons along the direction connecting the two. The resulting stabilization binds the two atoms strongly together, which is known as a chemical (covalent) bond. Bonded contributions to the total energy then include all interactions among atoms that are connected *via* such chemical bonds. This already introduces a significant limitation to conventional force fields: The bonding characteristics or *connectivity* has to be predetermined and cannot change during the simulation (for the purpose of clarity we will not discuss so-called reactive force fields, which partly alleviate this limitation). The mathematical form of bonded interactions then typically follows a many-body expansion of the total energy (*body* is mostly used synonymously to *atom* in this thesis). The energy and interatomic forces are then described in terms of a linear combination of potential energy contributions from local degrees of freedom including bond distances (two-body), angles (three-body) and dihedral angles (four-body). The non-bonded interactions, finally, describe all interactions and relevant energy contributions not covered so far. Given the considerable size and complex, intertwined structure of biomolecular systems, these play an essential role for an accurate description. In protein folding, as in a majority of biomolecular processes, for example, chemical bonds remain stable. The structure and dynamics of the system are then governed by non-bonded interactions, while bonded terms mostly only provide given constraints to the accessible conformations. In fact, non-bonded interactions are responsible for the condensation of molecular systems. In MM potentials, non-bonded terms are mostly based on a cumulative, phenomenological formalism of pairwise-additive, effective potentials. Finally, the parameters defining the individual energy contributions are optimized to provide satisfactory agreement with reference results from experiment or less-approximate methods. As a result of

this parameterization, different force fields may reproduce the same key quantities like native structures and folding rates of proteins, the predicted mechanistic pathways, however, can vary substantially among different parameterizations.[23,24] Given that MM force fields are usually designed to predict the *free energy landscape* of the system, they include the quantum-mechanical (electronic) energy as well as the effect of thermodynamic conditions such as finite temperature and ambient pressure. As a result, a given force field for biomolecular simulations is usually extremely limited in its applicability and is mostly designed to perform well for (near-)native proteins in aqueous solution at ambient temperature and pressure. This emphasizes the considerable lack of completeness of the physical description they provide. In order to identify potential shortcomings and so-far neglected interaction mechanisms, it is thus desirable to investigate the relevant interactions in biomolecular systems on a more fundamental level. This particularly concerns non-covalent contributions given their pivotal role for biomolecular systems and that these are treated on the arguably least-physical basis in MM approaches.

Non-covalent Interactions

The fundamental basis for non-covalent interactions is the same as for all phenomena contributing to the non-relativistic energy of a molecular system: the Coulomb interaction among electrons and nuclei. From the framework of intermolecular quantum-mechanical perturbation theory, they are typically classified into individual terms with a specific, physical interpretation. The main (*i.e.* leading-order) contributions are *electrostatics*, *polarization* (or *induction*), *van der Waals (vdW) dispersion* and *exchange repulsion*.[25] As schematically represented in Fig. 1.1, electrostatics describes the simple, Coulombic interaction among static (average) charge distributions. In the case of electrons, this average distribution can then be modified due to the presence of one (or more) additional molecular entities. The resulting change in the energy represents polarization/induction interactions. Due to their quantum character, the distribution of electrons is not static, however. As a simple result of Heisenberg's uncertainty principle,[26,27] the instantaneous electronic charge distribution shows intrinsic, quantum-mechanical fluctuations. These fluctuations give rise to instantaneous electronic multipoles with non-zero net interaction. This is what is referred to as vdW dispersion interaction (for a more detailed introduction and discussion, see Sec. 2.3). Exchange repulsion, finally, arises from the *Pauli exclusion principle*,[28]

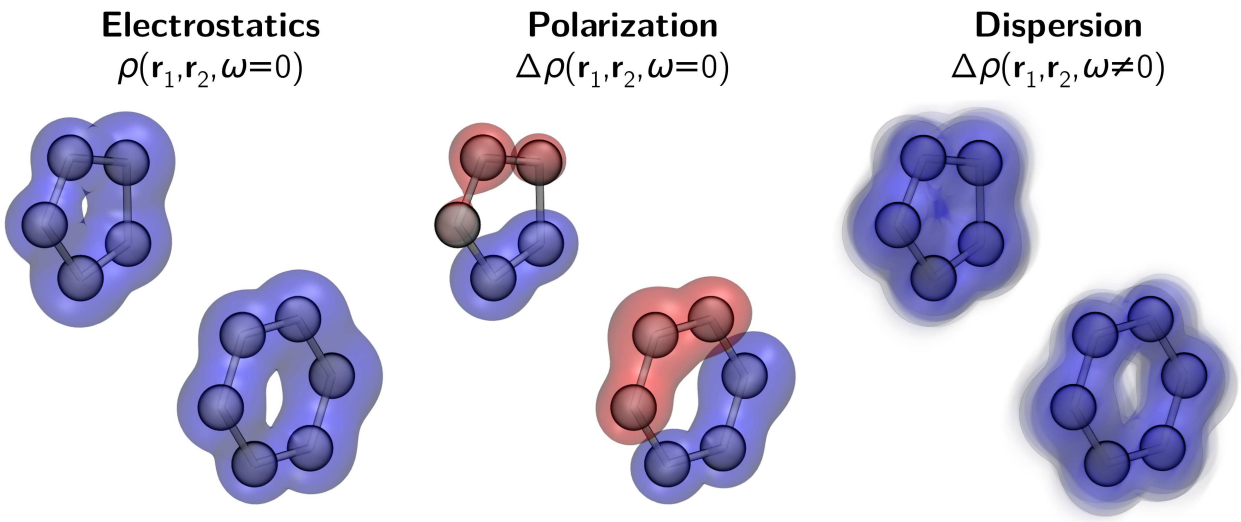

Fig. 1.1: Schematic illustration of non-covalent interactions: Electrostatics represents the interaction between unperturbed charge densities, polarization (also known as induction) the interaction due to induced static multipoles (*i.e.* changes in the charge density with zero frequency). van der Waals dispersion describes the interaction among intrinsic quantum-mechanical fluctuations in the instantaneous electronic charge distribution. Exchange interaction, the fourth major type of non-covalent interactions, is not shown for the sake of clarity.

which leads to an additional, repulsive interaction (see Sec. 2.1). Considering the interplay of the above phenomena, various higher-order terms arise. These are typically believed to be negligible, however. Note that, while the original formulation of this classification and illustration in Fig. 1.1 are given for intermolecular forces, the same effects and phenomena appear among parts of extended molecules or materials leading to intramolecular non-covalent interactions, which can be interpreted and classified in the same way. In the language of classical MM force fields, non-bonded interactions are typically described by Lennard-Jones-type, pairwise potentials (dispersion + exchange) together with the cumulative terms steric effects (electrostatics + exchange) and hydrogen-bonding (electrostatics + polarization + dispersion).

Total Energy Description and Force Field Design

Designing the final MM force field is usually done within a *top–down* philosophy: A given parametric expression of the total energy is optimized to reproduce a select set of reference properties. The functional form of this expression is thereby typically motivated by the knowledge on simple, small systems. However, the far-from-trivial origin of the majority of non-covalent interactions outlined above gives rise to pronounced non-additive and collective effects. This renders simple pairwise-additive potentials based on the interaction and scaling laws in small systems incomplete and often insufficient for describing non-bonded energy contributions in large-scale systems. This applies in particular to coarse-grained models, where the interactions between multiple atoms are combined into effective potentials. As we shall see below, especially vdW dispersion is highly collective and many-body in nature. For molecular crystals and supramolecular complexes, which show similar bonding characteristics as biomolecular systems, this has been shown previously to be of utmost importance for an accurate and reliable description.[29–35] So, establishing a reliable functional space for top–down parametric optimization of atomistic force fields and the construction of effective potentials for coarse-grained models requires knowledge of the emergent effects in many-atom systems and their energetic implications. As an alternative to traditional MM force fields, recent years have witnessed an ever-growing interest in machine learning potentials for molecules and materials. While avoiding the limitations of fixed functional expressions in the total energy, the employed descriptors and architectures still suffer from a lack of physical design rules, which is due to the limited knowledge of the energetics and dynamical behaviors in large-scale systems. Such insights can be gained by a more *bottom–up* approach using a more complete, quantum-mechanical description. A more fundamental theoretical treatment in addition provides access to properties on an electronic level, which is completely absent in the more phenomenological MM force fields. These properties are essential for the accurate description of charge transport and conductivity, optical response and many other paramount aspects for technological applications and the understanding of biomolecular processes.

Aim and Outline

Following the above-mentioned bottom–up philosophy, the focus of this thesis is the investigation of collective electronic behaviors and the emergent (many-body) vdW physics in biomolecular systems including solvation and intramolecular aspects to advance our general understanding of vdW forces in complex systems. One step in this regard is to identify effects on the vdW interaction beyond current state-of-the-art methods and to highlight their potential role for biomolecular processes and function. As a second important step, this thesis develops practical methodologies towards an efficient, yet more complete, description of complex, large-scale systems such as (solvated) biomolecules. Given that most of our current understanding of biomolecular systems is predominantly based on conclusions from simple classical potentials, such insights and more complete theoretical methods hold the key to also advance our conceptual understanding and interpretation of complex phenomena as relevant to biological processes. In turn, this understanding can provide essential guidelines for the construction of more accurate and reliable MM force fields, coarse-grained models or machine learning potentials.

This thesis is organized as follows: Chapter 2 provides a brief introduction to the theoretical background and methodological approaches that the remainder of the work is build on. This introduction focuses on the theory of vdW interactions in the context of atomistic modeling and the density-functional tight-binding approach as used in this work. Unless explicitly noted with the according references, the derivations in chapters 3 to 6 then represent the outcome of work done during this doctorate. Chapter 3 provides first insights into the vdW physics and electronic behaviors of liquids and solvated systems. In Chapter 4, a detailed study of the vdW energetics in solvated proteins as obtained in a dipolar many-body treatment is given. The results and underlying interaction mechanisms are discussed in the context of biomolecular processes and function. Chapter 5 presents a practical approach to capture many-body vdW forces beyond the usual dipolar or random phase approximation and highlights their role in complex systems. In Chapter 6, a combined approach of density-functional tight-binding and machine learning potentials is developed, which substantially improves upon state-of-the-art semi-empirical methods and aims at providing a reliable and transferable quantum-mechanical description of large-scale systems. The chapter concludes with a perspective on promising steps for advancing the presented methodology towards this aim. Chapter 7, finally, provides a brief summary of the results given in this thesis and a general outlook on future work and projected implications of the presented findings. In general, the thesis is written in a semi-cumulative manner, where each chapter features a short, more specific introduction and discussion of the respective contents.

Chapter 2

Theoretical Background

Parts of Sec. 2.3 in this chapter have been reproduced from
M. Stöhr, T. Van Voorhis, and A. Tkatchenko, *Chem. Soc. Rev.* **48**, 4118 (2019)
with permission from the Royal Society of Chemistry.

Sec. 2.1 provides a brief introduction to the basics of quantum mechanics in the context of electronic-structure theory and Kohn-Sham (KS) density-functional theory (DFT). For the sake of brevity and clarity, this section only covers a few of the essential steps towards the methodologies used in this thesis. For a more in-depth introduction and overview of the theory and practical methods, see the books by, *e.g.*, Atkins & Friedman,[36] Levine,[37] Szabo & Ostlund,[38] Cramer[39] or Parr & Yang,[40] which this section is based on. All derivations provided in this thesis are thereby given in atomic units.

Sec. 2.2 summarizes the methodological aspects of the density-functional tight-binding[41,42] (DFTB) formalism as an efficient semi-empirical approximation to (semi-local) DFT. In particular, the third order variant[43] (DFTB3) and how it originates from KS-DFT will be introduced. The section concludes with a brief discussion of some of the major limitations of the basic DFTB formalism. For further details, the reader is referred to the original publications or reviews provided in Refs. 44–47, for example.

Sec. 2.3 provides an introduction to van der Waals (vdW) dispersion interactions in particular in the context of DFT and its approximate formulation as represented by DFTB. The section includes an exact formulation based on the adiabatic-connection fluctuation-dissipation theorem as well as an approximate reduction to pairwise-additive interatomic interaction potentials. As examples of state-of-the-art atom-pairwise vdW methods also used in this thesis, the vdW(TS)[48] and Grimme's D3 approach[49] are summarizes in detail. This is followed by an overview on beyond-pairwise effects and the non-additivity of vdW forces — both from theoretical considerations and recent experimental observations. The section concludes with an introduction to the many-body dispersion (MBD) formalism[50,51] as a practical many-body approach to vdW interactions including a discussion of the coupled quantum harmonic oscillator model of (non-local) electronic response and intrinsic quantum electrodynamic behaviors.

2.1 Basics of Electronic-Structure and Density-Functional Theory

Quantum mechanics is based on a series of postulates formulated at the turn of the 20$^{\text{th}}$ century. It is stated that any closed quantum-mechanical (QM) system is fully described by a *wave function*, Ψ. Any physically-acceptable wave function thereby has to be continuous, single-valued and square-integrable (with the exception of continuum states).[36,39] The wave function can be interpreted as *probability amplitude* and its absolute square as measure for the probability density of finding a particle at a given coordinate.[52] Any observable of a quantum system is represented by a linear, hermitian operator.[36] The observable can be obtained as an expectation value of the operator acting on the wave function:

$$\langle O \rangle = \frac{\int \Psi^* \hat{O} \Psi \, d\tau}{\int \Psi^* \Psi \, d\tau} \equiv \frac{\langle \Psi | \hat{O} | \Psi \rangle}{\langle \Psi | \Psi \rangle} \,, \tag{2.1}$$

where the integrals go over all coordinates and the full space on which Ψ is defined. Expanding Ψ in the complete basis of eigenfunctions of $\hat{O}$ together with the hermiticity of $\hat{O}$ shows that any physical observable according to Eq. (2.1) (given by a linear combination of eigenvalues) is real-valued. In the case of discrete eigenspectra, this leads to a quantization of observables. The fundamental relation connecting the definition of a (non-relativistic) quantum system and its wave function is the so-called *Schödinger equation* (SE).[53–56] For a system of nuclei and electrons it is given by,

$$i \frac{\partial}{\partial t} \left| \Psi(\mathbf{R}, \mathbf{r}, \mathbf{s}, t) \right\rangle = \hat{\mathcal{H}} \left| \Psi(\mathbf{R}, \mathbf{r}, \mathbf{s}, t) \right\rangle \,, \tag{2.2}$$

where the (total) wave function Ψ depends on the spatial coordinates of all nuclei, $\mathbf{R}$, and electrons, $\mathbf{r}$, and the electrons' spin, $\mathbf{s}$, as well as time, t, and $\hat{\mathcal{H}}$ is the (QM) Hamiltonian containing all kinetic and potential energy terms. For the following paragraphs, we shall combine the electronic position and spin into the electronic coordinate, $\mathbf{q} = (\mathbf{r}, \mathbf{s})$. For what concerns a majority of chemical physics, we can limit our description to *stationary states*. After separating the above differential equation into a time-dependent phase factor oscillating in the complex plane and a stationary part of the wave function, one arrives at the *time-independent SE*,

$$\hat{\mathcal{H}} \left| \Psi(\mathbf{R}, \mathbf{q}) \right\rangle = E_{\text{tot}} \left| \Psi(\mathbf{R}, \mathbf{q}) \right\rangle \tag{2.3}$$

with E_{tot} being the observable of the Hamiltonian operator. So, the energy of the system. As proposed by Born and Oppenheimer,[57] one can further factorize the above into a nuclear and an electronic part. This approximation, commonly referred to as *Born-Oppenheimer* or *adiabatic approximation*, is motivated by the large separation between the masses of nuclei and electrons. Instead of treating electrons and nuclei simultaneously as in Eq. (2.3), we write

$$\hat{\mathcal{H}} \left| \Psi(\mathbf{R}, \mathbf{q}) \right\rangle \approx \left(\hat{\mathcal{H}}_{\text{nuc}} + \hat{\mathcal{H}}_{\text{el}} \right) \left| \Psi_{\text{nuc}}(\mathbf{R}) \right\rangle \left| \Psi_{\text{el}}(\mathbf{q}; \mathbf{R}) \right\rangle = \left(E_{\text{nuc}} + E_{\text{el}} \right) \left| \Psi_{\text{nuc}}(\mathbf{R}) \right\rangle \left| \Psi_{\text{el}}(\mathbf{q}; \mathbf{R}) \right\rangle \tag{2.4}$$

$$\text{with} \qquad \hat{\mathcal{H}}_{\text{nuc}} \left| \Psi_{\text{nuc}}(\mathbf{R}) \right\rangle = E_{\text{nuc}} \left| \Psi_{\text{nuc}}(\mathbf{R}) \right\rangle \tag{2.5}$$

$$\text{and} \qquad \hat{\mathcal{H}}_{\text{el}} \left| \Psi_{\text{el}}(\mathbf{q}; \mathbf{R}) \right\rangle = E_{\text{el}} \left| \Psi_{\text{el}}(\mathbf{q}; \mathbf{R}) \right\rangle \,, \tag{2.6}$$

where the *electronic wave function*, Ψ_{el}, depends only parametrically on the nuclear positions.[38] The time-independent SE for nuclei (2.5) is thereby usually solved in the classical limit, where E_{nuc} simplifies to the classical Coulomb repulsion between point-like nuclei. It is worth noting, however, that this is strictly speaking not part of the Born-Oppenheimer approximation. Eq. (2.6) describes the electronic SE, which provides a spectrum of possible solutions referred to as electronic states. The lowest-energy solution is thereby known as *ground-state*, $\Psi_{\text{el},0}$. When two or more electronic states are close in energy, the adiabatic approximation breaks down.[36] For the systems considered in this thesis, however, Eq. (2.4) is expected to be a viable approximation.

Considering a system of multiple indistinguishable particles, it is obvious that swapping particles in Ψ_{el} describes the same state as before. This means Ψ_{el} has to be an eigenfunction of the exchange (or permutation) operator, $\hat{\mathcal{P}}$, which for two particles has the eigenvalues ± 1.[37] So, the wave function has to be symmetric $(+1)$ or anti-symmetric (-1) with respect to particle-exchange. In detail, the wave function has to be symmetric with respect to the interchange of any two indistinguishable particles with integral spin (*bosons*) and anti-symmetric for the exchange of two particles with half-integral spin (*fermions*) such as electrons. This statement, known as *spin-statistics theorem*, has important consequences: For an N-electron system, it requires

$$\hat{\mathcal{P}}_{ij}\left|\Psi_{\text{el}}(\mathbf{q}_1,...,\mathbf{q}_i,...\mathbf{q}_j,...,\mathbf{q}_N)\right\rangle = \left|\Psi_{\text{el}}(\mathbf{q}_1,...,\mathbf{q}_j,...\mathbf{q}_i,...,\mathbf{q}_N)\right\rangle \overset{(!)}{=} -\left|\Psi_{\text{el}}(\mathbf{q}_1,...,\mathbf{q}_i,...\mathbf{q}_j,...,\mathbf{q}_N)\right\rangle , \tag{2.7}$$

which further implies that there is zero probability of finding any two (or more) electrons with the same $\mathbf{q}$ at the same time[37,52] (known as *Pauli exclusion principle*[28]). This condition leads to a form of repulsion between electrons with equal spin called *exchange* or *Pauli repulsion*. This "repulsion" is not a physical force *per se*, but a manifestation of the spin-statistics theorem affecting the non-relativistic electronic energy. Solving Eq. (2.6) under the constraint of Eq. (2.7) is at the center of theoretical chemistry and chemical physics as, together with a classical treatment of nuclei, it dictates the total energy and enables the theoretical treatment of molecules and materials. Unfortunately, a closed-form, analytical solution of the electronic SE is only possible for a hydrogen or hydrogen-like atom. Over the past century, numerous methods have been devised to solve Eq. (2.6) for many-electron systems at various levels of approximation.

2.1.1 Hartree-Fock Theory and Beyond

The probably most decisive step enabling the QM treatment of molecules and materials is to map the N-electron problem to N effective one-particle problems.[38] The corresponding single-particle states, ψ_a, are also known as spin orbitals and can be written in the non-relativistic case as $\psi_a(\mathbf{q}_i) = \phi_a(\mathbf{r}_i)\sigma(s_i)$ with spatial part ϕ and spin part σ. The notation $\psi_a(\mathbf{q}_i)$ refers to particle i in state a. Representing the basis for most of quantum chemistry, the *Hartree-Fock (HF) method* treats the individual single-particles *via* a set of coupled equations, where each electron is subject to the average potential generated by the remaining electrons and all nuclear charges. The total electronic wave function, Ψ_{el}, is thereby expressed as an anti-symmetrized product of single-particle states referred to as *Slater determinant*. For N electrons,

$$\Psi_{\text{el}}(\mathbf{q}_1,\mathbf{q}_2,...,\mathbf{q}_N) = \frac{1}{\sqrt{N!}}\begin{vmatrix} \psi_1(\mathbf{q}_1) & \psi_2(\mathbf{q}_1) & \cdots & \psi_N(\mathbf{q}_1) \\ \psi_1(\mathbf{q}_2) & \psi_2(\mathbf{q}_2) & \cdots & \psi_N(\mathbf{q}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(\mathbf{q}_N) & \psi_2(\mathbf{q}_N) & \cdots & \psi_N(\mathbf{q}_N) \end{vmatrix} , \tag{2.8}$$

which by construction fulfills the spin-statistics theorem[38] and naturally gives rise to exchange repulsion. As the electrons are treated within the average field of each other (known as *mean-field approach*), the potential for particle i depends on the solution for particle j and *vice versa*. Hence, the HF equations have to be solved iteratively until self-consistency is reached. One significant shortcoming of the mean-field approach is the neglect of the instantaneous, Coulomb-coupled motion of electrons. The resulting lack in energy is referred to as (electron) correlation energy. While this energy contribution is typically small in magnitude, it is essential for the description of chemical and biomolecular systems as well as materials. In order to recover the missing correlation energy, numerous *post*-HF methods have been developed in the quantum chemistry community. The most renowned approaches are *configuration interaction*,[58] which expresses the total wave function as a linear combination of (virtually) excited determinants, *coupled-cluster theory*[59] using a cluster expansion of determinants, or perturbational treatments such as *Møller-Plesset perturbation theory*.[60] For further details, see Refs. 36–39 and Sec. 2.3.

2.1.2 Basis Set Representation and Variational Principle

Thus far, we have not defined the shape of the spatial (orbital) part of the single-particle states. As its mathematical form is generally not known, one typically employs a finite-elements approach. In the context of electronic-structure theory, this is referred to as *basis set expansion*, where

$$\phi_a(\mathbf{r}) = \sum_i c_{ia}\,\varphi_i(\mathbf{r}) \tag{2.9}$$

with c_{ia} as expansion coefficient and φ_i as basis functions. By representing the single-particle wave functions in a given basis one bypasses the problem of explicitly solving complex wave equations. Instead, the problem can be represented in terms of coefficients, which ultimately allows efficient numerical solutions. The optimal choice of c_{ia} and thus the shape of the single-particle orbitals is then obtained *via* the *Rayleigh-Ritz variational principle*, which states that for any normalized, physically-acceptable trial wave function, Ψ_{trial},

$$E_{\text{trial}} = \left\langle \Psi_{\text{trial}} \middle| \hat{\mathcal{H}}_{\text{el}} \middle| \Psi_{\text{trial}} \right\rangle \geq \left\langle \Psi_{\text{el},0} \middle| \hat{\mathcal{H}}_{\text{el}} \middle| \Psi_{\text{el},0} \right\rangle = E_{\text{el},0}\ . \tag{2.10}$$

So, the expectation value of the energy for any trial wave function is greater or equal to the true ground-state electronic energy, $E_{\text{el},0}$. Furthermore, for non-degenerate ground-states, the equality holds if and only if $\Psi_{\text{trial}} = \Psi_{\text{el},0}$. Thus, optimal coefficients can be obtained by (numerically) minimizing the energy of the system with respect to c_{ia}.

2.1.3 Density-Functional Theory

The basic idea behind DFT is to express the total energy as a functional of the three-dimensional electronic charge density, $\rho(\mathbf{r})$, instead of the complex $4N$-dimensional wave function (spatial coordinate plus spin). For systems with non-degenerate ground-state, Hohenberg and Kohn [61] have proven the mathematical existence of such an energy functional as well as the validity of a variational principle analogous to Eq. (2.10). While significantly reducing the complexity of the problem, this seminal reformulation still does not allow for a straightforward solution of the many-electron problem. In lieu thereof, Kohn and Sham formulated a corresponding effective independent-particle system. [62] This formulation is termed KS-DFT. Exchange and electron correlation effects, which are not covered by the mean-field description in the independent-particle framework, are thereby collected into the so-called *exchange-correlation (xc) functional*, E_{xc}. [37,39] All other terms as well as the (classical) nuclear repulsion energy are well-known and can be determined at arbitrary accuracy. The mathematical form of E_{xc}, on the other side, is generally unknown. Current applications rely on approximative density functionals, for which several formulations at various levels of sophistication are available. To a first approximation, one may assume a uniform electron distribution, for which E_{xc} can be determined in an efficient and straightforward manner. The corresponding approach is known as *local density approximation* (LDA). A more profound *ansatz* to account for the non-uniform electron distribution inherent to all finite systems is to also include the gradient of the density, which is then called *generalized gradient approximation* (GGA). In some more detail, the generic KS energy functional reads

$$E_{\text{tot}}\left[\rho(\mathbf{r});\mathbf{R}\right] = \sum_a f_a \left\langle \psi_a \middle| -\frac{1}{2}\nabla^2_{\mathbf{r}} + \mathcal{V}_{\text{ext}}(\mathbf{R}) + \mathcal{V}_{\text{H}}\left[\rho(\mathbf{r})\right] \middle| \psi_a \right\rangle + E_{\text{xc}}\left[\rho(\mathbf{r}), \nabla_{\mathbf{r}}\rho\right] + E_{\text{nuc}}(\mathbf{R})\ , \tag{2.11}$$

where f_a denotes the occupation number of the single-particle wave function, ψ_a, and $\mathcal{V}_{\text{ext}}$ and $\mathcal{V}_{\text{H}}$ the external (electron–nuclei) and Hartree (mean-field electron–electron) potential. Going beyond the typically insufficient description of exchange interactions within the LDA or GGA, hybrid DFT functionals such as the popular PBE0 [63] include a given fraction of exact (HF-like) exchange repulsion, which can substantially improve upon some of the shortcomings of pure GGA functionals such as the description of hydrogen-bonding, for instance. [64,65]

2.2 Density-Functional Tight-Binding

KS-DFT represents the state-of-the-art approach for electronic-structure calculations of molecules and materials and has been successfully applied in various computational studies. Despite the ever-growing availability of computing resources and high-performance implementations, the intrinsic scaling of DFT with the system size limits its applicability to complex biomolecular systems or nanostructures. In this regard, semi-empirical methods represent a more efficient alternative, while still offering an explicit QM treatment of electrons and a reasonable, intermediate level of accuracy. As such, they pose a promising tool for pushing the boundaries of current computational studies and the exploration of larger length and time scales as relevant to biomolecular or other nanoscale systems, for example.

Among other approaches,[66–71] the DFTB formalism[41–43] is one of the most popular representatives of semi-empirical QM methods. The DFTB method is thereby based on a perturbation expansion of the KS energy functional (2.11) in a minimal basis representation of atomic orbitals. This perturbation expansion is performed around a given non-interacting reference electron density, $\rho_0(\mathbf{r})$, given as the superposition of (effective) atomic densities, $\rho_A(\mathbf{r})$. The true density of the interacting system, $\rho(\mathbf{r})$, and the corresponding energy functional are then represented in terms of perturbations to this reference, $\delta\rho(\mathbf{r})$.[44] In an expansion series up to third order, the corresponding approximate energy functional is given by,

$$\begin{aligned} E_{\text{tot}} = &\sum_a f_a \left\langle \psi_a \middle| -\frac{1}{2}\nabla^2 + \mathcal{V}_{\text{ext}}(\mathbf{R}) + \mathcal{V}_{\text{H}}[\rho_0'] + \left.\frac{\delta E_{\text{xc}}[\rho]}{\delta\rho}\right|_{\rho_0} \middle| \psi_a \right\rangle \\ &+ \frac{1}{2}\iint_{\mathbb{R}^6} \left(\frac{1}{\|\mathbf{r}-\mathbf{r}'\|} + \left.\frac{\delta^2 E_{\text{xc}}[\rho]}{\delta\rho\,\delta\rho'}\right|_{\rho_0} \right) \delta\rho\,\delta\rho'\,d\mathbf{r}d\mathbf{r}' \\ &+ \frac{1}{6}\iiint_{\mathbb{R}^9} \left.\frac{\delta^3 E_{\text{xc}}[\rho]}{\delta\rho\,\delta\rho'\,\delta\rho''}\right|_{\rho_0} \delta\rho\,\delta\rho'\,\delta\rho''\,d\mathbf{r}\,d\mathbf{r}'\,d\mathbf{r}'' \\ &+ \sum_{A<B} \frac{Z_A Z_B}{\|\mathbf{R}_A - \mathbf{R}_B\|} - \frac{1}{2}\int_{\mathbb{R}^3} \mathcal{V}_{\text{H}}[\rho_0']\,\rho_0\,d\mathbf{r} + E_{\text{xc}}[\rho_0] - \int_{\mathbb{R}^3} \left.\frac{\delta E_{\text{xc}}[\rho]}{\delta\rho}\right|_{\rho_0} \rho_0\,d\mathbf{r} + \mathcal{O}\left((\delta\rho)^4\right), \end{aligned} \tag{2.12}$$

where f_a denotes the occupation number of the single-particle state ψ_a, $\rho \equiv \rho(\mathbf{r})$, $\rho' \equiv \rho(\mathbf{r}')$ and $\rho'' \equiv \rho(\mathbf{r}'')$. $\mathcal{V}_{\text{ext}}$ is the external or nuclear–electron potential, $\mathcal{V}_{\text{H}}$ is the Hartree potential and E_{xc} the xc energy. $\delta/\delta\rho$ denotes the functional derivative with respect to ρ and Z_A corresponds to the atomic number of atom A. The first line in Eq. (2.12) is the so-called band structure term, which equals the DFT energy of the starting density, ρ_0. The second line is the Coulomb energy, or second-order term, the third line the third-order contribution and the last line is collectively referred to as the repulsive energy.[44,46] As mentioned above, DFTB then represents the single-particle states as a linear combination of atomic orbitals (LCAO) according to Eq. (2.9) using a minimal basis set of atomic valence orbitals.[47] The orbitals are thereby obtained by solving the atomic DFT problem,

$$\left(-\frac{1}{2}\nabla^2 + \mathcal{V}_{\text{eff}}[\rho_{\text{atom}}] + \mathcal{V}_{\text{conf}}\right)|\varphi_i\rangle = \tilde{\varepsilon}_i\,|\varphi_i\rangle\,, \tag{2.13}$$

where $\mathcal{V}_{\text{eff}}[\rho_{\text{atom}}]$ is the effective potential for an isolated atom. In order to improve numerical stability in the long-distance limit and effectively model the confining effect of neighboring atoms, an additional confinement potential, $\mathcal{V}_{\text{conf}}$, is added. Typically, a simple harmonic potential is chosen,[44] but other forms such as the less-invasive Woods-Saxon potential have been proposed.[72]

2.2.1 Band Structure Energy

The band structure energy, E_{BS}, can be written in compact form as

$$E_{\mathrm{BS}} = \sum_a f_a \left\langle \psi_a \middle| -\frac{1}{2}\nabla^2 + \mathcal{V}_{\mathrm{ext}}(\mathbf{R}) + \mathcal{V}_{\mathrm{H}}[\rho_0] + \left.\frac{\delta E_{\mathrm{xc}}[\rho]}{\delta\rho}\right|_{\rho_0} \middle| \psi_a \right\rangle = \sum_a f_a \left\langle \psi_a \middle| \hat{\mathcal{H}}^{(0)} \middle| \psi_a \right\rangle . \quad (2.14)$$

Within the basis set representation described above and c_{ia} as LCAO coefficients, the band structure energy from Eq. (2.14) simplifies to

$$E_{\mathrm{BS}} = \sum_a f_a \sum_{i,j} c_{ia}^* \hat{\mathcal{H}}_{ij}^{(0)} c_{ja} \quad (2.15)$$

Using a two-center approximation, the Hamiltonian elements $\hat{\mathcal{H}}_{ij}^{(0)}$ can be precalculated using KS-DFT. With the help of Slater-Koster transformation rules,[73] they can further be tabulated for varying interatomic distances. At run time, the appropriate matrix element can then be obtained by interpolation. In order to guarantee correct convergence towards isolated atoms the diagonal elements $\hat{\mathcal{H}}_{ii}^{(0)}$ are approximated by the valence energies of the corresponding unconfined, neutral atom, $\varepsilon_{A,i}$. So,

$$\hat{\mathcal{H}}_{ij}^{(0)} = \begin{cases} \varepsilon_{A,i} & \text{if } i = j \ \wedge\ (i,j) \in A \\ \left\langle \varphi_i \middle| -\frac{1}{2}\nabla^2 + \mathcal{V}_{\mathrm{eff}}[\rho_A + \rho_B] \middle| \varphi_j \right\rangle & \text{if } i \in A \ \wedge\ j \in B \neq A \\ 0 & \text{else} \end{cases} , \quad (2.16)$$

where ρ_A denotes the ground state electron density of the isolated, neutral atom A and $\mathcal{V}_{\mathrm{eff}}$ corresponds to the effective KS-DFT potential for the diatomic molecule AB.[45]

2.2.2 Beyond Tightly-Bound Electrons: Second- and Third-Order Contributions

In the above, electrons were considered to be fixed at (*i.e.*, tightly-bound to) their respective atoms as assumed by the reference density. As such, the theory thus far represents a classic tight-binding formalism. This, however, neglects important effects arising from the relaxation or transfer of charge upon the interaction among atoms. This charge rearrangement as represented by $\delta\rho$ gives rise to additional electrostatic interactions. This effect is captured to within first-order by the coulombic term, E_{C}. Expressing $\delta\rho$ as a sum of atomic contributions with single-exponential decay from the respective nucleus and employing an atomic monopole approximation allows to express the second-order term as

$$E_{\mathrm{C}} = \frac{1}{2}\sum_{A,B} \Delta q_A \gamma_{AB}(R_{AB}) \Delta q_B , \quad (2.17)$$

where Δq_A is the (partial) charge of atom A obtained *via* Mulliken analysis:[74]

$$\Delta q_A = n_{\mathrm{val}}^{(A)} - \sum_a f_a \left[\sum_{i\in A} \left(c_{ia}^* c_{ia} + \frac{1}{2} \sum_{j\notin A} c_{ia}^* S_{ij} c_{ja} \right) \right] \quad (2.18)$$

with $n_{\mathrm{val}}^{(A)}$ as the number of valence electrons on atom A and $S_{ij} = \langle \varphi_i | \varphi_j \rangle$ as (precomputed and tabulated) overlap integrals. For on-site terms, one finds $\gamma_{AA} = U_A$ corresponding to the Hubbard U parameter (*i.e.*, twice the chemical hardness of atom A). With the neglect of xc effects, one can also identify the off-site terms,[42]

$$\gamma_{A\neq B}(R_{AB}) = \frac{1}{R_{AB}} - e^{-(\kappa_{AB}+\kappa_{BA})R_{AB}} \quad (2.19)$$

$$\text{with} \qquad \kappa_{AB} = \frac{16}{5} U_A \left(\frac{16}{10} \frac{U_A U_B^4}{\left(U_A^2 - U_B^2\right)^2} + \frac{3U_A^2 U_B^4 - U_B^6}{\left(U_A^2 - U_B^2\right)^3 R_{AB}} \right) . \quad (2.20)$$

U can be approximated by the difference between the (first) ionization potential and the (first) electron affinity, which can be obtained by reference DFT calculations for the corresponding cation and anion. In cases where species do not have stable ions in the corresponding reference calculation, one can also relate U to the orbital energy of the highest occupied molecular orbital *via* Janak's theorem. [45,75]

Considering the inconsistent correlation between the atomic size and chemical hardness of hydrogen, a modification of the γ-function was introduced for H–X contributions. [76] In the modified form, $\gamma_{AB}(R_{AB})$ is damped at intermediate ranges with a U-dependent Gaussian if A or B correspond to a hydrogen atom. [43,76] To account for the changes in the chemical hardness with changes in Δq_A, the third-order term introduces the derivative of the above γ-function with respect to the (partial) atomic charges,

$$\Gamma_{AB} = \left.\frac{\partial \gamma_{AB}}{\partial q_A}\right|_{q_A^{(\mathrm{ref})}} = \left(1 - \frac{\delta_{AB}}{2}\right) \frac{\partial \gamma_{AB}}{\partial U_A} \left.\frac{\partial U_A}{\partial q_A}\right|_{q_A^{(\mathrm{ref})}} , \tag{2.21}$$

where δ_{AB} is the Kronecker delta. The corresponding energy contribution can be written as, [43]

$$E_{3\mathrm{rd}} = \frac{1}{6} \sum_{A,B} \Delta q_A \left(\Delta q_A \Gamma_{AB} + \Gamma_{BA} \Delta q_B\right) \Delta q_B = \frac{1}{3} \sum_{A,B} \Delta q_A^2 \Gamma_{AB} \Delta q_B \, . \tag{2.22}$$

The derivative of γ with respect to U as it enters Γ can thereby be done analytically, [43] while the derivative of the Hubbard U with respect to the atomic charge is provided as input in practical calculations.

2.2.3 Solution of the Electronic DFTB Problem

The optimal set of LCAO coefficients is finally obtained *via* the Rayleigh-Ritz variational principle (2.10). The resulting secular equations representing the DFTB equivalent of the KS equations in DFT are given by

$$\sum_j c_{ja} \left(\hat{\mathcal{H}}_{ij} - \epsilon_a S_{ij}\right) = 0 \quad \forall \, (i,a) \, , \tag{2.23}$$

where with definition (2.16) the full electronic Hamiltonian for DFTB3 is given by, [43,46]

$$\hat{\mathcal{H}}_{ij} = \hat{\mathcal{H}}^{(0)}_{ij} + S_{ij} \sum_C \Delta q_C \left[\frac{\gamma_{AC} + \gamma_{CB}}{2} + \frac{\Delta q_A \Gamma_{AC} + \Gamma_{CB} \Delta q_B}{3} + \frac{\Gamma_{AC} + \Gamma_{CB}}{6} \Delta q_C\right] \quad \forall \, i \in A, j \in B \, .$$

So, the electronic Hamiltonian and thus the secular equations depend on the (partial) atomic charges, which, in turn, depend on the LCAO coefficients. Solving Eq. (2.23) requires a self-consistent procedure — a given set of c_{ia} determines the charges Δq_A and Eq. (2.23) yields a new set of c_{ia} until a certain threshold in the change of Δq_A in subsequent iterations is reached. In practical applications, Hamiltonian elements and overlap integrals are interpolated from precalculated values, while atomic charges can be obtained solely on the basis of the system's density matrix. As a result, DFTB avoids the time- and memory-consuming construction of the real-space electron density and numerical integration. This allows DFTB to achieve a remarkable speed-up and favorable scaling with system size in comparison to full DFT, yet it still provides access to most electronic-structure properties.

2.2.4 The Repulsive Energy

With the electronic DFTB problem defined and solved above, we are left with the last line of Eq. (2.12): the repulsive energy, E_{rep}, formally defined as

$$E_{\mathrm{rep}} = \sum_{A<B} \frac{Z_A Z_B}{\|\mathbf{R}_A - \mathbf{R}_B\|} - \frac{1}{2} \int_{\mathbb{R}^3} \mathcal{V}_{\mathrm{H}}[\rho_0'] \, \rho_0 \, d\mathbf{r} + E_{\mathrm{xc}}[\rho_0] - \int_{\mathbb{R}^3} \left.\frac{\delta E_{\mathrm{xc}}[\rho]}{\delta \rho}\right|_{\rho_0} \rho_0 \, d\mathbf{r} \, . \tag{2.24}$$

By similar division into atomic contributions, as was also done for the coulombic term above, and a two-center approximation the (formal) repulsive energy can be divided into pairwise repulsive terms (see Sec. 6.3.2). In practice, the repulsive energy is represented by such an atom-pairwise formalism,

$$E_{\text{rep}} \approx \sum_{A<B} \mathcal{V}_{\text{rep}}^{(AB)}\left(\|\mathbf{R}_A - \mathbf{R}_B\|\right) . \tag{2.25}$$

The corresponding two-body potentials, $\mathcal{V}_{\text{rep}}$, are then chosen in order to reproduce (semi-)local DFT or other reference methods. So, all in all, the repulsive energy can be seen as an analogue to the xc functional in DFT in some sense. In addition to its formal contributions given in Eq. (2.24), it incorporates all the complicated many-body effects not considered otherwise. From a practical point of view, E_{rep} is given by the difference of some total energy reference (E_{ref}) and the terms obtained so far,

$$E_{\text{rep}} = E_{\text{ref}} - E_{\text{BS}} - E_{\text{C}} - E_{\text{3rd}} . \tag{2.26}$$

In practice, the nuclear derivative of E_{rep}, the repulsive force (F_{rep}), is determined by interpolating the difference in forces as obtained by full DFT (or other reference methods) and the force contribution from the electronic DFTB problem for a suitable set of reference systems. Integration of the resulting force, in the end, yields the repulsive potential, which is then tabulated as a function of interatomic separation and interpolated at run time.

2.2.5 Limitations of the basic DFTB Formalism and Adaption of post-DFT Methods

While the above introduction is presented for the simpler spin-collinear case, the DFTB method has equally-well been formulated and applied including account for spin polarization.[77,78] Due to the approximate character of DFTB, one clearly has to balance a certain trade-off of accuracy for the sake of the much increased efficiency. The empirically fitted repulsive potential thereby often complicates a clear-cut disentanglement of the (main) reasons behind individual failures. One major source of inaccuracy and reduced transferability is the minimal basis set of confined atomic wave functions chosen in DFTB. As a result of the minimal basis, polarization effects, which require more diffuse, delocalized basis function, are often poorly described. This is particularly affecting the description of hydrogen bonds, the hybridization states of nitrogen or phosphorous, or conduction bands in solids.[47,79,80] While improving general performance and stability, the employed confinement of atomic wave functions for the LCAO *ansatz*, can further impair the description of charge polarization and often leads to a considerable underestimation of exchange (Pauli) repulsion.[81,82] As a result of the two-center integral approximations and the minimal, confined basis set, it has been shown that the optimal pairwise repulsive potential for a given DFTB parametrization can strongly depend on the system or property under consideration. For example, current parametrizations seldom allow simultaneous accuracy in energetic as well as vibrational properties.[47,79] First steps towards beyond-pairwise repulsive potentials with higher accuracy and transferability are discussed in Chapter 6.

Being based on and often parameterized to semi-local density functionals, DFTB also inherits the shortcomings known for these functionals. In the context of DFT, various strategies have been developed to address such limitations. Many such *post*-DFT methods have also been adapted to the DFTB framework, such as +U corrections[83] and long-range corrected hybrid functionals[84,85] as well as time-dependent DFTB[86,87] and the description of excited states[88] and electron transport.[89,90] The most relevant deficiency in the context of this work, is the lack of (dynamic) long-range electron correlation that DFTB inherits from semi-local DFT. Due to their complex non-linear scaling with the system size, long-range correlation forces (most prominently including vdW dispersion) represent an essential extension to the DFTB framework. The following section reviews the theoretical background of vdW interactions and their description within electronic-structure theory. Reference to methods adapted to the DFTB framework are given where applicable.

2.3 van der Waals Dispersion Interactions

As mentioned in Sec. 2.1, theoretical and computational modeling of molecules and materials is largely contingent on mapping the N-electron problem to N effective one-electron problems. This seminal reformulation already captures, depending on the system, 99 % or more of the total electronic energy.‡ Unfortunately, the remaining fraction of the total electronic energy can be crucial for various observables and properties of interest, such as relative energies, [34,91–93] binding properties, [33,35,94–96] or structural features [34,93,97,98] as well as the mechanical, [99,100] thermodynamic, [34,101] kinetic [102–104] and electronic [105,106] signatures of a given system. In the simple case of an Argon dimer for instance, KS-DFT with the hybrid PBE0 functional captures about 99.95 % of the *total* energy, but it gives no more than roughly 15 % of the *interaction* energy. HF does not even bind an Argon dimer. The major part of the missing electronic energy is due to the correlated motion of electrons, or correlated quantum-mechanical fluctuations of the average electron distribution in the DFT picture. It is thus referred to as electron correlation energy.§ In particular, the long-range correlation energy represents a challenging task in electronic-structure calculations due to its highly non-local character. The main component of this long-range contribution is what is known as vdW dispersion interaction. As such, vdW interactions are inherently quantum-mechanical and many-body ("collective") in nature and, moreover, they are ubiquitous in molecular systems and materials. The strongly non-linear scaling with size in polarizable systems [92,110] presents further challenges for modeling such long-range correlation forces.

Describing (long-range) electron correlation has been a central topic in the quantum chemistry community, which since the early days mainly focused on wave function-based methods typically starting from the HF mean-field picture. Thanks to extensive methodological developments a number of asymptotically correct and to some extent practical methods have been devised. Among those the *coupled-cluster* technique has established itself as one of the most prevalent *post*-HF methods. Coupled-cluster theory starts from a Slater determinant based on the mean-field HF orbitals and includes (virtual) excitations by the use of the exponential cluster operator. Such virtual electronic excitations represent the perturbation picture analogue of fluctuations in the electronic charge distribution. Accounting for up to double excitations together with a perturbative treatment of triple excitations, labeled as CCSD(T), is usually referred to as the *gold standard* and often relied upon as a reference method for more approximate models. However, CCSD(T) and comparably accurate methods are still limited to small- and medium-sized systems (typically less than ~200 atoms) due to the immense computational costs characterized by a scaling of the computation time with the number of electrons to the power of 7. A quite different, yet similarly accurate, approach is quantum Monte-Carlo (QMC). Here, one solves the many-electron SE in a stochastic manner. The most relevant flavors of QMC in the context of modeling molecules and materials are: Variational Monte-Carlo, Green's function Monte-Carlo and diffusion Monte-Carlo, which exploits the similarity between the SE and a diffusion equation in imaginary time. Thanks to its stochastic character one can even estimate the expected deviation from the *exact* solution. Parallelization of this approach is straightforward and tractable system sizes have reached a few hundreds of atoms, [111] which has boosted its use as a benchmark method in recent years. In the end, both CCSD(T) and QMC are typically only used to benchmark (interaction) energies based on a given structure, as force evaluation can become extremely intricate as a result of their perturbational or stochastic character. In contrast to accurate quantum-chemical methods, density-functional approximations (DFAs) require less computational workload and offer access

‡ Although the KS equations in DFT are, in principle, exact the universal xc functional remains so-far unknown and the (semi-)local approximations to it based on the uniform electron gas give rise to a similar coverage of the energy.

§ Electron correlation is often divided into dynamic and static (or non-dynamic) correlation. [107,108] VdW interactions are part of the dynamic correlation energy, representing the lack in energy due to a mean-field approximation. [107] Static correlation becomes particularly relevant near electronic degeneracies and requires so-called multi-reference methods, see *e.g.* Ref. 109. In this thesis, "correlation energy" shall refer to dynamic correlation only.

to atomic forces. Since the first successful applications of DFT, however, the lack of explicit electron correlation has proven itself an important issue when modeling molecular systems and the same applies to all derivative methods including DFTB. In the context of DFT, a vast number of possible remedies has been proposed, many of which have also been adapted for semi-empirical methods by now. Thereby, an *a posteriori* inclusion of long-range correlation forces is the most widely used approach. Nevertheless, it is worthwhile to point out that long-range correlation is, in principle, part of the electronic Hamiltonian and can thus also affect the solution of the self-consistent field procedure, [105,106] which leads to dispersion-polarization effects.

The electron correlation energy, E_{corr}, is typically defined as the difference between the exact (non-relativistic) solution of the electronic SE and the effective mean-field description such as the HF or KS reference. Hence, it depends on the definition of the mean-field description and can be rigorously formulated in multiple ways. [108] The adiabatic-connection fluctuation-dissipation (ACFD) theorem provides an exact formulation of the non-relativistic (non-retarded) xc energy of a system in terms of the Coulomb-coupled density response on top of an independent-particle framework such as the HF or KS picture. Relativistic effects such as retardation as well as thermal field fluctuations can play an important role for extended, mesoscopic systems (*cf.* Casimir forces), but will not be covered in this thesis. For a deeper discussion and review of this topic, see Refs. 112 or 113, for instance.

2.3.1 Exact Formulation from Adiabatic-Connection Fluctuation-Dissipation Theorem

The ACFD formula originates from linear response theory‡ and relies on the non-local, time-dependent density-density response function, $\chi(\mathbf{r},\mathbf{r}',t,t')$, which describes the response of an electron density at point $\mathbf{r}$ and time t to a perturbation at position $\mathbf{r}'$ at time t'. Under the assumption of time invariance, which holds for stationary states in quantum mechanics, the time-dependent density response can be Fourier-transformed to the frequency domain (u) yielding the non-local, frequency-dependent, and complex-valued density response, $\chi(\mathbf{r},\mathbf{r}',u)$. Thereby, the imaginary part describes the contribution due to dissipation. [115] Hence, integrating over the Coulomb-coupled imaginary part of $\chi(\mathbf{r},\mathbf{r}',u)$ gives the energy due to dissipation of a (scalar) perturbation.

In quantum mechanics, any charge distribution is subject to instantaneous fluctuations, which gives rise to intrinsic fluctuations of the electric field within the system (or *vice versa*). The electron correlation energy is the dissipation energy of this fluctuating electric field. According to the ACFD theorem the energy due to dissipation of such internal perturbations is the same as for external perturbations and can thus be calculated from the imaginary part of $\chi(\mathbf{r},\mathbf{r}',u)$. Evaluation of the correlation energy on top of an independent-particle formalism is then carried out by means of the adiabatic theorem, [116,117] meaning *via* coupling parameter integration from the non-correlated system to the fully correlated density response:

$$E_{\text{corr}} = -\frac{1}{2\pi}\int_0^\infty\int_0^1\iiint_{\mathbb{R}^6}\Big[\chi_\lambda(\mathbf{r},\mathbf{r}',iu)-\chi_{\lambda=0}(\mathbf{r},\mathbf{r}',iu)\Big]\mathcal{V}_{\text{Coul}}(\mathbf{r},\mathbf{r}')\,d\mathbf{r}\,d\mathbf{r}'\,d\lambda\,du \tag{2.27}$$

with λ as the coupling constant, where $\lambda=1$ corresponds to the fully correlated system and $\lambda=0$ to the non-correlated system of independent electrons (*e.g.*, the KS (or HF) reference system), for which χ can be calculated based on the corresponding single-particle states *via* the Adler-Wiser formula. [118,119] For the integration of the imaginary part of $\chi(\mathbf{r},\mathbf{r}',u)$ we have used: [115]

$$\int_0^\infty \text{Im}\Big\{\chi(\mathbf{r},\mathbf{r}',u)\Big\}\,du = \int_0^\infty \chi(\mathbf{r},\mathbf{r}',iu)\,du\,. \tag{2.28}$$

‡ While the response can in general be highly non-linear, the electron correlation energy can be fully recovered solely based on linear response functions, which allows linear response theory and the ACFD theorem to be exact. [114]

To further simplify the derivation and explanation of the practical approaches outlined below, we may also reformulate the ACFD formula (2.27) in terms of the non-local, frequency-dependent polarizability tensor $\boldsymbol{\alpha}(\mathbf{r},\mathbf{r}',iu)$, which is connected to χ by $\chi(\mathbf{r},\mathbf{r}',iu) = \langle \nabla_{\mathbf{r}} | \nabla_{\mathbf{r}'} \boldsymbol{\alpha}(\mathbf{r},\mathbf{r}',iu) \rangle$. Introducing the dipole coupling tensor $\mathbf{T}(\mathbf{r},\mathbf{r}') = -\nabla_{\mathbf{r}} \otimes \nabla_{\mathbf{r}'} \mathcal{V}_{\text{Coul}}(\mathbf{r},\mathbf{r}')$, allows to rewrite the ACFD correlation energy (2.27) as

$$E_{\text{corr}} = \frac{1}{2\pi} \int_0^\infty \int_0^1 \iint_{\mathbb{R}^6} \text{Tr}\left\{ \left[\boldsymbol{\alpha}_\lambda(\mathbf{r},\mathbf{r}',iu) - \boldsymbol{\alpha}_{\lambda=0}(\mathbf{r},\mathbf{r}',iu) \right] \mathbf{T}(\mathbf{r},\mathbf{r}') \right\} d\mathbf{r} d\mathbf{r}' \, d\lambda \, du \,, \tag{2.29}$$

where Tr{·} denotes the trace operator over Cartesian components.[114] The non-local polarizability within the independent-particle framework ($\boldsymbol{\alpha}_{\lambda=0}$) can then be used to calculate the polarizability tensor of the correlated system *via* the self-consistent Dyson equation,

$$\begin{aligned} \boldsymbol{\alpha}_\lambda(\mathbf{r},\mathbf{r}',iu) &= \boldsymbol{\alpha}_{\lambda=0}(\mathbf{r},\mathbf{r}',iu) - \iint_{\mathbb{R}^6} \boldsymbol{\alpha}_{\lambda=0}(\mathbf{r},\mathbf{r}'',iu)\, \lambda \mathbf{T}_{\text{xc},\lambda}(\mathbf{r}'',\mathbf{r}''',iu)\, \boldsymbol{\alpha}_\lambda(\mathbf{r}''',\mathbf{r}',iu)\, d\mathbf{r}'' d\mathbf{r}''' \\ &\equiv \boldsymbol{\alpha}_0 - \left\langle \lambda \boldsymbol{\alpha}_0 \mathbf{T}_{\text{xc},\lambda} \boldsymbol{\alpha}_\lambda \right\rangle_{\mathbf{r}'',\mathbf{r}'''} = \sum_{n=0}^{\infty} \left\langle \boldsymbol{\alpha}_0 \left(-\lambda \mathbf{T}_{\text{xc},\lambda} \boldsymbol{\alpha}_0 \right)^n \right\rangle_{\mathbf{r}'',\mathbf{r}'''} \,, \end{aligned} \tag{2.30}$$

where we have introduced the shorthand $\langle\cdot\rangle_{\mathbf{r}'',\mathbf{r}'''}$ for the integration over spatial coordinates, $\mathbf{r}''$ and $\mathbf{r}'''$, and skipped the explicit notion of the variables of $\boldsymbol{\alpha}$ and $\mathbf{T}_{\text{xc},\lambda}$ for clarity. The coupling tensor $\mathbf{T}_{\text{xc},\lambda}$ is defined for each coupling strength λ as,[114]

$$\mathbf{T}_{\text{xc},\lambda}(\mathbf{r}'',\mathbf{r}''',u) = \mathbf{T}(\mathbf{r}'',\mathbf{r}''') - \frac{1}{\lambda} \nabla_{\mathbf{r}''} \otimes \nabla_{\mathbf{r}'''} f_{\text{xc},\lambda}(\mathbf{r}'',\mathbf{r}''',u) \,. \tag{2.31}$$

In practice, the exact xc kernel, $f_{\text{xc},\lambda}$, in Eq. (2.31) is not known. Thus, direct evaluation of the ACFD formula, both in terms of χ and $\boldsymbol{\alpha}$, is not possible. Practical approaches that directly rely on the ACFD formulation therefore involve the neglect of the explicit dependence on λ and additional approximations for the xc kernel. The most prominent method among those relies on the random phase approximation (RPA).

For the purpose of discussing vdW interactions, *i.e.* long-range correlation forces, and especially for comparing the various approaches to describe vdW interactions, it is worthwhile to further separate the above exact formulation of the electron correlation energy into short- and long-range contributions as detailed in Refs. 114 and 120. For this, we split the coupling tensors in Eqs. (2.29) and (2.30) by means of a range-separating function $g_{\text{rs}}(\|\mathbf{r}-\mathbf{r}'\|)$, which satisfies $g_{\text{rs}}(0) = 0$ and $g_{\text{rs}}(\|\mathbf{r}-\mathbf{r}'\| \to \infty) = 1$. This separates the total electron correlation energy into a short- and a long-range contribution, where the latter can be seen as an analogue to the original definition of vdW dispersion interactions based on a perturbative picture of intermolecular interactions.[114] In the Dyson equation (2.30), we may range-separate the full coupling tensor, $\mathbf{T}_{\text{xc},\lambda}$, into a short- and a long-range screening tensor ($\mathbf{T}^{(\text{sr})}_{\text{xc},\lambda}$ and $\mathbf{T}^{(\text{lr})}_{\text{xc},\lambda}$, respectively) according to

$$\begin{aligned} \mathbf{T}_{\text{xc},\lambda} &= \underbrace{\left[1 - g_{\text{rs}}(\|\mathbf{r}-\mathbf{r}'\|)\right] \mathbf{T}_{\text{xc},\lambda}}_{\mathbf{T}^{(\text{sr})}_{\text{xc},\lambda}} + \underbrace{g_{\text{rs}}(\|\mathbf{r}-\mathbf{r}'\|) \mathbf{T}_{\text{xc},\lambda}}_{\mathbf{T}^{(\text{lr})}_{\text{xc},\lambda}} = \\ &= \mathbf{T}^{(\text{sr})}_{\text{xc},\lambda} + \mathbf{T}^{(\text{lr})}_{\text{xc},\lambda} \,, \end{aligned} \tag{2.32}$$

which account for short- and long-range screening of the non-local polarizability, respectively. Inserting this into Eq. (2.30) and subsequently contracting all short-range screening components lets us define an effective, non-local polarizability, $\boldsymbol{\alpha}^{(\text{sr})}$, which already includes short-range screening. By the use of this definition, Eq. (2.30) becomes

$$\boldsymbol{\alpha}_\lambda = \sum_{n=0}^{\infty} \left\langle \boldsymbol{\alpha}^{(\text{sr})} \left(-\lambda \mathbf{T}^{(\lambda)}_{\text{xc,lr}} \boldsymbol{\alpha}^{(\text{sr})} \right)^n \right\rangle_{\mathbf{r}'',\mathbf{r}'''} \,. \tag{2.33}$$

Finally, combining Eq. (2.33) and the long-range part of the ACFD formula (2.29) gives the long-range electron correlation energy,

$$E_{\mathrm{corr}}^{(\mathrm{lr})} = -\sum_{n=1}^{\infty} \frac{1}{2\pi} \int_0^{\infty} \int_0^{1} \left\langle \mathrm{Tr}\left\{ \left[\left\langle \boldsymbol{\alpha}^{(\mathrm{sr})} \left(\lambda \mathbf{T}_{\mathrm{xc},\lambda}^{(\mathrm{lr})} \boldsymbol{\alpha}^{(\mathrm{sr})} \right)^n \right\rangle_{\mathbf{r}'',\mathbf{r}'''} \right] \mathbf{T}_{\mathrm{lr}}(\mathbf{r},\mathbf{r}') \right\} \right\rangle_{\mathbf{r},\mathbf{r}'} d\lambda\, du\,, \tag{2.34}$$

where $\mathbf{T}_{\mathrm{lr}}$ is the long-range part of the bare dipole coupling tensor and the $n = 0$ term cancels out. Note that we have not introduced any approximations up to this point and the sum of Eq. (2.34) and its short-range analogue still equals the exact total correlation energy as defined by Eq. (2.29). Many of the nowadays widely-used vdW models can be traced down to this form of the ACFD formula, where each flavor involves a given approximation for the effective polarizability after short-range screening, $\boldsymbol{\alpha}^{(\mathrm{sr})}$, the (dipolar) xc kernel in $\mathbf{T}_{\mathrm{xc},\lambda}^{(\mathrm{lr})}$ or combinations thereof.

2.3.2 Approximate Reduction to Atom-Pairwise Interaction Potentials

As can be seen from the definitions above, the polarizability and the electron correlation energy have a highly complex, non-local character. As of today, numerous experimental and theoretical works have clearly shown the many-body nature of dispersion forces. Nonetheless, one of the most common approaches to model vdW interactions is by the use of pairwise-additive potentials. In this section, we sketch the approximations and basic steps leading to the fundamental form of pairwise potentials for long-range correlation forces based on the long-range ACFD formula (2.34). At this point, we would like to note that the functional form derived below can be, and was obtained, in multiple ways including (many-body) perturbation theory and other approximations to the ACFD formula.

One of the most successful and common approximations is the so-called random phase approximation (RPA), which corresponds to the neglect of the contributions from the unknown xc kernel ($\nabla_{\mathbf{r}} \otimes \nabla_{\mathbf{r}'} f_{\mathrm{xc},\lambda} = \mathbf{0}$). Within the range-separated ACFD formula, we may apply that approximation only in the long-range part, where $f_{\mathrm{xc},\lambda}$ indeed barely contributes, such that $\mathbf{T}_{\mathrm{xc},\lambda}^{(\mathrm{lr})}$ in Eq. (2.34) no longer explicitly depends on λ and reduces to $\mathbf{T}_{\mathrm{lr}}$. This allows us to analytically carry out the integration over the coupling strength, which leads to the series

$$E_{\mathrm{corr}}^{(\mathrm{lr,RPA})} = -\sum_{n=2}^{\infty} \frac{(-1)^n}{n} \frac{1}{2\pi} \int_0^{\infty} \left\langle \mathrm{Tr}\left\{ \left\langle \left(\boldsymbol{\alpha}^{(\mathrm{sr})} \mathbf{T}_{\mathrm{lr}} \right)^n \right\rangle_{\mathbf{r}'',\mathbf{r}'''} \right\} \right\rangle_{\mathbf{r},\mathbf{r}'} du\,, \tag{2.35}$$

where the index n is shifted by +1 due to integration over λ. Next, we approximate the non-local polarizability by a sum of polarizabilities situated at the N atomic positions, $\{\mathbf{R}_A\}$. Inserting this into the long-range RPA correlation energy gives,

$$E_{\mathrm{corr}}^{(\mathrm{lr,RPA})} \approx -\sum_{n=2}^{\infty} \frac{(-1)^n}{n} \frac{1}{2\pi} \int_0^{\infty} \left\langle \mathrm{Tr}\left\{ \left\langle \left(\sum_{A=1}^{N} \boldsymbol{\alpha}_A^{(\mathrm{sr})} \mathbf{T}_{\mathrm{lr}} \right)^n \right\rangle_{\mathbf{r}'',\mathbf{r}'''} \right\} \right\rangle_{\mathbf{r},\mathbf{r}'} du\,. \tag{2.36}$$

It can be seen that we get $\boldsymbol{\alpha}_A^{(\mathrm{sr})} \mathbf{T}_{\mathrm{lr}} \boldsymbol{\alpha}_B^{(\mathrm{sr})} \mathbf{T}_{\mathrm{lr}}$ for $n = 2$, $\boldsymbol{\alpha}_A^{(\mathrm{sr})} \mathbf{T}_{\mathrm{lr}} \boldsymbol{\alpha}_B^{(\mathrm{sr})} \mathbf{T}_{\mathrm{lr}} \boldsymbol{\alpha}_C^{(\mathrm{sr})} \mathbf{T}_{\mathrm{lr}}$ for $n = 3$, and so on. Thus, the expansion series (2.36) is a series of all the n^{th}-order correlation terms. As such, the order n does not represent a pure n-body (*body* referring to *atom*) vdW interaction term, as for instance defined in the perturbational approach. For example, $n = 3$ contains non-vanishing terms with $C = A$, which correspond to screened two-body interactions. $n = 2$, on the other side, only involves non-vanishing terms with two different polarizability centers A and B and is therefore a pure (yet incomplete) two-body vdW interaction. If we limit ourselves to this second-order term, $E_{\mathrm{corr}}^{(2)}$, we can carry out the integration over spatial coordinates to arrive at,

$$E_{\mathrm{corr}}^{(2)} = -\frac{1}{2} \frac{1}{2\pi} \int_0^{\infty} \mathrm{Tr}\left\{ \sum_{A \neq B} \boldsymbol{\alpha}_A^{(\mathrm{sr})} \mathbf{T}_{AB}^{(\mathrm{lr})} \boldsymbol{\alpha}_B^{(\mathrm{sr})} \mathbf{T}_{BA}^{(\mathrm{lr})} \right\} du\,, \tag{2.37}$$

where $\mathbf{T}_{AB}^{(\mathrm{lr})} \equiv \mathbf{T}_{\mathrm{lr}}(\mathbf{R}_A, \mathbf{R}_B)$. As a final approximation, we assume the polarizabilities to be isotropic, *i.e.* $\boldsymbol{\alpha}_A^{(\mathrm{sr})} = \alpha_A^{(\mathrm{sr})} \mathbb{1}_3$, with $\mathbb{1}_3$ being (3×3) unity. As a result, the polarizabilities and dipole tensors commute and

$$E_{\mathrm{corr}}^{(2)} = -\frac{1}{2} \sum_{A \neq B} \frac{3}{\pi} \int_0^\infty \alpha_A^{(\mathrm{sr})} \alpha_B^{(\mathrm{sr})} \, du \, \frac{1}{6} \mathrm{Tr} \left\{ \mathbf{T}_{AB}^{(\mathrm{lr})} \mathbf{T}_{BA}^{(\mathrm{lr})} \right\} . \tag{2.38}$$

The integral in the above equation is known as the *Casimir-Polder integral*[121] and corresponds to the so-called C_6-interaction coefficients (*Hamaker constant* for macroscopic spherical bodies[122]). Noting that $\mathbf{T}_{AB}^{(\mathrm{lr})} = g_{\mathrm{rs}}(\|\mathbf{R}_A - \mathbf{R}_B\|) \mathbf{T}_{AB}$ and $\mathrm{Tr}\{\mathbf{T}_{AB} \mathbf{T}_{BA}\} = 6/\|\mathbf{R}_A - \mathbf{R}_B\|^6$ leads to

$$E_{\mathrm{corr}}^{(2)} = -\frac{1}{2} \sum_{A \neq B} \underbrace{\frac{3}{\pi} \int_0^\infty \alpha_A^{(\mathrm{sr})} \alpha_B^{(\mathrm{sr})} \, du}_{C_{6,AB}^{(\mathrm{eff})}} \frac{g_{\mathrm{rs}}(\|\mathbf{R}_A - \mathbf{R}_B\|)^2}{\|\mathbf{R}_A - \mathbf{R}_B\|^6} \tag{2.39}$$

$$= -\frac{1}{2} \sum_{A \neq B} C_{6,AB}^{(\mathrm{eff})} \frac{f_{\mathrm{damp}}(R_{AB})}{R_{AB}^6} , \tag{2.40}$$

where we have used $R_{AB} = \|\mathbf{R}_A - \mathbf{R}_B\|$ and introduced the damping function $f_{\mathrm{damp}}(R_{AB}) = g_{\mathrm{rs}}(R_{AB})^2$. This is the well-known formula for the vdW dispersion interaction between two microscopic bodies within the dipole approximation as first derived by London.[123] As of today, a manifold of pairwise-additive vdW models has been devised and widely used.[48,49,124–130] The general difference between those models lies in the damping function, f_{damp}, and how effective, short-range screened interaction coefficients are obtained. It is worthwhile to point out that the same functional form can be derived based on a full-range RPA of Eq. (2.29), by invoking the so-called full potential approximation, *i.e.* $\boldsymbol{\alpha}_\lambda = \boldsymbol{\alpha}_1$,[131] from a model system of quantum Drude oscillators[25] or in various ways from (second-order) perturbation theory invoking a multipole expansion and subsequent dipole approximation for the interaction potential.[25]

2.3.3 Pairwise-Additive van der Waals Models

Augmenting (semi-)local electronic-structure calculations *a posteriori* with a London-type vdW term, as first put forward by Wu and Yang[132] and popularized as a general framework by Grimme,[124] represents an early and efficient approach to correct for the lack of long-range electron correlation. As detailed in Sec. 2.3.2, the fundamental mathematical form can be derived from a coarse-grained ACFD/RPA formula. Note that from our derivation of (2.40), the C_6-coefficients are defined by the Casimir-Polder integral based on isotropic, static atomic polarizabilities, which should already include short-range screening. However, almost none of the pairwise-additive approaches devised to date explicitly accounts for the electrodynamic screening. Instead, most methods rely on effective polarizabilities or C_6-coefficients, which are meant to implicitly include such screening effects. The various pairwise models we have today basically differ in the way those effective vdW parameters are determined. These typically also involve an atom pair-dependent reference distance, which parametrically defines the switching behavior of the damping function. Thereby, the actual mathematical form of this damping function has been shown to have a minor effect on the final vdW energetics.[130]

The Electronic-Structure-Based Atom-Pairwise vdW(TS) Method

One very successful way to model the polarizability of the KS reference system without recourse to the computationally demanding Adler-Wiser formalism,[118,119] is by incorporating information on the (ground-state) electronic structure. This represents an approximate, yet reliable and efficiently method to account for the effect of the (local) chemical environment on polarizabilities and interaction coefficients. A variety of successful schemes in this spirit has been devised to date, such as the LRD model,[128] the non-local density functional for interaction coefficients

by Tao *et al.*,[133,134] the vdW-WF method[135,136] or the exchange-hole dipole moment (XDM) model[125,126] and the related density-dependent dispersion correction (dDsC) scheme.[129] This thesis focuses on the vdW(TS) approach[48] as electronic-structure-based pairwise vdW model. As it shares a common basis with the many-body dispersion method (*vide infra*), their comparison is expected to provide an optimal estimate for beyond-pairwise effects.

The vdW(TS) scheme starts from the leading Padé approximant of the dynamic polarizability[137] based on an effective static atomic polarizability, $\alpha_{A,0}^{(\mathrm{TS})} \equiv \alpha_A^{(\mathrm{TS})}(iu=0)$. This polarizability is then used to approximate the short-range screened polarizability entering Eq. (2.39):

$$\alpha_A^{(\mathrm{sr})}(iu) \approx \alpha_A^{(\mathrm{TS})}(iu) = \alpha_{A,0}^{(\mathrm{TS})}\left[1+\left(\frac{u}{\eta_A}\right)^2\right]^{-1}, \tag{2.41}$$

where η_A corresponds to an effective excitation frequency.[48] Inserting this into the Casimir-Polder integral in Eq. (2.39) yields the London formula,[138] from which we can define the C_6-interaction coefficients entirely based on effective static atomic polarizabilities according to

$$C_{6,AB}^{(\mathrm{eff})} \approx C_{6,AB}^{(\mathrm{TS})} = \frac{2\,C_{6,AA}^{(\mathrm{TS})}\,C_{6,BB}^{(\mathrm{TS})}}{\frac{\alpha_{B,0}^{(\mathrm{TS})}}{\alpha_{A,0}^{(\mathrm{TS})}}C_{6,AA}^{(\mathrm{TS})} + \frac{\alpha_{A,0}^{(\mathrm{TS})}}{\alpha_{B,0}^{(\mathrm{TS})}}C_{6,BB}^{(\mathrm{TS})}} \qquad \text{and} \qquad C_{6,AA}^{(\mathrm{TS})} = \frac{3}{4}\eta_A\left[\alpha_{A,0}^{(\mathrm{TS})}\right]^2. \tag{2.42}$$

Hence, the key quantity is the effective static atomic polarizability. To obtain this polarizability, one takes advantage of the (near-)linear correlation between the atomic volume, V_A, and the (static) atomic polarizability, *i.e.*, $\alpha_A(iu=0) = \kappa_A \cdot V_A$ with κ as proportionality constant.[139] This allows the definition

$$\alpha_{A,0}^{(\mathrm{TS})} = \frac{\kappa_{\mathrm{eff}}^{(A)}\,V_{\mathrm{eff}}^{(A)}}{\kappa_{\mathrm{free}}^{(A)}\,V_{\mathrm{free}}^{(A)}}\,\alpha_{A,0}^{(\mathrm{free})} = \frac{\kappa_{\mathrm{eff}}^{(A)}}{\kappa_{\mathrm{free}}^{(A)}}\cdot x_V^{(A)}\cdot\alpha_{A,0}^{(\mathrm{free})}, \tag{2.43}$$

where $\alpha_{A,0}^{(\mathrm{free})}$ is the static polarizability of the corresponding atom *in vacuo*.[48] The atomic volume can be determined as the expectation value of the cube of the electron-nucleus distance, r, based on the atomic density of the atom in its chemical environment or of the corresponding isolated atom, respectively. The effective atomic density is conventionally obtained *via* Hirshfeld analysis,[140] from which the rescaling factor, x_V, is given by

$$x_V^{(A)} = \frac{V_{\mathrm{eff}}^{(A)}}{V_{\mathrm{free}}^{(A)}} = \frac{\int r^3\,w_A(\mathbf{r})\,\rho(\mathbf{r})\,\mathrm{d}\mathbf{r}}{\int r^3\,\rho_{\mathrm{free}}^{(A)}(\mathbf{r})\,\mathrm{d}\mathbf{r}}, \quad w_A(\mathbf{r}) = \frac{\rho_{\mathrm{free}}^{(A)}(\mathbf{r})}{\sum_B \rho_{\mathrm{free}}^{(B)}(\mathbf{r})}, \tag{2.44}$$

where w_A is the Hirshfeld weighting factor and $\rho(\mathbf{r})$ is the total electron density of the molecule or material. All densities, including the *in vacuo* atomic density, are evaluated at run time. Finally, inserting the effective atomic polarizability into the second part of Eq. (2.42), together with an equivalent consideration of (2.42) for an isolated atom, gives

$$C_{6,AA}^{(\mathrm{TS})} = \frac{\eta_A}{\eta_A^{(\mathrm{free})}}\left[\frac{\kappa_{\mathrm{eff}}^{(A)}}{\kappa_{\mathrm{free}}^{(A)}}\right]^2\left[x_V^{(A)}\right]^2 C_{6,AA}^{(\mathrm{free})} \simeq \left[x_V^{(A)}\right]^2\cdot C_{6,AA}^{(\mathrm{free})}, \tag{2.45}$$

where, upon closer inspection, the two prefactors involving η and κ together have been found to be well approximated by unity.[48] Relying on accurate reference data for the C_6-coefficients of the corresponding isolated atoms, $C_{6,AA}^{(\mathrm{free})}$, this approach has been shown to yield accurate effective interaction coefficients within 5.5 % from values derived from experimental dipole oscillator strength distributions.[48] As final ingredient of the energy expression (2.40), a Fermi-type damping function was proposed,

$$f_{\mathrm{damp}}^{(\mathrm{TS})}\left(R_{AB};R_{\mathrm{vdW}}^{(AB)}\right) = \left\{1+\exp\left[-d\left(\frac{R_{AB}}{s_R\cdot R_{\mathrm{vdW}}^{(AB)}}-1\right)\right]\right\}^{-1}, \tag{2.46}$$

where the steepness of the damping, d, has been found to have a negligible effect on binding energies and is therefore fixed to $d = 20$. The onset of the range-separation, finally, is determined by the scaling parameter s_R (typical values: 0.94 for PBE, 0.96 for PBE0, 0.84 for B3LYP) and an effective vdW distance, $R_{\mathrm{vdW}}^{(AB)}$, given by the sum of the corresponding effective vdW radii of atoms A and B. Based on the definition of the vdW radius by Pauling and considerations from classical physics, the vdW radius of an atom is proportional to the cube-root of its volume.‡ This allows to define an effective vdW radius of an atom in a similar manner from its *in vacuo* counterpart:

$$R_{\mathrm{vdW}} \propto \sqrt[3]{V} \quad \Longrightarrow \quad R_{\mathrm{vdW}}^{(A)} = \sqrt[3]{x_V^{(A)}} \cdot R_{\mathrm{vdW}}^{(A,\mathrm{free})} \,. \tag{2.47}$$

Inserting the definitions in Eqs. (2.44)–(2.47) into the London-type vdW formula (2.40) then fully defines the vdW dispersion energy in vdW(TS) as given by

$$E_{\mathrm{vdW}}^{(\mathrm{TS})} = -\frac{1}{2} \sum_{A \neq B} f_{\mathrm{damp}}^{(\mathrm{TS})}(R_{AB}) \frac{C_{6,AB}^{(\mathrm{TS})}}{R_{AB}^6} \,. \tag{2.48}$$

Effective, electronic-structure-based vdW parameters can also be obtained by an alternative, yet similarly accurate and reliable, approach, which relies on *net atomic populations* instead of the real-space representation of the electron density as used in the Hirshfeld scheme outlined above. Atomic populations as initially classified by Mulliken, can be calculated in *Fock space, i.e.*, from the density-matrix in an atom-centered basis set representation. The alternative rescaling factor $x_{\mathbf{P}}$ is defined as,[142]

$$x_{\mathbf{P}}^{(A)} = \frac{h_A}{Z_A} \quad ; \qquad h_A = \sum_a f_a \sum_{i \in A} \left\| c_{ia} \right\|^2 = \sum_{i \in A} \mathbf{P}_{ii} \tag{2.49}$$

with Z_A as nuclear charge (atomic number) of atom A corresponding to h_A for an atom *in vacuo*, f_a as occupation number of the electronic state a and c_{ia} as LCAO coefficient. Note that h_A, being the atom-projected trace of the Mulliken population matrix $\mathbf{P}$, does not involve off-diagonal (mixed) terms of the density matrix. As such, it does not suffer from the ambiguity of partitioning the electron population of overlap regions, which represents the main and fundamental pitfall of Fock-space charge partitioning schemes. This approach yields interaction coefficients *en par* with the original scheme[142] and allows for the usage of the vdW(TS) model in conjunction with electronic-structure methods without real-space representation of the electron density, such as the semi-empirical DFTB method or other density matrix-based approaches. Optimized damping parameters for select DFTB parametrizations are given in Appendix A1. Similar in spirit, yet neglecting some hybridization effects and relying on the not well-defined *full* Mulliken charge, is the dDMC vdW model by Petraglia and co-workers.[143]

The vdW(TS) scheme can also be used to investigate the effect of dispersion interactions on the electronic structure and derived properties.[106] As all parameters entering Eq. (2.48) are represented as a functional of the electron density or density matrix, the effective potential arising from long-range correlation forces can be derived. Inclusion of this term in the self-consistency procedure, termed self-consistent vdW(TS), has been shown to affect the work function of metals, for instance,[106] and can be expected to play an increasingly important role with increasing system size and polarizability. The main approximations of the vdW(TS) approach for describing dispersion forces are the pairwise formulation and the neglect of explicit screening effects, which appear in higher-order terms of the (coarse-grained) RPA correlation energy (2.36). A considerable limitation beyond these effects in vdW(TS) is the usage of Hirshfeld analysis to capture effects of the local chemical environment on the polarizability. Ordinary Hirshfeld partitioning tends to underestimate charge transfer[144,145] and in line with that the volume ratios tend to

‡ In a side project of this doctorate (Ref. 141), we indicate that such classical considerations may be insufficient and that, in quantum systems, different scaling laws can apply.

underestimate the corresponding effect on the atomic polarizability. This can lead to considerable deficiencies in the description of the vdW parameters of ionic systems.[146,147] Significant improvements can be achieved when relying on the computationally more demanding, but much more accurate, iterative Hirshfeld scheme[145,147] or when using a charge-dependent reference state for the polarizability.[148] For the simulation of hybrid organic-inorganic interfaces, an adapted version *vdW^{surf}* has been devised,[149] which accounts for the metallic screening in the substrate according to *Lifshitz-Zaremba-Kohn theory*.[150,151] The vdW$^{\text{surf}}$ model significantly improves upon the original scheme and provides an accurate description of the binding properties of metal surface-adsorbed organic molecules.[149,152–155]

(Semi-)Empirical Pairwise-Additive Approaches: The D3 Method

The first widely used vdW model in the context of DFT was the DFT-D approach by Grimme, which followed the form of Eq. (2.40) and featured effective, but fixed C_6-interaction coefficients and a Fermi-type damping function.[124] The applicability was later extended by deducing effective interaction coefficients from atomic properties (DFT-D2).[127] It is worthwhile to mention that both approaches did not account for any effects of the (local) chemical environment nor did they yield the correct asymptotic behavior.[49] These obsolete methods can thus not be recommended for use in electronic-structure calculations today. After careful numerical investigation of the effect of the local chemical environment, a new semi-empirical variant, termed D3, was devised. The scheme is based on atom pair-specific C_6 coefficients and includes local information in the form of geometry-motivated, fractional coordination numbers,[49]

$$\mathrm{CN_A} = \sum_{B\neq A} \left\{ 1 + \exp\left[-p_1 \cdot \left(p_2 \cdot \frac{R_{\mathrm{cov}}^{(A)} + R_{\mathrm{cov}}^{(B)}}{R_{AB}} - 1 \right) \right] \right\}^{-1} , \tag{2.50}$$

where the parameters $p_1 = 16$ and $p_2 = 4/3$ have been chosen based on a set of organic molecules, R_{AB} is the distance between atoms A and B, and $R_{\mathrm{cov}}^{(A)}$ is the (scaled) covalent radius of atom A. The final procedure has been shown to yield chemically sensible coordination numbers for a variety of organic and non-organic systems.[49] The interaction coefficient for atoms A and B is then calculated for a number of different coordination numbers, which is achieved by considering the corresponding hydrides, and approximately decomposed to provide reference values for $C_{6,\mathrm{AB}}$. This collection of coordination number-dependent C_6-coefficients then serves as a reference database and the final effective interaction coefficient, $C_{6,AB}^{(\mathrm{D3})}$, which enters Eq. (2.40), is obtained from interpolation of the reference coefficients *via*

$$C_{6,AB}^{(\mathrm{D3})}(\mathrm{CN}_A, \mathrm{CN}_B) = \frac{1}{\mathcal{L}} \sum_{A_{\mathrm{ref}}} \sum_{B_{\mathrm{ref}}} C_{6,A_{\mathrm{ref}}B_{\mathrm{ref}}} L(A_{\mathrm{ref}}, B_{\mathrm{ref}}) , \tag{2.51}$$

$$\text{where} \quad L(A_{\mathrm{ref}}, B_{\mathrm{ref}}) = e^{-p_3\left[\left(\mathrm{CN}_A - \mathrm{CN}_{A_{\mathrm{ref}}}\right)^2 + \left(\mathrm{CN}_B - \mathrm{CN}_{B_{\mathrm{ref}}}\right)^2\right]} \tag{2.52}$$

and $\mathcal{L}$ is the sum of all Gaussian distances $L(A_{\mathrm{ref}}, B_{\mathrm{ref}})$, The last global *ad hoc* parameter $p_3 = 4$ to assure smooth behavior at integer coordination numbers.[49] Thus, the effective interaction coefficients are interpolated from reference values based on their local coordination. The general procedure for the definition of coordination numbers and the interpolation scheme is thereby, in principle, completely arbitrary and was motivated by numerical results.[49] The geometry-based D3 model neglects any electronic-structure and explicit screening effects, but at the same time allows for a vdW correction for any given total energy method including MM or DFTB.[156] As showcased by Ehrlich *et al.* strong electronic-structure effects like far-from-neutral species, can be incorporated by a suitable choice of reference systems for the interpolation scheme.[157] For general applications however, such an approach introduces a certain degree of empiricism and requires a careful choice and testing. Recently, also a more straightforward approach to include such effects by rescaling of interaction coefficients based on atom-in-a-molecule charges was

proposed (D4) and shown to significantly improve transferability and general applicability.[158] The D3 scheme also involves an additional term for pairwise dipole–quadrupole vdW interactions, which scales as $1/R^8$ (derived from perturbation theory). The C_8-interaction coefficients, *i.e.* the equivalent of C_6 for dipole–quadrupole vdW interactions, are computed recursively[159–161] based on the corresponding C_6-coefficients.[49] For the damping, the original scheme employed a formulation proposed by Chai and Head-Gordon.[162] Including the quadrupolar interaction term, this defines the vdW energy in D3 as

$$E_{\text{vdW}}^{(\text{D3})} = -\frac{1}{2}\sum_{A\neq B}\sum_{m=6,8}\frac{p_4^{(m)}}{1+6\left(p_5^{(m)}\cdot R_{\text{D3}}^{(AB)}/R_{AB}\right)^{2n_b+m+4}}\frac{C_{m,AB}^{(\text{D3})}}{R_{AB}^m}\,. \tag{2.53}$$

Above, $n_b = 2$ for two-body interaction and $m = 6,8$ denotes dipole–dipole and dipole–quadrupole interaction. p_4 is a rescaling factor for quadrupolar interactions ($p_4 = 1$ for $m = 6$), while p_5 is optimized for a given electronic-structure method and, together with the atom pair-dependent cutoff radius, R_{D3}, determines the onset of the vdW correction. The cutoff radius R_{D3} is determined from the attenuation of the (semi-)local interaction energy below a certain threshold for the corresponding dimer.[49] Such a choice of cutoff parameters instead of vdW radii in the damping function can conceptually be justified as the appropriate range-separation is not necessarily a function of vdW radii, but depends on the range of electron correlation captured by the underlying (semi-)local electronic-structure method. This, however, is highly system-dependent and a rigorous scheme for arbitrary systems has not been derived so far. As an alternative range-separation, it was proposed to employ a rational damping as developed for the XDM model also in D3.[125,126,130] The damping in this widely-used D3-BJ approach thereby relies on a switching radius defined by the square-root of the ratio of effective C_8- and C_6-interaction coefficients.[130] A common extension of the D3 framework is to also include beyond-pairwise terms in the form of the three-body term according to Axilrod and Teller[163] and Muto (ATM),[164]

$$E_{\text{vdW}}^{(\text{ATM})} = \sum_{A,B,C}\left[\cos(\varphi)\cos(\vartheta)\cos(\theta)+1\right]\frac{C_{9,ABC}}{(R_{AB}R_{BC}R_{CA})^3}\,, \tag{2.54}$$

where $\varphi, \vartheta, \theta$ are the angles of the triangle spanned by the positions of atoms A, B, and C. In the context of D3, the effective three-body C_9-interaction coefficients are approximated as minus the square-root of the product of the corresponding two-body C_6-coefficients. In order to use the ATM three-body energy together with (semi-)local electronic-structure calculations, the ATM term has to be damped at short distances. Within the D3 framework, this damping is the same as the first term in Eq. (2.53) using $p_4 = 4/3$, $n_b = 3$ and $m = 6$. The definition of an effective distance for the damping thereby introduces an inevitable ambiguity, which lead to the formulation of various forms.[91,165–167] Together with intricate error cancellations, this ambiguity can give rise to a sometimes unpredictable performance of adding the ATM three-body term.[29,91,168,169]

2.3.4 Non-Additive Aspects of van der Waals Interactions

Theoretical Considerations

The above derivations introduce the theoretical foundations and select practical formulations of the approximate, pairwise vdW (interaction) energy. However, no rigorous and seamless way of obtaining the damping function, the short-range screened polarizabilities, or the resulting C_6-interaction coefficients has been put forward to date. In addition, the expansion (or corresponding perturbation) series is truncated at second order and one has to invoke an additional approximation for the — at least long-range part of the — coupling potential, $\mathbf{T}_{\text{xc}}$, to arrive at the pairwise formula (2.40). As classified by Dobson,[170] effects beyond this pairwise-additive expression for vdW interactions can, in general, be understood in terms of three different aspects of non-additivity.

- **Type A**: The effect of the local chemical environment on the polarizability apart from short-range screening. One often relies on partitioning the system into its constituent atoms for the construction of the polarizability of the (sub)system. Type A non-additivity can be understood by the fact that the polarizability in the KS reference system does not correspond to a superposition of $\boldsymbol{\alpha}_0$ of isolated atoms. This type of non-additivity is accounted for in almost all modern vdW models.
- **Type B**: Electron correlation and screening are defined by multi-center integrals. This enters the ACFD formalism both in form of the electrodynamic screening in the Dyson-like equation (2.30) as well as in the expansion series of the (long-range) correlation energy to finite orders of n in Eq. (2.34), *i.e.*, many-body interactions and higher-order correlation terms. The difference of coupled N-center interactions and a sum of pairwise terms is illustrated in Fig. 2.1: In the pairwise formula, the energy arises as the sum of the interaction of ideally oscillating pairs of dipoles. However, all oscillations are coupled simultaneously (multi-center interaction), which does not necessarily correspond to a sum of ideally correlated dipoles [171] as shown by select collective density fluctuation patterns for the simple example of an Argon trimer in Fig. 2.1 (right). This type of non-additivity manifests itself particularly in (sub)systems with strong anisotropy, complex geometrical arrangements, or reduced symmetry (1D, 2D materials) and can substantially alter the scaling laws of the vdW interaction. [50,92,110,114,170,172]
- **Type C**: Assigned to systems with extremely large delocalization lengths, basically corresponding to intrinsic electron hopping between atomic centers. Such phenomena cannot be fully described within an atom-centered framework of polarizability and electronic fluctuations. Understandably, type C non-additivity almost exclusively appears in systems with (near-)zero band (or HOMO-LUMO) gap, which allow for quantum-mechanical fluctuations to cause instantaneous electron hopping. [170]

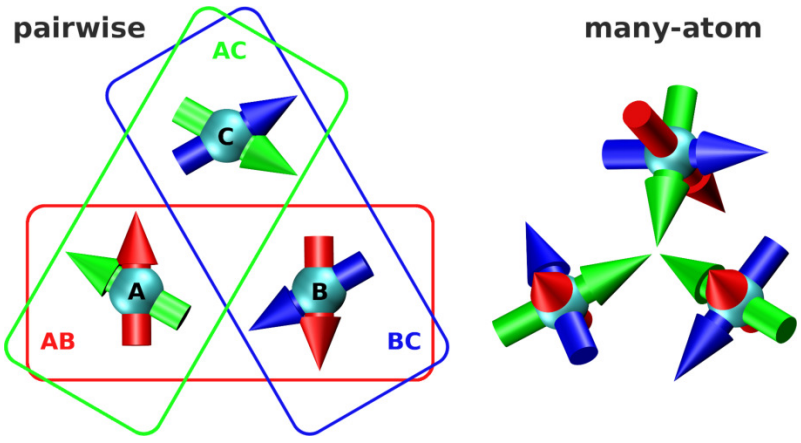

Fig. 2.1: Illustrative comparison of the (assumed) physics behind pairwise-additive (left) and many-atom (right) van der Waals methods in an atomistic picture and dipole limit for an Argon trimer. The arrows of a given color each depict a mode of simultaneous electronic fluctuations (represented by instantaneous fluctuations around atomic centers). Argon atoms shown in cyan. Reproduced from Ref. 1 with permission from the Royal Society of Chemistry.

Experimental Indication and Proof of Non-Additivity

Recent years have also brought forth a growing number of experimental observations of the non-additive nature of vdW dispersion forces. One of the most well-known deviations from pairwise additivity thereby appears when a single atom or molecule is interacting with a metallic surface. Since the early theoretical works by Lifshitz [150] and Zaremba and Kohn [151] it is known that, at larger separations, the interaction energy follows a D^{-3} power law, where D is the distance of the atom or molecule to the surface. Atomic force microscopy (AFM) measurements by Wagner *et al.* [173] confirmed this scaling law and quantified the non-additivity. Also *between* adsorbed molecules, several experiments observed strongly non-additive long-range interactions. [174–176]

In a study on the adsorption of gold nanoparticles on multi-walled carbon nanotubes, Rance *et al.* showed that the adsorption affinity scales quadratically with the accessible surface area of the nanotubes and is highly non-linear for more complex nanostructures.[177] In contrast, pairwise-additive vdW models, neglecting molecular anisotropy and collective behavior, predict a simple linear dependence in those cases. Batista *et al.* emphasize that the non-additivity of interactions, including dispersion forces, arise particularly at the nanoscale[178] due to complex geometrical arrangements and the resulting polarizability anisotropy. Such behavior beyond pairwise additivity, however, can also extend to the meso- and macroscale as shown by the interaction range of proteins, bacteria, and gecko feet with bulk silicon. By separating the respective adhesive partner and the silicon substrate with an increasing layer of silicon dioxide, Loskill and co-workers showed that the interaction extends up to a separation of 10–20 nm,[179,180] while a pairwise formalism predicts only 1 nm. On the other side, covering dielectric bulk materials with strongly anisotropic monolayers can also screen the vdW interaction between the surface and an adsorbed molecule. Using AFM, it has been shown that the D^{-3}-dependence predicted by Lifshitz-Zaremba-Kohn theory holds for the interaction of the metallic AFM tip with pristine silicon dioxide. When the surface is covered by a 2D-material, such as graphene or molybdenum disulfide, the tip seems to only interact with the adsorbed 2D-material.[181] This unexpected behavior could be explained by in-plane electronic fluctuations within the 2D-material being decoupled from the bulk and with that screening electronic fluctuations perpendicular to the surface, *i.e.*, those responsible for the interaction of the AFM tip with the surface through the adsorbed monolayer.[182]

As most of our experience and understanding of vdW interactions is based on rather small systems, where a pairwise approximation tends to be qualitatively sufficient, many of the phenomena arising at larger length scales are still not entirely understood. This and the growing interest in nanostructured and low-dimensional materials motivate on-going studies including a quantum-mechanical many-body treatment of vdW interactions. The ability to reliably model and understand the interactions in such systems is of utmost importance for the design of composite nanostructures[177] and future (nano)technological developments as well as a better understanding and exploitation of biomolecular processes. This calls for a more advanced description of vdW dispersion beyond the traditional pairwise picture.

2.3.5 Interatomic Many-Body Method: The Many-Body Dispersion Formalism

A very successful and efficient approach to beyond-pairwise vdW interactions is the many-body dispersion (MBD) formalism.[50,51] As most practical *post*-DFT vdW models, it is based on a dipole approximation or RPA in an interatomic framework. The latter can be interpreted as coarse-graining the response functions entering the long-range ACFD/RPA formula (2.35), which we will repeat here for reasons of clarity:

$$E_{\mathrm{corr}}^{(\mathrm{lr,RPA})} = -\sum_{n=2}^{\infty} \frac{(-1)^n}{n} \frac{1}{2\pi} \int_0^{\infty} \left\langle \mathrm{Tr}\left\{ \left\langle \left(\boldsymbol{\alpha}^{(\mathrm{sr})}\, \mathbf{T}_{\mathrm{lr}}\right)^n \right\rangle_{\mathbf{r}'',\mathbf{r}'''} \right\} \right\rangle_{\mathbf{r},\mathbf{r}'} du \,. \tag{2.55}$$

The coarse-graining is usually chosen such that the spatial integration in Eq. (2.55) can be performed analytically, which significantly reduces the computational cost. In Sec. 2.3.2, we already introduced such a coarse-grained polarizability in terms of atomic polarizabilities, see Eq. (2.36). In MBD formalism, the total polarizability is contracted to a sum of effective isotropic atomic (dipole) polarizabilities.‡ Such atomic/molecular response properties have been shown to be accurately described by a quantum harmonic oscillator (QHO) model.[50,51,183–186] In fact, the leading Padé approximant of the dynamic isotropic atomic dipole polarizability[137] follows the

‡ The MBD formalism does not fundamentally exclude anisotropic polarizabilities on the atomic scale. The choice of *isotropic* atomic polarizabilities, however, allows for an efficient, analytical evaluation of the dipole coupling.

same formula as the dynamic dipole polarizability of an isotropic QHO,

$$\alpha_A^{(\mathrm{QHO})}(iu) \equiv \alpha_A^{(\mathrm{TS})}(iu) = \alpha_{A,0}^{(\mathrm{TS})}\left[1+\left(\frac{u}{\eta_A}\right)^2\right]^{-1} , \tag{2.56}$$

where $\alpha_{A,0}^{(\mathrm{TS})} \equiv \alpha_A^{(\mathrm{QHO})}(0)$ is the effective static QHO polarizability and η_A is the characteristic excitation frequency of QHO A. Hence, the remaining step is the parametrization of such QHOs to model atoms in molecules. As in vdW(TS), α_A and η_A are obtained from atomic reference data taking into account the local chemical environment (type A non-additivity) according to

$$x_A \approx \frac{\alpha_{A,0}^{(\mathrm{TS})}}{\alpha_{A,0}^{(\mathrm{ref})}} \approx \sqrt{\frac{C_{6,AA}^{(\mathrm{TS})}}{C_{6,AA}^{(\mathrm{ref})}}} \qquad \text{and} \qquad \eta_A = \frac{4}{3}\frac{C_{6,AA}^{(\mathrm{TS})}}{\left(\alpha_{A,0}^{(\mathrm{TS})}\right)^2} , \tag{2.57}$$

where the rescaling factor x is derived from the electronic structure, conventionally as the ratio of the volumes of the atom in the system and the reference atom as obtained *via* Hirshfeld analysis [140] (this is further detailed in Sec. 2.3.3 above). Equivalently to the case of vdW(TS), the MBD formalism can also be parameterized solely based on the system's density matrix with the use of Eq. (2.49), which allows for its use in conjunction with, *e.g.*, the DFTB framework. It is worthwhile to mention, that a QHO has a natural width and thus goes beyond point-like dipoles, [187] while the short-range dipole coupling tensor between QHOs with overlapping densities can still be evaluated analytically. [188] This short-range dipole tensor, $\mathbf{T}_{\mathrm{QHO}}^{(\mathrm{sr})}$, is then used to explicitly account for short-range screening on top of the effects of the local environment according to the inverted coarse-grained Dyson equation, [187,189]

$$\alpha_A^{(\mathrm{sr})}(iu) = \frac{1}{3}\mathrm{Tr}\left\{\sum_C \mathbf{B}_{AC}\right\} ; \quad \mathbf{B} = \left(\mathbf{A}_u^{-1} + \mathbf{T}_{\mathrm{QHO}}^{(\mathrm{sr})}\right)^{-1} , \tag{2.58}$$

where $\mathbf{A}_u = \mathrm{diag}\{\alpha_A^{(\mathrm{TS})}(iu)\cdot\mathbb{1}_3\}$ is a diagonal matrix containing three times $\alpha_A^{(\mathrm{TS})}(iu)$ for each atom A (*i.e.*, the *xx*, *yy*, and *zz* component of the corresponding isotropic atomic polarizability tensor). The summation over all atoms C corresponds to the integration over the whole space in the Dyson equation and the factor $\frac{1}{3}$, together with the trace operator, restores an isotropic effective polarizability. Explicit account for short-range screening already significantly improves the description of the polarizability compared to the superposition of effective atomic polarizabilities [50,147] and then enters a coarse-grained ACFD/RPA formula for the long-range correlation energy of the form of Eq. (2.36). For the long-range coupling there is a negligible overlap between the QHOs. Therefore, the bare dipole potential is applied. So, to a very good approximation, the long-range ACFD/RPA formula for an N atom system can be evaluated based on a set of N dipole-coupled QHOs. Such a set of N three-dimensional QHOs can be described in terms of mass-weighted displacements, $\boldsymbol{\zeta}_A = \sqrt{m_A}\,(\mathbf{r}_A - \mathbf{R}_A)$, and the Hamiltonian,

$$\mathcal{H}_{\mathrm{MBD}}(\boldsymbol{\zeta}) = \sum_{A=1}^{N} -\frac{1}{2}\nabla_{\boldsymbol{\zeta}_A}^2 + \sum_{A=1}^{N}\frac{\eta_A^2}{2}\|\boldsymbol{\zeta}_A\|^2 + \sum_{B=1}^{N}\frac{\eta_A\eta_B}{2}\sqrt{\alpha_{A,0}^{(\mathrm{sr})}\alpha_{B,0}^{(\mathrm{sr})}}\,\boldsymbol{\zeta}_A^\dagger\mathbf{T}_{AB}^{(\mathrm{lr})}\boldsymbol{\zeta}_B = \mathcal{T}_{\boldsymbol{\zeta}} + \frac{1}{2}\boldsymbol{\zeta}^\dagger\boldsymbol{\mathcal{V}}\boldsymbol{\zeta} , \tag{2.59}$$

$$\text{with} \quad \boldsymbol{\mathcal{V}}_{AB}^{(i,j)} = \eta_A\eta_B\left(\delta_{ij} + \sqrt{\alpha_{A,0}^{(\mathrm{sr})}\alpha_{B,0}^{(\mathrm{sr})}}\,\mathbf{T}_{AB,\mathrm{lr}}^{(i,j)}\right) ,$$

where the collective variable $\boldsymbol{\zeta}$ is the direct sum of all $\boldsymbol{\zeta}_A$ and (i,j) denotes the Cartesian components of the AB-sub-blocks of the potential matrix $\boldsymbol{\mathcal{V}}$ and the long-range dipole coupling tensor $\mathbf{T}_{AB}^{(\mathrm{lr})}$. Similar models to describe (many-body) dispersion interactions within the dipole limit were already known and used earlier. [183,190–196] These methods, however, were typically based on simpler model polarizabilities and did not offer general parametrization and applicability for realistic systems.

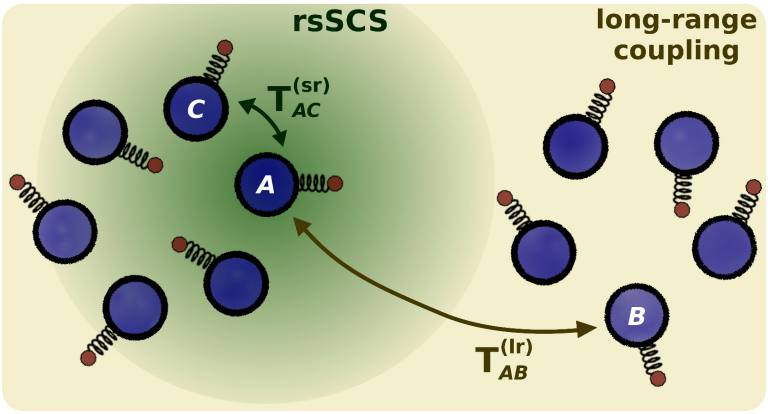

Fig. 2.2: Schematic illustration of the MBD model with range-separated self-consistent screening (rsSCS): Effective atomic polarizabilities are obtained from electrodynamic screening using the short-range part of the range-separated dipole tensor for quantum harmonic oscillators. The interaction between the oscillators is then evaluated using the long-range part of dipole coupling tensor. Reproduced from Ref. 1 with permission from the Royal Society of Chemistry.

As a mathematically equivalent, yet much more efficient, alternative approach, it has been shown that the long-range RPA correlation energy of this dipole-coupled set of QHOs equals its (zero-point) interaction energy.[51,131] Thanks to the bilinear form (2.59), this can be obtained numerically exact *via* unitary transformation to a new collective variable, $\boldsymbol{\xi} = \mathbf{C}\boldsymbol{\zeta}$, where the transformation matrix $\mathbf{C}$ diagonalizes the potential matrix:

$$\mathbf{C}\boldsymbol{\mathcal{V}}\mathbf{C}^{\dagger} = \mathrm{diag}\{\omega_i^2\} \ . \tag{2.60}$$

With the kinetic energy operator, $\mathcal{T}$, being invariant under unitary rotations, $\mathbf{C}$ transforms the MBD Hamiltonian into an uncoupled set of $3N$ one-dimensional QHOs,

$$\mathcal{H}_{\mathrm{MBD}}(\boldsymbol{\xi}) = \mathcal{T}_{\boldsymbol{\xi}} + \frac{1}{2}\boldsymbol{\zeta}^{\dagger}\mathbf{C}^{\dagger}\mathbf{C}\boldsymbol{\mathcal{V}}\mathbf{C}^{\dagger}\mathbf{C}\boldsymbol{\zeta} \ = \ \mathcal{T}_{\boldsymbol{\xi}} + \frac{1}{2}\boldsymbol{\xi}^{\dagger}\,\mathrm{diag}\{\omega_i^2\}\,\boldsymbol{\xi} = \sum_{i=1}^{3N}\mathcal{T}_{\boldsymbol{\xi}_i} + \frac{\omega_i^2}{2}\|\boldsymbol{\xi}_i\|^2 \ . \tag{2.61}$$

This set of QHOs can be solved according to textbook and its total energy is given by half the sum of its characteristic frequencies ω_i. The (zero-point) interaction energy, ultimately corresponding to the RPA long-range correlation energy within the QHO response model, is given by

$$E_{\mathrm{vdW}}^{(\mathrm{MBD})} = E_{\mathrm{QHO}}^{(\mathrm{coupled})} - E_{\mathrm{QHO}}^{(\mathrm{non\text{-}interacting})} = \frac{1}{2}\sum_{i=1}^{3N}\omega_i - \frac{3}{2}\sum_{A=1}^{N}\eta_A \ . \tag{2.62}$$

Finally, the range-separating function to define $\mathbf{T}_{\mathrm{sr}}$ and $\mathbf{T}_{\mathrm{lr}}$, is chosen of Fermi-type,

$$g_{\mathrm{rs}}^{(\mathrm{MBD})}\left(R_{AB}; R_{\mathrm{vdW}}^{(AB)}\right) = \left\{1 + \exp\left[-a\left(\frac{R_{AB}}{\beta \cdot R_{\mathrm{vdW}}^{(AB)}} - 1\right)\right]\right\}^{-1} , \tag{2.63}$$

where $a = 6$ and the effective vdW distance, $R_{\mathrm{vdW}}^{(AB)} = \sqrt[3]{x_A}R_{\mathrm{vdW}}^{(A,\mathrm{ref})} + \sqrt[3]{x_B}R_{\mathrm{vdW}}^{(B,\mathrm{ref})}$, where one relies on a rescaling of accurate reference data of vdW radii (for further details, see Sec. 2.3.3). The range-separation parameter, β, is an empirical parameter fitted to provide optimal results for small molecular dimers.[50] As the range-separation parameter also represents a measure of when the long-range correlation model has to be included, it gives an estimate of the range of correlation already captured by the underlying (semi-)local electronic-structure method.[197] Optimized range-separation parameters for select DFTB parametrizations are given in Appendix A1.

Practical Aspects and Related Models

Being formally equivalent to the full long-range RPA correlation energy for a set of QHOs, the MBD formalism includes many-body interactions up to N atoms and incorporates correlation effects up to infinite order. The two main differences are the assumption of isotropic polarizabilities and that those can be modeled *via* a single QHO per atom. Relying on the QHO model

polarizability, on the other side, fundamentally binds the charge fluctuations to a given atom (no electron hopping), which limits the validity and applicability of MBD for metallic systems (no account for type C non-additivity, see Sec. 2.3.4). For a variety of non-metallic systems in different chemical environments, on the other side, the two approximations have been found to be very reliable and accurate. In fact, in various cases the MBD formalism yields the same results as the RPA-approach, while requiring only a fraction of the computational workload: The MBD method scales as $\mathcal{O}(N^3)$ with the number of atoms N for inversion and diagonalization, see Eqs. (2.58) and (2.60), and has a comparably small prefactor thanks to avoiding integrations at run time and requiring only a few inversions and a single diagonalization. After its original formulation for the vdW energetics of finite-gap molecules, the applicability and efficiency of MBD was further boosted by the derivation of the analytical gradient expression [51,189] and a reciprocal space formulation, [198] which allows for efficient simulations within periodic boundary conditions and advanced the applicability of the MBD formalism to molecular crystals and layered materials.

As mentioned above, for a set of QHOs, the MBD formalism is mathematically equivalent to the full long-range RPA correlation energy. For this to yield accurate energies for realistic systems, the set of QHOs has to accurately model the response properties of the system. For that, the MBD model relies on the procedure originally proposed in the vdW(TS) scheme (*vide supra*), which is based on the rescaling of accurate reference data according to Hirshfeld volume ratios. As a result, MBD can also suffer from the common shortcomings of the Hirshfeld partitioning scheme such as to underestimate charge transfer effects on the atomic polarizability. [144–147] It should be stressed, however, that this is a shortcoming of the underlying (Hirshfeld) partitioning scheme and not the MBD framework itself. As for the vdW(TS) model, iterative Hirshfeld partitioning or using a charge-dependent reference state for atomic polarizabilities can substantially improve the limitations in the parametrization of the QHO model of electronic response. [145,147,148]

Modeling electronic fluctuations and their interactions within the so-called *Drude approximation* (*i.e.*, *via* negatively charged pseudo-particles harmonically oscillating around atomic centers) has already been known and used in the context of vdW interactions by London in the 1930s. [123,199,200] Based on this picture, Whitfield and Martyna [201] proposed a more general approach to model (many-body) induction and dispersion: the quantum Drude oscillator (QDO) model, which also largely motivated the development of the MBD framework. In the QDO model, the oscillating pseudo-particles interact *via* the full Coulomb potential, with that going beyond the typically invoked RPA or dipole approximation. The model is defined by the (effective) charge and mass of the pseudo-particles and the characteristic frequency of their oscillation. With an appropriate choice of these three parameters, the QDO model can accurately describe the response properties, many-body induction and dispersion interactions of a given system up to infinite order. [185] Direct derivation of effective parameters for realistic systems, however, represents a challenging task. Also the evaluation of the interaction energy, which is typically done *via* imaginary-time path integration [201] or diffusion quantum Monte-Carlo, [202] limits its applicability in terms of system size. Recently, this model has been used to showcase the relevance of many-body and multipolar vdW interactions in water and at its surfaces. [186,203]

SCS and MBD as Model for Electronic Response and Polarization Fluctuations

As presented above, introducing the coupled oscillator model allows to coarse-grain and invert the Dyson-like screening equation. This allows to obtain effective polarizabilities, which include the effect of the surrounding field of (finite-width) dipoles. [187] Starting from a set of unscreened, local polarizabilities as represented by $\mathbf{A}_u = \mathrm{diag}\{\alpha_A(iu)\}$, the so-called *relay matrix*,

$$\mathbf{B} = \left(\mathbf{A}_u^{-1} + \mathbf{T}_{\mathrm{QHO}}\right)^{-1} \tag{2.64}$$

introduces non-local terms through the dipolar coupling between the oscillators. Note that for a single atom or a set of non-interacting oscillators, $\mathbf{B} = \mathbf{A}_u$. In the framework of MBD@rsSCS,

one only applies the short-range part of $\mathbf{T}_{\mathrm{QHO}}$ in the above equation, because the RPA formula and thus the MBD model implicitly include long-range screening effects as shown above or more explicitly and in further detail in Ref. 51. For obtaining polarizabilities with full-range account of screening effects, however, one has to use the full dipole–dipole tensor in the above equation. The individual elements of the relay matrix, finally, describe the non-local *intra*-system polarizability,

$$\boldsymbol{\alpha}\left(\mathbf{r},\mathbf{r}',iu\right) = \sum_{A,C} \delta^3(\mathbf{r}-\mathbf{R}_A)\,\delta^3\left(\mathbf{r}'-\mathbf{R}_C\right)\,\mathbf{B}_{AC}\ . \tag{2.65}$$

This procedure does not provide polarizabilities as accurate as, *e.g.*, time-dependent DFT calculations, but represents a reliable approximation [114] and thanks to its simplicity allows to study the electronic response properties in complex systems as considered in this thesis. Summing over rows or columns of $\mathbf{B}$ then provides local (atomic) polarizabilities as used for the definition of the MBD Hamiltonian above. It should be noted that using a diagonal matrix $\mathbf{A}_u$ implies that there are no polarizability anisotropies introduced in the preceding parametrization of the QHOs as assumed in the ordinary vdW(TS) procedure. Nonetheless, the above formalism and equations are equally valid with other definitions of $\mathbf{A}_u$ and thus also allow to introduce anisotropy already at the level of the local chemical environment or the chosen parametrization procedure. Through the (anisotropic) dipolar coupling, Eq. (2.64) finally introduces anisotropy into the polarizability both on an atomic and molecular level. Early on, this has already been shown to considerably improve the polarizability anisotropy in N_2, for example. [50] The SCS formalism can also be used in conjunction with the vdW(TS) approach. In order to account for electrodynamic screening in addition to the Hirshfeld-rescaling-procedure, the pairwise C_6-interaction coefficients entering Eq. (2.48) can be calculated from atomic polarizabilities as obtained from the relay matrix, *i.e.*,

$$C_{6,AB}^{(\mathrm{TS@SCS})} = \frac{3}{\pi}\int_0^\infty \alpha_A^{(\mathrm{SCS})}\,\alpha_B^{(\mathrm{SCS})}\,du \qquad \text{with} \qquad \alpha_A^{(\mathrm{SCS})} = \frac{1}{3}\mathrm{Tr}\left\{\sum_C \mathbf{B}_{AC}\right\}\ . \tag{2.66}$$

Such an approach can be used to dissect the role of screening and multi-center interactions for vdW dispersion (further sub-classifying type B non-additivity as described in Sec. 2.3.4).

The MBD formalism then builds upon this QHO model of electronic response. The vdW (interaction) energy in MBD is defined as the difference between the ground-state energy of the dipole-coupled (DC) state and the uncoupled set of oscillators. The vdW energy can thus be seen as the (change in the) "zero-point vibrational energy" of the electron density, which resembles the intrinsic electronic quantum fluctuations the ACFD theorem is based on. In the same spirit, the solution of the DC set of oscillators resembles "zero-point vibrations" in the electron density and thus a model for the intrinsic electronic behavior within dipolar coupling. In its transformed form (2.61), the Hamiltonian for the DC state,

$$\mathcal{H}_{\mathrm{DC}}(\boldsymbol{\xi}) \equiv \mathcal{H}_{\mathrm{MBD}}(\boldsymbol{\xi}) = \sum_i \mathcal{T}_{\xi_i} + \frac{1}{2}\omega_i^2\,\|\xi_i\|^2\ , \tag{2.67}$$

represents a simple uncoupled set of one-dimensional harmonic oscillators. Following the above remarks, the collective coordinates ξ_i, provide an eigenmode representation of the intrinsic electronic behavior and have been used to characterize the vdW bonding of nanostructures and π–π-stacked or supramolecular complexes. [35,172] Within the coupled oscillator model of electronic response, the eigenfrequencies obtained from the eigenmode transformation represent dissipation channels for incident (oscillating) electric fields. As such, there is a strong relation between the eigenfrequency and the electronic (dipole) polarizability of a QHO as given by

$$\alpha(u) = \frac{q^2}{m\left(\omega^2-u^2\right)}\ , \tag{2.68}$$

where q and m represent the effective charge and mass of the oscillator. [36] This also allows to define mode-specific polarizabilities, α_i, providing detailed insights into electrodynamic

response properties. For a system of identical oscillators, this further gives rise to $\alpha_i(0)\cdot\omega_i^2 = \text{const.}$ due to charge conservation.[172] With the simple form (2.67), it is also trivial to obtain the wave function of the DC ground-state according to textbook as the product state of one-dimensional QHOs along eigenmode coordinates,

$$\Psi_{\text{DC}}(\boldsymbol{\xi}) \equiv \Psi_{\text{MBD}}(\boldsymbol{\xi}) = \prod_{i=1}^{3N} \left(\frac{\omega_i}{\pi}\right)^{1/4} e^{-\frac{1}{2}\omega_i \xi_i^2} \ . \tag{2.69}$$

This can be used to obtain general observables, O, for the DC state as the expectation value of the corresponding operator,

$$\langle O \rangle_{\text{DC}} = \langle \Psi_{\text{DC}} | \hat{O} | \Psi_{\text{DC}} \rangle \ . \tag{2.70}$$

As an example, one can obtain the density of (Drude) pseudo-particles (*i.e.*, the density of electronic or polarization fluctuations) and its change upon interaction. The latter thereby resembles charge-polarization due to vdW dispersion and has been used to characterize the interaction in supramolecular guest-host complexes.[35] In the following chapters, we will use the coupled oscillator SCS approach and MBD formalism as a model for intrinsic electronic fluctuations as well as the eigenmode and wave function representation of the DC state to further enhance our understanding and description of vdW forces and electronic behaviors in complex systems.

Chapter 3

Van der Waals Interactions and Electronic Behaviors in Solvents

The molecular dynamics simulations, on which parts of the results in this chapter are based on, were performed by Dr A. Hakami-Zanjani and the group of Prof S. Varma.

A system's properties strongly depend on its environment — be it the thermodynamic conditions or the phase in which the system resides. To accurately model and understand realistic systems, it is imperative to account for these parameters during simulation. A very prevailing example ubiquitous in chemistry, materials science and especially biology is a liquid or solvated system. Yet, this setting arguably represents the most intricate case for simulation due to the mostly ill-defined system boundaries and because of the various types of interaction that need to be considered. In particular, non-covalent intermolecular interactions play an important role for the electronic, thermodynamic, structural and response properties of a solute–solvent system and its dynamics. For this purpose, various explicit and implicit (solvation) models have been put forward, which usually perform well in describing the electrostatic and polarization, or induction, effects of the solvent. This, however, usually holds true only for certain conditions (typically designed for the liquid phase at room temperature and ambient pressure), which shows that the physical description is incomplete. The role of vdW dispersion interactions with and within a liquid phase has so far not been fully explored and understood on a fundamental level.

As detailed in Sec. 2.3, vdW interactions arise from long-range electron correlation, which makes them inherently quantum-mechanical and many-body in nature. This fact strongly limits an in-depth investigation of dispersion forces in solvation: Due to the solvent such systems involve several thousands of atoms, but the majority of sophisticated electronic-structure methods, which would accurately capture (long-range) electron correlation, is limited to a few hundred atoms due to the associated computational costs. As an additional complication, vdW interactions show a highly complex scaling with the system size, which renders an extrapolation based on small- to medium-size systems insufficient. Typically, vdW forces are modeled *via* atom-pairwise potentials of the well-known London-type form, which is based on a second-order truncation of

the many-body expansion for vdW interactions. In molecular mechanics (MM) force fields, which as of today represent the prevalent technique for modeling (large-scale) solvated systems, the vdW component is usually combined with an effective treatment of exchange (Pauli) repulsion into a phenomenological description in the form of pairwise Lennard-Jones potentials. The associated parameters, in the end, are usually obtained in a global fit procedure to reproduce a select variety of observables. As a result there is only very limited direct access or physical rational behind the parameters defining the effective potentials. The lack of treating the various (intermolecular) interactions at their full complexity also poses a major reason behind the typically strongly limited transferability and validity of MM potentials. Despite the tremendous success of these simplifications, our current description of vdW dispersion interactions in solvated systems is far from the quantum-mechanical many-body character of long-range electron correlation and neglects essential parts of the underlying physics. Thus, it is worthwhile investigating the vdW physics of molecular solvents such as water on a more fundamental level before considering (solvated) biomolecular systems.

3.1 van der Waals Energetics of Water and Cavity Formation

As a first step in this regard, one can consider the solvation of biomolecules in water on a more qualitative level. In protein folding, for instance, a biomolecule transitions from a more extended, disordered state to a more globular shape. Furthermore, hydrophilic residues get exposed to the surrounding water, while hydrophobic residues get buried inside the protein core, which optimizes the (short-range) interaction with the aqueous environment and minimizes the disruption of the dynamic hydrogen bond network of the surrounding water.[204–206] So, as a first model system to study vdW interactions during protein folding from the perspective of the solvent, we consider simple, prismatic or spherical "placeholders" described by fixed, coarse-grained methyl units. In order to investigate the effect of surface hydrophobicity, a given portion of methyl units is replaced by methanol units. Classical molecular dynamics (MD) simulations of the placeholders submerged in water are used to obtain ensembles to analyze the effects of thermal fluctuations. The solvent's vdW energy of the MD snapshots is calculated using the atom-pairwise vdW(TS) approach and using a many-body treatment as given by the MBD formalism. This allows to investigate beyond-pairwise effects on the vdW energetics of water for the qualitative model of unfolded versus folded proteins. Effective polarizabilities as used to parameterize the vdW(TS) and MBD models are obtained from DFTB calculations *via* Eq. (2.49).

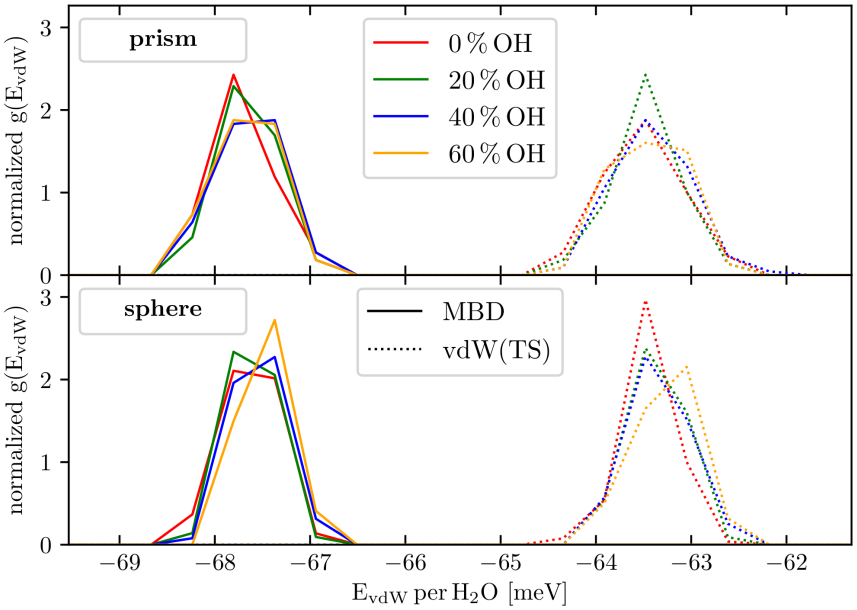

Fig. 3.1: Distribution of van der Waals energy during thermal sampling of bulk water with prismatic or spherical cavity and different interface hydrophobicity. Distribution of MBD energies shown as full lines, vdW(TS) energies as dotted lines. *top:* prismatic cavity with varying interface hydroxylation. *bottom:* spherical cavity.

Fig. 3.1 shows the (normalized) distributions of the vdW energy of water per solvent molecule obtained from the MD simulations of the individual placeholders. As can be seen from the distributions for the prismatic (Fig. 3.1 top) and spherical placeholder (Fig. 3.1 bottom), neither the shape of the resulting cavity, nor the degree of surface hydroxylation (OH percentage) substantially alter the thermal sampling of vdW energies. This holds true for both, the MBD method (solid lines) as well as for the vdW(TS) approach (dotted lines). It is worth pointing out that the vdW(TS) model yields consistently lower vdW contributions per water molecule than the MBD formalism. Fig. 3.2 then shows the distribution of beyond-pairwise vdW energies per solvent molecules as given by the difference between the many-body and pairwise treatments. It shows that beyond-pairwise effects lead to an increase in the vdW energy by about 4.25 meV per H_2O ($\hat{=}$ 0.1 kcal/mol). This difference, however, is consistent among all considered cases. With the exception of a spherical cavity (Fig. 3.2 bottom) with 0 or 40 % surface hydroxylation, the distribution of beyond-pairwise contributions shows a narrow spread. The two exceptions are thereby likely only due to an insufficient thermal sampling and expected to converge to the narrow distributions of the remaining cases with increased numbers of samples.

Given the close resemblance between our results for spherical and prismatic cavities in the distribution of the vdW energy as well as of beyond-pairwise effects, the formation of differently-shaped cavities does not affect the total vdW energetics of water. As the similarity in the distributions applies equally-well throughout different percentages of interfacial hydroxy groups, also the surface hydrophobicity does not give rise to different solvation states from the perspective of the total or beyond-pairwise vdW energy. This allows to conclude a considerable robustness of the *intra*-water vdW interactions with respect to thermal fluctuations. This can be explained by the finite width of the polarizability, which renders long-range dispersion forces less sensitive to the precise atomistic structure and thus thermal fluctuations in such dense systems. In addition, the thermal sampling was obtained from MD simulations using classical MM force fields. The systems studied here thus lack long-range structural ordering, which introduces a consistent stochastic structure at the spatial ranges relevant to long-range vdW dispersion. This lack of structural long-range order is supported by various nuclear magnetic resonance studies.[206,207] It is important to note, however, that those experiments as well as the vdW *energy* merely probe local structural characteristics and do not preclude dynamical long-range ordering or collective nuclear dynamics. In this regard, recent terahertz experiments indicate considerable long-range effects on the dynamics and polarization of water.[208–211] Obviously, this requires a long-range interaction mechanism, which is not present in the traditional view of water being purely governed by short-range interactions.

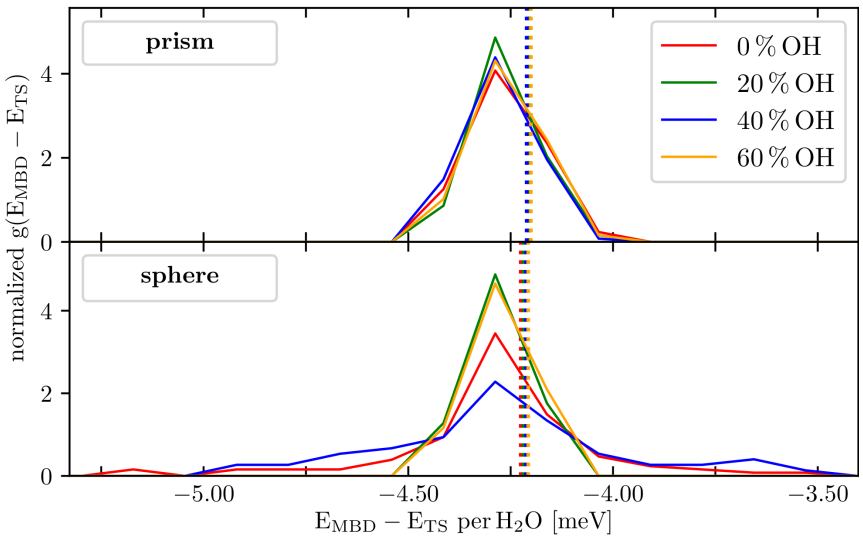

Fig. 3.2: Distribution of beyond-pairwise contributions to van der Waals energy of bulk water with different cavities and interface hydrophobicity. *top:* prismatic cavity with varying interface hydroxylation. *bottom:* spherical cavity. Mean values of distribution shown as dotted lines.

3.2 Electronic Behaviors in Molecular Liquids

In a next step, we analyze the characteristics of intrinsic electronic behaviors as described within the MBD framework. For comparison, we extend the study to a mixed-polar (octanol) and a non-polar solvent (*n*-hexane). The structures of the pristine liquids have been obtained as snapshots from classical MD simulations as above. The first characteristic is the spectrum of effective frequencies, *i.e.*, the density of states (DOS) of eigenfrequencies, ω_i, as obtained within the MBD formalism. As introduced in Sec. 2.3.5, the individual ω_i can be interpreted as the excitation frequencies of collective "zero-point vibrations" of the electron density. The eigenfrequencies therefore represent absorption or dissipation channels for incident energy within the QHO model of electronic response and are closely related to optical and (vacuum) ultraviolet absorption — comparable to plasmon absorption in metallic systems. However, a clear comparison to absorption spectra is heavily limited due to superimposed presence of several spectral features of the solvent in the low-petahertz region, such as the onset of the Rydberg series and individual electronic excitation and ionization processes. The eigenspectrum further has a close connection to the polarizability *via* Eq. (2.68). Note that such insights to electronic response properties are completely absent in phenomenological, atom-pairwise vdW models.

Fig. 3.3 shows the DOS for bulk water (top), octanol (middle) and *n*-hexane (bottom). The spectra of the corresponding monomers are shown in black. It can be seen that the DOS of bulk water is substantially changed in comparison to the monomer spectrum. This is not only due to the (restricted) sampling of different molecular conformations, but especially due to screening effects and the emergence of complex collective electronic fluctuations, which span over several molecules and are ultimately responsible for intermolecular vdW interactions. One example that appears very prominently in the spectrum of electronic fluctuations is hydrogen bond formation as illustrated for water in the top graph of Fig. 3.3: Upon condensation, the high-frequency peak of the monomer gets renormalized into a bimodal distribution, producing considerable spectral weight at lower frequencies. In the atomistic framework of MBD, the high-frequency gas-phase peak can be mapped to the oxygen atom, while the emergent lower-frequency signal

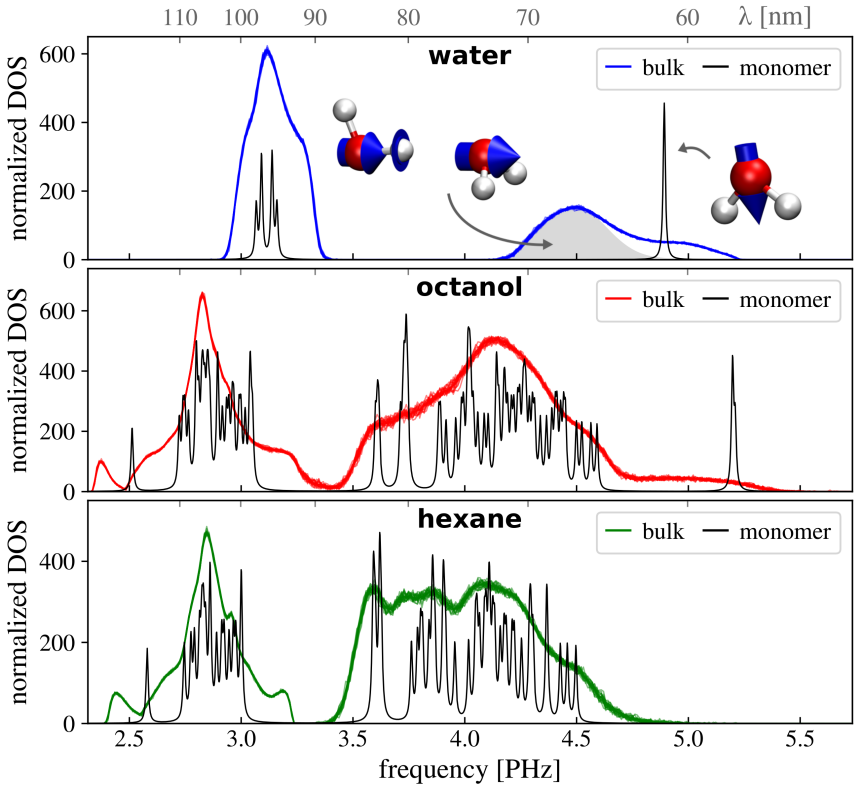

Fig. 3.3: Density of states (DOS) of collective charge fluctuations in bulk solvents and corresponding monomers. *top:* MBD DOS for bulk water (blue). Renormalization of high-frequency peak due to hydrogen bond formation and O–O coupling upon condensation highlighted in grey. *middle:* octanol (red) and *bottom: n*-hexane (green). Respective monomer DOS shown in black.

(colored in grey in the top plot of Fig. 3.3) can be assigned to collective fluctuations linking several oxygen atoms *via* hydrogen bonds. Comparing this behavior to octanol and *n*-hexane, we see a similar but much less pronounced effect upon hydrogen bond formation in the alcohol, while hydrogen-bond formation and the corresponding feature in the spectrum of eigenfrequencies is absent in the pure hydrocarbon *n*-hexane. For octanol and *n*-hexane, one further observes the lowest-frequency signal at larger wavelengths than in the case of water as well as more pronounced satellite peaks in the lower-frequency band. This can be attributed to H–C coupling within the monomer and between molecules in the condensed phase.

Next, we consider the change introduced in the DOS upon embedding different cavities into water, given by the difference of the DOS of pristine water and the systems studied in Sec. 3.1 above. Fig. 3.4a shows the range covered by plus and minus one standard deviation of the eigenspectrum of bulk water in blue. In addition, the difference between the average spectrum of bulk water and the DOS of water with a prismatic or spherical cavity is shown in red and green, respectively. In both cases, this difference (ΔDOS) shows a transition from positive to negative values in the low-frequency band and from negative to positive in the higher-frequency region. This represents a considerable red-shift at lower frequencies and a blue-shift at higher frequencies. Given that the MBD eigenfrequencies represent possible absorption channels, the same shifts apply to the dynamic polarizability, $\alpha(u) = q^2/m(\omega^2 - u^2)$. This change in the eigenspectrum can be attributed to the introduction of a sizable vacuum inside the cavity. It should be noted that the observed shifts are characterized by a high sensitivity to the water density and water model used in the MD simulations. The shape of the cavity, on the other side, turned out to have a negligible effect on the eigenspectrum, *cf.* agreement of ΔDOS for the prismatic and spherical cavity in the bottom right inset of Fig. 3.4a. The same applies to the interface polarity. As shown in Fig. 3.4b, the differences between different degrees of hydroxylation and a non-polar interface (0 % OH) are well within the corresponding standard deviation. All (differential) spectra shown in Figs. 3.3 and 3.4 show the superposition of 100 snapshots of thermal sampling at room temperature. The almost complete lack of deviations reflects the robustness of the frequency fingerprints of intrinsic electronic fluctuations with respect to thermal fluctuations in the atomistic structure.

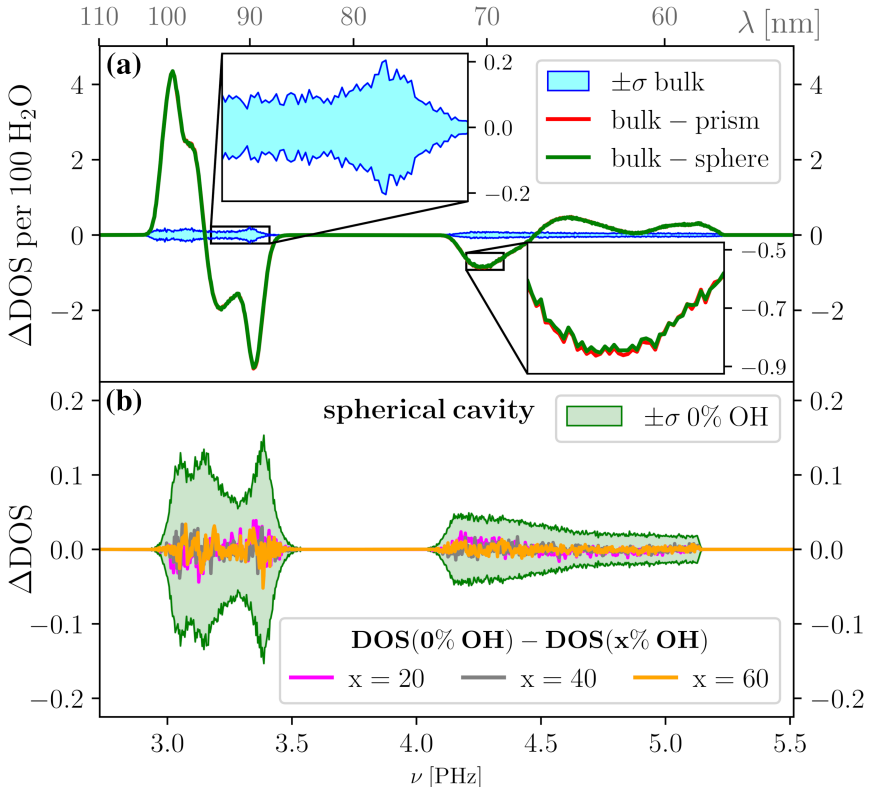

Fig. 3.4: Change in density of states (ΔDOS) of collective electronic fluctuations in bulk water upon cavity formation. (a) Range covered by one standard deviation of bulk water DOS (blue) obtained from thermal sampling and DOS upon formation of prismatic (red) or spherical (green) cavity. (b) Range covered by one standard deviation of water with a spherical cavity corresponding to non-hydroxylated interface (green) and difference of DOS for cavities corresponding to hydroxylated interfaces (magenta: 20 % OH, grey: 40 % OH, orange: 60 % OH).

3.3 Polarization Response in Solvents and Solvated Proteins

To further investigate electronic behaviors in liquids and molecular solvents, we study the non-local (*intra*-system) polarizability. In Sec. 2.3.5, we have introduced the SCS formalism as an efficient approach to obtain a reliable approximation to the non-local polarizability *via* the relay matrix,

$$\boldsymbol{\alpha}(\mathbf{r},\mathbf{r}',iu) = \sum_{A,C} \delta^3(\mathbf{r}-\mathbf{R}_A)\,\delta^3(\mathbf{r}'-\mathbf{R}_C)\;\mathbf{B}_{AC} = \sum_{A,C} \delta^3(\mathbf{r}-\mathbf{R}_A)\,\delta^3(\mathbf{r}'-\mathbf{R}_C)\left(\left(\mathbf{A}_u^{-1}+\mathbf{T}_{\mathrm{QHO}}\right)^{-1}\right)_{AC}. \quad (3.1)$$

Ref. 114 presents the static non-local polarizability ($iu = 0$) for small organic molecules, which shows an approximately-exponential decay with the distance between $\mathbf{r}$ and $\mathbf{r}'$. To study the scaling behavior of the intermolecular polarizability in the considered solvents, we here first coarse-grain the relay matrix as obtained from Eq. (3.1) to molecular polarizability centers located at the centers of mass, $\overline{\mathbf{R}}$, of molecules $\mathcal{I}$ and $\mathcal{J}$ according to

$$\alpha(\mathbf{r},\mathbf{r}',iu) = \sum_{\mathcal{I},\mathcal{J}} \delta^3(\mathbf{r}'-\overline{\mathbf{R}}_{\mathcal{J}})\,\delta^3(\mathbf{r}-\overline{\mathbf{R}}_{\mathcal{I}})\;\frac{1}{3}\mathrm{Tr}\left\{\sum_{A\in\mathcal{I}}\sum_{B\in\mathcal{J}}\mathbf{B}_{AB}\right\}, \quad (3.2)$$

where we further simplify to isotropic polarizabilities. Finally, we integrate over frequencies to account for the dissipation at all frequencies. According to the fluctuation–dissipation theorem, this is proportional to the correlation between polarization fluctuations.[115,212] Dividing by the Hartree energy, E_h, restores the units of (non-local) polarizability and allows to interpreted the result as *polarization response*. Fig. 3.5 shows the absolute value of the integrated non-local polarizability for water (a) and hexane (b) as a function of intermolecular distances. In both cases, there is no unique relation between the polarization response and the distance $d = \|\mathbf{r}-\mathbf{r}'\|$. This means, the coupling of molecular polarization can vary significantly for a given distance. This can be explained with the broad sampling of (relative) molecular orientations, local screening effects and the response through a heterogeneous medium. In the very long-distance limit, the polarization coupling can be represented by an effectively point-to-point-like response through an average, approximately homogeneous medium, which causes a transition to a very narrow spread. Point-to-point-like response through a homogeneous medium is a standard approach to describe solvents in implicit models. Yet, the present results suggest that such a simplification is only valid on length scales larger than ~60 Bohr (30 Å) in water and ~70 Bohr (37 Å) in hexane.

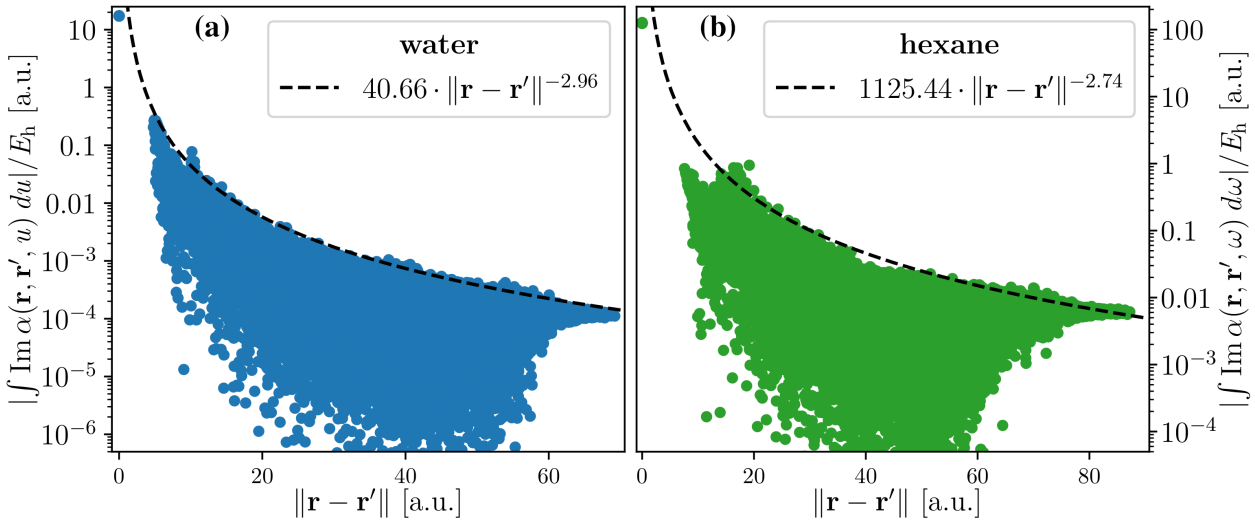

Fig. 3.5: Non-local polarization response in select molecular liquids as a function of intermolecular distance. Results shown for bulk water (a) and hexane (b). Upper bound shown as dashed line.

Despite the absence of a unique relation, there appear clear limits for the maximum polarization response. As represented by the dashed lines in Fig. 3.5, the upper bound for water decays with d^{-3}, which corresponds to the well-known trend of (dipolar) pairwise polarization. For *n*-hexane a slightly slower decay with $\sim d^{-2.7}$ is observed. The prefactor of the scaling law also incorporates the necessary conversion factors and has units of polarizability per length$^{6-\gamma}$, where $-\gamma$ is the exponent of the scaling law. For a d^{-3}-decay, this yields units of polarizability per volume.

Let us now turn to the response of a solvated biomolecule. To this end, we study the 35-residue Fip35 Hpin1 WW domain (Fip35-WW). The structure of Fip35-WW in explicit water was taken from a previous protein folding study by Shaw and co-workers.[14] To understand the different aspects of electronic polarizability in such a multi-component system, we study the polarization response within the aqueous environment and the protein as well as between the protein and the surrounding water molecules individually. In the case of the aqueous environment, we observe the same behavior as for pristine water. The polarization response among water molecules is not notably altered by the presence of the protein. The *intra*-protein response is studied in atomistic detail and on the level of residues coarse-grained according to Eq. (3.2). When considered on an atomistic level, as depicted in Fig. 3.6a), the *intra*-protein polarization response shows a similar behavior as observed for water above (see Fig. 3.5). In contrast to water, however, the large spread at intermediate distances is due to the fact that each polarizability center can represent a different atom and screening effects of the heterogeneous environment. As a result of the former, we also do not observe a collapse onto a narrow spread at large separations as it was the case for water. Coarse-graining the *intra*-protein response to a residue-based model then naturally leads to an overall increase of the polarization response as the non-local polarizability scales with the size of the considered moieties. In addition, the decay of the upper bound is slightly more long-range with a $\sim d^{-2.8}$-trend similar to the one observed for *n*-hexane. This phenomenon can be explained by collective features emerging from combining individual atomic non-local polarizabilities in the integration over the residue volume. This also explains the decay found for *n*-hexane, where the polarizability is integrated over a molecular volume that is similar to the one occupied by the individual residues.

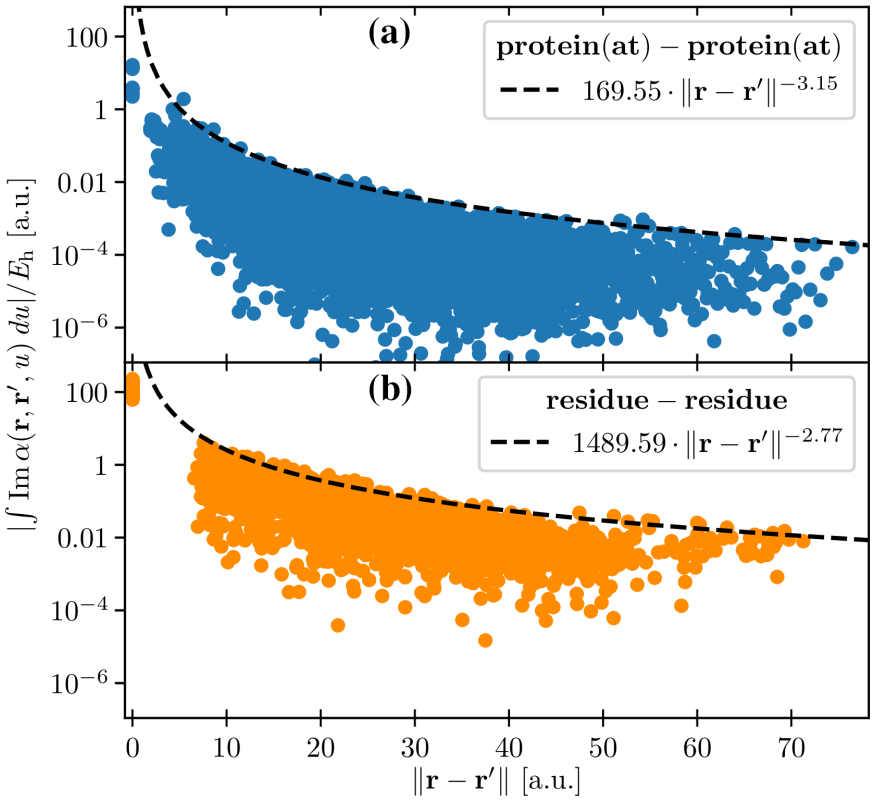

Fig. 3.6: Polarization response in the protein subspace of solvated Fip35 Hpin1 WW domain. (a) Response on atomistic level as a function of interatomic distance (blue). (b) Non-local polarization coupling among protein residues as a function of their separation (orange). Polarizability given on logarithmic scale and in atomic units. Upper limit to polarization response shown as dashed line including corresponding power-law decay.

The above behaviors of the *intra*-protein response as a function of the distance, thereby, apply to unfolded as well as folded states. The polarization coupling between individual regions of the protein, on the other side, does change significantly during the folding process. This can be explained by the change in conformation and relative orientation of the residues. Fig. 3.7 compares the (integrated) polarizability matrix for an unfolded and folded state of Fip35-WW on a per-residue-level. Light colors show weak coupling and dark regions a more pronounced response between the residues. For the disordered state shown on the left of Fig. 3.7, the relay matrix consists of two fairly-decoupled parts at both termini. This resembles the conformation of the corresponding structure with two compact moieties as shown in the inset. For the folded state depicted on the right of Fig. 3.7, a much more structured pattern emerges in the polarization coupling. During the folding process, the WW domain forms a the triple-β-sheet among the residue groups Pro6–Arg14, Gly16–Ile24 and Asn26–Arg32. The corresponding submatrices of the non-local polarization response are marked in blue in Fig. 3.7. As can be seen from this analysis, the distinct features of increased polarization coupling arise in exactly these areas and can thus be assigned to the formation of the triple-β-sheet during folding. Interestingly, there is also a sizable response between the first and third sheet. The last pronounced difference between the relay matrices of the unfolded and folded state is the considerable coupling between the two termini (marked in green). This can be explained by their close proximity in the native conformation compared to the disordered state. So, while the decay of the *intra*-protein polarization response does not change notably during folding, the coupling between individual residues/moieties does change substantially and the emergence of secondary structure elements has a marked effect on the non-local polarization response. With ordering the polarizability matrix according to sequence, substructures orthogonal to the diagonal can be assigned to the coupling between anti-parallel strains like neighboring β-sheets or hairpins, while features parallel to the diagonal represent coupling between parallel strains (*e.g.*, non-neighboring β-sheets or helices).

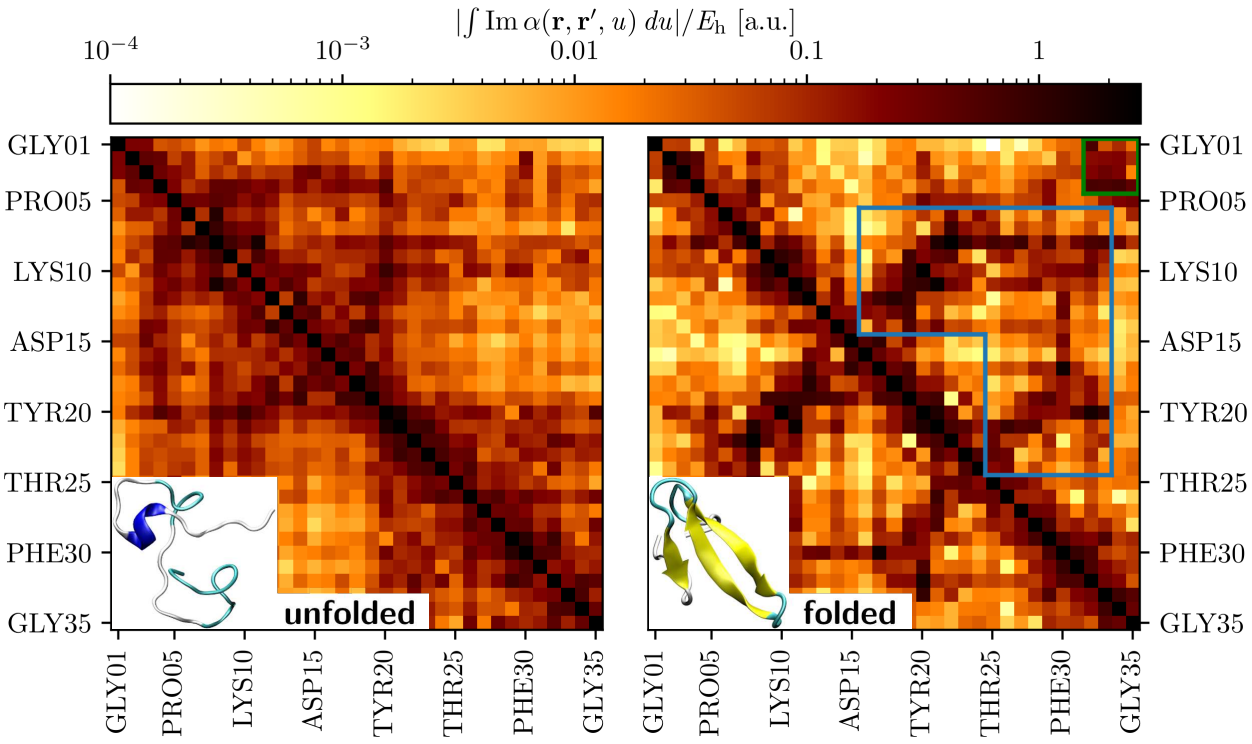

Fig. 3.7: Non-local polarization response within the Fip35 Hpin1 WW domain in an unfolded (left) versus folded conformation (right). During the folding process, a clear pattern of strong polarization coupling emerges. This can be assigned to the formation of the triple-β-sheet (submatrix of the corresponding residues marked in blue) and the proximity of the two termini in the folded state (top right corner, marked in green).

For the polarization coupling between the protein and the surrounding water molecules, finally, we consider three levels of coarse-graining: atomistic resolution of the protein, coarse-graining to individual residues and the water molecules' response to the *full* protein represented by a single polarizability center. The aqueous environment is represented by one polarizability center per water molecule as used for pristine water above. As such, the response is affected by the different atoms or residues on the protein side, screening effects of the heterogeneous medium as well as the molecular orientation of the individual water molecules, which leads to a large spread in the intermediate region. Fig. 3.8 shows the absolute value of the non-local polarization response for the three approaches as a function of distance. For the protein–water response within atomistic treatment of the protein, we again observe the conventional decay with the third power of the distance. As a result of the different polarizability centers on the protein side, the response does not converge to a point-to-point-like response at large distances. When coarse-graining the protein to individual residues, we find a slightly slower decay of the upper limit following $d^{-2.7}$ (*cf.* response among residues or *n*-hexane molecules above). The coarse-graining procedure furthermore involves a slight alleviation of the differences between the polarizability centers on the protein side (individual residues are more alike than individual atoms in terms of their response properties), which leads to a much decreased spread in the large-distance limit. Turning to the polarization response between the full protein and individual water molecules shown in Fig. 3.8c), finally, we observe a substantial divergence from the conventional behavior. When considering the polarization coupling between water and the full protein, the long-range response is increased more strongly than the short-range regime. This leads to a much slower

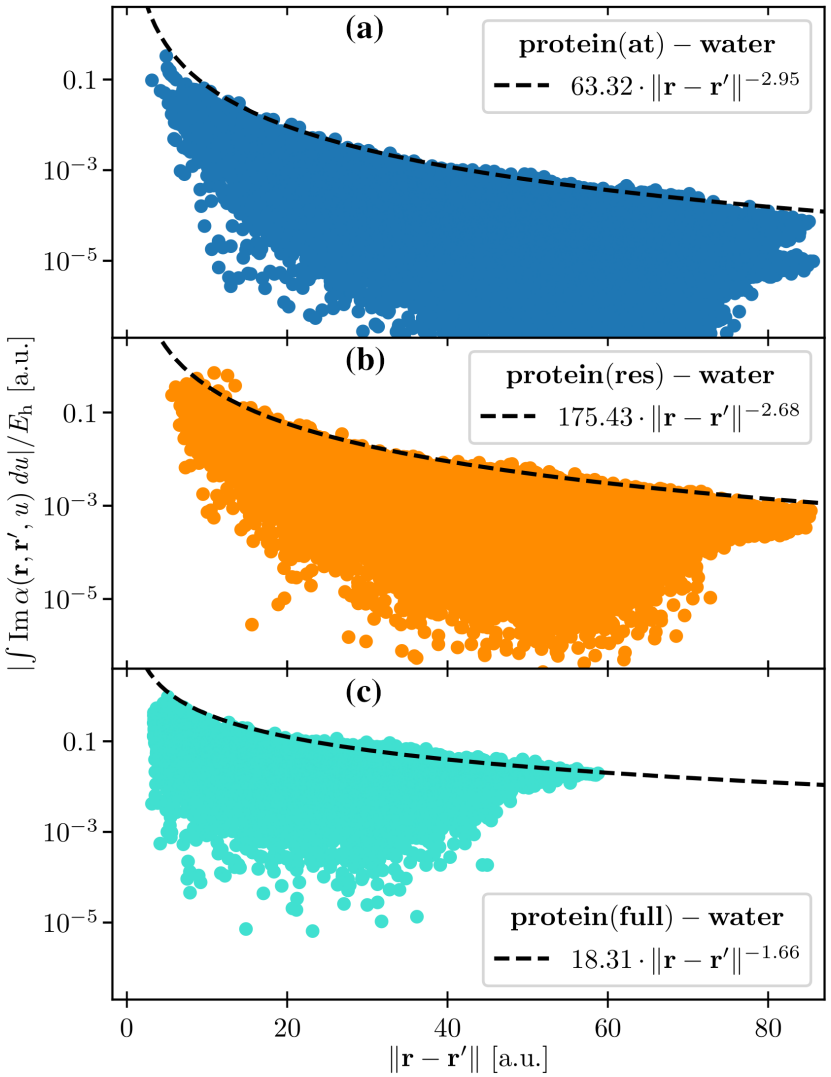

Fig. 3.8: Polarization response between Fip35 Hpin1 WW domain and surrounding water molecules. (a) Polarization response of coarse-grained aqueous environment to atomically resolved protein (blue). (b) Coarse-grained representation of molecule on a per-residue-level (orange). (c) Full contraction of protein to a single polarizability center (turquoise). Polarizability given on logarithmic scale and in atomic units. Upper limit to polarization response shown as dashed line including corresponding power-law decay.

decay of the upper bound following a power law of $\sim d^{-1.7}$ with a considerable decrease of the prefactor in order to also incorporate the weaker effect on the shorter-range response. Integrating over the response to the full protein, finally, also lifts the strong heterogeneity of the protein as a response partner. As a result, we observe a clear collapse to a narrow spread at a distance of ~48 Bohr (25 Å) to the protein–water interface. In this regard it should be noted that the distance of a water molecule to the protein surface is inherently shorter than the one to individual protein atoms or residues. The above behaviors have thereby been found for folded as well as unfolded protein states.

3.4 Discussion of vdW Energetics and Response in Solvated Systems

Overall, the vdW energetics and response within solvents appear to be well described by pairwise models and conventional scaling laws on an atomistic level. This includes the effects introduced by cavity formation and interface hydrophobicity. For a final conclusion and confirmation, it should be noted that the results presented in this chapter are based on classical MD simulations using conventional MM force fields. As such the thermal sampling corresponds to the low-energy states when using a pairwise model, which can differ from the thermally-accessible states within many-body treatment of dispersion forces. Recent work in this regard has shown that structural relaxation of conformations with a more complete treatment can bring out pronounced many-body effects, when departing from the manifold of equilibrium/low-energy structures predicted by pairwise vdW schemes.[213] This can be interpreted in terms of a "generalized variational principle": Many-body effects can provide additional stabilization to the system and relaxation on the corresponding energy landscape maximizes these effects in order to achieve a minimum energy. As a thermal ensemble is governed by Boltzmann population of available states, a similar interplay of nuclear conformations and the effects of correlated electronic motion can appear for the thermal sampling considered here.

In contrast to the bare *intra*-solvent vdW physics, the analysis of the *intra*-protein response and protein–water coupling indicates emergent behaviors when considering larger moieties or the spatial/conformational aspects of biomolecular processes. The much slower decay for the non-local polarizability between the protein and water is thereby in line with the experimental finding of long-range polarization response of water[208–211] and can likely be attributed to a general collectivity in the non-local polarizability. The emergent many-body effects accumulate with more rigorous coarse-graining and finally lead to the slow decay when representing the response of the full protein by a single polarizability center. This much slower decay can eventually even surpass electrostatic coupling, which is strongly limited by the pronounced electro*static* screening in water. On the contrary, the polarization response discussed in Sec. 3.3 is concerned with the correlation of polarization fluctuations, which are subject to electro*dynamic* screening as accounted for by Eq. (3.1). This corroborates the open question about the dominant form of longer-range interaction in aqueous solvation. The considerable changes in the scaling laws upon coarse-graining also bear an important fact for the construction of coarse-grained models or our general understanding on the more macroscopic scale: *In complex systems and environments, the conventional characteristics of response (and interaction) appear to still be valid on an atomistic scale. Yet, on a larger or coarse-grained scale, non-additivity and the emergence of many-body effects can lead to much different characteristics and complex behaviors.* This is well-known as a major challenge in the construction of coarse-grained descriptions of the free energy landscape.[214,215] Based on the present results, we can conclude the same for the polarization response in condensed systems (and the derivative electrodynamical coupling and interactions). Note that this not only applies to the much altered long-range power-law decay, but also to the fundamental validity of simple, effective models. As can be seen from the analysis of the polarization coupling as a function of distance, large parts of the systems are characterized by a wide range of possible results without unique relation. From Fig. 3.8c), for example, it

can be expected that the effect of the full protein on a second solute cannot be described as point-to-point-like response through a homogeneous medium unless the inter-solute separation is beyond 25 Å or more depending on the size and polarizability of the second solute. Below these distances, the response properties and (electrodynamic) interaction strongly depends on the heterogeneous environment, which renders a simple effective-medium model insufficient. It is important to point out that the electrodynamic coupling not only plays a central role for vdW dispersion, but also represents a key factor for spectroscopic response. As such, the emergence of collective effects and alteration of conventional trends also affects electronic and optical response properties. This can play an important role for the construction of coarse-grained oscillator models and (exciton) Hamiltonians for the description of spectroscopic properties. Such models are frequently used in the context of circular dichroism or electronic absorption in complex systems, for instance. [216–220]

Apart from the construction of coarse-grained models, the above results can also have important implications for biomolecular processes. In the majority of cases, biophysics is not concerned with the interaction or response between individual atoms or small molecules, but rather involves larger-scale phenomena such as structural transitions and the cumulative response between larger moieties, which can be substantially affected by the emergent behaviors presented above. This can be seen from the characteristic effects of structure formation on the *intra*-protein coupling (Fig. 3.6) as well as the much slower decay of the polarization response of water molecules to a full protein and, *vice versa*, the response and coupling of protein residues to the entirety of the surrounding aqueous environment instead of to individual water molecules. To highlight the characteristics and emergent effects on *intra*-solute and solute–solvent vdW physics during such a larger-scale biomolecular process, the following chapter provides on a more detailed study of the *intra*-protein and protein–water vdW interaction during protein folding including spatial analysis of the collective electronic behaviors.

Chapter 4

The Collectivity of van der Waals Forces in Biomolecular Systems

Parts of this chapter have been published in this or similar form in
M. Stöhr and A. Tkatchenko, *Sci. Adv.* **5**, eaax0024 (2019), distributed under CC-BY NC.

When considering biomolecules under physiological conditions, it is well-known that water provides the environment in which the biomolecular machinery can exist and function. By screening and stabilizing static electronic multipoles, water significantly alters the structure, stability, and dynamics of biomolecules.[206,221,222] As mentioned in the previous chapter, the favorable exposure of moieties with static electronic multipoles to water and the corresponding burying of non-polar residues into a hydrophobic core, is also an important driving force for protein folding.[204–206] While the importance of this *hydrophobic effect* and the pivotal role of water for biomolecular systems is under no dispute,[204,205] the underlying fundamental physics of solvated (bio)molecular systems is still not fully explored and understood.

In particular, here we focus on the quantum-mechanical nature of solute–solvent interactions. It has already been shown that polarization effects and the many-body character of bonded interactions and hydrogen bond networks play an important role for solvated systems,[206,223–225] but also long-range vdW dispersion interactions form an essential component for water and for both *intra*-protein and protein–water interactions. This vdW component has not been investigated in full detail nor on a fundamental level up to now. In this chapter, we thus more specifically address the protein–water vdW interaction and find that these quantum-mechanical interactions can account for up to 30 % of the total solvation energy. Together with their essential role for *intra*- and inter-protein interactions, this calls for a more complete microscopic understanding of vdW dispersion forces under physiological conditions, which is imperative to shed light on the physics of proteins in aqueous solvation. This is further motivated by the fact that conventional classical potentials show limited transferability with respect to the conformational space and results can vary widely among different force fields, when trying to cover both folded and unfolded states or intrinsically disordered proteins.[24,226–228] It has already been shown that the

quantum-mechanical, non-local vdW interaction is described insufficiently and inconclusively by conventional approaches.[229,230] As such, they can represent a major origin for the limited transferability of classical potentials. In this context, previous studies have pointed out that traditional MM potentials and water models likely provide an unbalanced description of vdW interactions for proteins in water,[231] which usually results in an over-compaction of unfolded states.[226–228] Typically, this unbalanced description is approached by adapting the pairwise vdW interaction coefficients for the *intra*-protein, water–water, or protein–water interaction.[227,228,231] However, such adaptions still can be highly system-dependent especially when also considering disordered proteins.[24]

In this work, we seek to understand the fundamental basis for the failure of the traditional models in a bottom-up approach. Such understanding is essential to pave the way towards a more accurate and balanced description of the relevant conformational space of biomolecules, which in the end plays a key-role for predicting folding mechanisms, barriers and potential misfolds. In fact, an *ab initio* description of vdW interactions is the only way to study the role of the solvent, as force field methods are typically strongly limited in their transferability between gas and liquid phase due to their high degree of parameterization. In our study, we focus on a comparison of a pairwise-additive description of vdW interactions and an accurate, quantum-mechanical many-body treatment. The pairwise vdW models are represented by the vdW(TS)[48] as well as Grimme's D2[127] and D3[49] approaches. Such an approximate, pairwise formalism represents the basis for the standard phenomenological description of long-range correlation in biomolecular simulations *via* Lennard-Jones potentials. For comparison, we study the vdW interaction within the MBD formalism,[50,51,131] which accounts for the many-body character of vdW dispersion interactions to infinite order in perturbation theory within an interatomic framework and has been proven to provide quantitative improvements and a better qualitative understanding compared to the pairwise-additive approximation in numerous studies.[31,33–35,50,92,93,230] In the case of polarizable supramolecular complexes and complex molecular crystals, which show similar bonding features as biomolecular systems, MBD allows to reach a quantitative agreement to within 1 and 0.5 kcal/mol, respectively, when compared to (near-)exact quantum-mechanical methods and experiment.[33,35,232] Yet, its computational efficiency together with modern implementations and the ever-growing availability of computational resources, allows for treatment of systems consisting of several thousands of atoms as required for studying biomolecular systems such as solvated proteins.

As introduced in Sec. 2.3.5, the MBD formalism also represents a model for intrinsic electronic fluctuations[35,172,233] — a molecular analogue to the plasmon pseudo-particle in metallic systems — which we will use to further characterize the protein–water interaction. Such investigations can provide essential insights to comprehend and conceptually understand vdW interactions as recently illustrated for hybrid and nanostructured systems[172,233] or π–π stacked molecules.[35] Considering the electrodynamical view on vdW dispersion offered by this model, it is important to note that the dielectric permittivity of water has an expected value around 2.3[234] at the frequencies of the electronic fluctuations, which are responsible for dispersion interactions, *i.e.*, at a few petahertz (PHz). In contrast to static electronic multipoles, vdW interactions are therefore not strongly screened by aqueous environments and, thus, can give rise to long-range interactions also in solvated systems. As such, long-range correlation forces may play an important role for the long-range ordering often observed in biological systems or form the quantum-mechanical basis for the emergence of coherent molecular vibrations.[31,235] Such collective nuclear behavior has been proposed to play an important role in long-distance recognition among biological macromolecules.[236–239] Within the conventional view of solvated proteins, however, the basis for long-range recognition under physiological conditions is still controversially discussed. Recent studies also suggest connections between collective electronic fluctuations — the basis of vdW dispersion interactions — and enzymatic action on DNA[240] or pharmaceutical activity.[241]

We exemplify our findings in detail for the Fip35 Hpin1 WW domain (Fip35-WW) and further showcase their general validity for the *de novo* Chignolin variant "cln025" and the fast-folding Nle/Nle double mutant of the villin headpiece (HP35-NleNle). The folding trajectories of the three proteins have been obtained in atomistic detail and explicit solvent in previous MD simulations by Shaw *et al.*,[14] Lindorff-Larsen and co-workers,[242] and Ensign *et al.*,[13] respectively. All vdW(TS) and MBD calculations reported below have been parameterized from DFTB calculations with recent *mio* (Fip35-WW, HP35-NleNle) or *3ob* parameters (cln025) using the DFTB+ software package.[47] Consistency of the obtained polarizabilities for the two parameterizations has been tested and confirmed based on Fip35-WW. MBD calculations have been performed using a self-written implementation as now available in the libMBD package.[243]

4.1 *Intra*-Protein van der Waals Interaction

We start out by investigating the Fip35-WW trajectory artificially removing the surrounding solvent from a MD trajectory in explicit water to focus on *intra*-protein interactions, where dispersion forces represent one of the main sources of interaction within the protein core. Accordingly, we observe an increased magnitude of the vdW energy while this core is being formed and particularly during the hydrophobic collapse in all applied dispersion models (see Fig. 4.1). Notably, in comparison to the results obtained within the pairwise approaches, many-body dispersion effects consistently decrease the relative stability of the native state for the isolated protein by 6 kcal/mol on average, *cf.* Fig. 4.1(bottom). The outliers of this general behavior observed around 15 and 26 µs correspond to transient, partially-folded intermediates. The relative destabilization by beyond-pairwise contributions can be explained by an overestimation of the *intra*-core vdW interactions ("over-correlation") in the pairwise approximation. By reducing the interaction to pairwise-additive potentials, a two-body formulation assumes ideal correlation between all pairs of atoms (see Fig. 2.1) and with that, neglects the complex geometrical arrangement within the protein core. Such geometrical constraints limit the emergence of correlated fluctuating dipole patterns and thus lower the interaction energy as already observed for a wide variety of systems.[33,35,92,93,244] For small peptides such effects have been found to be mostly negligible.[229,245] Our findings show that for larger biomolecules, however, a many-body treatment of vdW interactions is indeed essential, which is in line with the findings of Schubert *et al.* for 20-residue peptides.[230]

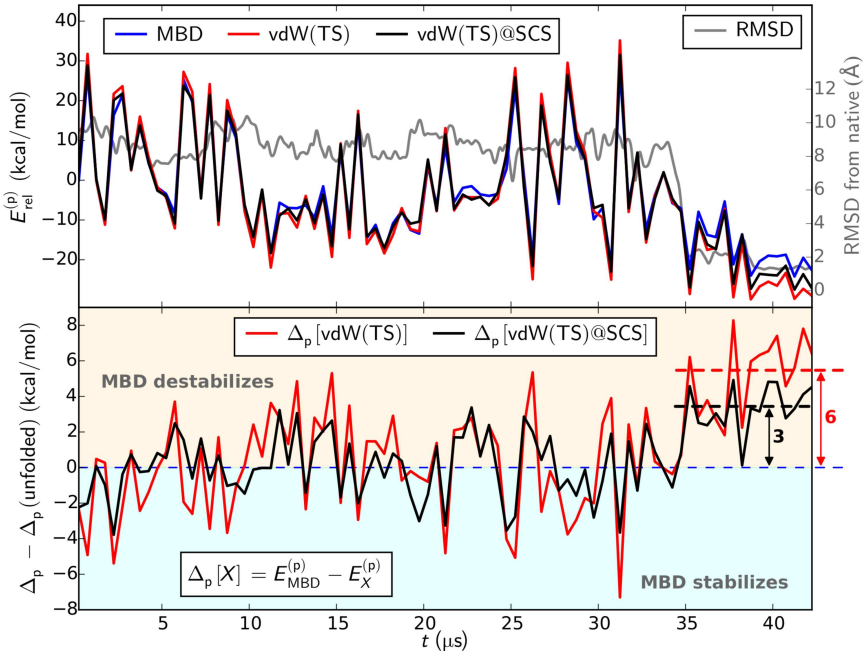

Fig. 4.1: *Intra*-protein vdW interaction energy along the folding trajectory of Fip35 Hpin1 WW-domain in solvated geometry. *top*: vdW energetics as obtained with MBD and the pairwise approaches vdW(TS) and vdW(TS)@SCS, i.e., vdW(TS) with self-consistent screening. RMSD from native state shown in grey. *bottom*: Beyond-pairwise contributions, as given by the difference between many-body and pairwise treatment. Reproduced from Ref. 2 (distributed under CC-BY NC).

In the MBD formalism, we make use of a two-step procedure: We obtain effective, screened atomic polarizabilities from self-consistent electrodynamic screening to account for the presence of multiple fluctuating dipoles in the system and then solve a many-body Hamiltonian, which is defined in terms of these polarizabilities, to capture many-body vdW interactions (see Sec. 2.3.5). To study the effect of each step, we combined vdW(TS) with the self-consistent screening procedure. In this variant, which we refer to as vdW(TS)@SCS, screened interaction coefficients enter the pairwise-additive potentials instead of the hybridized chemical analogue used in ordinary vdW(TS), see Eq. (2.66). In this way, we account for the effects on atomic polarizabilities due to the field of the surrounding dipoles, but do not include long-range many-body interactions. With vdW(TS)@SCS we already capture some part of the destabilization of native states amounting to 3 kcal/mol (*cf.* Fig. 4.1). Thus, half of the overstabilization in vdW(TS) is from neglecting the presence of multiple dipoles in the system and half from the many-body character of dispersion interactions. This also implies that for a proper description of gas-phase proteins, one has to account for the screening of polarizabilities and the many-body nature of vdW interactions.

4.2 van der Waals Solvation Energy

Fig. 4.2(top) shows the vdW contribution to the solvation energy obtained with MBD and the vdW(TS) model, as defined by $E(\text{sol}) = E_{\text{vdW}}[\text{ps}] - E_{\text{vdW}}[\text{p}] - E_{\text{vdW}}[\text{s}]$ with ps referring to Fip35-WW in solvation, p to Fip35-WW in gas phase, and s to the pristine solvent. As an artifact of the above-mentioned over-correlation within the pairwise approach, we find a consistent overestimation of the dispersion contribution in vdW(TS). In terms of the relative solvation energy, however, pairwise and many-body treatment show the same general trend, which qualitatively follows the inverse root-mean-square deviation (RMSD) from the native state with a step coinciding with the collapse of the protein into the native, more globular shape. This finding can be explained by the removal of hydrophobic residues from the protein–water interface and thus decreasing their interaction with the solvent. The average dispersion contribution to the solvation energy drops by 29 kcal/mol (15 %) at the hydrophobic collapse. The step-like behavior of the vdW solvation energy along the trajectory is even more pronounced than observed for the *intra*-protein vdW interaction in the gas phase and almost resembles a two-state model of folded and unfolded states. As such, the vdW solvation energy strongly correlates with the protein's deviation from the folded conformation. This feature has been found for all dispersion models considered here.

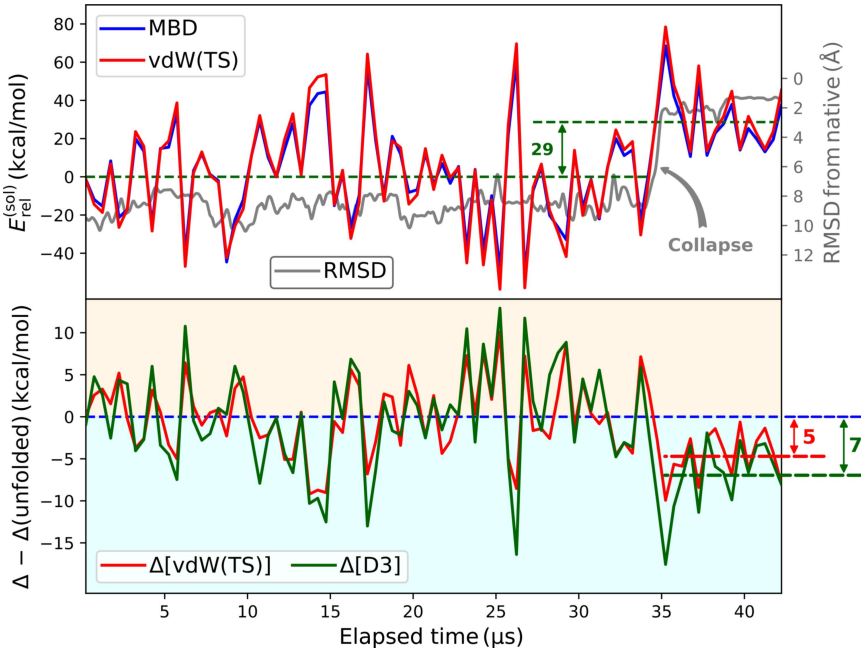

Fig. 4.2: Relative vdW solvation energy, $E_{\text{rel}}^{(\text{sol})}$, during the folding process of the Fip35 Hpin1 WW-domain. *top*: backbone root-mean-square deviation from final conformation illustrating the hydrophobic collapse around 35 μs (gray). The vdW contribution to the relative solvation energy is shown for the pairwise vdW(TS) model (red) and MBD (blue). *bottom*: Difference in the relative stabilization by the solvent between MBD and the pairwise vdW(TS) and D3 referenced to the unfolded state. Reproduced from Ref. 2 (distributed under CC BY-NC).

Comparing many-body and pairwise treatment of dispersion interactions, Fip35-WW does no longer feature a consistent change in the relative stability of native versus non-native conformations once embedded in an aqueous environment. Thus, beyond-pairwise effects in the protein–water vdW interaction stabilize folded conformations. Correspondingly, we see a clear increase in the relative vdW solvation energy for native and native-like states, when comparing the pairwise models to MBD (5 kcal/mol for vdW(TS), 7 kcal/mol for D3). This shift is due to the lack of a systematic many-body (de-)stabilization in the total vdW energy of solvated Fip35-WW and the water box during the whole folding trajectory (as concluded in Chapter 3), combined with an inversion of the behavior observed for the isolated protein shown in Fig. 4.1. This implies that the protein–water interaction compensates for the destabilization of native states through many-body dispersion effects, observed *in vacuo*. In summary, besides screening permanent electronic multipoles, water also provides the necessary environment to stabilize native conformations *via* beyond-pairwise vdW interactions, which counteracts the destabilizing effect that such many-body terms have on the *intra*-protein interaction.

4.3 Plasmon-like Solute–Solvent Interactions

As has been shown previously, MBD also provides a model for the intrinsic electronic fluctuations:[35,172,233] In analogy to nuclear quantum effects, the electronic fluctuations, which ultimately give rise to vdW interactions, can be understood as the zero-point fluctuations in the instantaneous electronic charge distribution. The MBD formalism gives access to an orthonormal decomposition of this zero-point fluctuation, which can be interpreted as "eigenmodes" of the electron density. A detailed analysis of these electronic eigenmodes reveals that the number of very localized high-frequency fluctuations, formerly mainly located on the solute, significantly decreases upon coupling to the surrounding water. This implies a delocalization of electronic fluctuations and an increase of the collectivity of electronic behavior. This plasmon-like character and the delocalization over protein and solvent form the fundamental reason for the stabilization of native states through many-body dispersion effects in the protein–water interaction. The role of the surrounding solvent can be seen as providing weakly structured polarizable matter, which counteracts the destabilizing many-body effects observed *in vacuo* for native and partially folded states. To gain further insight into the characteristics of protein–water vdW interactions, we additionally obtain the contribution of individual electronic fluctuations to the solvation energy. As originally proposed in Ref. 35, the vector of mode-projected interaction energies, ε_{int}, is obtained *via*

$$\boldsymbol{\varepsilon}_{\text{int}} = \frac{1}{2}\mathcal{U}^{\dagger}[\mathbf{C}_{\text{ps}}]\left\{\mathcal{U}[\mathbf{C}_{\text{ps}}]\boldsymbol{\omega}_{\text{ps}} - \boldsymbol{\omega}_{\text{p/s}} - \mathcal{U}[\mathbb{1}]\left(\boldsymbol{\eta}_{\text{ps}} - \boldsymbol{\eta}_{\text{p/s}}\right)\right\}, \tag{4.1}$$

where $\mathbf{C}$ is the eigenvector matrix of the MBD potential matrix and $\boldsymbol{\omega}$ and $\boldsymbol{\eta}$ represent vectors of all MBD eigenfrequencies and uncoupled frequencies, respectively. ps denotes the solvated protein and p/s the system without protein–water interaction. The transformation matrix, $\mathcal{U}$, thereby transforms the individual eigenfrequencies to the same eigenspace and is given by

$$\mathcal{U}[\mathbf{X}] = \left[\left(\mathbf{C}_{\text{p}} \oplus \mathbf{C}_{\text{s}}\right)^{\dagger}\mathbf{X}\right]^{*} \circ \left[\left(\mathbf{C}_{\text{p}} \oplus \mathbf{C}_{\text{s}}\right)^{\dagger}\mathbf{X}\right], \tag{4.2}$$

where $\circ$ denotes an element-wise product. The above definition yields a doubly-stochastic matrix, which upon multiplication with a vector preserves the sum of its elements representing the MBD (interaction) energy. For a more detailed discussion, see Appendix A2. Fig. 4.3 shows the mode-projected vdW solvation energy for an unfolded, partially-folded and native conformation of Fip35-WW. Overall, the contributions to the vdW solvation energy arise from the frequency-regions with a low DOS of MBD eigenfrequencies, which is dominated by the contributions of the aqueous environment. As shortly discussed in Sec. 3.2, interaction between the individual QHOs leads to the a broadening of the non-interacting DOS and the emergence of satellite peaks in the

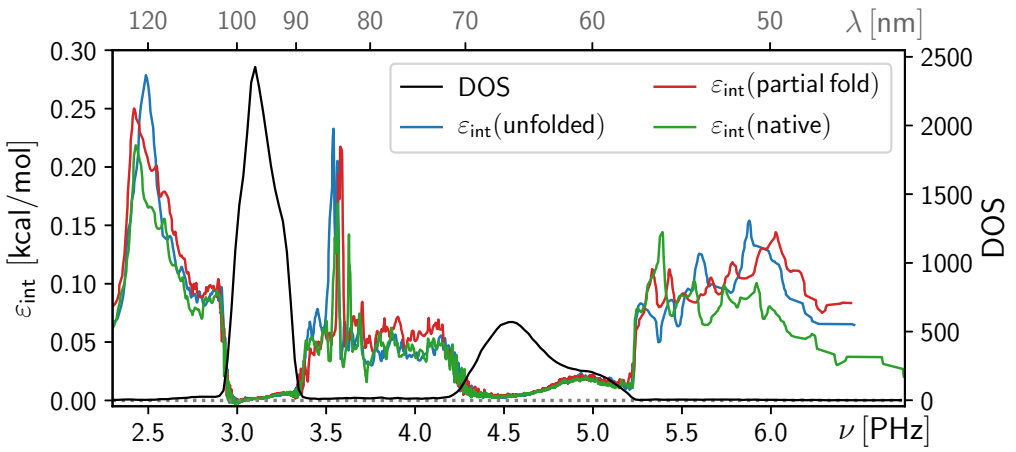

Fig. 4.3: Mode-projected van der Waals solvation energy (ε_{int}) of Fip35 Hpin1 WW domain. Contribution per "electronic eigenmode" as a function of its characteristic frequency for unfolded state (blue), partially-folded state (red) and native conformation (green). Density of states of eigenfrequencies shown in black.

spectrum. As such, the regions neighboring the main peaks of the DOS contain the emergent collective electronic eigenmodes responsible for the protein–water interaction, which is perfectly captured by our mode-projection analysis. The strongest contributions to the vdW solvation energy can thereby be found for highly collective, wavelike electronic fluctuations around 2.5 PHz ($\hat{=}$ 120 nm). The contribution of this low-frequency domain to the protein–water interaction adds up to 15 % ($\hat{\approx}$ 41 kcal/mol) and about 20 % ($\hat{\approx}$ 6.2 kcal/mol) for relative solvation energies. The second largest contributions are from the region around 3.6 PHz. Also the high-frequency end of the spectrum shows a notable interaction energy per mode, but due to the minuscule DOS in this region, the overall contribution is negligible (*e.g.*, ~1 kcal/mol for relative solvation energies). Regarding the magnitudes of the results shown in Fig. 4.3, note that the mode-projected interaction energies are normalized to yield the total vdW solvation energy by construction.

Fig. 4.4A shows a real-space illustration of one of the low-frequency modes, which commonly feature large charge displacements along the polarizable protein backbone coupled to electronic fluctuations throughout the surrounding solvent. While a number of these wavelike fluctuations remain largely on the protein, many reach from the protein backbone inside the hydrophobic core far into the aqueous environment. To analyze this spatial extent, we combine the spatial information of a given eigenmode with the corresponding contribution to the solvation energy. This allows to define a radial interaction distribution as given by

$$G_{\text{int}}(R) = \sum_i \varepsilon_i \sum_A \delta_{R,R_A} \sum_{j \in A} \|\mathbf{C}_{\text{ps}}^{(i,j)}\|^2 \,, \tag{4.3}$$

with δ as Kronecker-delta generalized to $\mathbb{R}$. Conceptually, the above defines a radial distribution of the atomic contributions to an individual eigenmode and then combines the distributions with ε_{int} as weighting factor. Note that integrating $G_{\text{int}}(R)$ yields the full protein–water vdW interaction energy thanks to $\mathbf{C}$ being a unitary matrix and the definition of ε_{int} (see Appendix A2). Comparing the radial distribution of the contributions to the vdW solvation energy between the pairwise vdW(TS) and MBD models, as shown in Fig. 4.4B, reveals a striking difference in the interaction range within the two treatments: In the pairwise model, the contribution of solvent atoms to the vdW solvation energy subsides beyond 6 Å, so roughly twice the sum of the vdW radii of carbon and oxygen. Accounting for many-body dispersion, on the other hand, shows that electronic correlation between the protein and solvent atoms up to 25 Å from the protein–water interface is still relevant for the protein–water interaction. This reflects the weakness of the screening of dispersion forces by the solvent and is in evident contrast to the often assumed locality of vdW interactions in solvated systems. The slow decay of contributions to the protein–water interaction

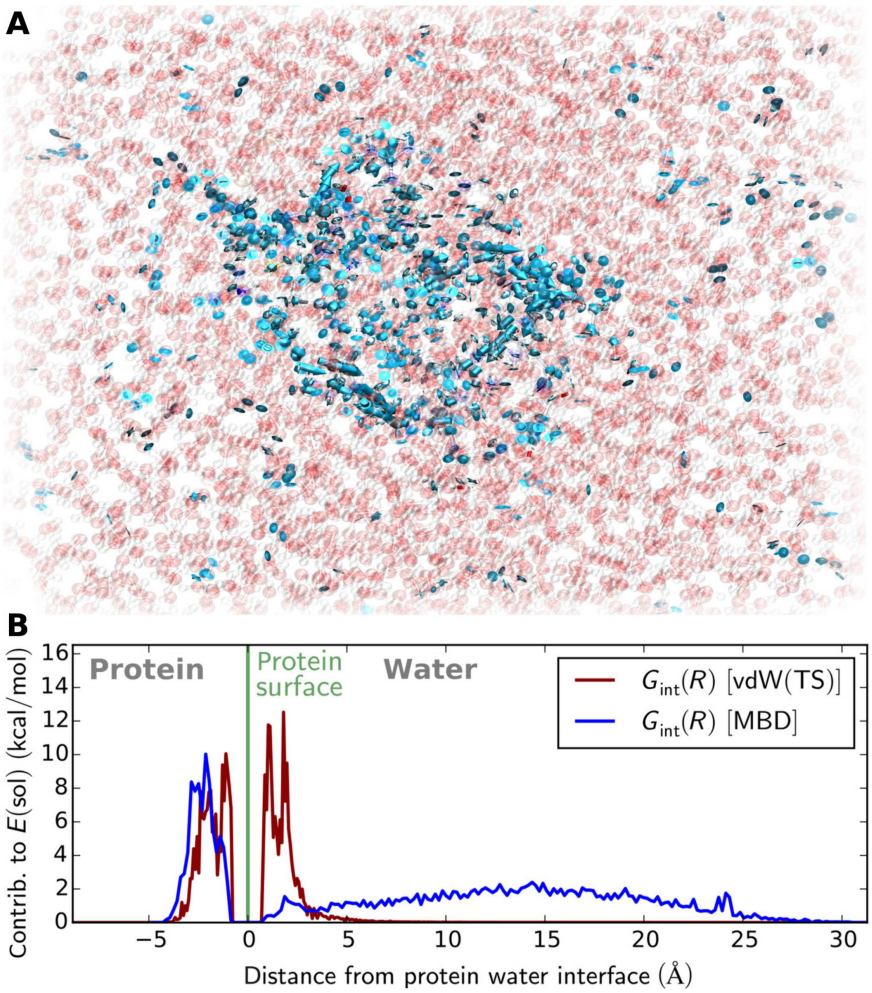

Fig. 4.4: Characteristics of correlated electronic fluctuations in the WW-domain of Fip35 Hpin1. (A) Illustration of low-frequency, plasmon-like fluctuations in solvated Fip35 Hpin1 WW domain, which show the largest contribution to the protein–water interaction (solvent shown in atomistic detail; oxygen: red, hydrogen: white). The arrows (blue) depict the direction of simultaneous electron density deformations (eigenmode of the electron density). If no arrow is shown, the given atom does not contribute significantly to the eigenmode. (B) Contributions to the vdW solvation energy within the pairwise vdW(TS) approach and the many-body dispersion formalism (MBD) as radial distribution functions. Reproduced from Ref. 2 (distributed under CC BY-NC).

can be further rationalized by the comparably slow decay of the polarization response between water and the full protein as shown in Fig. 3.8. While such ranges are unprecedented in the context of solvated systems, similar and larger interaction ranges have already been found for molecular crystals [34] or nanostructures. [92] From a different point of view, Fig. 4.4B represents a radial analysis of the change in the distribution and frequency of electronic fluctuations introduced by embedding the protein in water. It thus demonstrates that, while the atomistic structure and *local* dynamics of water typically remain largely unperturbed beyond a few solvation layers, [206,207] electronic behaviors can indeed indicate the presence of a protein over large distances. Such long-range collective behavior and the ensuing *non-local* nuclear dynamics could be experimentally probed thanks to recent advances in terahertz spectroscopy, [208–210] for instance.

4.4 Effect of Secondary Structure

Fip35-WW is a showcase example for the formation of β-sheets. To test the general validity of our hypotheses, we carried out the same analysis for the modified villin headpiece, HP35-NleNle, (formation of α-helical entities) and the cln025 variant of the *de novo* protein Chignolin (plain β-hairpin formation). Our analysis confirmed our early findings for Fip35-WW. The vdW contribution to the solvation energy reflects the trend of the inverse RMSD from the native structure with a drop of 15 % at the hydrophobic collapse. Again, the protein–water interaction counteracts the destabilizing many-body dispersion effects observed in gas phase and increases the relative stabilization of native states with respect to unfolded structures. Also, collective plasmon-like electronic fluctuations have been found to show a major contribution to the total and relative solvation energy ($\Delta E_{\mathrm{sol}}^{(\mathrm{low}\,\omega)}$) for both, the hairpin-forming cln025 and the helix-forming HP35-

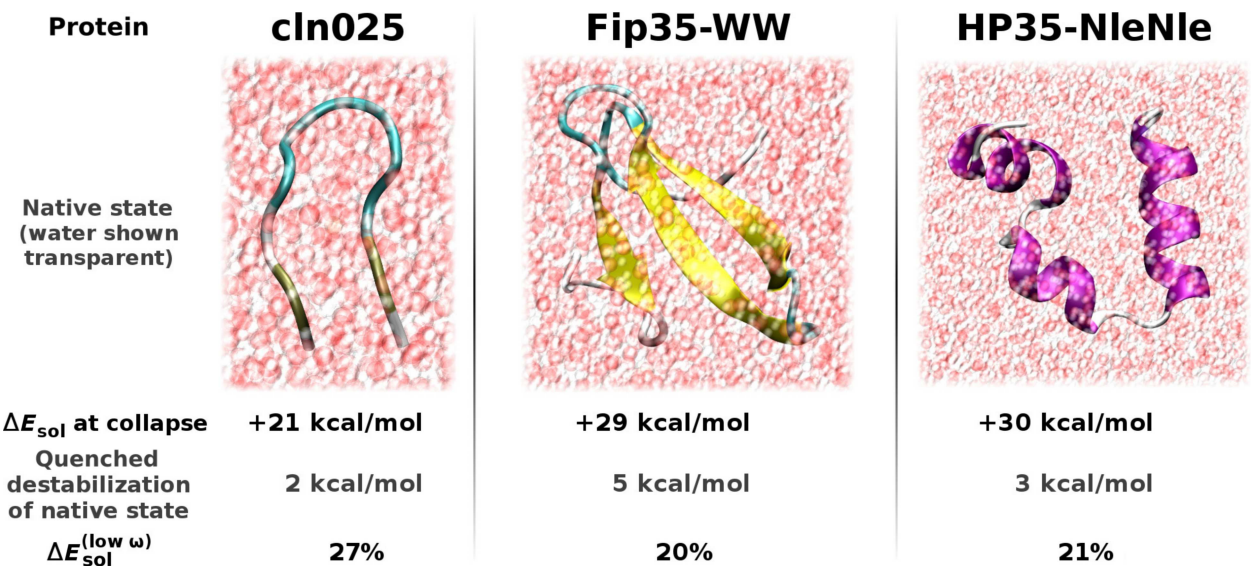

Fig. 4.5: Characteristics of protein–water dispersion interactions: Independent of the secondary structure, the van der Waals solvation energy captures the hydrophobic collapse in form of a 20–30 kcal/mol jump ("ΔE_{sol} at collapse") and many-body protein–water van der Waals interactions consistently stabilize native states in solvation. Low frequency, collective electronic fluctuations contribute significantly to the relative solvation energy ($\Delta E_{sol}^{(low\,\omega)}$) in all cases. Reproduced from Ref. 2 (distributed under CC BY-NC).

NleNle. The above conclusions also remain unaltered, when studying a structural sampling obtained with the more recent and accurate a99SB-*disp* force field and the TIP4P-D water model, which avoids the spurious over-compaction of unfolded protein states [227,228] (results shown in Appendix A3). Fig. 4.5 summarizes the above-mentioned features and highlights the general validity of the present findings in the biomolecular context. Independent of the secondary structure to be formed, the vdW solvation energy captures the hydrophobic collapse in the form of a sizable jump in stabilization ("ΔE_{sol} at collapse") for all considered proteins. Also, the consistent increase of the relative stability of native states due to the many-body character of protein–water vdW interactions and the significance of low-frequency, wavelike electronic fluctuations (as characterized by their contribution to relative solvation energies, $\Delta E_{sol}^{(low\,\omega)}$) turned out to be independent of the final secondary structure. We note that the magnitudes of the different quantities are not necessarily representative for secondary structure elements as also the system size varies from 6,000 to 14,000 atoms. The systematic investigation of the relation between system size, secondary structure elements and the magnitude of the present observations or characteristic features of fluctuation patterns, as shown in Fig. 4.4A, is beyond the scope of this thesis.

4.5 Discussion of Protein–Water van der Waals Interactions

In conclusion, we have shown that many-body dispersion effects lead to a considerable relative destabilization (≈ 4.5 kcal/mol for the proteins studied here) of the native state of solvated proteins in terms of the *intra*-protein interaction. We find that the screening of the instantaneous dipoles due to the surrounding dipole field and many-body interactions contribute in similar parts to the destabilization. Notably, this effect is of a comparable order of magnitude as estimates for the zero-point vibrational and entropic contribution found for the folding of isolated polypeptides. [246,247] The destabilization through many-body dispersion effects can play an important role in explaining why proteins often do not adopt the same folded conformation in the gas phase and in solvation. It also indicates how the neglect of the inherent many-body character of dispersion interactions in traditional vdW approaches (and derivative MM potentials) can lead to a spurious description of *intra*-protein interactions in general.

In aqueous solvation, the vdW contribution to the solute–solvent interaction of (small) proteins closely tracks the conformational changes of the folding process. The hydrophobic collapse of the protein is accompanied by a jump of about 15 % (20 to 30 kcal/mol) in the vdW solvation energy. The total electronic energy of solvation, for comparison, does not provide such clear insight — only the free energy of solvation does. The beyond-pairwise contributions to the protein–water vdW interaction favor folded states, and thus, the many-body aspect of solvation leads to a considerable stabilization of native conformations. So, a pairwise vdW model overestimates the internal interaction as a driving force of folding (pairwise additive potentials in general favor minimal atom–atom distances and thus globular shapes), while the protein–water vdW interaction should have a larger stabilizing effect on the native state than the pairwise formalism suggests. Careful measurements of the solvation enthalpy of folded and disordered proteins would allow one to further study the correctly balanced description of *intra*-protein and protein–water interactions and to estimate the effect of beyond-pairwise vdW interactions *via* complementary simulations. The lack of stabilization in the pairwise approach is the result of a distinct many-body character of the protein–water dispersion interaction in the form of delocalization and a high degree of collectivity of electronic fluctuations across protein and solvent. This long-range mediation of vdW interactions in solvated systems can also be pivotal for other quantities, as demonstrated for the protein–water interaction range in Fig. 4.4B. Collective electronic petahertz fluctuations should also directly manifest in the system's (response) properties. As such, the plasmon-like fluctuations are, in principle, directly accessible by (vacuum) ultraviolet spectroscopy. However, this requires careful disentanglement from ionization and other excitation processes in this spectral range, *e.g.*, the Rydberg series of water.

Our study shows that the findings summarized above can be generalized for helix-, β-sheet-, or hairpin-forming proteins and are, thus, independent of secondary structure motifs. So, an accurate description of solvated proteins generally requires capturing the subtle balance between beyond-pairwise effects on the *intra*-protein vdW interaction (destabilizing native states) and the highly collective character of protein–water interactions (stabilizing native states). With increasing system size and complexity, finding this balance without explicit account for the quantum-mechanical many-body nature of vdW interactions is an intricate task and failure to do so can contribute to the fundamental origin of the previously reported [24,226–228,231] unbalanced description of vdW forces by traditional MM potentials and water models. The previously proposed remedies for this shortcoming effectively involve a general adaptation of the relative magnitudes of protein–protein and protein–water dispersion interactions. [226–228] In contrast to these "static fixes", our findings suggest that conventional pairwise potentials actually lack a conformation-dependent adaptation of *intra*-protein and protein–water vdW interactions due to the neglect of many-body effects. In the spirit of a static fix, rescaling of pairwise vdW energetics considerably improves the obtained protein–water interaction (see Appendix A4). Nevertheless, this still shows considerable deviations and is highly-dependent on the system size and conformation (no rescaling provides optimal performance for small systems). This renders the simple rescaling without system-dependent adaptations insufficient for treating (bio)molecules on different length scales, as required for describing assembly or docking processes, for example. Further analysis of the relation between structure and the many-body (de)stabilization represents a promising avenue towards more transferable and first-principles-motivated "fixes" to conventional potentials. Last, we have also performed our analysis based on a new sampling of folded and unfolded states of the Chignolin variant cln025 using the a99SB-*disp* force field and the TIP4P-D water model. [227,228] This approach has been designed and shown to provide a more balanced description of *intra*-protein and protein–water interactions and thus avoids the spurious over-compaction of unfolded states. The additional analysis confirmed the results reported above and shows that the present conclusions are not an artifact of an unphysical sampling of protein states, but truly represent the effect of many-body vdW physics (see Appendix A3).

In a broader perspective, our findings imply that an effective pairwise-additive treatment of vdW interactions and derivative MM potentials can provide accurate energetics for a particular application, but only explicitly quantum-mechanical models, such as the one used here, allow one to attain an unbiased microscopic understanding of biomolecular interactions in a more general context. On the basis of Fig. 4.4B, for instance, we expect that for several solutes or a larger solute, the approximation of pairwise additivity can fail on a fundamental level. The persistence and collectivity of electronic fluctuations through the solvent can mediate long-range correlation forces between individual solutes or moieties of solvated macromolecules and an effective pairwise description might not be able to reproduce the subtle balance between long-range correlation on all these scales. As such, collective electrodynamical interactions can also substantially affect molecular assembly and the formation of tertiary structures. In addition, complex long-range fluctuations of the electronic charge distribution are less sensitive to the instantaneous solvent structure and thus to thermal fluctuations, which makes them an ideal contender for biomolecular recognition. In this form of recognition, the solvent provides electron density that serves as a mediator for long-range interaction, while the actual atomistic structure and the nuclear dynamics of the solvent do not necessarily have to be altered in the process, which has been concluded from a number of experiments.[206] [207] It is worthwhile to mention, however, that most of these experiments probed rather local interactions and dynamics.

Recent terahertz spectroscopy experiments, for instance, show that the presence of a solute can have a considerable effect on the long-time scale dynamics and long-range polarization of water.[208–211] Obviously, such correlations require a long-range interaction mechanism, which is not present in the traditional classical view of biomolecular systems, but represented by the highly-collective many-body vdW forces shown in this work. From previous studies, we expect the observed long-range persistence of electron correlation through aqueous environments to mainly manifest in the longer-time scale nuclear dynamics of the system. Such behavior has been observed in crystalline molecular systems, where many-body dispersion effects particularly affect low-frequency ("slow") phonon modes.[31,235] Long-range electronic correlation between (solvated) biomolecules can also form the quantum-mechanical basis for correlated collective nuclear motion within the respective partners. Such concerted motion is essential for the emergence of coherent molecular vibrations, a promising explanation for long-range recognition through electrodynamic interaction of the resulting oscillating molecular dipoles,[238,239] or coordinated enzymatic action.[240] Likewise, the presented long-range interaction mechanism can also provide a qualitatively new explanation of allosteric regulation as it is observed in various biomolecular systems. Together, our findings apply in a broader context of biomolecular interactions — not just in the case of protein folding as exemplified here.

Obviously, the stability and functionality of biomolecules are ultimately determined by their free energy. Hence, this work represents a first step towards a more fundamental understanding of the physics of proteins in water, but to accurately address the implications of plasmon-like features within biological systems, we need to extend our study to free energy at finite temperature. It is already known that the many-body character of vdW dispersion interactions can give rise to low-frequency, high-entropy vibrational modes in organic matter.[31,235] Experimentally, a wealth of such vibrations has been observed also for proteins. While classical simulations often do predict a notable vibrational eigenmode density in the low-terahertz domain, they still lack several qualitative features, and a final confirmation by correctly reproducing actual experimental spectra is rarely provided.[248] Given that the quantum-mechanical interaction mechanism presented in this work is strongly delocalized over many atoms including solvent, we suggest that it represents an essential component for the emergence of collective vibrations and the feature-rich terahertz signal of (solvated) biomolecules. In the case of crystalline aspirin, for instance, the lowest-frequency phonon band arises only because of many-body dispersion interactions. These low-frequency vibrations then lead to a selective, relative entropic stabilization of one of the poly-

morphs by 0.6 kcal/mol per molecule at room temperature.[31] This can be seen as an estimate for the effect on a protein's relative entropy per residue. As the dynamics and functionality of a biomolecule can be strongly related to its eigenmodes,[249] the impact on collective vibrational modes also hints at an unrevealed role of collective electrodynamical interaction mechanisms for the functionality and coordination in the biomolecular machinery. In summary, our work presents the quantum-mechanical basis for a long-range interaction mechanism in (solvated) biomolecular systems, which is proposed to play important roles in long-distance recognition, enzymatic action,[240] and pharmaceutical activity.[241] Our DFTB+MBD framework provides a robust formalism for the investigation of such further-reaching implications as it allows for a fully quantum-mechanical many-body treatment of large-scale systems in atomistic detail.

4.6 Perspective: van der Waals Allostery & Biomolecular Regulation

Parts of this section have been produced in collaboration with Dr M. Gori and Dr P. Kurian.

The emergence and preservation of well-defined, coordinated processes are fundamental prerequisites for the successful exploitation of the biomolecular machinery and for life itself. How such ordered and robust mechanisms can exist amidst the crowded and noisy interior of living cells is one of the most fascinating yet at the same time most puzzling questions in molecular biology and biophysics. Not only must biomolecular partners coordinate their interactions in order to find each other in the right sequence, at the right time, in the right location, but once these partners are bound, electrostatic and electrodynamic interactions must be controlled precisely to produce the intended outcome of the biochemical event. The classical "lock-and-key" picture of biomolecular regulation and allostery,[250] which has motivated biological intuition for more than a century, takes the dominant electrostatic view of rigid, mesoscopic biomolecules and does not incorporate the physics of fluctuating networks of atoms coupled by charge (and spin) into its vision of living systems. This aspect became well-recognized in the 1980s thanks to the seminal extension of the classical picture of rigid mesoscopic bodies to include the dynamic behavior of the atoms forming them.[251] Regulation and allosteric control were then formulated based on the modulation of vibrational/phononic thermal fluctuations and the resulting entropic contribution to the free energy landscape.[252,253] The advent of this so-called *thermal fluctuations allostery* or *dynamic allostery* allowed to overcome many of the limitations of the previous picture and since then enabled a better explanation and understanding of various examples of biomolecular regulation such as the allosteric uptake of oxygen by hemoglobin,[254,255] signaling through intrinsically disordered proteins,[256] the kinetic isotope effect in enzymes,[257] the olfactory system,[258] drug-receptor interactions,[259] microtubule deformations,[260] protein–water coupling,[208,261] and inelastic electron tunneling-mediated activation of G-protein-coupled receptors.[262]

4.6.1 Regulation and Allostery in the View of Long-Range van der Waals Dispersion

The findings on the long-range persistence of electron correlation in biomolecules (see Chapter 4) and the previously reported selective entropic stabilization of molecular crystal polymorphs through many-body vdW interactions[31,93] motivate to expand our view of allostery and regulation even further and to consider not only the dynamical (vibrational) behavior of nuclei, but also the one of the electronic charge density. Extending the picture of dynamic allostery to the level of electrodynamic quantum fluctuations propels us into a new regime of observation and begs the question of how these divergent scales can overlap and interconnect. Long-range vdW interactions arising from such electronic quantum fluctuations may, in fact, play a two-fold role in biomolecular regulation. Being inherently non-local in nature, the wavelike propagation and collective character of vdW dispersion can lead to far-reaching effects for the potential energy surface. By means of modulating the energetic landscape it can affect relative stabilities and thermodynamic populations, but also the dynamic behavior of nuclei, which induces a

renormalization of the vibrational/phononic spectrum. As such, long-range electron correlation can provide the basis for both, enthalpic and entropic, contributions to regulation, allostery and signaling, while it naturally incorporates long-range cooperativity rather than merely local structural modifications.

On a macroscopic scale, many mechanisms can be organized *via* signal-and-response cascades and the underlying, more holistic picture might fade, but simple macroscopic explanations rarely apply on the nano- to micro-scale of individual biomolecules such as enzymes. Nevertheless, many enzymes are known to perform very precise and coordinated actions. One enzyme category of interest is the class of restriction endonucleases. As a part of so-called restriction modification systems, they are ubiquitous in bacteria or archaea and represent a major part of the organism's defense against foreign DNA as introduced to the cell by engaging bacteriophages and other viruses, for example.[263,264] In this defense system, modification enzymes introduce a specific methylation pattern to host DNA and restriction enzymes cleave and incapacitate DNA without this specific pattern. As such, restriction endonucleases are able to perform two intertwined functions: (1) scan and recognize a given DNA sequence together with its specific methylation pattern, and (2) perform a repeatable and coordinated double-strand DNA cleavage. This capability is particularly interesting for the sub-class of type-II endonucleases, which do not require ATP or other external chemical energy sources and are known to cut both DNA strands of the recognition sequence in a concerted manner.[265] As the canonical representative of type-II restriction enzymes, Fig. 4.6 illustrates the chirally-symmetric homodimer structure of *Eco*RI, which reflects the palindromic symmetry of its target DNA substrate. The widely accepted mechanism of DNA cleavage by *Eco*RI is based on the recognition of the palindromic 5'-GAATTC-3' sequence together with the host's characteristic methylation pattern by the residues Met137–Arg145 (highlighted in green in Fig. 4.6). This sequence is referred to as recognition (R) loop. Enzymatic cleavage of the phosphodiester backbones is then initiated by residues Asp91, Glu111, and Lys113 (highlighted in red), which capture a water molecule from the inner solvation shell and transport it to the cleavage site through a rotational motion. The hydrolytic cutting then proceeds *via* divalent metal cations, likely magnesium or manganese cofactors[265] (shown in brown). How the recogni-

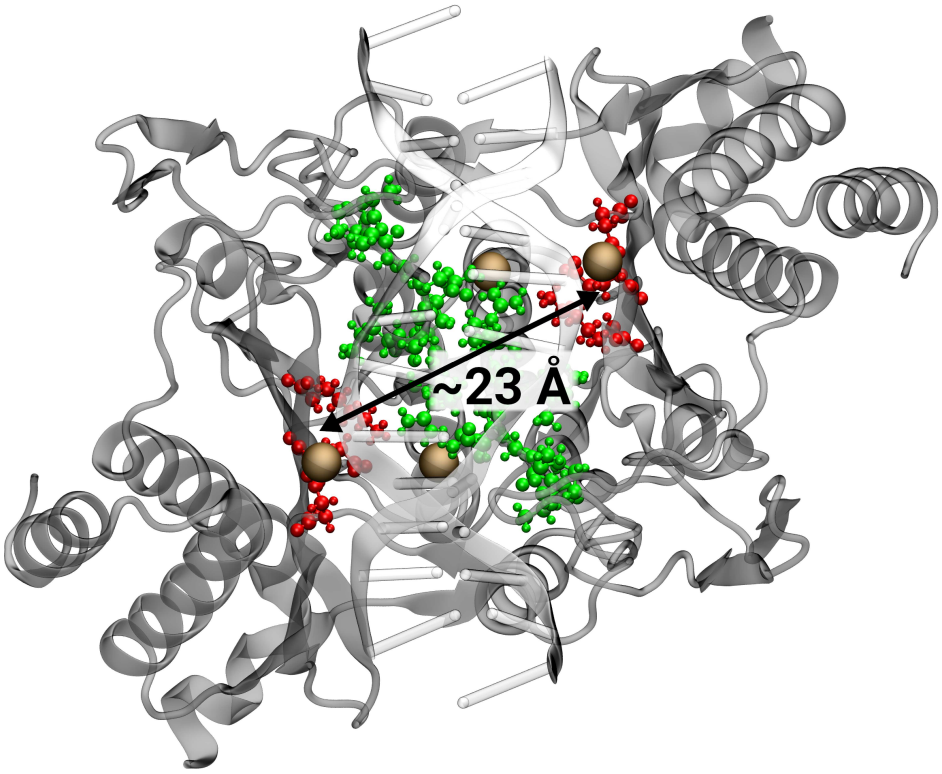

Fig. 4.6: Illustration of the *Eco*RI restriction endonuclease in complex with double-strand DNA. Recognition loops highlighted in green, catalytic centers in red and DNA in white. The catalytic centers perform the hydrolysis of the DNA double-strand *via* Mg^{2+} cofactors (brown). The average distance between the active sites is 23 Å.

tion mechanism is coupled to catalytic action and how DNA cleavage is coordinated between the active sites, however, is still one of the least understood aspects of (type-II) endonucleases.[265]

Based on a simplified QM model of oscillating dipoles, coherent electronic behavior in the π-stacked DNA bases has been proposed as a possible mechanism for the synchronization of the double-strand cleavage.[240] Such collective electronic fluctuations form the basis of long-range vdW forces.[1,171] In the context of biomolecular systems, vdW interactions are typically considered rather "short-ranged". This notion roots from the R^{-6}-decay between two atoms. As we have seen in the above sections and Chapter 3, however, pairwise additivity and atomistic scaling laws do not necessarily apply to the effective coupling/interaction between extended moieties as represented by a set of catalytically-active residues. In contrast to the previously proposed role of entangled π-electrons on DNA basepairs, vdW dispersion is ubiquitous in nature and does not suffer from the renowned limitations of preserving (quantum) coherence within noisy environments. When accounting for their collective character and non-additivity, vdW interactions have been shown to substantially alter energetics,[33,35] interaction scaling laws[92,170,172,266] and collective nuclear behavior[31,213] in comparison to the limited traditional view of pairwise interaction potentials. The results presented in the previous sections, further confirm that the underlying wavelike electronic behavior persist on the length scales and amidst the dizzying complexity of biomolecular systems. For a brief perspective on the characteristics of collective electronic behavior and potential role of long-range vdW interactions in *Eco*RI, we chose the crystal structure of the pre-reactive complex available from X-ray diffraction (PDB entry 1CKQ with two Mg^{2+} cofactors per sub-unit, as depicted in Fig. 4.6).[267] The actual number of cofactors present under physiological conditions is thereby still under debate.[265] For the purpose of this perspective, this structure represents a viable model for the physiological structure. In contrast to previous studies, this work is not limited to the complexed DNA, but investigates long-range vdW interactions within the full DNA-enzyme complex.

4.6.2 Collective Electronic Behavior as Basis for Long-Range Coordination

To understand the potential role of collective electronic behavior for enzymatic regulation, we first studied the characteristics of the intrinsic electronic behavior. In addition to providing accurate dispersion forces, the MBD formalism represents a model for the intrinsic electronic fluctuations, which give rise to vdW interactions[2,35,172] (see also Sec. 2.3.5). The eigenmode transformation of the MBD Hamiltonian can serve as a harmonic approximation to intrinsic electronic quantum fluctuations, which we can conceptually interpret as "zero-point vibrations" of the electron density. Fig. 4.7 shows one such "electronic eigenmode" for the *Eco*RI-DNA complex, where the arrows depict simultaneous electron density deformations. One recurring feature is the emergence of a wavelike propagation of electronic fluctuations over large parts of the DNA-enzyme complex. Of particular interest here are the ones extending from one active center through the R-loop and the complexed DNA to the active center in the other sub-unit. With such plasmon-like modes, the intrinsic electronic behavior and thus vdW interactions at the two catalytic center are connected. This represents the basis for any potential long-range regulation through dispersion forces. In this regard, we can also estimate the vdW interaction potential energy. In the traditional pairwise view of vdW interactions, this is simply given by the sum of all C_6/R^6-potential terms between the atoms of the catalytic centers of each sub-unit. With an average interatomic distance of ~23 Å, the R^{-6}-dependency leads to a nearly-negligible interaction between the active centers in such pairwise-additive approaches. As we have seen in the previous sections and Sec. 3.3, such scaling laws do not necessarily apply between larger-scale moieties such as the set of catalytically-active residues considered here. Within a many-body state, individual fragment interactions are not naturally available, however. In order to get an estimate for the (effective) interaction potential energy within the many-body framework, we can define an interaction operator given by the sum of dipole coupling tensors between all atoms of

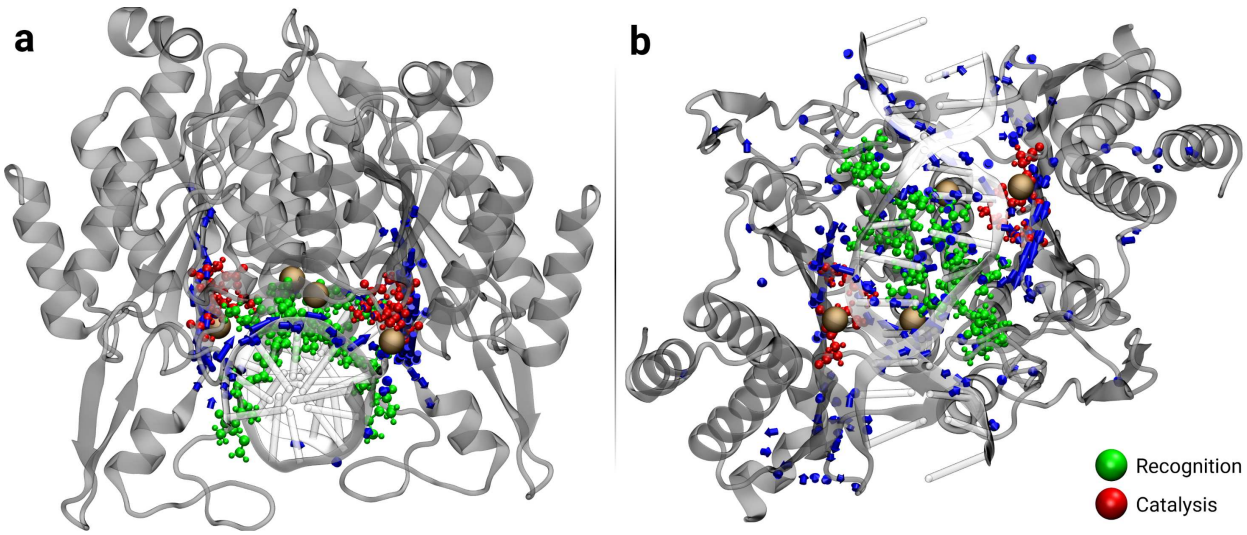

Fig. 4.7: "Eigenmode representation" of intrinsic electronic behavior in *Eco*RI. *Eco*RI–DNA complex shown in grey, recognition loops highlighted in green and catalytic centers in red. Blue arrows depict direction of coupled deformations of the electron density. Top (**a**) and front (**b**) view of eigenmode spanning active regions and DNA as basis for plasmon-like long-range interaction mechanism.

the catalytic residues of each sub-unit defined by the set of atoms $\mathcal{C}_1$ and $\mathcal{C}_2$. Then, we evaluate the expectation value of this (dipolar) interaction operator for the MBD wave function, Ψ_{MBD},

$$E_{\mathrm{dip}}^{(\mathrm{int})}\left[\Psi_{\mathrm{MBD}}\right] = \sum_{A\in\mathcal{C}_1}\sum_{B\in\mathcal{C}_2}\left\langle \Psi_{\mathrm{MBD}}\left| g_{\mathrm{rs}}(R_{AB})\,\eta_A\eta_B\sqrt{\alpha_{A,0}^{(\mathrm{sr})}\,\alpha_{B,0}^{(\mathrm{sr})}}\,\zeta_A^{\mathrm{T}}\,\mathbf{T}_{AB}\,\zeta_B\right|\Psi_{\mathrm{MBD}}\right\rangle . \tag{4.4}$$

The interaction potential is thereby defined in accordance with the MBD formalism: η_A and $\alpha_{A,0}^{(\mathrm{sr})}$ are the effective excitation frequency and polarizability of atom A, ζ_A is the mass-weighted displacement of the QHO representing atom A and $\mathbf{T}_{AB}$ is the bare dipole–dipole interaction tensor for QHOs A and B. For the interaction potential energy between fragments, we define the $(3N\times 3N)$-dimensional coupling tensor $\tilde{\mathbf{T}}_{\mathcal{C}_1\mathcal{C}_2}$ with

$$\tilde{\mathbf{T}}_{\mathcal{C}_1\mathcal{C}_2}^{(AB)} = \begin{cases} g_{\mathrm{rs}}(R_{AB})\,\eta_A\eta_B\sqrt{\alpha_{A,0}^{(\mathrm{sr})}\,\alpha_{B,0}^{(\mathrm{sr})}}\,\mathbf{T}_{AB} & \text{if } A\in\mathcal{C}_1, B\in\mathcal{C}_2 \;\vee\; B\in\mathcal{C}_1, A\in\mathcal{C}_2 \\ 0 & \text{else.} \end{cases} \tag{4.5}$$

Following the more detailed derivation of the expectation value of a dipolar interaction potential for dipole-coupled QHOs given in Sec. 5.1.2 and Appendix A6, we obtain

$$E_{\mathrm{dip}}^{(\mathrm{int})}\left[\Psi_{\mathrm{MBD}}\right] = \sum_i \frac{\left(\mathbf{C}^{\mathrm{T}}\tilde{\mathbf{T}}_{\mathcal{C}_1\mathcal{C}_2}\mathbf{C}\right)_{ii}}{4\,\omega_i} , \tag{4.6}$$

where $\mathbf{C}$ is the eigenvector matrix of the MBD potential matrix and ω_i are the MBD eigenfrequencies. The above procedure yields the dipolar potential energy contribution to the vdW dispersion interaction between the fragments, which can then serve as an estimate for the effective vdW interaction energy within the many-body state. The resulting estimate of the many-body vdW coupling between the active sites is substantially larger than in a pairwise approach. It should be noted, however, that this estimate was found to be very sensitive to the molecular structure and placement of Mg-cofactors. Compared to a pairwise treatment, the increase in interaction energy can thereby range from five- to almost twenty-fold. So, in all cases, we observe a considerable increase of the coupling between active sites due to many-body effects, but reliable quantitative results require further careful study of physiologically-representative structural ensembles.

In order to gain a qualitative understanding, we compared our estimates for the interaction potential energy in the full DNA-enzyme complex to the corresponding systems without R-loop. In all cases, removal of the R-loop gives rise to a decrease in the interaction between the catalytic sites. This can be understood by considering the interaction mechanism outlined above: Instead of a simple point-to-point interaction through vacuum — as assumed in the traditional pairwise picture — the interaction is the result of the wavelike propagation of electron correlation through the enzymatic complex. The emergence of such wavelike patterns has already been reported for simple model systems such as carbyne wires.[172] In a very similar fashion, peptide backbones, residues and DNA basepairs can serve as polarizable medium mediating dynamic electron correlation over large distances. This wavelike propagation and the mediation of electron correlation is highly-dependent on the presence and structure of the interjacent protein — in this case, the R-loop. Experiments have shown that the R-loop indeed plays a crucial role for the catalytic process and cleavage activity of type-II endonucleases.[265,268] Obviously, the traditional atom-pairwise view of vdW interactions is completely agnostic to such modifications and yields the same results independent of the presence and structure of the R-loop or even when considering the two catalytic sites in absence of the remaining enzymatic complex. Our qualitative analysis within a many-body framework of vdW interactions, on the other side, indicates that the non-local effect of such modifications can be attributed to a direct link between recognition and catalysis within a single, non-local mechanism, where the R-loop mediates long-range electron correlation forces between the active centers.

Interestingly, the well-defined domains of directed collective electronic behavior, which form the basis for such an interaction mechanism, coexist orthogonally with a variety of rather noisy and undirected patterns of electronic fluctuations. In contrast to the directed "electronic eigenmodes" shown in Fig. 4.7, these mostly do not extend over the active sites for recognition and catalysis, but are situated towards peripheral regions (*cf.* Fig. 4.8). A noisy, thermal environment or other external perturbations would primarily couple to and affect these modes simply due to spatial proximity. This can provide a natural shielding mechanism, where peripheral electronic modes preserve the internal domains of electron correlation and thus the non-local interaction mechanism responsible for the regulation of enzymatic activity in thermal environments.

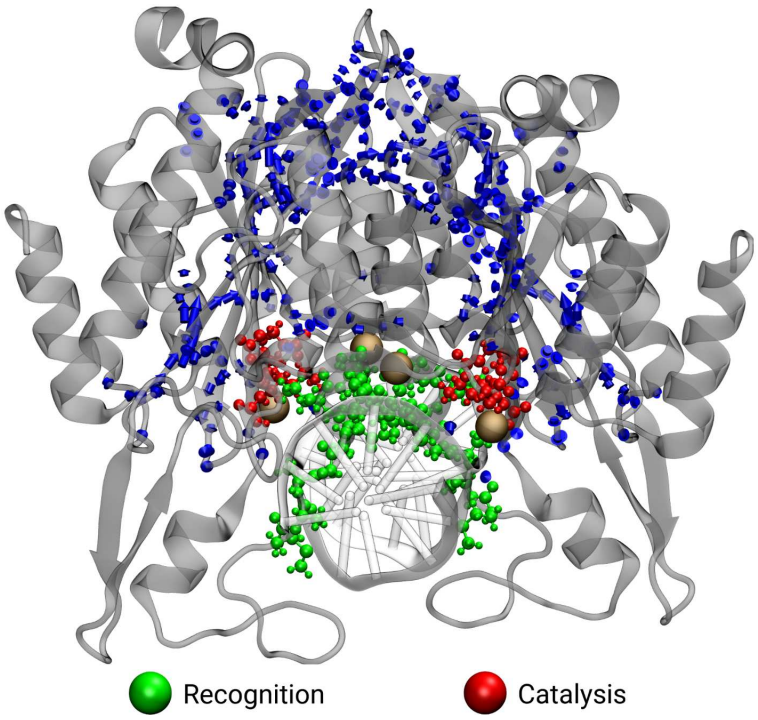

Fig. 4.8: Example of disordered, peripheral electronic "eigenmode" providing shielding mechanism for correlation among active centers. Recognition loops highlighted in green, catalytic centers in red and DNA in white.

Chapter 5

Many-Body Dispersion beyond the Random Phase Approximation

Parts of this chapter have been published in
M. Stöhr, M. Sadhukhan, Y. S. Al-Hamdani, J. Hermann, and
A. Tkatchenko, *Nat. Commun.* accepted (2020).
The presented methodology represents the practical extension of Ref. 269 to atomistic modeling.

As introduced in Chapter 2, the majority of practical vdW methods are based on either an atom-pairwise approximation, the interatomic dipole limit, or the Random Phase Approximation (RPA). With that, the current approaches applicable to practically-relevant systems fail to treat the full complexity of vdW dispersion arising from Coulomb-coupled quantum-mechanical fluctuations in the electronic charge density. Nevertheless, effects beyond the above approximations are rarely investigated and usually *ad hoc* considered to be negligible. Lately, this notion is increasingly disputed, however, as the missing physics in the most widely used computational methods continues to stand in the way of understanding a growing number of state-of-the-art experimental phenomena. For example, the experiments of Pollice *et al.* reveal a considerable impact of the solvent on the *intra-* and intermolecular dispersion interactions in proton-bound dimers, [270] whilst their computational study using implicit solvents in combination with methods based on atom-pairwise dipolar vdW interactions fails to capture the effect. In another related experiment, Secchi *et al.* found that water flows ultra-fast through narrow carbon nanotubes (CNTs), but not through boron nitride nanotubes. [271] In this regard, theorists are still working towards satisfactory modeling of this effect and capturing the underlying physical interactions in van der Waals materials. [111,272–274] In a similar vein, the spatial separation and ordering of large polarizable molecules on metal surfaces, [174,275] salient in organic thin films for organic electronics, highlights gaps in common modeling approaches. For instance, Wagner *et al.* showed that large aromatic molecules organize into highly ordered arrays at high coverage on Au(111). [174] Such puzzling experimental observations and the various phenomena emerging under nanoscale confinement [276–280] challenge our current understanding of intermolecular interactions in complex

systems. Given the frequent presence of confinement in (bio)molecular systems and materials, these phenomena have broad implications across biology, chemistry, and materials science. In biology, for example, cells provide a crowded environment featuring processes and dynamics under membrane confinement or within (ion) channels, and non-covalent protein–protein, protein–DNA/RNA, or virus–host interactions are often mediated by transient, interfacial water molecules.[281–286] In chemistry and materials science, nanostructured materials are at the forefront of technological developments such as (selective) catalyst systems, sensing and filter applications as well as energy materials.[287–293] The disagreement of experimental observations with our current understanding and methodological approaches calls for a more complete, yet sufficiently efficient, description of intermolecular interactions. With the important role of vdW interactions for the dynamics and diffusion at the nanoscale,[294] this also includes an accurate many-body treatment of vdW forces beyond the common interatomic dipole limit or RPA.

In a previous work, Sadhukhan and Tkatchenko have introduced a formalism to account for vdW interactions beyond dipolar coupling and exemplified the ground-breaking effect it can have for two oscillators with reduced dimensionality representing a model system for confined atoms or molecules.[269] This chapter presents the applicable extension of this formalism to atomistic modeling, which provides a consistent and practical approach to incorporate higher-order multipolar coupling while retaining a full many-body treatment of vdW forces based on the MBD framework.[50] In Sec. 5.1, the theoretical foundation of this extended formalism is outlined and discussed in comparison to state-of-the-art electronic-structure methods. Sec. 5.2 highlights the role of the obtained energy contributions and how they give rise to non-trivial behaviors in the long-range interaction inside nanoscale structures. Sec. 5.3 summarizes the presented formalism and discusses the role of such higher-order contributions for the physical interactions under nanoscale confinement in the context of the above-mentioned experiments.

5.1 Coulomb Interaction Between Dipolar Quantum Fluctuations

The starting point for the proposed approach to beyond-RPA many-atom vdW interactions is the MBD formalism, which models instantaneous electronic fluctuations as a set of QHOs coupled *via* the dipole–dipole potential (see Sec. 2.3.5). The Hamiltonian for this set of dipole-coupled (DC) oscillators can be written as

$$\hat{\mathcal{H}}_{\mathrm{DC}} = \hat{\mathcal{T}} + \hat{U} + \hat{\mathcal{V}}_{\mathrm{dip}} \equiv \hat{\mathcal{H}}_0 + \hat{\mathcal{V}}_{\mathrm{dip}} \tag{5.1}$$

where $\hat{\mathcal{T}}$ and $\hat{U}$ are the kinetic energy and harmonic potential operators, respectively (*cf.* MBD Hamiltonian (2.59) as detailed in Sec. 2.3.5). This Hamiltonian lends itself to a closed-form solution *via* eigenmode transformation and the vdW energy is obtained as the difference in the ground-state energy between the DC system and its non-interacting variant (described by $\hat{\mathcal{H}}_0$),

$$E_{\mathrm{MBD}} \;=\; E_{\mathrm{DC}} - E_0 \;=\; \sum_i \frac{\omega_i}{2} - \sum_i \frac{\eta_i}{2} \tag{5.2}$$

where ω_i and η_i are the effective frequency of the i^{th} oscillator mode in the DC and uncoupled system, respectively.[50] This work presents a consistent extension to go beyond the dipole approximation in $\hat{\mathcal{H}}_{\mathrm{DC}}$. Given that the equivalent of Eq. (5.1) with the full Coulomb potential does not allow for a straightforward closed-form solution, we instead use the correction towards fully Coulomb-coupled oscillators,

$$\hat{\mathcal{V}}' = \sum_{A<B} g_{\mathrm{rs}}\big(R_{AB}\big)\left[\hat{\mathcal{V}}_{\mathrm{Coul}}(\mathbf{r}_A,\mathbf{r}_B) - \hat{\mathcal{V}}_{\mathrm{dip}}(\mathbf{r}_A,\mathbf{r}_B)\right] , \tag{5.3}$$

where $\mathbf{r}_A$ and $\mathbf{r}_B$ denote the position of the (pseudo-)particles of oscillators centered at $\mathbf{R}_A$ and $\mathbf{R}_B$, respectively. With full Coulomb coupling the vdW dispersion energy is well-behaved in all

cases. Using only dipolar or beyond-dipolar coupling individually, however, leads to a divergence of the vdW energy at short distances. The perturbing potential in the proposed formalism, $\hat{\mathcal{V}}'$, is therefore given by the long-range beyond-dipolar potential with g_{rs} as range-separating function. The corresponding correction in the (interaction) energy can then be derived and evaluated within first-order perturbation theory starting from the DC state, Ψ_{DC}. Note that the interaction energy of a set of oscillators beyond dipolar coupling also includes mean-field effects.[25,295] Such contributions are already accounted for in (semi-)local electronic-structure methods such as DFT(B). In order to only describe electron correlation and correlation-induced phenomena, we need to subtract the mean-field contributions, which can be achieved with the help of the expectation value of the beyond-dipolar potential for the uncorrelated state, Ψ_0. Thus,

$$E_{\mathrm{DCS}} = \left\langle \Psi_{\mathrm{DC}} \middle| \hat{\mathcal{V}}' \middle| \Psi_{\mathrm{DC}} \right\rangle - \left\langle \Psi_0 \middle| \hat{\mathcal{V}}' \middle| \Psi_0 \right\rangle , \tag{5.4}$$

where Ψ_{DC} and Ψ_0 represent the ground-state of the DC and non-interacting system as described by $\hat{\mathcal{H}}_{\mathrm{DC}}$ and $\hat{\mathcal{H}}_0$, respectively. In the spirit of the terminology of quantum-chemical expansion series such as coupled-cluster theory, we will refer to the first-order full Coulomb correction over the MBD energy as *Dipole-Correlated Coulomb Singles* (DCS). Given that the zeroth-order (MBD) Hamiltonian already represents a correlated state within dipolar coupling, the present *Singles* term is to be distinguished from those in *post*-Hartree-Fock methods or the RPA, where the zeroth-order theory corresponds to a mean-field, uncorrelated state. A more detailed discussion of DCS in the context of correlated electronic-structure methods is given in Sec. 5.1.4 below. Inserting Eq. (5.3) into Eq. (5.4), yields four individual terms entering E_{DCS}:

$$E_{\mathrm{Coul}}[\Psi_{\mathrm{DC}}] = \left\langle \Psi_{\mathrm{DC}} \middle| \sum_{A<B} g_{\mathrm{rs}}(R_{AB}) \frac{q_A q_B}{\|\mathbf{r}_A - \mathbf{r}_B\|} \middle| \Psi_{\mathrm{DC}} \right\rangle \tag{5.5}$$

$$E_{\mathrm{Coul}}[\Psi_0] = \left\langle \Psi_0 \middle| \sum_{A<B} g_{\mathrm{rs}}(R_{AB}) \frac{q_A q_B}{\|\mathbf{r}_A - \mathbf{r}_B\|} \middle| \Psi_0 \right\rangle \tag{5.6}$$

$$E_{\mathrm{dip}}[\Psi_{\mathrm{DC}}] = \left\langle \Psi_{\mathrm{DC}} \middle| \sum_{A<B} g_{\mathrm{rs}}(R_{AB}) \mathcal{V}_{\mathrm{dip}}(\mathbf{r}_A, \mathbf{r}_B) \middle| \Psi_{\mathrm{DC}} \right\rangle \tag{5.7}$$

$$E_{\mathrm{dip}}[\Psi_0] = \left\langle \Psi_0 \middle| \sum_{A<B} g_{\mathrm{rs}}(R_{AB}) \mathcal{V}_{\mathrm{dip}}(\mathbf{r}_A, \mathbf{r}_B) \middle| \Psi_0 \right\rangle = 0 , \tag{5.8}$$

where q_A and q_B are the (effective) charge of the oscillators A and B, respectively, and the last term (dipolar interaction in uncorrelated state, Ψ_0) is identical to zero.[295] In accordance with Sec. 2.3.5, the MBD wave function can be written as

$$\begin{aligned} \Psi_{\mathrm{DC}} &= \prod_{i=1}^{3N} \left(\frac{\omega_i}{\pi}\right)^{1/4} \exp\left(-\frac{\omega_i}{2}\xi_i^2\right) &&= \left(\prod_{i=1}^{3N} \frac{\omega_i}{\pi}\right)^{1/4} \exp\left(-\frac{1}{2}\sum_{j,l}\overbrace{\sum_i \mathbf{C}_{ji}\omega_i\mathbf{C}_{li}}^{\boldsymbol{\Omega}_{jl}}\zeta_j\zeta_l\right) \\ &= \left(\prod_{i=1}^{3N} \frac{\omega_i}{\pi}\right)^{1/4} \exp\left(-\frac{1}{2}\sum_{j,l}\zeta_j\boldsymbol{\Omega}_{jl}\zeta_l\right) &&= \left(\prod_{i=1}^{3N} \frac{\omega_i}{\pi}\right)^{1/4} \exp\left(-\frac{1}{2}\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\zeta}\right) , \end{aligned} \tag{5.9}$$

where ω_i are the MBD eigenfrequencies (*i.e.*, the square-root of the eigenvalues of the MBD potential matrix), $\xi_i = \sum_j \mathbf{C}_{ji}\zeta_j$ is the transformed eigenmode coordinate of the MBD Hamiltonian, ζ_i an individual mass-weighted oscillator displacement, $\mathbf{C}$ is the eigenvector matrix of the MBD potential matrix and $\boldsymbol{\Omega} = \mathbf{C}\,\mathrm{diag}\{\omega_i\}\,\mathbf{C}^{\mathrm{T}}$. $\boldsymbol{\zeta}$, finally, is the generalized coordinate of all mass-weighted displacements in the N oscillators corresponding to the vector,

$$\boldsymbol{\zeta} = \bigoplus_i \zeta_i = \left(\zeta_1, \zeta_2, \ldots, \zeta_{3N}\right)^{\mathrm{T}} = \bigoplus_A \left(\boldsymbol{\zeta}_A^{(x)}, \boldsymbol{\zeta}_A^{(y)}, \boldsymbol{\zeta}_A^{(z)}\right)^{\mathrm{T}} = \bigoplus_A \boldsymbol{\zeta}_A = \bigoplus_A \sqrt{m_A}\left(\mathbf{r}_A - \mathbf{R}_A\right) . \tag{5.10}$$

In terms of notation, ζ_i refers to the individual scalar elements of the generalized coordinate, while $\boldsymbol{\zeta}_A$ denotes the three-dimensional vector containing the elements corresponding to oscillator A (*i.e.*, the x, y, z-components of the mass-weighted displacement of the QHO representing atom A). Ψ_0 can be written in the same compact form as Eq. (5.9), when using the replacements $\omega_i = \eta_i$ and $\mathbf{C} = \mathbb{1}_{3N}$. The following subsections show the derivations of the non-vanishing expectation values for wave functions of the general form (5.9).

5.1.1 Coulomb Interaction Energy

With the above notations, the expectation value of the Coulomb interaction energy becomes

$$E_{\text{Coul}}[\Psi_{\text{DC}}] = \left(\prod_{i=1}^{3N}\sqrt{\frac{\omega_i}{\pi}}\right)\int\cdots\int \sum_{A<B} g_{\text{rs}}(R_{AB})\,\frac{q_A\,q_B}{\|\mathbf{r}_A-\mathbf{r}_B\|}\,e^{-\boldsymbol{\zeta}^{\text{T}}\boldsymbol{\Omega}\boldsymbol{\zeta}}\prod_{i=1}^{3N} d\xi_i\,. \tag{5.11}$$

Following the detailed derivations presented in Appendix A5, the Coulomb interaction energy can be simplified to

$$E_{\text{Coul}}[\Psi_{\text{DC}}] = \sum_{A<B}^{N}\frac{2\,q_A q_B}{\sqrt{\pi}}\left(m_A\,m_B\right)^{3/2} g_{\text{rs}}(R_{AB})\sqrt{\frac{\det\{\boldsymbol{\Omega}\}}{\det\{\boldsymbol{\Omega}''_{AB}\}}}\int_0^\infty \frac{e^{-\mathbf{R}_{AB}^{\text{T}}\boldsymbol{\Theta}_{AB}\mathbf{R}_{AB}}}{\sqrt{\det\{\boldsymbol{\Gamma}_{AB}+\mathbf{S}_2\}}}\,ds\,, \tag{5.12}$$

where we have introduced $\boldsymbol{\Omega}''_{AB}$ corresponding to the submatrix of $\boldsymbol{\Omega}$ obtained by removing all rows and columns that correspond to oscillators A or B. Using $\boldsymbol{\Omega}'_{AB}$ to denote the submatrix of $\boldsymbol{\Omega}$ obtained by removing all columns corresponding to QHOs A or B, the remaining matrices in the integral are given by

$$\boldsymbol{\Gamma}_{AB} = \left[\left(\sqrt{m_A}\,\mathbb{1}_3\right)\oplus\left(\sqrt{m_B}\,\mathbb{1}_3\right)\right]\left(\boldsymbol{\Omega}_{AB}-\boldsymbol{\Omega}'^{\text{T}}_{AB}\boldsymbol{\Omega}''^{-1}_{AB}\boldsymbol{\Omega}'_{AB}\right)\left[\left(\sqrt{m_A}\,\mathbb{1}_3\right)\oplus\left(\sqrt{m_B}\,\mathbb{1}_3\right)\right] \tag{5.13}$$

$$\text{and}\qquad \boldsymbol{\Theta}_{AB} = \boldsymbol{\Gamma}_{AB}-\boldsymbol{\Gamma}_{AB}^{\text{T}}\left(\boldsymbol{\Gamma}_{AB}+\mathbf{S}_2\right)^{-1}\boldsymbol{\Gamma}_{AB}\qquad \text{with}\qquad \mathbf{S}_2 = s^2\begin{pmatrix}\mathbb{1}_3 & -\mathbb{1}_3\\ -\mathbb{1}_3 & \mathbb{1}_3\end{pmatrix}. \tag{5.14}$$

The remaining integrand represents a rapidly converging function and numerical integration can be carried out efficiently *via* Gauß-Legendre quadrature. It should be noted, however, that $\boldsymbol{\Gamma}_{AB}$ and thus $\boldsymbol{\Theta}_{AB}$ involve the inverse of $\boldsymbol{\Omega}''_{AB}$. Inverting this $(3N-6)\times(3N-6)$-dimensional matrix represents the main computational bottleneck. Considering that the above has to be evaluated for all pairs AB, the calculation of E_{Coul} scales as $\mathcal{O}(N^5)$ with the number of atoms N. With increasing system size, the removal of six rows and columns from $\boldsymbol{\Omega}$ to obtain $\boldsymbol{\Omega}''_{AB}$ represents a low-rank update of the original matrix. The required inverse can then be obtained more efficiently as a low-rank update of the inverse of the original $\boldsymbol{\Omega}$-matrix. With the use of the Woodbury Identity, we can reduce this to simple matrix multiplications and low-dimensional matrix operations (see Appendix A7). While the scaling with the number of atoms remains the same, such an *exact* reformulation considerably reduces the prefactor in the computational costs and allows for a more efficient evaluation of E_{Coul} and thus E_{DCS}.

From Eq. (5.12), we can also obtain the expectation value for the uncoupled state, Ψ_0. With $\omega_i = \eta_i$, where $\eta_i = \eta_A$ for $i \in A$, and $\mathbf{C} = \mathbb{1}_{3N}$, we get

$$E_{\text{Coul}}[\Psi_0] = \sum_{A<B}^{N}\frac{2\,q_A q_B}{\sqrt{\pi}}\left(m_A\,m_B\,\eta_A\,\eta_B\right)^{3/2} g_{\text{rs}}(R_{AB})\int_0^\infty \frac{e^{-\mathbf{R}_{AB}^{\text{T}}\boldsymbol{\Theta}_{AB}\mathbf{R}_{AB}}}{\sqrt{\det\{\boldsymbol{\Gamma}^{(0)}_{AB}+\mathbf{S}_2\}}}\,ds\,. \tag{5.15}$$

The definitions of $\boldsymbol{\Theta}_{AB}$ and $\mathbf{S}_2$ remain the same as given in Eq. (5.14), while $\boldsymbol{\Gamma}_{AB}$ simplifies to

$$\boldsymbol{\Gamma}^{(0)}_{AB} = \left[\left(\sqrt{m_A}\,\mathbb{1}_3\right)\oplus\left(\sqrt{m_B}\,\mathbb{1}_3\right)\right]\left[\left(\eta_A\,\mathbb{1}_3\right)\oplus\left(\eta_B\,\mathbb{1}_3\right)\right]\left[\left(\sqrt{m_A}\,\mathbb{1}_3\right)\oplus\left(\sqrt{m_B}\,\mathbb{1}_3\right)\right]. \tag{5.16}$$

5.1.2 Dipole Interaction Energy

Next, we deal with the term (5.7), the expectation value for the dipolar interaction potential. The dipolar interaction potential as used in the MBD formalism is given by

$$\hat{\mathcal{V}}_{\text{dip}} = \frac{1}{2}\sum_{A,B} g_{\text{rs}}(R_{AB})\,\eta_A\eta_B\sqrt{\alpha^{(\text{sr})}_{A,0}\,\alpha^{(\text{sr})}_{B,0}}\,\boldsymbol{\zeta}_A^{\text{T}}\mathbf{T}_{AB}\,\boldsymbol{\zeta}_B = \frac{1}{2}\sum_{A,B}\boldsymbol{\zeta}_A^{\text{T}}\tilde{\mathbf{T}}_\zeta^{(AB)}\,\boldsymbol{\zeta}_B = \frac{1}{2}\boldsymbol{\zeta}^{\text{T}}\tilde{\mathbf{T}}_\zeta\boldsymbol{\zeta}\,, \tag{5.17}$$

where in the second step we have contracted the range-separating function, excitation frequencies (η), polarizabilities and $\mathbf{T}_{AB}$ into $\tilde{\mathbf{T}}_\zeta$ and in the last step we used the definition of the generalized $3N$-dimensional coordinate $\boldsymbol{\zeta}$ as used above. A detailed derivation of the expectation value of the dipolar interaction potential is given in Appendix A6. The final dipole interaction energy between a set of QHOs is given by

$$E_{\text{dip}}[\Psi_{\text{DC}}] = \sum_i \frac{\left(\mathbf{C}^{\text{T}}\tilde{\mathbf{T}}_\zeta\mathbf{C}\right)_{ii}}{4\omega_i} = \sum_{i,k,l}\frac{\mathbf{C}_{ki}\,\tilde{\mathbf{T}}_\zeta^{(kl)}\mathbf{C}_{li}}{4\omega_i}\,, \tag{5.18}$$

where $\mathbf{C}$ again denotes the eigenvector matrix of the MBD potential matrix and ω_i are the MBD "eigenfrequencies". It should be mentioned that the expectation value of the dipole interaction potential based on the DC state is not equal to the MBD energy (even when using the same range-separation). This can be seen by re-writing

$$E_{\text{dip}}[\Psi_{\text{DC}}] = \langle\Psi_{\text{DC}}|\,\hat{\mathcal{V}}_{\text{dip}}\,|\Psi_{\text{DC}}\rangle = \langle\Psi_{\text{DC}}|\,\hat{\mathcal{H}}_{\text{DC}}\,|\Psi_{\text{DC}}\rangle - \langle\Psi_{\text{DC}}|\,\hat{\mathcal{H}}_0\,|\Psi_{\text{DC}}\rangle\,. \tag{5.19}$$

We know that $\hat{\mathcal{H}}_0$ has a non-degenerate ground-state, $|\Psi_0\rangle$. Moreover, we know $\langle\Psi_{\text{DC}}|\Psi_{\text{DC}}\rangle = \langle\Psi_0|\Psi_0\rangle$ and $|\Psi_{\text{DC}}\rangle \neq |\Psi_0\rangle$ and that both represent proper physical wave functions. Therefore, the variational principle tells us

$$\langle\Psi_{\text{DC}}|\,\hat{\mathcal{H}}_0\,|\Psi_{\text{DC}}\rangle > \langle\Psi_0|\,\hat{\mathcal{H}}_0\,|\Psi_0\rangle = E_0 \tag{5.20}$$

$$\Longrightarrow \qquad \langle\Psi_{\text{DC}}|\,\hat{\mathcal{V}}_{\text{dip}}\,|\Psi_{\text{DC}}\rangle < \langle\Psi_{\text{DC}}|\,\hat{\mathcal{H}}_{\text{DC}}\,|\Psi_{\text{DC}}\rangle - E_0 = E_{\text{MBD}}\,. \tag{5.21}$$

Taking into account that E_{MBD} is negative (attractive), this means that the contribution of the dipolar potential energy to the interaction is more attractive than the MBD energy. This can be explained as the introduction of the interaction potential in the Hamiltonian leads to an increase in the kinetic and on-site (harmonic) potential energy contributions, which is not accounted for by the bare expectation value of the interaction potential.

5.1.3 Conceptual Interpretation and Separation of Dipole-Correlated Coulomb Singles

Typically, the electronic Coulomb energy is divided into its classical and correlation parts in electronic-structure theory. Likewise, we also divide the DCS energy into its classical component, $J[\rho]$, and the correlation energy, $E_{\text{corr}}[\Psi]$. Since $E_{\text{corr}}[\Psi_0] = 0$ and $E_{\text{dip}}[\Psi_0] = 0$, the first-order full Coulomb contribution can be written as

$$E_{\text{DCS}} = \left(J[\rho_{\text{DC}}] - J[\rho_0]\right) + \left(E_{\text{corr}}[\Psi_{\text{DC}}] - E_{\text{dip}}[\Psi_{\text{DC}}]\right). \tag{5.22}$$

The first term on the right is the change in the electrostatic energy of the oscillators caused by charge density polarization, which is itself induced by vdW dispersion, and the second term is a Coulomb correction to the dipole correlation energy. To first order in $\Delta\rho = \rho_{\text{DC}} - \rho_0$, the first term can be expressed as $J[\rho_0, \Delta\rho]$, *i.e.*, the electrostatic interaction energy of the non-interacting oscillator densities with the density polarization induced by the dispersion. E_{DCS} can therefore also be seen as a dispersion-polarization interaction plus the beyond-dipolar correlation energy in the coupled oscillator model. The beyond-dipolar contribution is thereby evaluated on the density

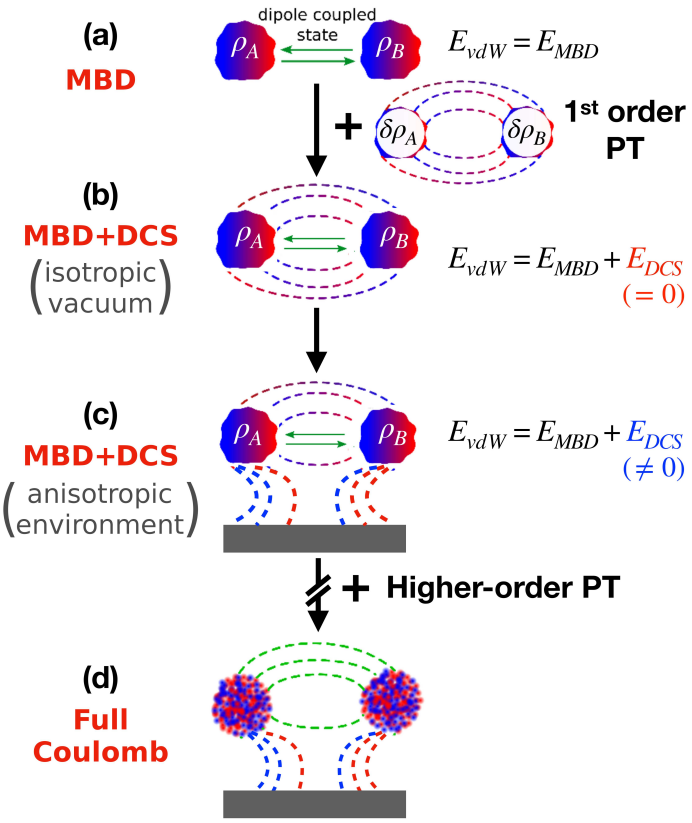

Fig. 5.1: Schematic representation of *Dipole-Correlated Coulomb Singles*: (a) Green arrows represent dipole coupling between electronic fragments. First-order perturbation theory (PT) captures the interaction energy, E_{DCS}, between $\delta\rho_{\mathrm{A}}$ and $\delta\rho_{\mathrm{B}}$, depicted by field lines. (b) E_{DCS} vanishes in 3D isotropic vacuum because of symmetry. (c) Under rotational symmetry-breaking confinement, electric field-lines between electronic fragments deform, which leads to $E_{\mathrm{DCS}} \neq 0$. (d) Further inclusion of higher-order terms leads to full Coulomb-coupled vdW interaction.

of dipolar quantum fluctuations and thus in contrast to conventional higher-order multipolar interatomic vdW interactions such as C_8/R^8-terms, for example. Fig. 5.1 provides a schematic representation of the DCS energy. We note that in the presence of a polarizable environment, the DCS interaction between two bodies can decay asymptotically slower than the zeroth-order MBD energy. For example, the dipole correlation in the MBD ground-state wave function between a *finite* body and its environment induces permanent quadrupole moments on the oscillators, causing the resulting interaction to decay as R^{-5}. In contrast, the MBD interaction energy between two bodies in such a system (including the environment) decays asymptotically as R^{-6}. Hence, the DCS term can be of leading order and comparable to or even more important than the renowned contribution from higher-order atomic multipoles [25] to the vdW dispersion energy.

5.1.4 Dipole-Correlated Coulomb Singles and Current Electronic-Structure Methods

From the above discussion it is evident that, in order to fully capture E_{DCS}, an electronic-structure method has to describe its dispersion-polarization-like term and correlation beyond interatomic dipole–dipole interactions (see Eq. 5.22). In the language of coupled-cluster theory, the former requires a self-consistent coupling between singles and doubles. As such, CCSD and beyond captures both components. Only treating doubles amplitudes as in CCD or perturbative treatment of doubles does not. QMC in principle provides a full solution of the many-electron SE and thus fully includes DCS. Also the quantum Drude oscillator model with Coulomb coupling [185,201,202] captures E_{DCS}. Symmetry-adapted perturbation theory includes the correlation component of DCS for intermolecular interaction, but dispersion–polarization contributions only appear beyond the typical limitation to second order. Given that the polarization induced by long-range correlation leads to a delocalization of electron density, [35,106] all of the above require sufficiently large basis sets, which further increases their already high computational costs. From the approximate electronic-structure methods applicable to larger systems, ordinary RPA captures the full Coulomb interaction, but neglects the singles-like effect of the long-range electron correlation on the one-electron orbitals. This can further be seen from the equivalency of RPA and CCD within

ring-diagram approximation.[296] In second-order Møller-Plesset perturbation theory (MP2), the effect of long-range correlation on the wave function is not reflected in the energy and thus MP2 does not cover E_{DCS}. Evaluating singles(-like) contributions on top of a long-range correlated wave function as obtained from MP2, however, does allow to recover E_{DCS}. Accounting for single excitation contributions within RPA as presented by Ren and co-workers,[297] on the other side, does not as it is based on a (mean-field) DFT wave function. Conventional (semi-)local electronic-structure methods neglect long-range vdW interactions entirely and the standard *post*-DFT(B) vdW approaches, including those that go beyond interatomic dipole–dipole interactions, such as Grimme's D3[49] or the XDM approach,[298–301] do not account for E_{DCS}. Incorporating vdW functionals into DFT(B) in a fully self-consistent fashion[106] may recover the dispersion-polarization contribution of E_{DCS}, but will not capture the full Coulomb component, complementary to RPA.

5.2 Dipole-Correlated Coulomb Singles in Realistic Systems

The accurate prediction of interaction energies for non-covalently bound molecules and materials is an ongoing challenge for researchers and workhorse DFT methods are at the forefront of development efforts. The complex balance of intermolecular interactions is particularly challenging to predict and what is more, the target accuracy is generally in the range of a few kJ/mol in interaction energies. Our approach to better accuracy is to improve the physical basis of theoretical methods — here, by incorporating the DCS contributions. First, we combine DCS with MBD to compute interaction energies of small molecular dimers. Following this, MBD+DCS is used to compute the binding energies of supramolecular host–guest complexes and confined Xe dimers in CNTs. These more complex systems reveal the impact and importance of DCS, as well as the length scale and character of the emergent changes to long-range interactions.

5.2.1 Computational Details

In accordance with Eq. (5.4), the DCS contribution to the vdW energy can be calculated from the "beyond-dipolar potential" and the wave functions of DC and uncoupled quantum (Drude) oscillators, respectively. These wave functions can be obtained directly from solving the MBD Hamiltonian (5.1), see above and Sec. 2.3.5. The range-separating function for E_{DCS} has been chosen to follow the same Fermi-like functional form as in MBD, see Eq. (2.63), where the parameters a and β have been set to 10.12 and 1.4, respectively. The ratio of the chosen parameters is thereby the same as for the parameters used for MBD. Keeping this ratio constant assures the same switching behavior and avoids skewing of the range-separating function, see Fig. 5.2. This choice of the parameters provided robust results for all systems studied in this work. Optimal tuning of the range-separation, however, requires an increased availability of reference data, which accurately includes the effects covered by DCS. Calculation of E_{DCS} was carried out within the libMBD software package.[243]

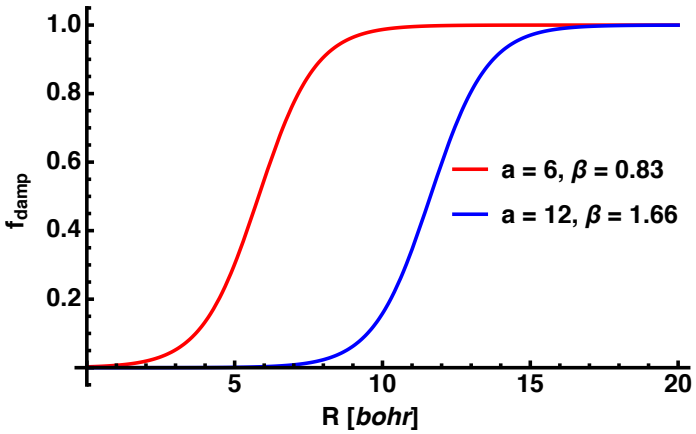

Fig. 5.2: Shift of original MBD range-separating function (red) by keeping a fixed ratio of range-separating parameters. For *Dipole-Correlated Coulomb Singles* a shift to larger distances using a = 10.12 and β = 1.4 was used.

All DFT calculations have been carried out within the FHI-aims package using a basis set of numerical atomic orbitals.[302] For the calculations employing the PBE0 hybrid functional, highly-converged results at the "really tight" level of settings (meV-level of convergence in energies, see Ref. 302) have been extrapolated based on PBE0 with "tight" settings and PBE results with "tight" and "really tight" settings. While significantly reducing computational costs, this scheme has been proven to provide an excellent estimate for PBE0 at the "really tight" level.[232] The same extrapolation scheme was used to correct the atomic Hirshfeld volumes, which form the basis for all MBD and DCS calculations. The results reported for Xe inside CNTs were obtained using the PBE functional with the "tight" level of settings.

5.2.2 Dipole-Correlated Coulomb Singles in small Molecular Dimers

We begin by applying the first-order perturbation term (5.4) to the S66 data set[303] of small, unconfined molecular dimers. For such systems, semi-local or hybrid density-functional approximations in conjunction with the MBD formalism provide excellent agreement with accurate reference results from coupled-cluster calculations with single, double, and perturbative triple excitations (CCSD(T)).[50] The S66 set contains non-covalently interacting dimers in 3D isotropic vacuum and in accordance with Fig. 5.1 we expect minuscule first-order Coulomb corrections in this case. Indeed, we find that the DCS contributions for all systems in S66 are very small and they can have both positive as well as negative values (see Fig. 5.3). As a result, the accuracy of vdW-inclusive DFT remains equally good upon account for DCS contributions to the interaction energies of small molecular dimers as contained in the S66 database.

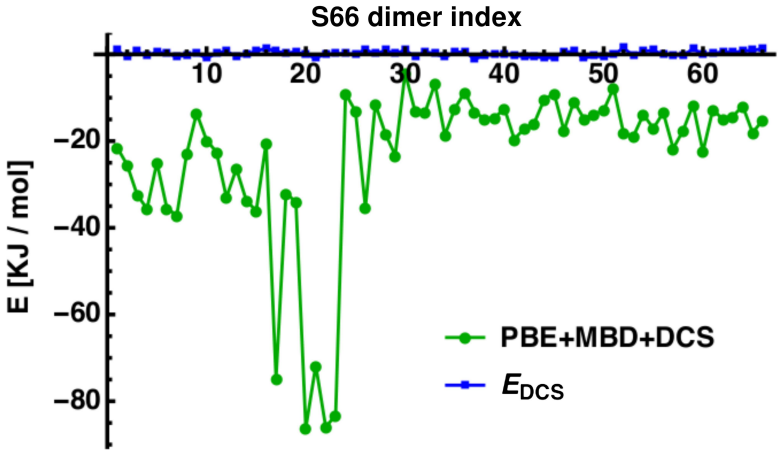

Fig. 5.3: Binding energies of S66 dimers and *Dipole-Correlated Coulomb Singles* contributions. PBE0+MBD+DCS results (green) and bare *Dipole-Correlated Coulomb Singles* contribution (blue) for all dimers contained in the S66 dataset.

5.2.3 Coulomb Corrections for Host–Guest Complexes

Host–guest molecular systems are significantly more complex than the S66 dimers, but are still tractable with accurate benchmark methods such as diffusion quantum Monte-Carlo (DQMC). Here, we first ascertain the impact of DCS on the binding energies of host–guest complexes and demonstrate the accuracy achieved by including DCS contributions into the PBE0+MBD approach in comparison to DQMC reference interaction energies from previous works.

Fig. 5.4 shows two examples of host–guest complexes. For both systems, the guest molecule is a C_{70}-fullerene (buckyball) while the host is either [6]-cycloparaphenyleneacetylene (6-CPPA, Fig. 5.4 left) or the "catcher" molecule (Fig. 5.4 right).[304] In the framework of Fig. 5.1, the host molecule serves as both confinement and the interaction partner. The 6-CPPA and catcher molecules provide a different confining environment for the buckyball by virtue of their geometry. We therefore focus on these systems to showcase the contribution of E_{DCS}. Within the dipole approximation, it has already been shown that many-body effects play an important role for the description of binding energies in such host–guest complexes.[33,35] Here, we show that also

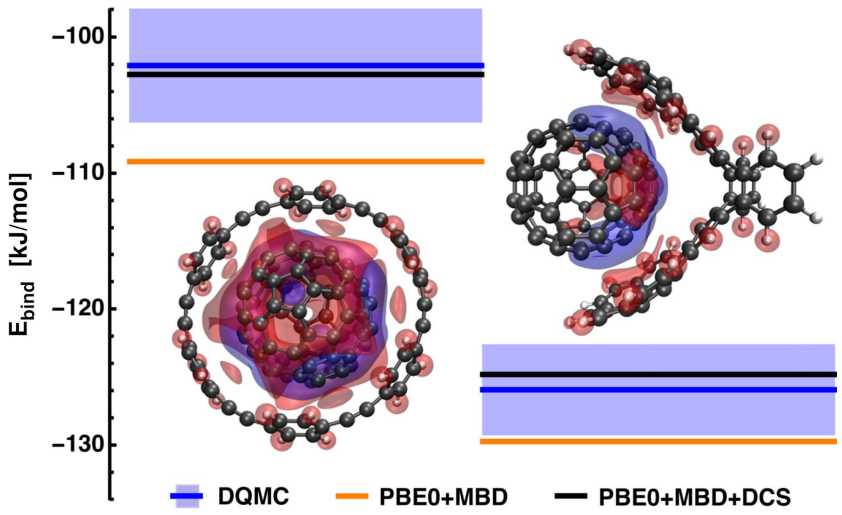

Fig. 5.4: Binding energies of C_{70} in 6-CPPA (left) and in the "buckyball-catcher" (right): PBE0+MBD results (orange), DQMC reference (blue line, error bars shown as boxes), PBE0+MBD including *Dipole-Correlated Coulomb Singles* contribution (black line). DQMC reference data were taken from Ref. 35. The depiction of the complexes includes isosurfaces at ±0.003 (a.u.) of the change in the density of electronic fluctuations with respect to the isolated monomers (red: decrease, blue: increase).

beyond-dipole interactions, in the form of DCS, have a significant effect, and that inclusion of DCS to PBE0+MBD yields excellent agreement with the reference results from DQMC. The DCS contribution for C_{70} in 6-CPPA and in the buckyball-catcher is 6.4 kJ/mol and 4.9 kJ/mol, respectively. Considering our findings for the S66 data set above, this clearly highlights the importance of DCS corrections once the 3D isotropy of vacuum is substantially perturbed.

The relative contribution of DCS to the total binding energy for C_{70} in 6-CPPA and in the buckyball-catcher is 6.2% and 3.9%, respectively. As can be seen from Fig. 5.4 (and more clearly from Fig. 5.5), the DCS contribution does not correlate with the system size nor the vdW interaction within the dipolar approximation. However, the different E_{DCS} contributions can be interpreted in terms of the physics described in the DC state, which represents the starting point (unperturbed state) for the calculation of the DCS contribution by means of perturbation theory. One important factor is how the dipolar coupling changes the density of electronic fluctuations (*i.e.*, $\delta\rho$ shown in Fig. 5.1), which can be obtained as the expectation value of the charge density operator acting on the MBD wave function, Ψ_{DC}, of the DC and uncoupled set of QHOs. Analysis of the difference of $\delta\rho$ in the host–guest interaction with respect to the isolated monomers gives us a measure of how much the density deforms upon the host–guest interaction and, therefore, forms a connection between confinement and electronic fluctuations and polarizability. The density difference shown in Fig. 5.4 shows that, upon dipole coupling, the density of electronic fluctuations for C_{70} in 6-CPPA is more strongly deformed into the plane of 6-CPPA. Furthermore, we find that the overall displaced charge, *i.e.*, the integral over the absolute value of the density difference, can serve as a qualitative descriptor for the DCS contribution to the interaction energy: With increasing displaced charge, we observe an increased DCS contribution to the interaction energy. We point out that this also applies to the systems considered below. Hence, the electronic properties obtained for the DC state can serve as a qualitative rule-of-thumb to estimate relative magnitudes of E_{DCS}.

5.2.4 Relation to Asymmetry and Steric Effects

To explore the connection between E_{DCS} and confinement, we analyze a set of geometrically similar ring–C_{70} complexes depicted in Fig. 5.5: The four complexes are C_{70} hosted by four different conformations of 8-CPPA. In a previous work, PBE0+MBD has been shown to provide reasonably accurate binding energies with respect to DQMC.[35] As can be seen from Fig. 5.5A, the addition of the DCS contribution to PBE0+MBD further improves the binding energies of all four complexes. However, the individual DCS contributions vary significantly across these conformations (see relative E_{DCS} shown as f_e in Fig. 5.5B).

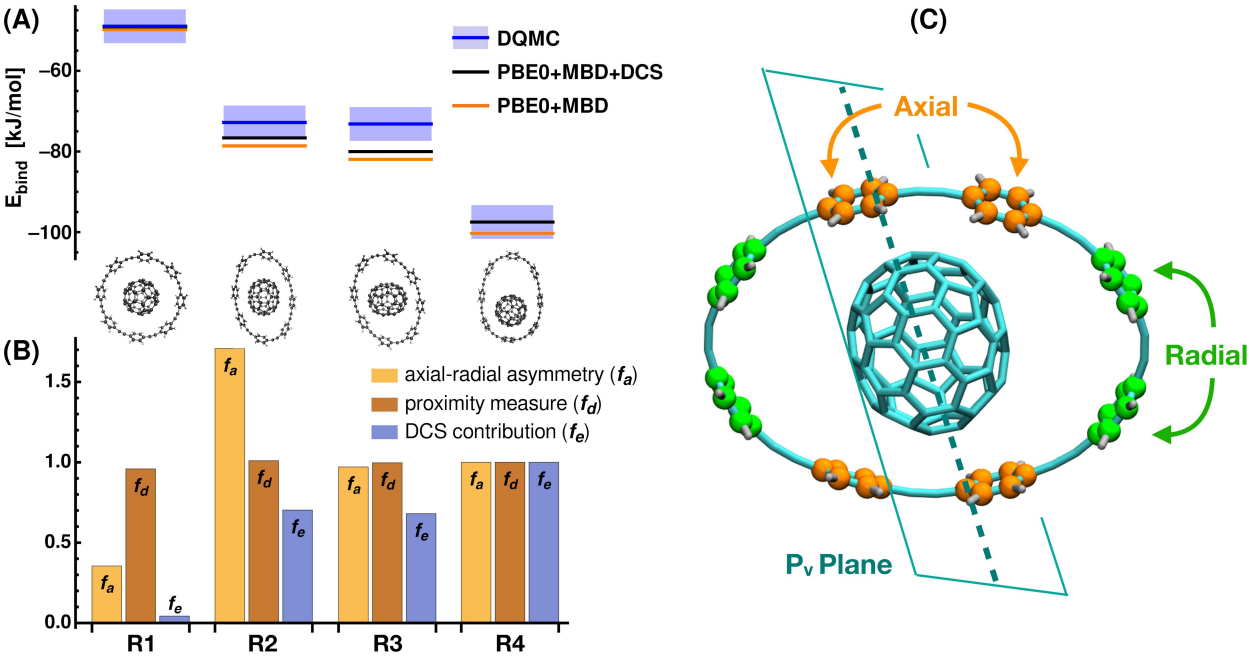

Fig. 5.5: *Dipole-Correlated Coulomb Singles* contributions to binding energies of ring–C_{70} complexes and correlation to structural features. (A) Binding energies for ring–C_{70} host–guest complexes (R1–R4): PBE0+MBD (orange), DQMC reference (blue line, error bars shown as boxes), PBE0+MBD with *Dipole-Correlated Coulomb Singles* contribution (black). DQMC reference data taken from Ref. 35. The hosts for R1–R4 are 8-CPPA rings. (B) Measure of axial-radial asymmetry (f_a) and proximity (f_d), and *Dipole-Correlated Coulomb Singles* contribution (f_e) to binding energy (all values normalized to the results for R4). Definition of "axial" and "radial" phenyl units of 8-CPPA by the P_v plane shown in (C).

In order to understand the potential relation between E_{DCS} and asymmetry or steric effects, we define two geometrical measures: One for proximity (f_d), which is given by the sum of inverse distances between the atoms of the fullerene guest molecule and the CPPA-host, and one for the asymmetry of the system (f_a). For the latter, we define a plane along the elongated axis of C_{70} that is perpendicular to the CPPA ring (labeled P_v plane in Fig. 5.5C). The four phenyl units closest to this plane are considered "axial" and the remaining four "radial". Based on this classification, we define "axial vicinity" ($A_\parallel$) and "radial vicinity" ($A_\perp$) by summing the inverse distances between all fullerene atoms and atoms of the "axial" and "radial" phenyl rings, respectively. Our measure of (axial-radial) asymmetry is then given by $f_a = (A_\parallel - A_\perp)/(A_\parallel + A_\perp)$. Fig. 5.5B summarizes the results for the proximity and asymmetry measures and the ratio of E_{DCS} of each system and that of R4 ($f_e = E_{\mathrm{DCS}}(R_i)/E_{\mathrm{DCS}}(R_4)$, $i = \{1,2,3,4\}$). It is clear that, in principle, proximity plays a role in electronic confinement (*cf.* proximity and DCS contributions for C_{70} in 6-CPPA and R1). As can be seen from the detailed analysis in Fig. 5.5B, however, purely geometric considerations do not correspond directly to the trends in E_{DCS}. First, the proximity measure, f_d, is insensitive to the different confining environments and remains almost constant among all four conformations, whereas the DCS contribution varies significantly. Furthermore, the asymmetry measure, f_a, has no correlation with E_{DCS} (see R2 versus R3 and R4 in Fig. 5.5B). Thus, also a pairwise description of asymmetry between atomic positions is insufficient to predict the qualitative trend of the contribution of DCS.

The failure to capture the behavior of E_{DCS} in terms of simple geometric characteristics stems from the fact that the DCS contribution is a quantum-mechanical effect arising from long-range electron correlation, which shows a non-trivial dependence on the geometrical features of a system. A considerable part of E_{DCS} represents charge polarization effects due to long-range electron correlation, *cf.* Eq. (5.22). As discussed for the previous complexes, the displaced charge within MBD (as depicted in Fig. 5.4) can provide a measure for the dispersion-polarization-like term and

indeed tracks the qualitative trend in the relative DCS interaction energies for all supramolecular complexes treated here. The best geometry-based metric found here is a sum of inverse distances to the power of five, which resembles an interaction of quadrupoles induced by long-range correlation (*vide supra*). Further information on qualitative descriptors can be found in Appendix A9.

5.2.5 Impact of Dipole-Correlated Coulomb Singles on vdW Interaction inside CNTs

Having established the importance of E_{DCS} for confined host–guest systems, we now employ the developed methodology to investigate the effects of confinement for a Xe dimer inside CNTs. In particular, we answer the question how the presence and strength of confinement changes the relative importance of the DCS correction to dipolar vdW interactions. Xe does not possess permanent multipoles and has a substantial polarizability. As a result, the long-range Xe–Xe interaction has pure vdW character. CNTs can be generally classified according to their chiral indices (n_1, n_2), where armchair CNTs with $n_1 = n_2$ are metallic in nature. Since the DCS contribution becomes especially important in the presence of metallic screening,[269] we analyze the Xe–Xe interaction inside two armchair, hydrogen-capped CNTs with $n_1 = n_2 = 5$ and $n_1 = n_2 = 6$. The diameters of the (5,5)- and (6,6)-CNTs are 6.78 Å and 8.13 Å, respectively. The length of each nanotube was chosen to be 30 Å, which is sufficient to avoid any significant edge effects. The binding energies of the Xe dimer inside the nanotubes are calculated as

$$E_{\mathrm{int}} = E_{\mathrm{Xe_2@NT}} + E_{\mathrm{CNT}} - E_{\mathrm{Xe_A@NT}} - E_{\mathrm{Xe_B@NT}} \,, \tag{5.23}$$

where $E_{\mathrm{Xe_2@NT}}$, $E_{\mathrm{Xe_{A/B}@NT}}$, and E_{CNT} are the energies of Xe_2 inside the CNT, the two single Xe atoms hosted by the nanotube and the bare CNT, respectively. Fig. 5.6 summarizes the effect of the confining potential of capped CNTs on the Xe dimer binding energy. We focus on the variation of the DC MBD and the corresponding DCS contributions as a function of the inter-Xe distance, R. In Fig. 5.6A, we show the effect of confinement on the individual contributions by comparing a Xe dimer inside the (6,6)-CNT and in gas phase. One clearly sees that both E_{MBD} as well as E_{DCS} become less attractive due to confinement. In the case of the MBD interaction energy this can be attributed to (i) decreased Xe polarizabilities due to the screening by the CNT and (ii) the restriction of electronic fluctuations on the Xe atoms due to correlation with fluctuations in the CNT. This reduction of the vdW interaction as a result of many-body correlation has been observed and detailed in a number of previous works.[2,33,35,50,305] It is notable that the bare presence of the confinement affects E_{MBD} more strongly than the DCS interaction energy. Fig. 5.6B then shows the MBD and DCS components for the two CNTs with different radii. In contrast to the bare presence of confinement, the type and strength of the confinement, as represented by the different nanotubes, has a larger effect on the DCS interaction than on the MBD contribution. We also note that the effect of the different environment on the MBD interaction is negligible after 6 Å and the binding curves follow the same behavior, while the effect is more long-ranged for E_{DCS}. DCS are more sensitive to the characteristics of the confinement compared to E_{MBD}. As expected, the destabilization of the Xe dimer due to screening and many-body correlation effects is less pronounced inside the (6,6)-CNT.

Fig. 5.6C shows the cumulative vdW binding energy, $E_{\mathrm{MBD}} + E_{\mathrm{DCS}}$. In total, confinement in a metallic CNT leads to a substantial decrease of the Xe–Xe vdW interaction. Mostly due to the DCS contribution, this destabilization is strongly dependent on the confining environment. This shows that the total long-range vdW interaction can be in fact substantially altered by (nano-)confinement, whereas the bare MBD treatment would predict the environment to have no effect beyond interatomic distances of 6 Å. For Xe_2 inside a (5,5)-CNT, the interplay of the repulsive DCS contribution and the attractive MBD interaction interestingly leads to a near-linear behavior for separations of 6 Å to ~8 Å. To explore the balance between the repulsive E_{DCS} and the attractive E_{MBD} more clearly, we show the absolute value of their ratio as a function of R in the inset of Fig. 5.6C. In all cases, the ratio, *i.e.*, the relative importance of the DCS contribution, increases

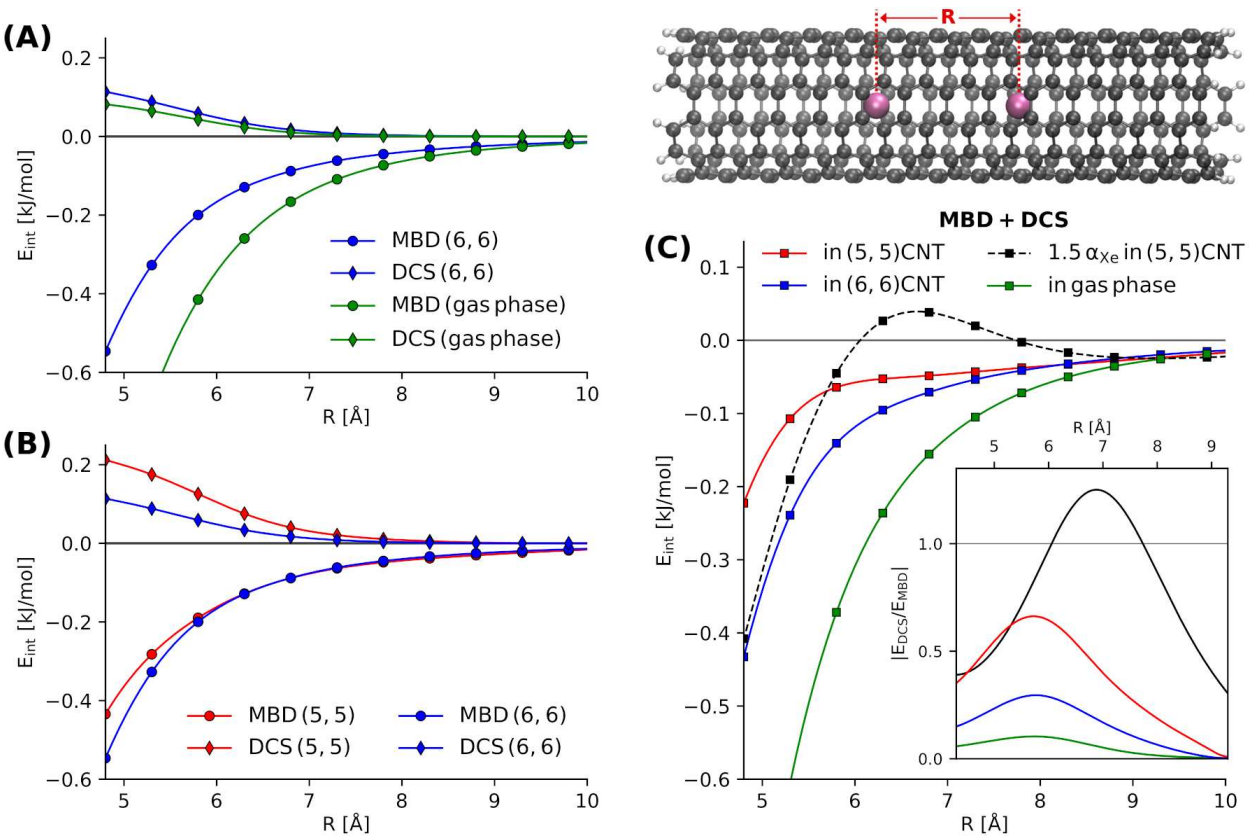

Fig. 5.6: MBD and *Dipole-Correlated Coulomb Singles* interaction inside carbon nanotubes. (A) MBD and DCS contributions inside a (6,6)-CNT and in gas phase as a function of the Xe–Xe separation. (B) Effect of the different confinements of a (5,5)- and (6,6)-CNT on E_{MBD} and E_{DCS}. (C) Two Xe atoms encapsulated in a CNT. Total van der Waals interaction energy given by the sum of E_{MBD} and E_{DCS} with results for increasing the Xe polarizability by 50 % (black). Inset: Absolute value of the ratio between E_{DCS} and E_{MBD}.

with larger inter-Xe distances and reaches a maximum around ~5.8 Å before converging to zero. The interatomic distance at which one observes the maximum is surprisingly independent of the presence and strength of the confinement. The ratio of E_{DCS} and E_{MBD} and its maximum value, on the other side, strongly depend on the environment of the interacting particles. In order to highlight the critical role played by the response properties of the objects interacting under confinement, we have performed the analysis for Xe_2 inside a (5,5)-CNT while increasing the Xe polarizability by 50% (black dashed curve). The results indicate a pivotal role of the polarizability in the total vdW interaction in confined systems: At shorter interatomic distances, the overall interaction is increased as the attractive MBD contribution is affected more strongly. In the very long-range limit the interaction converges to the same behavior as for "normal" Xe_2. In the intermediate region, however, the repulsive contribution from DCS increases more strongly than its attractive MBD counterpart, which leads to a substantial destabilization and eventually repulsive interaction energy. The interplay between E_{DCS} and E_{MBD} in this intermediate region gives rise to a maximum followed by a very shallow minimum in the binding curve creating a small barrier of ~0.1 kJ/mol in the binding curve. All this can be explained by a much higher sensitivity of E_{DCS} to the Xe polarizability compared to the MBD interaction energy. Accordingly, changing the polarizability has a strong effect on the ratio of E_{DCS} and E_{MBD} and its maximum (see Fig. 5.6C). The ratio surpasses one, meaning that E_{DCS} supersedes the MBD contribution in magnitude and introduces a region of repulsive interaction around 7 Å. The position of the maximum is thereby increased by almost 1 Å. Altogether, we can conclude that with increasing polarizability, the relative importance of the DCS contribution increases, becomes more long-ranged and can lead to non-trivial qualitative changes in the overall vdW interaction.

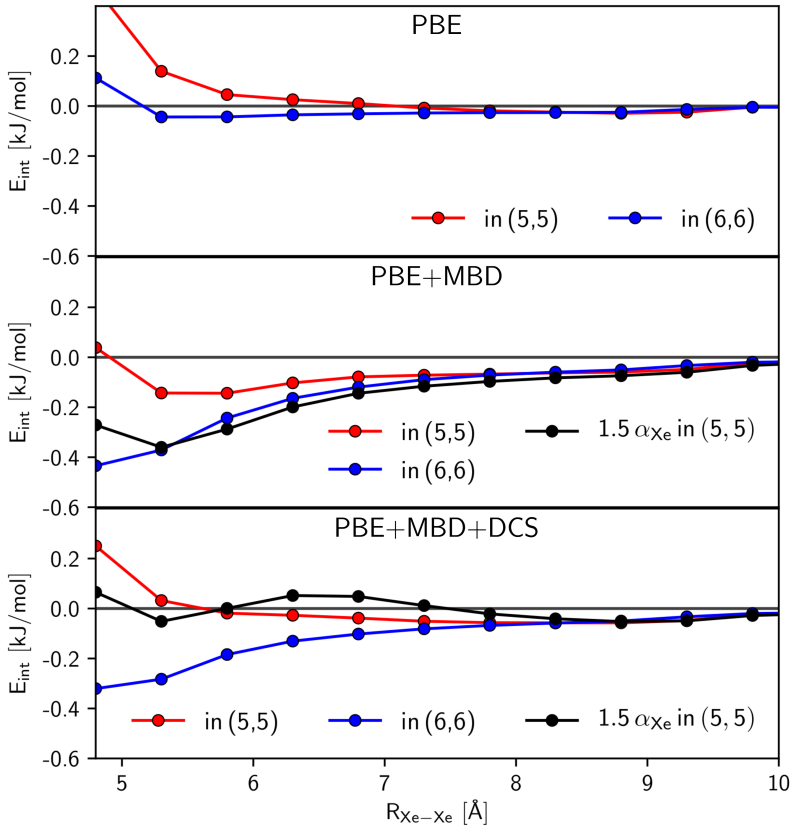

Fig. 5.7: Total Xe–Xe interaction energies in carbon nanotubes. (A) bare PBE-DFT interaction energies, (B) PBE+MBD energetics, (C) PBE+MBD with account for Dipole-Correlated Coulomb Singles.

For the considered system of Xe_2 inside CNTs, the vdW interaction thereby fully governs the total long-range interaction. The corresponding PBE-DFT interaction energies are of negligible magnitude at inter-Xe distances beyond ~6 Å and only introduce the well-known repulsive contributions at shorter separations. As a result, the qualitative changes due to DCS reported above remain unaltered by inclusion of contributions captured in semi-local DFT. One particularly interesting aspect of the total PBE+MBD+DCS interaction is that PBE contributions together with DCS cancel out the meta-stable state of Xe 2 in the (5,5)-CNT. On the level of PBE+MBD, one observes a local minimum around 5.5 Å, while PBE+MBD+DCS predicts only repulsive or negligible interaction at all distances. The corresponding PBE-DFT and total interaction energies are summarized in Fig. 5.7.

5.3 Conclusion of Many-Body Dispersion beyond the RPA

This chapter introduces *Dipole-Correlated Coulomb Singles* (DCS) as a distinct component of the interaction energy whose description is missing in state-of-the-art vdW-inclusive DFT(B). An explicit model based on the MBD framework is developed to account for DCS and the resulting (interaction) energy contributions are demonstrated to have a significant effect on vdW interactions in supramolecular systems and under nano-confinement. There are three main reasons why E_{DCS} has not been addressed before. First, the resolution and accuracy of experimental setups have not been sufficient to reveal the unusual behavior arising from E_{DCS} at the nano- to microscale. For example, the nano-fluidic techniques and manufacturing of nanotubes with desired properties, which helped reveal the phenomenon of accelerated water flow through carbon nanotubes, have become available only recently. Second, the prevalent conception about the universality of conventional long-range attraction between polarizable moieties has subdued explanations of the observed experimental phenomena that would accommodate long-range repulsive vdW contributions. Third, while *ab initio* electronic-structure methods such as coupled-cluster theory or QMC inherently describe DCS, the prohibitively high computational cost of such methods for larger systems did not allow for the fine analysis as enabled by the here-presented

efficient approach. The computational costs of the presented DCS formalism without approximations scale with the fifth power of the number of atoms. However, this is accompanied by a very small prefactor and as a result the computation of DCS produces negligible additional costs to semi-local or hybrid DFT calculations for systems of up to several hundred atoms. The present formalism further solely relies on the MBD wave function, which in turn is based on the definition of atomic polarizabilities within a molecule or material. So, the DCS formalism could equally-well be included in force field calculations as presented previously for the MBD model.[306] While the remaining computational costs limit its application in molecular dynamics simulations, DCS can be used to improve the description of structural ensembles via energy reweighting. In addition to such a posteriori corrections, the DCS formalism enables the determination of improved effective interatomic potentials for complex systems.

In summary, while a continuous treatment of DC oscillators at every point in space can essentially describe any response allowed by quantum field theory, a coarse-grained formalism of atomic response (like the MBD approach) has to account for beyond-dipolar couplings. The DCS formalism represents a consistent, unified methodology to incorporate this previously neglected part of the coupling between instantaneous electronic fluctuations within a quantum-mechanical many-body treatment of vdW interactions. It is shown that the inclusion of this contribution becomes significant for relatively larger molecular systems and can even change the qualitative nature of (long-range) intermolecular interactions. The negligible computational cost of the present methodology compared to benchmark electronic-structure methods allows to explore the emergent role of beyond-RPA, beyond-pairwise vdW interactions in large-scale systems. The surprising results for the interaction of a Xe_2 under the confinement of CNTs (Figs. 5.6 and 5.7) suggest a possible explanation for the high flow rate of water through nanotubes, by way of reducing the long-range vdW interaction and hence decreasing the density and microscopic "viscosity" inside nanostructured environments. Careful study of the mutual interplay of such effects as well as further accurate and well-defined reference data from methods that incorporate DCS may be necessary to fully explain such puzzling effects under nanoscale confinement, but the present work is a promising, first step in this direction.

CHAPTER 6

Machine-Assisted Density-Functional Tight-Binding

Parts of Sec. 6.1 and 6.2 have been reproduced with permission from
M. Stöhr, L. Medrano Sandonas, and A. Tkatchenko, *J. Phys. Chem. Lett.* **11**, 6835 (2020).
The data and figures in 6.2 have been produced in collaboration with Dr L. Medrano Sandonas.

The previous chapters highlight the potential role of collective long-range vdW forces and the wavelike behaviors of electrons in biomolecular and other nanoscale systems beyond the incomplete state-of-the-art treatment *via* phenomenological pair-potentials. As pointed out in the discussions of the previous chapters, the dynamics and ultimate importance of the above interactions is given by the free energy landscape of the system, which in turn also strongly depends on the vibrational/phononic properties (in form of zero-point energy and vibrational entropy contributions). From previous studies [31,213] it can be expected that the role of many-body vdW forces gets even more emphasized when considering vibrational/phononic contributions and nuclear relaxation. This can be explained by the inherently non-local and collective character of many-body dispersion, which mostly affects more collective, lower-frequency (or "slower") nuclear coordinates rather than local structural changes.

In order to describe vibrational/phononic properties and the free energy landscape, we need to go beyond the bare study of vdW forces and require an accurate and robust total energy method, which can account for the abovementioned phenomena. The main workhorse method in that regard is DFT. To gain the desired microscopic understanding of biomolecular and other nanoscale systems, current DFT approaches are facing considerable limitations, however. Despite the ever-growing availability of computational resources and high-performance implementations, they are still limited in terms of tractable system sizes due to the associated computational workload. Here, an intermediate level of theory as represented by semi-empirical QM methods can offer a promising alternative. Semi-empirical methods include an explicit QM treatment of electrons and can be combined with accurate (many-body) vdW models [142] while requiring only a fraction of the computational cost of DFT.

6.1 Hybrid QM/ML as Next-Generation Semi-Empirical Methods

Semi-empirical methods usually involve the formulation of an approximate, first-principles-based QM Hamiltonian together with an empirical "repulsive energy".[70,307] The definition of this repulsive energy can vary from method to method, but in all cases it is (considered to be) a more local contribution to the total energy, while the QM Hamiltonian to a certain degree accounts for non-locality and QM phenomena such as long-range electrostatics or charge transfer. As an example we will here treat the DFTB formalism[41–43] as used to parameterize *ab initio* vdW models in Chapters 3 & 4. As detailed in Sec. 2.2, DFTB is based on DFT using a superposition of (confined) atomic densities, $\rho_0 = \sum \rho_A$, extended to allow for the redistribution of electrons by expanding the density functional around the reference density ρ_0 in terms of changes in the electron distribution.[42–44] The resulting energy functional can then be written as a sum of an electronic energy, $E^{(\mathrm{el})}_{\mathrm{DFTB}}$, and the repulsive energy, E_{rep}. The electronic part of the total DFTB energy is evaluated from a tight-binding Hamiltonian parameterized through reference DFT calculations of (confined) atoms and diatomic molecules, which limits the amount of parameters for the electronic DFTB energy to a minimum.[44] The repulsive energy, on the other side, is usually obtained in a much less straightforward manner. As given in Sec. 2.2.4, it is formally defined as,

$$E_{\mathrm{rep}}[\rho_0] = E_{\mathrm{xc}}[\rho_0] - \int \mathcal{V}_{\mathrm{xc}}[\rho_0]\,\rho_0\,d\mathbf{r} - \frac{1}{2}\int \mathcal{V}_{\mathrm{H}}[\rho_0']\,\rho_0\,d\mathbf{r} + E_{\mathrm{nuc}}\,, \tag{6.1}$$

with $\rho \equiv \rho(\mathbf{r})$ and $\rho' \equiv \rho(\mathbf{r}')$ as in Sec. 2.2 and E_{xc} being the xc energy, $\mathcal{V}_{\mathrm{xc}}$ the corresponding potential, E_{nuc} the (effective) nuclear repulsion energy, and $\mathcal{V}_{\mathrm{H}}$ the Hartree potential.[44,47] As it only depends on the reference density, ρ_0, and atomic positions, E_{rep} can further be considered a more local property (see also Fig. 6.1). In lieu of the above formal definition, E_{rep} is in practice represented by (short-ranged) atom-pairwise potentials fitted to DFT reference calculations. So,

$$E_{\mathrm{rep}} = E_{\mathrm{DFT}} - E^{(\mathrm{el})}_{\mathrm{DFTB}} \qquad \text{and} \qquad \mathbf{F}_{\mathrm{rep}} = \mathbf{F}_{\mathrm{DFT}} - \mathbf{F}^{(\mathrm{el})}_{\mathrm{DFTB}}\,. \tag{6.2}$$

The parameterization of repulsive potentials thus represents a complex multidimensional fitting problem of atomic pair-potentials to DFT reference data, which renders it the most intricate step in the development of DFTB parameterizations. Various (semi-)automated approaches have been proposed to tackle this task,[44,308–314] but in practice many cases still do require inevitably subjective manual adjustments. An optimal DFTB parameterization should finally provide access to accurate electronic as well as energetic and structural properties. Optimization of the parameters governing the electronic DFTB Hamiltonian is thereby often done separately based on electronic properties as obtained with DFT or other reference methods. The final performance for energetic and structural properties then largely depends on the repulsive energy.

To test the hypothesis of locality for the repulsive potential, we performed a vast sampling of the conformational space of two select hydrocarbons. The sampling was obtained from high-temperature MD simulations, where a region with varying radius around the central atom was fixed. Following the procedure suggested in Ref. 315, we then obtain the spread of the repulsive force on the central atom according to Eq. (6.2) as a function of the radius of the fixed region. This spread serves as locality measure for the repulsive force. Fig. 6.1 shows the locality measure for a 2 nm-sized graphene flake and a complex, flexible alkane (see insets). In both cases, the sampling of changes outside a fixed region of ~4.7 Å have a negligible effect of less than 1 kcal·mol^{-1}·Å^{-1} on the repulsive force on the central atom. The slower decay in the case of the alkane can be explained by the much increased flexibility and the resulting increased sampling of the conformational space compared to the graphene flake. From the above analysis, we can conclude that all effects relevant for the repulsive potential are included within a range of 5 Å, which confirms the rather local character of the repulsive energy and force. The local environment within 5 Å thereby includes the first three nearest neighbors of the central atom. The internal coordinates within this local environment are what we will refer to as local degrees of freedom in the following.

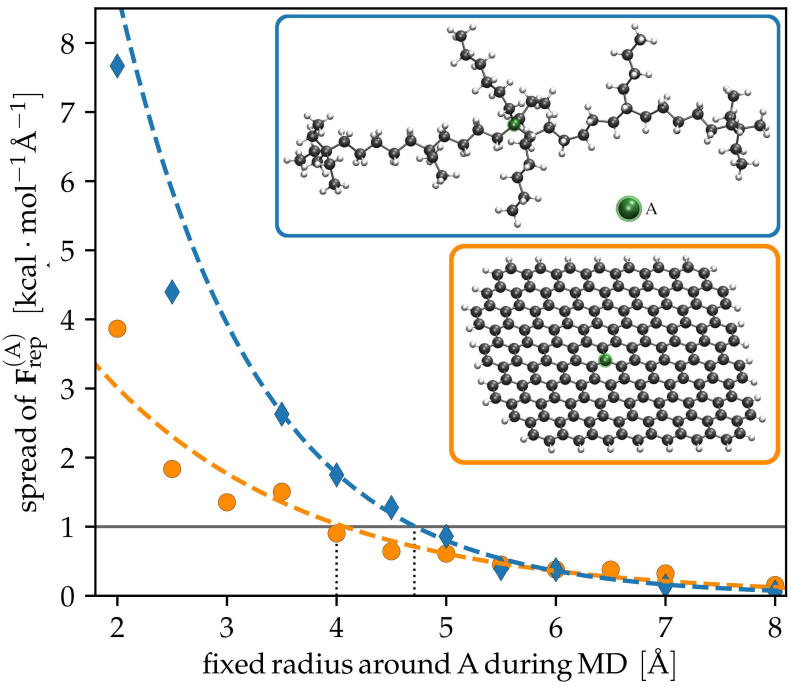

Fig. 6.1: Locality test of the repulsive force in a graphene flake and a complex, flexible alkane: The graphs show the spread in the repulsive force on the central atom as given by the difference of PBE0-DFT and electronic DFTB contributions obtained in constrained, high-temperature molecular dynamics simulations with a fixed region of varying radius around the central atom.

Fitting repulsive potentials according to Eq. (6.2) allows DFTB to often provide results at the DFT level of accuracy, but inevitably introduces empirical contributions to the repulsive potential to correct for the approximations in the remaining formalism. These contributions also introduce a beyond-pairwise character to E_{rep}. Consequently, the inherent limitations of the traditional pairwise formulation have proven a major pitfall for accuracy and general validity. For example, phonon band structures have been shown to be often poorly described by traditional DFTB and highly dependent on the employed repulsive potentials.[79,316,317] Independent of the formalism to treat van der Waals interactions, unit cell volumes of molecular crystals can be considerably underestimated.[317,318] Traditional repulsive potentials without empirical fixes have also failed to correctly describe the stability of Zundel ions and associated proton transfer barriers.[319] As observed for the different phases of ZnO,[320] optimal repulsive potentials can depend strongly on the chemical environment — a final confirmation of the breakdown of the pairwise-additive approximation.

Fig. 6.2 exemplifies this breakdown for the repulsive force between the two carbon atoms in acetylene, ethylene and ethane. The absolute value of $\mathbf{F}_{\mathrm{rep}}$ as defined in Eq. (6.2) can vary by more than 100 kcal·mol^{-1}·Å^{-1} around the equilibrium distance. All hydrogen atoms have thereby been allowed to relax, which leaves only the bare C–C repulsion. When going to larger-scale systems, approximating the C–C repulsion by an environment-independent pair-potential can then lead to an arbitrary accumulation of such errors per C–C bond. Tailoring parameterizations for a specific purpose can alleviate the shortcomings for individual properties or structural motives.[316,319–321] A multitude of applications requires simultaneous accuracy in a variety of properties and structural features, however. For example, we can conclude from Fig. 6.2 that even simple hydrocarbons containing both saturated and unsaturated regions (single and double/triple C–C bonds) cannot be accurately described by environment-independent pair-potentials. Using a specific repulsive potential for single, double and triple C–C bonds, on the other side, requires an *a priori* definition of bond types, which renders the description of transitions between bond types impossible and essentially leads to an expensive polarizable force field. Dynamical definition of bond types as proposed in Ref. 322, finally, can suffer from sudden transitions between bond orders and inevitably introduces further empiricism in form of switching functions. So, we require a more advanced description beyond the pairwise picture with a straightforward access to optimization in the corresponding high-dimensional functional space.

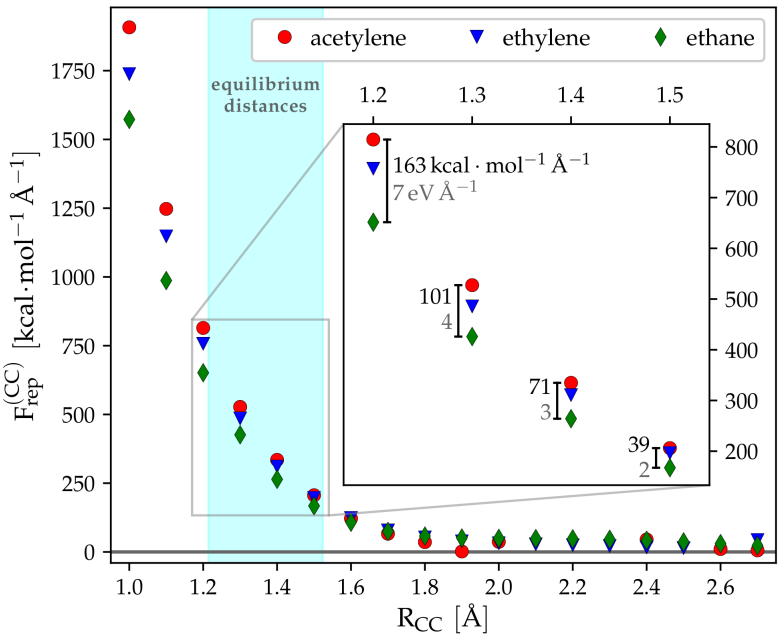

Fig. 6.2: Environment-dependent repulsive contributions to atomic forces in acetylene, ethylene and ethane. Depending on the local chemical environment, the repulsive force as given by the difference between PBE0-DFT and electronic DFTB contributions can vary by more than 100 kcal·mol^{-1}·Å^{-1} around the equilibrium distance.

A very successful tool for exploring the space of high-dimensional functions and mappings are Machine Learning (ML) approaches. Since early applications of "learning machines", such ML-based methods have recently gained tremendous interest in the context of modeling molecular systems and materials. Given sufficient reference data, modern ML techniques aided by chemical and physical knowledge have been proven to show great success in predicting physico-chemical properties of molecules and materials. [323–327] In many cases the data-driven predictions reach virtually the same level of accuracy as the provided reference data. Nowadays, several methods ranging from Neural Network (NN) potentials to Kernel Ridge Regression approaches using diverse molecular descriptors have been put forward, substantially accelerating high-accuracy atomistic modeling and advancing our understanding of chemical compound space. [328–334] Albeit aided by chemical intuition, these approaches are still mostly based on data-driven inference instead of physical laws and typically rest on an assumption of locality. Molecular properties including interatomic interactions, however, are governed by quantum mechanics and involve a variety of non-local characteristics and phenomena. As a result, the applicability of ML models is strongly limited due to the lack of decisive effects like long-range electrostatics‡ or van der Waals dispersion interactions. Approximate QM approaches like (electronic) DFTB together with augmenting models, on the other side, are able to capture such effects. [42,142,156,307,318]

Besides the limitations due to the locality of current ML models, data-driven prediction of global properties additionally suffers from the sheer depth and complexity of the global chemical space. The number of samples needed to infer a model for global properties of a given set of molecules scales with the number of compositional, configurational, and conformational degrees of freedom. Local (chemical) environments as required for the prediction of local properties, on the other side, span a drastically reduced and most importantly bound space. For the dataset and applications considered in this work, for example, we can estimate the number of training instances for local interactions to be three orders of magnitude smaller than to cover the global chemical space (see Appendix A11). A hybrid QM/ML formalism of ML potentials for (more) localized interactions, such as the repulsive energy, and approximate QM methods for non-local effects, thus, represents a very promising approach to obtain an accurate and transferable methodology for studying realistic and practically-relevant systems. For the case of DFTB, such combinations have been proposed and studied recently, providing a more straightforward access to atom-pairwise

‡ Based on physical models, very recent developments are able to provide substantial improvements in this regard. [335] Their general accuracy and validity remains to be fully assessed, however.

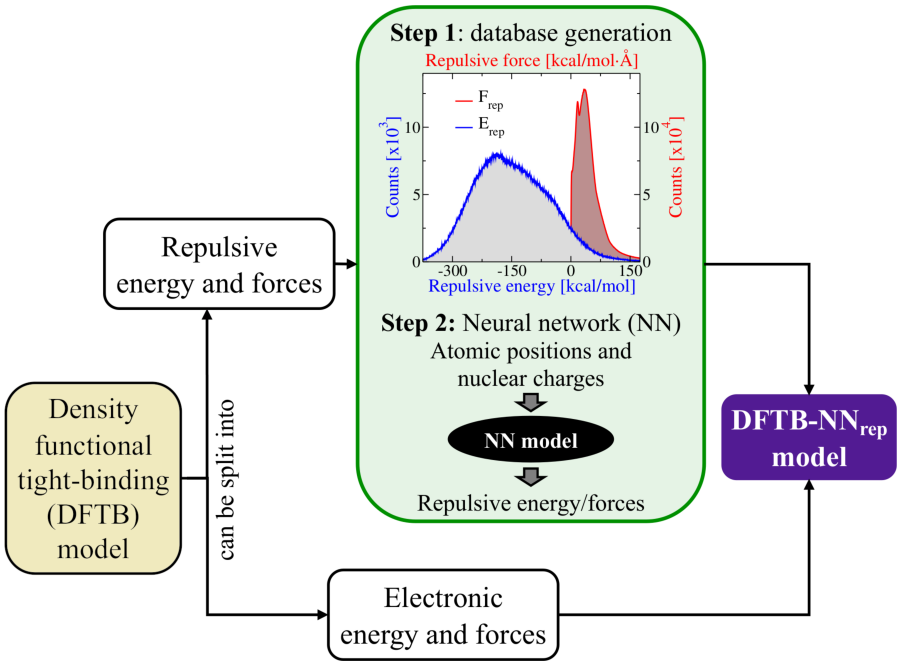

Fig. 6.3: Schematic representation of the DFTB-NN$_{rep}$ framework. DFTB repulsive energies and forces are obtained *via* a (deep tensor) neural network model based on PBE0 reference data, while electronic energies and properties are calculated within the density-functional tight-binding formalism. Reproduced in altered form with permission from Ref. 4. © 2020 American Chemical Society.

repulsive potentials from Gaussian process regression [307] or introducing partial non-additivity *via* bond-type-dependent pair-potentials. [322] In this work, we propose to further extend the latter idea to fully many-body ML potentials in order to account for beyond-pairwise repulsive contributions. This is to some extent similar in spirit to the Δ-ML approach, where it is proposed to use ML models to correct the final energy of semi-empirical methods. [336,337] As the final DFTB energy is already fitted to reproduce DFT, such corrections can involve a rather unspecific and noisy objective quantity, however. Targeting the full repulsive potential is expected to provide a well-behaved and smooth quantity with a more distinct mapping to molecular features.

6.2 Many-Body DTNN Potentials for Repulsive Energies in DFTB

As an example of such hybrid QM/ML approaches, we here use (electronic) DFTB in conjunction with a global deep tensor NN (DTNN) model. Hence, forming the DFTB-NN$_{rep}$ framework as described in Fig. 6.3. In particular, we used the deep learning toolbox SchNetPack. [325,333,338] The DTNN thereby infers the representation directly from the data instead of relying on fixed descriptors and metrics. With the limited knowledge about (geometrical) descriptors and physical constraints for E_{rep} this motivates the use of DTNN potentials. It is worthwhile to point out, however, that the presented methodology can be applied to various semi-empirical methods involving repulsive potentials and other ML approaches. The latter is briefly outlined in Appendix A10.

6.2.1 Computational Details

Repulsive energies and forces as given by Eq. (6.2) have been obtained for a large and diverse set of organic molecules. The basis for this dataset is the QM7-X database, [339,340] which contains molecular structures and QM properties of small organic molecules at the level of PBE0-DFT. For the purpose of this work, we have selected molecules containing the elements {C, N, O, H}, but the presented methodology is easily extended to other molecular compositions. In addition to equilibrium structures, QM7-X also features 100 non-equilibrium conformations per molecule. In this work, the QM7-X database has been further extended to cover less well-represented regions of configurational space, such as select pairwise O–O and H–H distances. The final set, which we shall use as the basis for all presented models contains ~4.1 million molecules. Repulsive contributions to energies and forces show a standard deviation of 94 kcal/mol and 46 kcal·mol^{-1}·Å^{-1}, respectively, and their distributions are shown in Fig. 6.3.

The employed NN architecture is based on atom-wise representations of molecular properties directly inferred from atomistic structures. In the layers of the DTNN, atoms are represented by a tuple of features $x_i^{(l)} \in \mathbb{R}^D$, where D is the dimension of the feature space and l denotes the layer. Interactions between atoms are modeled by a series of in total T-times refined pairwise interactions between all $x_i^{(l)}$ within a certain cutoff, which gradually introduces information about the chemical environment (*i.e.*, complex many-body terms). This procedure is carried out by using continuous-filter convolution layers with filter-generating networks. The final prediction is obtained after atom-wise updates of the feature representation and pooling of the resulting atom-wise property. Besides generating reliable NN-models for energy prediction, SchNetPack has been proven to also provide energy-conserving force models by differentiating the energy model with respect to the atomic positions.[333] Hence, we create a global DTNN-model for predicting repulsive energies and forces based on the QM7-X set. We used $T = 3$ interaction refinements, a cutoff of 5 Å and a 128-dimensional feature space. For all reported results, the DTNN was continuously trained with a descending learning rate from 10^{-4} to 10^{-6} and a decay factor of 0.5. The training has been performed using four Tesla P100 GPUs. SchNetPack and DFTB+ calculations as well as structural and vibrational analysis have been carried out within the Atomic Simulation Environment.[341] Calculation of repulsive and electronic contributions has thereby been combined into a single calculator instance *via* a locally modified QM/MM calculator. For the remainder of this work, we use DFTB-pw_{rep} to refer to conventional DFTB3 with pairwise repulsive potentials. DFTB-NN_{rep} refers to DFTB3 with the same electronic parameterization while using repulsive energies/forces from the DTNN-model. For comparisons of pw_{rep} and NN_{rep}, it should be noted that unlike NN_{rep} the *3ob* set was not designed to reproduce PBE0.

6.2.2 Overview and DTNN Training

First, we studied the influence of the training set selection on the learning process of repulsive energies. This is important for optimizing the training set size, accuracy and learning progress. To this end, we considered three cases: random selection from the whole dataset, farthest point sampling (FPS) with the relative repulsive energy as distance metric and a training set consisting of equilibrium structures together with a given number (X) of non-equilibrium structures. In the latter case, which we will refer to as eq+X in the following, non-equilibrium structures were selected at random from domains with low, middle or high values of E_{rep}. The corresponding learning curves are shown in Fig. A.7 of Appendix A10. On a log-log scale one expects a (near-)linear trend in the learning progress, which we can attribute conclusively only to the randomly selected training set. For eq+X, we find linear behavior only beyond ~200k training instances. When considering only equilibrium structures in the training set ($\hat{=}$ ~41k), for instance, the MAE is 19.5 kcal/mol and strongly deviates from a linear trend. With an increasing portion of non-equilibrium structures, eq+X approaches linear behavior and the results obtained with random selection. The minimum MAE achieved with FPS and eq+X are 0.59 kcal/mol and 0.55 kcal/mol, respectively. The lowest MAE of 0.47 kcal/mol is obtained for random selection of 800k training points. The rather unexpected finding that the more refined approaches starting from equilibrium structures do not improve over a random selection of the training set demonstrates the importance of non-equilibrium conformations and that their properties can seldom be inferred from the information on equilibrium structures. Regarding the required training set sizes, we would like to note that a purely data-driven, global model can be expected to require about $5 \cdot 10^6$ training points to cover the compositional, configurational and conformational space of QM7-X (see Appendix A11). The considerable reduction observed for NN_{rep} can be attributed to the partly included physics in the NN-model, the shared information among individual molecular degrees of freedom and the repulsive potential covering more local interactions. The latter also manifests itself in an improved performance of kernel ridge regression when using atomic descriptors (see Appendix A10) and a slower increase in MAEs towards higher-energy conformations.

As mentioned above, SchNetPack also allows to train NN-models on atomic forces. We employed the same protocol as above to develop a DTNN for the contribution of the repulsive potential to atomic forces, which also allows to predict energies since the force model is obtained by differentiating the energy model with respect to atomic positions.[333] The corresponding learning curves are shown in Fig. 6.4. At 500k randomly sampled training instances, the NN_{rep} model learned on forces produces a MAE of 0.51 kcal/mol and 0.70 kcal·mol^{-1}·Å^{-1} for energies and forces, respectively. So, while being only trained on forces, it also provides energies at an accuracy comparable to the model trained on energies, yet requires a smaller training set. This can be attributed to the increased information content in the gradient domain and the improved performance of resulting models thanks to the inherent incorporation of energy conservation.[331] Representing the best and most versatile model, all results reported for (DFTB-)NN_{rep} below were obtained with the DTNN-model trained on repulsive forces with randomly selected training instances. We found that for 89 % of the molecules in QM7-X the MAE in forces is lower than the threshold of 1 kcal·mol^{-1}·Å^{-1}. This indicates that our NN_{rep} model yields reliable energies and forces for the manifold molecules and conformations considered in QM7-X. The present model hence extends beyond previous works as reported by Zhu *et al.*[337] for instance, in which DFTB forces of a single molecule (glycine) have been corrected *via* Δ-learning using a NN. Finally, we would like to remark that better performances for repulsive energies and forces can be reached by further increasing the training set size as it was shown in previous ML studies. Given the overall very low MAEs with respect to PBE0-DFT, this is beyond the scope of the current work.

In a next step, we have compared the results of DFTB-NN_{rep} to a DTNN-model trained on the full PBE0 atomization energies of our QM7-X set, which we will refer to as NN-PBE0. The learning curve of the NN-PBE0 model is plotted in Fig. 6.4. NN-PBE0 follows a (near-)linear learning progress in log-scale, but shows a substantially larger MAE than NN_{rep} and cannot overcome the threshold of 1 kcal/mol within the considered training set sizes. The minimum MAE at 800k training instances amounts to 1.52 kcal/mol. The learning progress of NN-PBE0, as characterized by the slope of the corresponding learning curve, is comparable to the one of NN_{rep}, but thanks to the large QM prior in form of the electronic DFTB energy, NN_{rep} is able to reach a MAE below 1 kcal/mol at much smaller training set sizes. Finally, we have constructed a DTNN Δ-learning model for correcting the energy predicted by traditional DFTB-pw_{rep}. The standard deviation of the difference between PBE0 and DFTB-pw_{rep} is only 28 kcal/mol and hence much lower than for

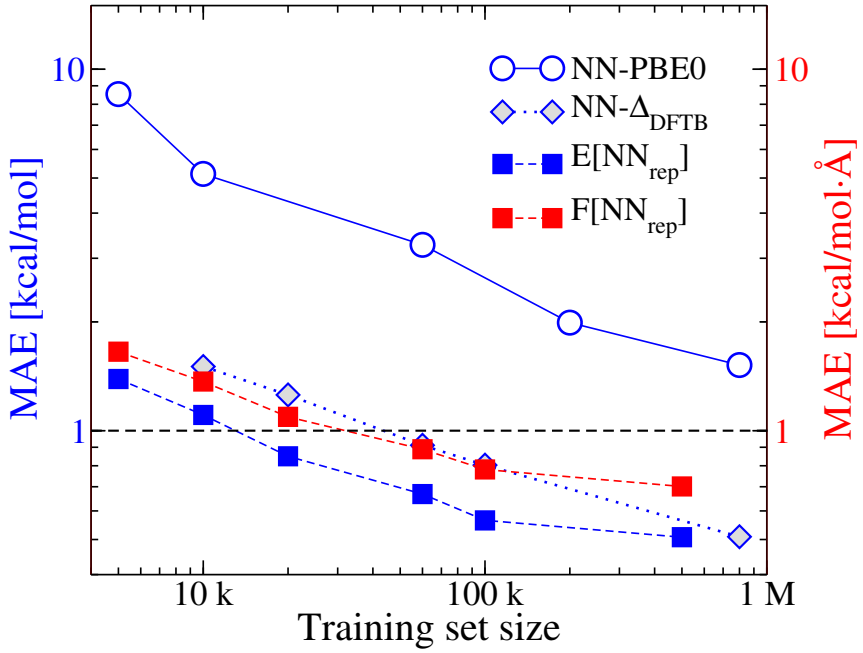

Fig. 6.4: Mean absolute error (MAE) as a function of training set size for the different neural network models. MAEs in target energies are plotted for a DTNN-model for PBE0 atomization energies (NN-PBE0) and a Δ-learning model (NN-Δ_{DFTB}). For our NN_{rep} model trained on repulsive forces, the MAEs in predicting energies (E[NN_{rep}]) as well as forces (F[NN_{rep}], right *y*-axis) is shown. Reproduced with permission from Ref. 4. © 2020 American Chemical Society.

E_{rep} (94 kcal/mol). However, the DTNN trained on this difference (NN-Δ_{DFTB}) does not provide a lower MAE (*cf.* Fig. 6.4). So, learning the traditional pair-repulsion in addition to the Δ-term (as done in NN_{rep}) does not pose an issue for the DTNN and does not affect the learning success. This allows to avoid the cumbersome construction of pair-potentials and to directly learn the full repulsive contribution without loss of accuracy, which strengthens the motivation for our NN_{rep} approach.

To further analyze the performance of DFTB-NN_{rep}, we have split the QM7-X set into four different subgroups depending on their composition: molecules containing only the element combinations {C,H}, {C,N,H}, {C,O,H}, or {C,N,O,H} and as a function of the number of non-hydrogen atoms (see Fig. 6.5). We then computed the atomization energy (E_{at}) for each equilibrium molecule and its respective non-equilibrium conformations using DFTB-NN_{rep}. As a rule of thumb, the atomization energies for molecules with fewer non-hydrogen atoms and correspondingly smaller size are predicted with higher accuracy. This can mainly be attributed to the higher number of degrees of freedom in larger molecules, which increases the complexity for describing their conformational space especially in regard to strongly distorted molecules. Overall, the MAE is 0.5 kcal/mol and 88 % of the molecules are predicted within an error 1 kcal/mol. The mean absolute relative error (MARE) in E_{at} for the whole set is 0.03 %. As such, the MAE and MARE are three times smaller than for NN-PBE0. This confirms the very good performance of DFTB-NN_{rep}. In particular, almost all hydrocarbons show very low errors thanks to the extensive sampling of C–C and C–H in the training set. Consequently, the least represented molecules, which at the same time show the highest configurational and conformational complexity (*e.g.*, formed by {C,N,O,H}) also show the largest errors due to the limited sampling. A more balanced sampling of the QM7-X reference can therefore further limit the obtained MAEs and is subject to ongoing investigations. In addition to the QM7-X database, we analyzed the performance of our DFTB-NN_{rep} model based on 2244 {C,N,O,H}-molecules as contained in the QM9 dataset. In comparison to hybrid DFT, the MAE in predicting E_{at} is 0.90 kcal/mol for molecules with up to seven heavy atoms and 1.6 kcal/mol for molecules with eight or nine heavy atoms. The relative error stays around 0.08 %. For DFTB with ML-based generalized pair-potentials, Kranz and co-workers have obtained a MAE of ~2.6 kcal/mol.[322] As a final test for atomization energies, we considered a set of linear and cyclic hydrocarbons as contained in the *NIST Standard Reference Database Number 69*.[342]

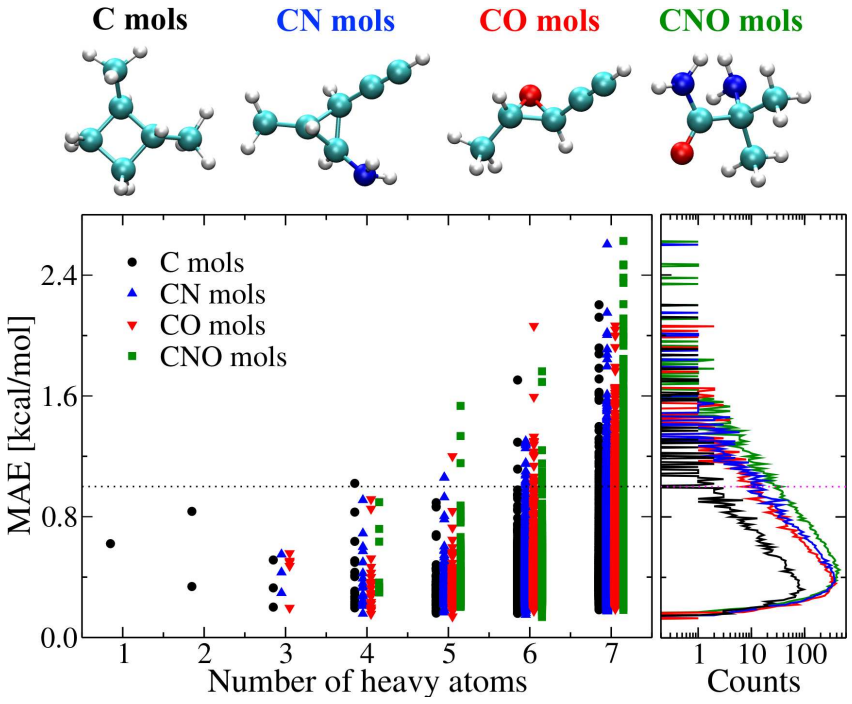

Fig. 6.5: Analysis of the mean absolute error (MAE) of DFTB-NN_{rep} in predicting atomization energies. *Left:* The MAE in atomization energies per molecule is shown as a function of the number of non-hydrogen atoms contained. The MAE is further separated into molecules containing only {C,H} (black circles), {C,N,H} (blue triangles up), {C,O,H} (red triangles down), and {C,N,O,H} (green squares). *Right:* Distribution of MAE color-coded according to contained non-hydrogen atoms. Reproduced with permission from Ref. 4. ©2020 American Chemical Society.

DFTB-NN_{rep} reproduces PBE0 atomization energies as tabulated in the HC160 set [343] within 0.60 kcal/mol for hydrocarbons with seven heavy atoms or less. When considering molecules with eight or nine carbon atoms, the MAE is 0.90 kcal/mol. This excellent performance confirms the transferability of DFTB-NN_{rep} for small organic molecules outside the training set.

Besides the (mean) absolute error for each conformation, the relative stabilities of each conformer represent an important and interesting performance measure. For instance, predicting the correct energetic ordering of different structures is of utmost importance in computational crystal structure prediction. To analyze the performance in predicting energy rankings, we have first determined the energetic ordering of all conformers for each molecule as given by PBE0. Based on this ordering, we can assign a rank to every structure. Fig. 6.6a shows a correlation plot of the energy ranks obtained with DFTB-NN_{rep} (red squares) and DFTB-pw_{rep} (green triangles) in comparison to PBE0, which clearly proofs the much superior performance of DFTB-NN_{rep}. In this regard, it has been shown previously that including relative energies in the fitting process of conventional DFTB-pw_{rep} can provide improvements. [344] For further analysis, we calculated the MAE in the rank predicted by the two DFTB formalisms with respect to PBE0. In Fig. 6.6b, this ranking MAE is subdivided into individual windows composed of five energy levels. It shows the much higher accuracy of DFTB-NN_{rep} holds true throughout the energy spectrum. The general behavior observed in the ranking MAEs with increased deviations in the lower third of the energy ranking can thereby be explained by the energy level density per kcal/mol for each window. The more densely the conformers populate the energy spectrum, the easier can already small errors in the energy prediction cause a reordering of energetic ranks, whereas the ordering of well-separated energy levels is typically well preserved.

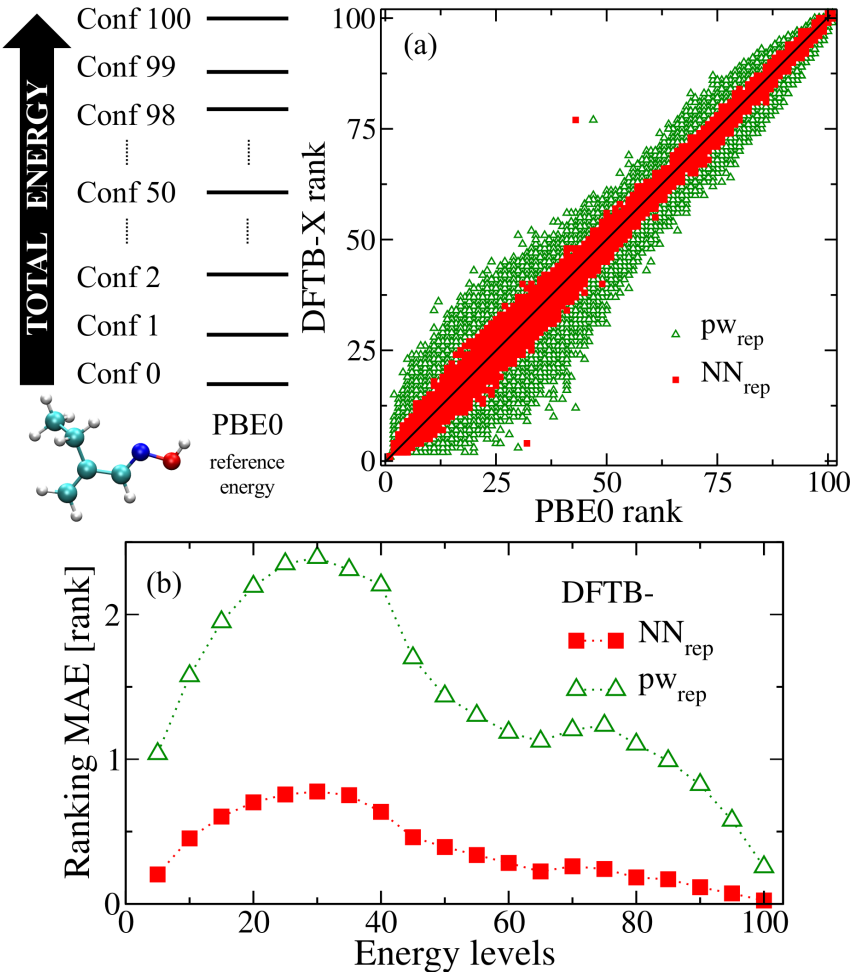

Fig. 6.6: Performance in predicting relative energy rankings of all 100 conformers of each molecule in QM7-X. (a) correlation plot of energy ranks predicted by DFTB-pw_{rep} (green triangles) and DFTB-NN_{rep} (red squares) in comparison to PBE0-DFT. (b) Mean absolute error (MAE) of the predicted rank in windows of 5 non-equilibrium conformations. Reproduced with permission from Ref. 4. ©2020 American Chemical Society.

6.2.3 Transferability and Validity Outside the Scope of the Training Set

We have further studied isomerizations of diverse neutral molecules containing the elements C, N, O and H. The so-called ISO34 dataset [345–347] and the considered isomers are detailed in Ref. 347. The experimental (and selectively computationally refined [347]) reference values of the corresponding isomerization energies have been widely used for benchmarking semi-empirical methods including DFTB. [348] We have tested the performance of PBE0-DFT and our DFTB-NN_{rep} model as well as for NN-PBE0, NN-Δ_{DFTB} and conventional DFTB-pw_{rep} in predicting the isomerization energies as summarized in Tab. 6.1. One can see that the accuracy of DFTB-NN_{rep} is much better than for NN-PBE0 and DFTB-pw_{rep} and is in fact close to the one obtained at the level of PBE0. As for the learning success and MAEs within the reference set, the Δ-learning model does not provide improvements over NN_{rep}.

Tab. 6.1: Performance of PBE0-DFT, NN-PBE0, NN-Δ_{DFTB}, DFTB-NN_{rep} and DFTB-pw_{rep} in predicting isomerization energies (in kcal/mol) in comparison to experimental and CCSD(T) reference data as considered in ISO34 dataset: [345–347] mean signed error (MSE), mean absolute error (MAE) and root-mean-square deviation (RMSD).

Method	MSE	MAE	RMSD
PBE0-DFT	−0.19	1.82	2.48
NN-PBE0	2.21	5.85	11.51
NN-Δ_{DFTB}	0.76	2.48	3.49
DFTB-NN_{rep}	−0.71	2.21	3.30
DFTB-pw_{rep}	1.33	3.57	5.05

In addition to the isomerization energies, we have investigated the performance of our DFTB-NN_{rep} model in predicting the equilibrium structures and vibrational mode frequencies of the isomers in ISO34. Tab. 6.2 compares the root-mean-square deviation of the optimized structures ($RMSD_s$) as well as the mean and maximum absolute error in vibrational frequencies (MAE_ν and MAX_ν, respectively) in reference to PBE0-DFT results for DFTB-pw_{rep} and DFTB-NN_{rep}. For optimized structures, both approaches perform similar and well on average. However, DFTB-NN_{rep} displays a better performance for a wider range of substrates and products. The only cases in which DFTB-pw_{rep} provides better results correspond to more complex structures composed of eight non-hydrogen atoms or unseen functional groups. In the case of vibrational frequencies, DFTB-NN_{rep} provides a much superior description for all isomers without exception. NN_{rep} reduces the MAEs by about 85 % and the maximum deviations by 60–90 % providing very good agreement with PBE0-DFT. It is worth mentioning that the vibrational calculations using DFTB-NN_{rep} took ~1.5 hours (on one GPU), while the PBE0-DFT calculations required ~8260 CPU hours. Previous studies have already shown that conventional DFTB parameterizations seldom allow for an accurate prediction of energetic, structural, and vibrational properties at the same time. As a result, special-purpose parameterizations have been devised for vibrational analysis, for instance. [47,79,321] The here developed DFTB-NN_{rep} framework, on the other side, indeed does enable accurate predictions of energetic as well as structural and vibrational properties. This is essential for, *e.g.*, consistent and seamless calculations of vibrational spectra, free energies and many other thermodynamic or transport properties.

Tab. 6.2: Structural root-mean-square deviation of equilibrium geometries ($RMSD_s$) and mean/maximum absolute error of vibrational frequencies (MAE_ν/MAX_ν) as predicted by DFTB-pw_{rep} and DFTB-NN_{rep}. All errors obtained in reference to PBE0 results.

contained elements	$RMSD_s$ [Å]		MAE_ν [cm^{-1}]		MAX_ν [cm^{-1}]	
	pw_{rep}	NN_{rep}	pw_{rep}	NN_{rep}	pw_{rep}	NN_{rep}
H,C	0.03	0.02	49.9	7.41	308	93
H,C,N	0.05	0.02	49.8	8.38	316	119
H,C,O	0.06	0.02	50.2	6.72	228	96
H,C,N,O	0.24	0.23	51.9	6.68	333	42
Overall	**0.05**	**0.03**	**50.0**	**7.47**	**333**	**119**

To finally highlight the advantages of our DFTB-NN$_{rep}$ formalism and its good transferability among small organic molecules, we have also studied the prediction of potential energy profiles upon dihedral rotation. As test cases, we chose *N*-methylacetamide — as a simple model for peptide bonds — and glycine. The corresponding torsional profiles are shown in Fig. 6.7. In the case of the peptide bond, DFTB-pw$_{rep}$ (green triangles) underestimates torsional barriers by 5.4 kcal/mol and overestimates the relative stability of the meta-stable intermediate by 2.8 kcal/mol in comparison to PBE0 (dashed black). DFTB-NN$_{rep}$ (red squares) considerably improves the performance only underestimating the barriers by 2.9 kcal/mol. As previously shown by Zhu *et al.*, the traditional *3ob* DFTB parameterization shows considerable shortcomings in describing the N–C–O–H dihedral rotation profile of glycine; underestimating torsional barriers by almost 8 kcal/mol.[337] In their study, the authors achieved substantial improvements and good agreement with DFT by supplementing DFTB within a Δ-learning framework. Our DFTB-NN$_{rep}$ model as well provides considerable improvements, yet no full agreement with hybrid DFT (see Fig. 6.7f). In this regard, it is important to note that the Δ-learning model by Zhu and co-workers was specifically designed for the glycine molecule, while our DFTB-NN$_{rep}$ model did not include glycine in its training set. For the N–C–C–O dihedral of glycine, finally, we investigated two pathways of dihedral rotation: full relaxation at fixed dihedral angle (6.7b) and rigid rotation of the N–C–C–O backbone (6.7c). In the case of the minimum-energy path in Fig. 6.7b, DFTB-pw$_{rep}$ again underestimates torsional barriers and overestimates the relative stability of the intermediate state at 180°, while DFTB-NN$_{rep}$ provides close agreement with PBE0-DFT. When considering the constrained pathway in Fig. 6.7c, DFTB-pw$_{rep}$ even lacks qualitative features of the rotational profile: Higher-level reference calculations (PBE0- and B3LYP-DFT as well as MP2) predict two low-energy barriers at 90 and 270° and a main barrier of ~3 kcal/mol at a dihedral of 180°. DFTB in conjunction with the traditional pairwise repulsion completely misses this most

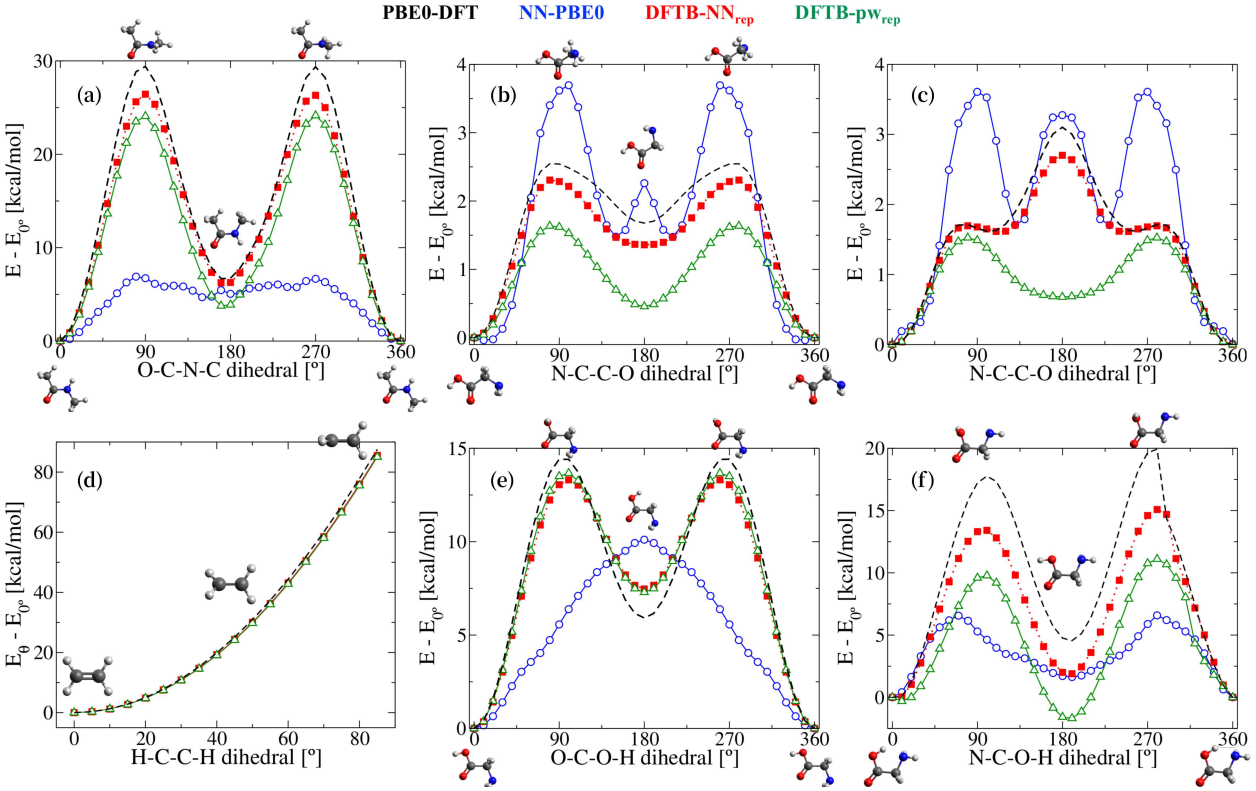

Fig. 6.7: Potential energy profiles upon dihedral rotation as predicted by PBE0-DFT, NN-PBE0, DFTB-NN$_{rep}$ and DFTB. (a) O–C–N–C dihedral of a peptide bond, (b) minimum-energy and (c) higher-energy pathway for N–C–C–O dihedral rotation in glycine, (d) ethylene, (e) O–C–O–H and (f) N–C–O–H dihedral of glycine. Reproduced with permission in altered form from Ref. 4. ©2020 American Chemical Society.

relevant torsional barrier and instead predicts a shallow minimum. Based on this, we can conclude that with the traditional atom-pairwise repulsion DFTB is unable to correctly describe the potential energy landscape around the transition state, which would be essential for calculating reaction rates. Our DFTB-NN$_{\text{rep}}$ model is able to correct for this shortcoming and provides very good agreement with the reference results. Among all considered models, DFTB-NN$_{\text{rep}}$ represents the only one that is able to give qualitatively and quantitatively good results for both pathways. These findings are particularly encouraging considering that neither the glycine molecule nor any rotational profiles were part of the training set. Further dihedral rotational profiles confirming the above observations for ethylene and the glycine O–C–O–H dihedral are shown in Figs. 6.7d and e. For comparison, we also computed dihedral energy profiles as predicted by NN-PBE0 (blue circles). As evident from Fig. 6.7, the DTNN trained on PBE0 is unable to predict meaningful rotational profiles. Together with the poor performance for the ISO34 set, we can conclude that the transferability of NN-PBE0 is very limited outside the immediate scope of its training set. The much increased transferability of our DFTB-NN$_{\text{rep}}$ approach can be attributed to the more local character of E_{rep} and further motivates exploring the synergy between (approximate) QM Hamiltonians and ML-potentials for localized many-body interactions.

6.3 Perspective: Tailor-made ML Model for Repulsive Potentials

The two main limitations of the DFTB-NN$_{\text{rep}}$ approach as presented above are the wealth of reference data needed to construct the model and a limited scalability. Both of these aspects can be (partially) attributed to the DTNN not fully exploiting the more local character of the target repulsive energies and forces. In this regard, it is important to note that while a cutoff of 5 Å has been employed for the definition of the atomic embeddings, the interaction refinements subsequently combine information from the individual embeddings and can thus considerably increase the range of effects included in the model. As one potential remedy, ongoing work is therefore devoted to identifying an optimal combination of cutoff and the number of interaction refinements. Ultimately, however, the original DTNN architecture is intended as a global model to describe total energies and forces. As for the broad majority of ML approaches the architecture and general approach is designed for such global properties. Given the much reduced part of the total energy covered by E_{rep}, it can be beneficial to employ a design that is more geared to repulsive potentials. One such idea is discussed in Sec. 6.3.2. Before going to tailor-made ML potentials for repulsive energies and forces, however, it is worthwhile to re-consider the selection of training instances. In the previous sections, we have seen that random training set selection outperforms algorithms based on energetic aspects (eq+X and FPS with energy metric). Considering the locality of the target quantity, a training set selection based on local characteristics such as the local degrees of freedom introduced in Sec. 6.1 and discussed in Appendix A11 can be a viable, more effective alternative. Fig. 6.1 gives a strong indication that local descriptors including up to three nearest neighbors (bond distances, angles and dihedral angles) suffice to capture the effects governing the repulsive potential. So, a balanced sampling of these local degrees of freedom holds the potential to substantially reduce the training set size and to enable models with broader validity.

6.3.1 Training Sets with Balanced Sampling of Local Chemical Environments

For the mathematical treatment of a balanced training set, we first define a sampling distribution. To this end, we may simply choose the histogram of the (local) degrees of freedom (currently) occurring in the training set, $S^{(n)}$. At each step of the training set selection, we then obtain the k most and least sampled points in the sampling space, X_{max} and X_{min}. At the n^{th} step of the selection algorithm, we define the respective maximum sampling distribution, $S^{(n)}_{\text{max}}(x)$, and minimum sampling distribution, $S^{(n)}_{\text{min}}(x)$, where $S^{(n)}_{\text{max}}(x)$ and $S^{(n)}_{\text{min}}(x)$ are defined on the k most and least sampled points of the sampling space (*i.e.*, the k lowest and highest values of the total

sampling distribution), respectively:

$$S^{(n)}_{\max}(x) = S^{(n)}(x_i) \;\mid\; x_i \in X_{\max} \qquad \text{and} \qquad S^{(n)}_{\min}(x) = S^{(n)}(x_i) \;\mid\; x_i \in X_{\min}\,. \tag{6.3}$$

A balanced, uniform sampling is characterized by the two distributions being the same. So, at each step of the training set generation, we want to select a new molecule that minimizes the differences between the two distributions. One way to quantify how different $S_{\max}$ is from $S_{\min}$ (or the divergence from $S_{\min}$ to $S_{\max}$), is the *Kullback-Leibler divergence*[349] as given by

$$\mathcal{D}\left(S^{(n)}_{\max} \| S^{(n)}_{\min}\right) = \sum_{i=1}^{k} S^{(n)}_{\max}(x_i) \log\left(\frac{S^{(n)}_{\max}(x_i)}{S^{(n)}_{\min}(x_i)}\right)\,. \tag{6.4}$$

The training set selection then proceeds as a consecutive, numerical minimization of $\mathcal{D}$. So, in the $(n{+}1)^{\text{th}}$ step we add the molecule that provides the maximum decrease in $\mathcal{D}$. The change in $\mathcal{D}$ is thereby given by its (discrete) derivative with respect to n,

$$\frac{d}{dn}\mathcal{D} = \sum_i \left[\log\left(\frac{S^{(n)}_{\max}(x_i)}{S^{(n)}_{\min}(x_i)}\right) + 1\right] \left.\frac{dS^{(n)}_{\max}(x)}{dn}\right|_{x_i} - \frac{S^{(n)}_{\max}(x_i)}{S^{(n)}_{\min}(x_i)} \cdot \left.\frac{dS^{(n)}_{\min}(x)}{dn}\right|_{x_i} \tag{6.5}$$

$$= \sum_i \left[\log\left(\frac{S^{(n)}_{\max}(x_i)}{S^{(n)}_{\min}(x_i)}\right) + 1\right] \Delta S_{\max}(x_i) - \frac{S^{(n)}_{\max}(x_i)}{S^{(n)}_{\min}(x_i)} \cdot \Delta S_{\min}(x_i)\,, \tag{6.6}$$

where $\Delta S(x_i) = S^{(n+1)}(x_i) - S^{(n)}(x_i)$. It should be noted that for all x_i, $\Delta S(x_i) \geq 0$ and $S_{\max}(x_i) > S_{\min}(x_i)$. So, the prefactors of both ΔS are strictly positive. To minimize the differences between the distributions of the most and least sampled regions, we thus require a minimal increase in $S_{\max}$ and a maximal increase in $S_{\min}$. Also, with $X_{\max}$ in descending and $X_{\min}$ in ascending order, the absolute minimum and maximum of the current sampling have the strongest impact on the selection criterion (6.6). In order to assure optimal sampling, we perform the minimization of Eq. (6.6) under the constraint that the new molecule has to provide an increase of the sampling at the current global minimum. The overall algorithm is outlined in Fig. 6.8.

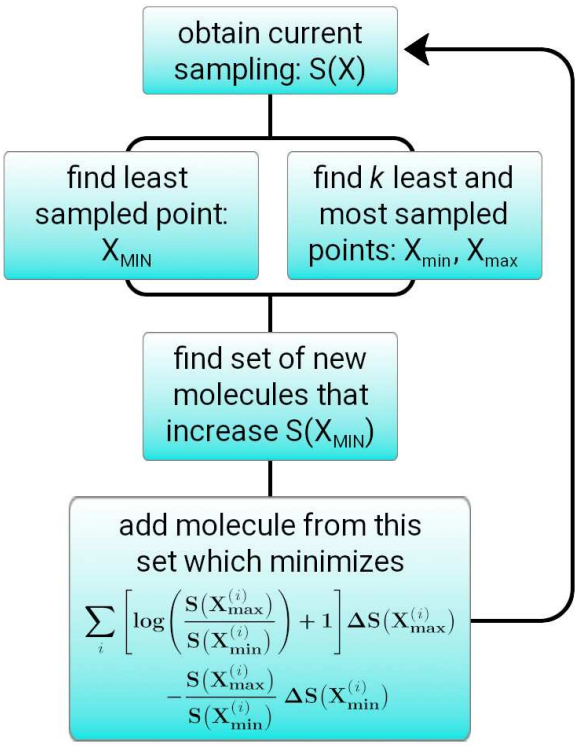

Fig. 6.8: Selection algorithm for training sets with balanced sampling. At each step one obtains the current sampling of the considered (local) chemical space, S(X). Then, the position of the global minimum (X_{MIN}) and the k least (X_{min}) and most sampled (X_{max}) points are identified. Among all molecules that increase the sampling at X_{MIN}, the molecule that minimizes the differences between the most and least sampled regions is added to the training set.

It is worth noting that the above procedure is not limited to local degrees of freedom. The proposed selection algorithm can be equally-well employed to obtain training sets, that provide a balanced sampling of an arbitrary choice of descriptors such as delocalized internal coordinates or molecular properties. Establishing a proper and reasonable sampling space and obtaining the corresponding sampling distributions can pose challenges in high-dimensional spaces, however.

6.3.2 DTNN Architecture with Two-Center Repulsion-Based Filter Generation

An optimized training set selection can allow to substantially reduce the amount of reference data needed or in other words allows to treat a broader set of elements/systems with the same number of training instances without loss of accuracy. Optimally balanced training sets also enable the construction of more general and robust models, which aids transferability and scalability. In order to achieve good accuracy at the desired length and time scales of biomolecular or nanoscale systems, however, a ML model which is more geared to exploiting the physical knowledge about repulsive potentials seems required. First tests of the $\mathrm{NN_{rep}}$ model as presented above on molecular dimers showed considerable deviations and thus the limited scalability, for example. In the current approach, the repulsive energy is obtained from atom-wise contributions, where each contribution depends on a respective atomic embedding. This embedding is represented *via* Gaussian expansion of the distance matrix, *i.e.*, by the tensor [338]

$$\tau_{ABk} = e^{-\kappa\left(\|\mathbf{R}_A-\mathbf{R}_B\|-\mu_k\right)^2} , \tag{6.7}$$

where κ represents a tuning parameter and μ_k resolves the distance matrix in consecutive neighbor shells. Taking the maximum value along k shows that, in principle, all atoms within the cutoff can contribute to equal amounts to the construction of the embedding and the information included from neighboring atoms is largest if their distance is near the average μ_k. From a physical perspective the atomic embeddings represent the interaction with neighboring atoms. So, τ resembles a basis set expansion of pairwise interactions within which the final non-linear model is constructed. In this regard we already know that all atoms within the cutoff will not contribute to equal amounts. Also, the most relevant information will probably not come from neighbors at intermediate distances. In the standard $\mathrm{NN_{rep}}$ approach, these two points have to be inferred during the learning process. Including our prior knowledge of the physical basis for our target quantity can thus considerably improve the learning progress and robustness of the final model. To this end, we can obtain an approximate pairwise representation of the repulsive potential, which we propose to incorporate into the definition of τ. The following paragraphs present such an approximate basis set for pairwise interaction using a Gaussian basis set expansion and two-center approximation of the formal definition of E_rep.

Formal DFTB Repulsion in Two-Center Approximation

As given in Sec. 2.2, the repulsive energy is formally defined as

$$E_\mathrm{rep} = \underbrace{-\frac{1}{2}\int\overbrace{\int\frac{\rho_0'}{\|\mathbf{r}-\mathbf{r}'\|}\,d\mathbf{r}'}^{\mathcal{V}_\mathrm{H}[\rho_0']}\rho_0\,d\mathbf{r}}_{E_\mathrm{rep}^{(\mathrm{H})}} + \underbrace{E_\mathrm{xc}[\rho_0]-\int\left.\frac{\delta E_\mathrm{xc}[\rho_0]}{\delta\rho}\right|_{\rho_0}\rho_0\,d\mathbf{r}}_{E_\mathrm{rep}^{(\mathrm{xc})}} + \tilde{E}_\mathrm{nuc} , \tag{6.8}$$

which we here subdivide into a Hartree(-like) term, $E_\mathrm{rep}^{(\mathrm{H})}$, an xc contribution, $E_\mathrm{rep}^{(\mathrm{xc})}$, and the (effective) nuclear–nuclear repulsion, $\tilde{E}_\mathrm{nuc}$. In order to obtain a pairwise-additive form of E_rep, we start by dividing the total volume into a sum of atomic volumes, V_A, such that $\int d\mathbf{r} = \sum_A \int_{V_A} d\mathbf{r}$.

With this, the first term in Eq. (6.8), $E_{\text{rep}}^{(\text{H})}$, can be written as

$$E_{\text{rep}}^{(\text{H})} = \frac{1}{2}\sum_{A\neq B}\int_{V_A}\int_{V_B}\frac{\rho_0\rho_0'}{\|\mathbf{r}-\mathbf{r}'\|}\,d\mathbf{r}d\mathbf{r}' + \frac{1}{2}\sum_{A}\iint_{V_A}\frac{\rho_0\rho_0'}{\|\mathbf{r}-\mathbf{r}'\|}\,d\mathbf{r}d\mathbf{r}' \;=\; \frac{1}{2}E_{\text{H}}^{(\text{val})}\;, \tag{6.9}$$

where we have divided the double sum into contributions of $A \neq B$ and $A = B$. Here, we would like to note that in DFTB one only treats valence electrons and hence that ρ_0 represents the superposition of atomic *valence* densities. As a result, the above expression also corresponds to (half) the Hartree energy of the valence shells. Within the tight-binding approximation natural to DFTB, we may now assume that only the density of atom A, ρ_A, contributes significantly to ρ_0 for integration over the corresponding volume V_A. This gives the two-center approximation to the repulsive energy components,

$$E_{\text{tcr}}^{(\text{H})} = \frac{1}{2}\sum_{A\neq B}\int_{V_A}\int_{V_B}\frac{\rho_A\rho_B'}{\|\mathbf{r}-\mathbf{r}'\|}\,d\mathbf{r}d\mathbf{r}' + \frac{1}{2}\sum_{A}\iint_{V_A}\frac{\rho_A\rho_A'}{\|\mathbf{r}-\mathbf{r}'\|}\,d\mathbf{r}d\mathbf{r}'\;. \tag{6.10}$$

For the xc contribution, $E_{\text{rep}}^{(\text{xc})}$, one can note that the second terms represents minus a first-order (*i.e.*, local) approximation to the first term. So, within the assumption of (semi-)local xc, we get

$$E_{\text{xc}}[\rho_0] \approx \int \left.\frac{\delta E_{\text{xc}}[\rho]}{\delta\rho}\right|_{\rho_0}\cdot\rho_0\,d\mathbf{r} \qquad \text{and thus} \qquad E_{\text{rep}}^{(\text{xc})} \approx 0\;. \tag{6.11}$$

Ultimately, we are left with $\tilde{E}_{\text{nuc}}$ from Eq. (6.8). Typically, this is thought of the simple nuclear repulsion term, but given that in DFTB one only treats valence electrons, $\tilde{E}_{\text{nuc}}$ in fact represents the effective (interaction) energy of nuclei plus core shells. This can be approached by using pseudo-potentials or by explicitly accounting for the individual contributions, which is what we shall do in this case. $\tilde{E}_{\text{nuc}}$ can be split into the nucleus–nucleus repulsion (E_{nuc}), the interaction of nuclei with core electrons (E_{NC}) and the interaction among core electrons (E_{CC}). It is noteworthy, that the last two terms are absent for hydrogen (and helium for that matter). Besides the potential contribution from p-type orbitals in the band structure and charge transfer terms, this can explain why repulsive potentials for H–X interactions are often qualitatively different (featuring a saddle or inversion point, for example). The nuclear Coulomb repulsion is given by

$$E_{\text{nuc}} = \sum_{A<B}\frac{Z_A Z_B}{R_{AB}}\;. \tag{6.12}$$

Within the two-center approximation, we can write the interaction of nuclei with core electrons and the interaction among core electron densities as

$$E_{\text{NC}} = -\sum_{A\neq B}\int\frac{Z_A\varrho_B}{\|\mathbf{r}-\mathbf{R}_A\|}\,d\mathbf{r} - \sum_{A}\int\frac{Z_A\varrho_A}{\|\mathbf{r}-\mathbf{R}_A\|}\,d\mathbf{r} \tag{6.13}$$

$$\text{and} \qquad E_{\text{CC}} = \sum_{A<B}\iint\frac{\varrho_A\varrho_B'}{\|\mathbf{r}-\mathbf{r}'\|}\,d\mathbf{r}d\mathbf{r}' + \sum_{A}E_{\text{H}}[\varrho_A]\;, \tag{6.14}$$

where ϱ_A is the core electron density of atom A. Collecting all the above terms, the total repulsive energy in tight-binding approximation can be written as

$$\begin{aligned} E_{\text{tcr}} = &-\frac{1}{2}\sum_{A\neq B}\iint\frac{\rho_A\rho_B'}{\|\mathbf{r}-\mathbf{r}'\|}\,d\mathbf{r}d\mathbf{r}' + \sum_{A<B}\frac{Z_A Z_B}{R_{AB}} \\ &-\sum_{A\neq B}\int\frac{Z_A\varrho_B}{\|\mathbf{r}-\mathbf{R}_A\|}\,d\mathbf{r} + \sum_{A<B}\iint\frac{\varrho_A\varrho_B'}{\|\mathbf{r}-\mathbf{r}'\|}\,d\mathbf{r}d\mathbf{r}' \end{aligned} \tag{6.15}$$

$$-\sum_{A}\left(\frac{1}{2}\iint\frac{\rho_A\rho_A'}{\|\mathbf{r}-\mathbf{r}'\|}\,d\mathbf{r}d\mathbf{r}' - \int\frac{Z_A\varrho_A}{\|\mathbf{r}-\mathbf{R}_A\|}\,d\mathbf{r} + E_{\text{H}}[\varrho_A]\right) + \underbrace{E_{\text{xc}}[\rho_0] - \int\left.\frac{\delta E_{\text{xc}}[\rho]}{\delta\rho}\right|_{\rho_0}\cdot\rho_0\,d\mathbf{r}}_{\approx 0}\;.$$

This can be seen as a justification of the traditional atom-pairwise representation of the repulsive potential. Keeping in mind that this result was only obtained by invoking a two-center approximation, it is also clear that it is unlikely to be able to correct for the shortcomings of the two-center approximation in the tight-binding Hamiltonian. The first two lines of the above expression involve pairs of atoms ("off-site" terms), while the last line only involves single atomic densities (geometry-independent "on-site" terms) and the approximately negligible xc term. Hence, this last line does not depend on the atomic positions and therefore does not affect relative energies, forces, or higher-order derivatives. Being part of the repulsive energy, it also does not affect electronic properties. The interaction between valence and core electron shells is implicitly accounted for by the orthogonalization procedure to obtain the effective single particle basis functions in DFTB.[45,46] In what is to follow, we will thus only consider the first two lines for the repulsive energy.

Two-Center Repulsion-based Filter Generation

In order to obtain an analytical expression and an estimate for the functional form of approximately-pairwise repulsive potentials, we introduce an *ansatz* for the atomic valence and core densities. To ease analytical derivations, we choose the simple, Gaussian-type representations

$$\varrho_A(\mathbf{r}) = n_A \cdot e^{-\frac{1}{\varsigma_A}\|\mathbf{r}-\mathbf{R}_A\|^2} \quad \text{and} \quad \rho_A(\mathbf{r}) = N_A \cdot \left(e^{-\frac{1}{\sigma_A}\|\mathbf{r}-\mathbf{R}_A\|^2} - e^{-\frac{1}{x_A\sigma_A}\|\mathbf{r}-\mathbf{R}_A\|^2}\right) \tag{6.16}$$

for core and valence densities, respectively. Thereby, n_A and N_A serve as normalization factors representing the number of electrons and ς_A, σ_A and $0 \leq x_A \leq 1$ determine the width and shape of the basis functions. Expanding Eq. (6.15) in the above basis set, we can express the two-center-approximated repulsive energy in generalized form (see Appendix A12) as

$$\mathcal{V}_{\text{tcr}}^{AB}(R_{AB}) = \frac{Z_A Z_B}{R_{AB}} + \sum_i p_i \cdot \frac{\operatorname{erf}(q_i \cdot R_{AB})}{R_{AB}} \,. \tag{6.17}$$

We may further extend the above expression by an additional exponentially-decaying repulsive term accounting for overlap-dependent, short-ranged exchange repulsion.‡ While such contributions should be accounted for by the electronic DFTB energy, it is usually described insufficiently. This can be attributed to the use of a minimal, confined basis set,[47,81,82] the neglect of xc effects in the second-order terms (see Sec. 2.2.2) and the parametrization through (semi-)local DFAs within the generalized gradient approximation, which are known to underestimate short-range exchange interactions.[82] The extended functional form then reads

$$\mathcal{V}_{\text{tcr}}^{AB}(R_{AB}) = \frac{Z_A Z_B}{R_{AB}} + \sum_i p_i \cdot \frac{\operatorname{erf}(q_i \cdot R_{AB})}{R_{AB}} + \sum_j v_j \cdot e^{-w_j \cdot R_{AB}} \,, \tag{6.18}$$

with the free fitting parameters p_i and $q_i, v_j, w_j \geq 0$. The above functional form now gives an estimate for the relative importance of neighboring atoms to atomic repulsive energies, which we can use for the construction of a (DT)NN-model for repulsive potentials. The approximate two-center potentials can thereby be used in several ways, for example:

(1) **Two-center repulsion as weighting factor:** The approximate pairwise repulsion can be used as a weighting factor in the Gaussian expansion, *i.e.*,

$$\tau_{ABk} = \mathcal{V}_{\text{tcr}}^{AB}(R_{AB}) \cdot e^{-\kappa(R_{AB}-\mu_k)^2} \quad \text{or} \quad \tau_{ABk} = \mathcal{V}_{\text{tcr}}^{AB}(\mu_k) \cdot e^{-\kappa(R_{AB}-\mu_k)^2} \,. \tag{6.19}$$

The parameters entering $\mathcal{V}_{\text{tcr}}$ can thereby be used as hyperparameters, where the optimal choice is inferred during the learning process. This way, the DTNN preserves a neighbor shell structure, while the interactions are weighted by an effective repulsive pair-potential.

‡ The more complex longer-ranged contributions can be captured explicitly by a range-separated formalism.[84,85]

(2) **Atomic embeddings in the basis of two-center repulsives:** One can alternatively use the obtained information on the (approximate) relevance of neighboring atoms in the basis set representation of the atomic embeddings. Instead of a Gaussian basis, we can use Eq. (6.18) to expand the distance matrix into the atomic embedding tensor,

$$\tau_{ABk} = \frac{Z_A^{(k)} Z_B^{(k)}}{R_{AB}} + \sum_i p_i^{(k)} \cdot \frac{\mathrm{erf}\left(q_i^{(k)} \cdot R_{AB}\right)}{R_{AB}} + \sum_j v_j^{(k)} \cdot e^{-w_j^{(k)} \cdot R_{AB}} \,, \tag{6.20}$$

where each k^{th} slice of τ represents a potential matrix of the above two-center approximation of the repulsive potential with a different set of parameters. As a first-order example, the DTNN can then adaptively create and mix potential matrices for different environments such as saturated and unsaturated C–C bonds. Subsequent interaction refinements then still introduce complex many-body terms, yet the relative importance of neighboring atoms for the individual embeddings is based on the physically-motivated pair-potential (6.18).

The above modifications to the NN architecture assure an inherently more local character, while aiding the model construction with (approximate) physical knowledge on the target quantity. This can improve the overall learning success as well as the robustness and transferability of the obtained NN potentials. The physically-motivated increase of locality can then substantially advance the scalability by avoiding artifacts from unknown global structure motifs. Together with the training set selection optimized to provide a balanced sampling of local degrees of freedom, this may pave the way to extend the accuracy of DFTB-$\mathrm{NN}_{\mathrm{rep}}$ towards larger length and time scales and eventually enable the simulation of biomolecular and nanoscale systems at a hybrid-DFT level of accuracy.

CHAPTER 7

Summary & Outlook

7.1 Synopsis

This section provides a brief and general summary of the results and insights gained in the course of this doctorate. More detailed and specific summaries and conclusions are given in each individual chapter above.

The electronic behaviors and polarization response in solvents are studied by means of a coupled QHO model. Condensation of molecular solvents is analyzed on the basis of emergent spectral characteristics of intrinsic electronic fluctuations. This is particularly pronounced for hydrogen-bond formation, which can be clearly seen from a considerable red shift of spectral weight. The polarization response in simple solvents is found to be characterized by an upper bound decaying with the third power of the distance, which resembles conventional polarization effects. The same behavior is observed for a multicomponent system of a solvated protein when treated in atomistic detail. Within more coarse-grained descriptions of the protein, however, emergent many-body phenomena can lead to substantially-different behaviors and much more long-ranged effects.

Thorough investigation of the (many-body) vdW energetics of protein folding in explicit water highlights the role of beyond-pairwise dispersion forces leading to a conformation-dependent decrease in the *intra*-protein interaction and increase in the protein–water vdW interaction. This increase arises from a remarkable persistence of long-range electron correlation through the aqueous environment, which can play an important role for structure formation and assembly of protein and the coordination of biomolecular processes such as enzymatic action.

A practical method for the description of beyond-pairwise, beyond-RPA dispersion forces is presented. The corresponding *Dipole-Correlated Coulomb Singles* allow to capture multipolar many-body vdW interactions as well as dispersion-polarization effects. Investigation of the binding in supramolecular complexes and under nanoscale confinement highlights the role of such contributions in the context of vdW-bound molecules and materials.

Finally, this thesis develops a hybrid approach of semi-empirical DFTB and localized many-body repulsive potentials from a DTNN-model. The resulting DFTB-NN_{rep} approach is shown to provide substantial improvements over conventional DFTB and bare ML potentials. Most prominently, DFTB-NN_{rep} enables the so-far unattainable accuracy in simultaneously describing energetic, structural and vibrational properties of small organic molecules.

7.2 Perspective and Future Work

In addition to the outlook given in the individual chapters above, this section summarizes projected future steps, challenges and implications in a broader context. This is done individually for the two main topics of this thesis, before concluding with the synergy of both as a reliable, yet efficient approach for the accurate simulation of complex (bio)molecules and materials.

van der Waals Forces in Dynamic (Bio)Molecular Systems

The biomolecular machinery is rich in large-scale structural changes that do not involve breaking covalent bonds. For such processes, vdW interactions represent a major source of interatomic forces and thus play an important role for the dynamics of the system. A well-known example is the folding process of peptides and proteins — especially when considered in the gas phase. In the absence of solvent effects, non-covalent interactions between the residues are responsible for the adaption of a secondary structure. Hence, (accurate) inclusion of vdW interactions is pivotal as even small errors can be propagated to qualitatively wrong results during the long-timescale dynamics. Inclusion of long-range correlation forces has thereby been shown to substantially contribute to the formation of helical entities in polypeptides, [229,230,246,350,351] for example. Another example is liquid water, where vdW forces considerably affect the obtained equilibrium radial distribution and diffusion coefficients. [352–356] Accurate treatment of dispersion interactions is also important for the dynamics of liquid water on 2D-materials, where minimal changes in microscopic geometrical features give rise to significant differences in macroscopic properties. [357]

Given the importance of an accurate treatment of vdW forces for the dynamical behavior of both, gas-phase proteins and water, it is evident that the same applies to proteins in water. The importance of beyond-pairwise vdW dispersion for the relative energetics of single protein structures presented in Chapter 4 further corroborate this conclusion. Locally, long-range interactions likely modify the potential energy landscape only slightly, but they do so on a larger-scale. The plethora of smallish local changes can then combine into substantial effects in the longer-timescale limit of slow processes such as protein folding or enzymatic function. In a similar vein, previous studies indicate that long-range vdW dispersion can have a considerable effect on the collective low-frequency vibrational/phononic behaviors of non-covalently bound systems as these are particularly sensitive to small, but large-scale changes to the curvature of the potential energy surface. [31] These low-frequency, high-entropy vibrational modes play a paramount role for the free energy landscape on which biomolecular processes ultimately occur.

Apart from free energies, the vibrational/phononic spectrum also governs various other phenomena. For instance, the heat capacity or thermal conductivity strongly depend on the phononic spectrum, *cf.* Debye model [358] and *e.g.* (phonon) Boltzmann transport equation. A decisive point for the correct description of such phenomena in condensed systems is thereby the coupling/interaction between moving entities (*i.e.* atoms or molecules). The basic idea is that collective vibrations/phonons as quanta of atomic motion are responsible for the storage and transport of energy supplied to the system in the form of heat. Given that the interaction among (sub)molecular moieties in systems such as layered materials, molecular crystals or biomolecules is largely due to vdW forces, their accurate (many-body) treatment is of utmost importance for the description of such phenomena. From a materials perspective, thermal stability and conductivity represent crucial parameters for technological applications. An interesting example case in the context of dispersion interactions is thereby heat transport in layered materials. In the context of biomolecules, the role of vdW dispersion for heat transport and storage is further strengthened by the fact that non-equilibrium conditions of constant flow and thermalization of energy are inherent in living systems. Obviously, the phononic DOS also governs infrared and vibrational Raman/circular dichroism spectra, where for the latter two also the polarizability and polarization response play an important role. The effect on such spectra is particularly interesting for THz modes, which represent more collective, large-scale motion (see discussion in Sec. 4.5).

The foremost step in future works is hence an extension of the above work on static systems to dynamic systems. This naturally includes zero-point vibrational and finite temperature contributions to the free energy. This also involves the longer-timescale dynamics and phenomena emerging from the effects reported in this thesis ranging from the collective, long-range character of intrinsic electronic fluctuations and vdW interactions as shown in Chapters 3 & 4 to the effect of beyond-dipolar many-body dispersion forces and dispersion-polarization presented in Chapter 5. A maximally-efficient methodology would thereby ideally depart from the atomistic framework used here towards an implicit or coarse-grained description. Such simplified descriptions, however, are vastly restricted and complicated by the inherently emerging collective effects. While the fundamental laws for the interaction and dipolar response between atoms can be fairly-well derived, the characteristics of electronic response and interaction between coarse-grained quantities is far from trivial (see Chapters 3 & 4) and highly dependent on the system and coarse-graining procedure. Obtaining effective interaction laws for coarse-grained moieties from atomistic simulation similar to the procedure introduced in Sec. 4.6 can thereby provide vital guidelines for the construction of effective interaction potentials. Another promising route is combining many-body interatomic approaches with a quantum field theoretical continuum treatment in the form of Casimir forces. First steps towards such a unified framework have been presented in Ref. 359, for instance. This can be of particular interest as the finite speed of light and thus retardation effects on the Coulomb coupling of electronic fluctuations become increasingly relevant when further extending the length scales beyond the systems considered in this thesis. One aspect neglected so far is the effect of charges on the vdW interaction. Solvated and complexed ions are omnipresent under physiological conditions or when describing DNA, enzymes with ionic cofactors or charge-signaling cascades. The field of electronic charges can thereby enhance or inhibit vdW dispersion interactions.[360] Together with the effects of Dipole-Correlated Coulomb Singles introduced in Chapter 5 and the interaction of ions with dispersion-polarized charge densities, this represents an important subject for further investigations and should be included for the accurate and realistic treatment of biomolecular systems.

Advancing Density-Functional Tight-Binding with Machine Learning

Realistic simulations and investigations of the above phenomena ultimately require a total energy method, which in this thesis is proposed to be provided by an efficient, vdW-inclusive electronic-structure method such as DFTB+MBD. In this regard, Chapter 6 shows how many-body repulsive ML potentials can offer a much increased accuracy and reliability of the energetic, structural and vibrational properties obtained from the tight-binding formalism. As already discussed in Sec. 6.3, the main limitations of the developed hybrid QM/ML approach are the large amount of reference data and the so-far limited scalability. Based on the analysis of the locality of repulsive forces (see Fig. 6.1), it can be expected that training sets designed to sample local degrees of freedom as introduced above allow to considerably reduce the amount of reference data needed. This also facilitates the inclusion of further elements and bonding motifs. For a more detailed perspective on this matter, see Sec. 6.3. The main concern of future developments is thus the improvement of scalability; in particular with respect to non-covalent interactions. While the dispersion component can be well captured by the vdW model, contributions from electrostatics, polarization and exchange interactions in the context of large-scale systems remain to be thoroughly addressed and possibly improved. The above analysis showed that the *intra*-molecular repulsive force is rather local and short-ranged. Any insufficiencies in the description of (non-dispersive) non-covalent interactions, however, can be considerably non-local and potentially longer-ranged. Recent work has shown, for example, that conventional DFTB lacks an accurate treatment of long-range electrostatics.[361] In a similar vein, other deficiencies such as the lack of a multipolar description of electrostatics and an accurate description of hydrogen-bonding represent potential pitfalls towards larger-scale systems and important aspects for future work. The implications of such shortcomings for the total energy and interatomic forces can and will be

included in the repulsive potential when constructed to reproduce higher-level reference data. This, however, can substantially reduce the applicability of the parametrization and does not account for the effects on electronic or optical properties. Besides the limitations of the DFTB formalism, the same applies to shortcomings of the underlying DFT description such as the treatment of systems with pronounced multi-reference character (*cf.* static correlation). Such qualitative deficiencies of the electronic DFTB Hamiltonian should not be accounted for by the repulsive potential, but should rather be addressed by the use of (augmenting) physical models as, for example, presented in this thesis for the inclusion of vdW forces. So, in addition to accurate many-body repulsive potentials, a reliable QM treatment of large-scale systems *via* DFTB requires the complementary advancement of the physical description of the electronic structure.

Combining van der Waals Models and Machine-Assisted DFTB

By means of the population matrix-based formalism of Eq. (2.49), DFTB can be directly combined with electronic-structure-based (many-body) vdW models such as MBD. Together with the methodological developments presented in Chapters 5 & 6, this enables the efficient and accurate fully quantum-mechanical treatment of complex systems and the investigation of emergent (quantum) phenomena at increased length and time scales. As in the case of (semi-)local DFT, this combination (maybe inevitably) introduces a certain degree of empiricism and system-specificity. Introducing a range-separation of the coupling tensor as shown in Sec. 2.3.1 provides a solid and, in principle, exact framework for the typically *ad hoc* employed damping function in vdW approaches. For practical applications then, short-range correlation forces are considered to be captured by the DFTB formalism, while the dispersion model is limited to the long-range regime. So, in order to obtain a seamless description of the total system, the range-separating function would need to describe the range of correlation that is captured by the DFTB parameterization and correspondingly switch on the vdW method beyond that. In the case of (semi-)local DFAs, this range is generally not known *a priori* and can vary strongly.[197] The approximations and parameterization in DFTB then blur any clear-cut separation even further. By fitting a combined DFTB+vdW method to total interaction energies, finally, the range-separation function is pushed to (unphysically) also correct for several non-vdW effects including contributions from electrostatics, exchange, induction, and different higher-order correlation terms.

In this regard, the combination of DFTB and ML potentials can offer a practical improvement. A given electronic parameterization can be combined with a reliable vdW method using a range-separating function of fixed form. This can be analogous to the so-called universal damping functions derived for interatomic and intermolecular vdW potentials by Tang and Toennies.[160] The construction of the ML potential for repulsive energies and forces can then be performed on top of the vdW-inclusive DFTB framework to reproduce high-level, correlated reference methods. This allows to include system-dependent corrections to the description of short- and intermediate-range terms without spuriously affecting the range-separation or damping function of the vdW model. As an alternative, one can also employ an adaptive, machine-predicted range-separation for combining DFT(B) with a method for dispersion interactions. Both approaches require the availability of a wealth of well-defined high-quality reference data for sufficiently diverse and large systems.

The open problems outlined above represent the major remaining steps in the more comprehensive effort towards a reliable, bottom-up understanding of complex (bio)molecular systems including a fundamental description of the underlying interaction mechanisms on a quantum-mechanical level. The theoretical insights and methodological advances provided by this thesis thus help to widen our view of the atomistic and molecular mechanisms behind the complex behaviors at increased length and time scales. This may ultimately allow to further exploit emergent nanoscale phenomena in biomolecular and materials science in a more efficient and reliable manner.

Appendices

Parts of Sec. A1, A3&A4, A9 and A10&A11 have been reproduced from the Supporting Materials of Refs. 47, 2, 3 and 4, respectively.

A1 MBD and vdW(TS) Range-Separation Parameters for DFTB

The recommended β (MBD) and s_R (vdW(TS)) values for the *mio* and *3ob* parameter sets are listed in Table A.1. Figures A.1 and A.2 show a summary of the mean absolute errors and mean absolute relative errors in interaction energies and predicted equilibrium distances plus the cost function for DFTB in conjunction with the vdW(TS) and MBD dispersion models as a function of the s_R- and β-parameter, respectively.

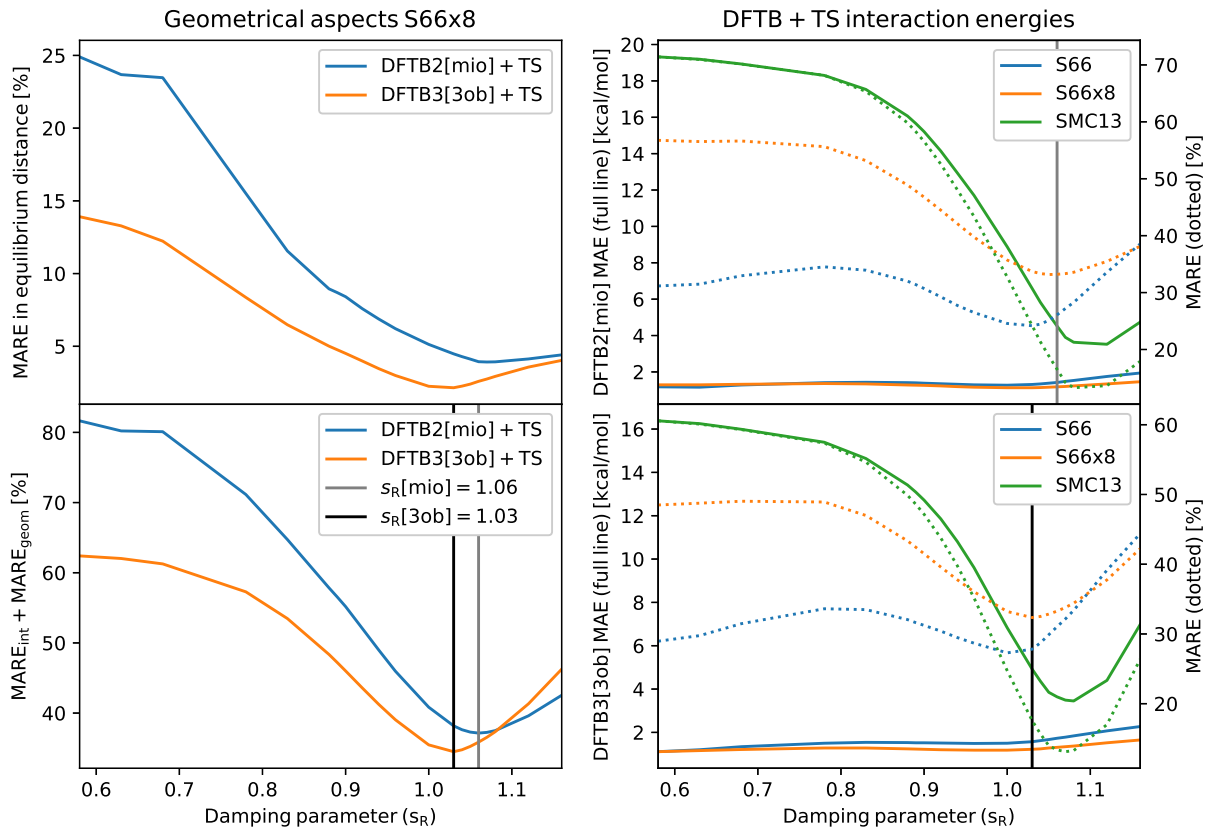

Fig. A.1: Performance of DFTB2[*mio*] and DFTB3[*3ob*] in conjunction with the vdW(TS) dispersion model as a function of the damping parameter s_R. top left: equilibrium binding distances for the S66x8[303] set. top right: DFTB2[*mio*]+vdW(TS) interaction energies for S66,[362] S66x8, and SMC13.[1,33,35] bottom left: cost function combining energetic and geometric aspects of S66x8. bottom right: DFTB3[*3ob*]+vdW(TS) interaction energies for the S66, S66x8, and SMC13.

Tab. A.1: Recommended MBD and vdW(TS) range-separation parameters for the *mio* and *3ob* parametrizations of DFTB as obtained by minimizing the sum of absolute relative errors in interaction energies and predicted equilibrium distances for the S66x8 set[303] of small molecular dimers.

	MBD (β)	vdW(TS) (s_R)
3ob	0.89	1.03
mio	0.95	1.06

Optimal damping and range-separation parameters have been obtained by minimizing the combined cost function

$$F(p_{rs}) = \mathrm{MARE}_{int}(p_{rs}) + \mathrm{MARE}_{geom}(p_{rs}) \tag{A.1}$$

for the S66x8 database[303], where $\mathrm{MARE}_{int}(p_{rs})$ is the mean absolute relative error in interaction energies at a range-separation (or damping) parameter p_{rs}, *i.e.*, β for MBD and s_R for vdW(TS), and $\mathrm{MARE}_{geom}(p_{rs})$ is the corresponding mean absolute relative error in equilibrium distances. The equilibrium distance for each dimer has thereby been estimated by the minimum of the spline-interpolated dissociation curve. In contrast to a previously reported optimization scheme,[318] which inspired the current procedure, a more balanced and straightforward combination of energetic and geometric aspects based on relative errors has been used here.

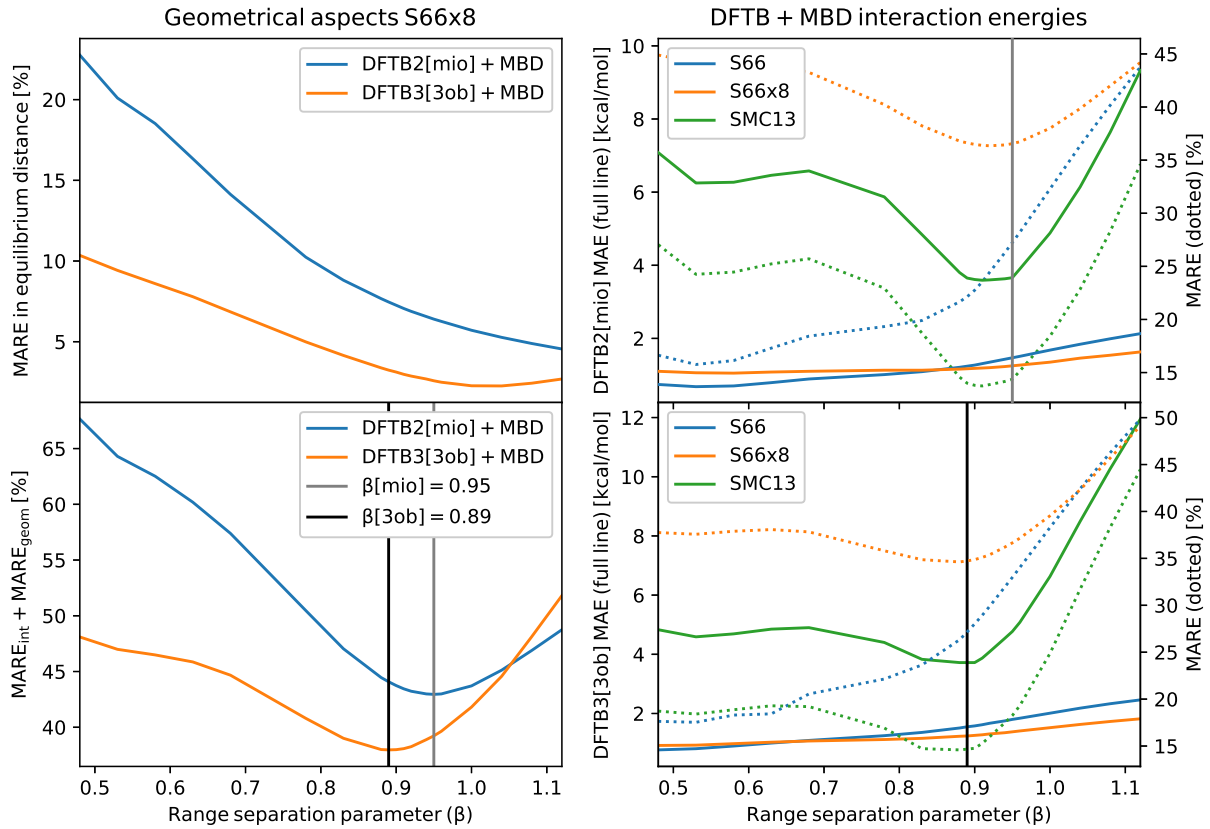

Fig. A.2: Performance of DFTB2[*mio*] and DFTB3[*3ob*] in conjunction with the MBD dispersion model as a function of the range-separation parameter β. top left: equilibrium binding distances for the S66x8[303] set. top right: DFTB2[*mio*]+MBD interaction energies for S66,[362] S66x8, and SMC13.[1,33,35] bottom left: cost function combining energetic and geometric aspects of S66x8. bottom right: DFTB3[*3ob*]+MBD interaction energies for the S66, S66x8, and SMC13.

A2 Mode-Projected van der Waals Interaction Energies

The below formulations represent a generalization of the procedure proposed in Ref. 35.

For the contribution of each mode to the interaction of individual subsystems/fragments (f) within a fully coupled (c) state, let us start with the expression for the total interaction energy:

$$E_{\mathrm{MBD}}^{(\mathrm{int})} = E_{\mathrm{MBD}}^{(\mathrm{system})} - \sum_I E_{\mathrm{MBD}}^{(\mathrm{subsystem}\,I)} = E_{\mathrm{MBD}}^{(\mathrm{c})} - E_{\mathrm{MBD}}^{(\mathrm{f})} \tag{A.2}$$

$$= \frac{1}{2}\left\{\left(\sum_i \omega_i^{(\mathrm{c})} - \sum_i \eta_i^{(\mathrm{c})}\right) - \left(\sum_i \omega_i^{(\mathrm{f})} - \sum_i \eta_i^{(\mathrm{f})}\right)\right\} \qquad \left|\begin{array}{l}\boldsymbol{\omega} = \bigoplus_i \omega_i \\ \boldsymbol{\eta} = \bigoplus_i \eta_i\end{array}\right. \tag{A.3}$$

$$= \frac{1}{2}\left\{\langle \mathbf{1}\,|\,\boldsymbol{\omega}_{\mathrm{c}}\rangle - \left\langle \mathbf{1}\,\middle|\,\boldsymbol{\eta}_{\mathrm{c}}\right\rangle - \langle \mathbf{1}\,|\,\boldsymbol{\omega}_{\mathrm{f}}\rangle + \left\langle \mathbf{1}\,\middle|\,\boldsymbol{\eta}_{\mathrm{f}}\right\rangle\right\} \tag{A.4}$$

$$= \frac{1}{2}\left\{\langle \mathbf{1}\,|\,\boldsymbol{\omega}_{\mathrm{c}}\rangle - \langle \mathbf{1}\,|\,\boldsymbol{\omega}_{\mathrm{f}}\rangle - \left(\left\langle \mathbf{1}\,\middle|\,\boldsymbol{\eta}_{\mathrm{c}}\right\rangle - \left\langle \mathbf{1}\,\middle|\,\boldsymbol{\eta}_{\mathrm{f}}\right\rangle\right)\right\}\ , \tag{A.5}$$

where $\mathbf{1}$ is a $3N$-dimensional vector with all ones and $\boldsymbol{\omega}$ and $\boldsymbol{\eta}$ collect all ω_i and η_i, respectively. Note that, before entering the MBD model, the polarizabilities and effective frequencies are subject to electrostatic screening based on the dipole–dipole interaction tensor for the fully coupled system ($\boldsymbol{\omega}_{\mathrm{c}}$) and for the uncoupled fragments ($\boldsymbol{\omega}_{\mathrm{f}}$), respectively. Hence, $\boldsymbol{\omega}_{\mathrm{c}} \neq \boldsymbol{\omega}_{\mathrm{f}}$.

Considering the above procedure of solving the MBD model the sets of eigenfrequencies $\boldsymbol{\omega}$ correspond to different collective coordinates (different eigenspaces) obtained from simple displacements *via* the transformation matrices $\mathbf{C}_{\mathrm{c}}$ ($\rightarrow \boldsymbol{\omega}_{\mathrm{c}}$), $\mathbf{C}_{\mathrm{f}} = \oplus_I \mathbf{C}_I$ ($\rightarrow \boldsymbol{\omega}_{\mathrm{f}}$), and $\mathbb{1}$ ($\rightarrow \boldsymbol{\eta}$), with $\mathbf{C}_I$ corresponding to subsystem I and $\mathbb{1}$ being the unit matrix. To combine the energies we project all vectors of eigenfrequencies to the same collective coordinates (basis) and for that we choose the fluctuation modes of the isolated subsystems. For the projection of energies corresponding to modes obtained from the transformation matrix $\mathbf{X}$ to the basis obtained via $\mathbf{C}_{\mathrm{f}}$ we choose the element-wise product of the matrix-product $\mathbf{C}_{\mathrm{f}}^{\dagger}\mathbf{X}$ and its complex conjugate:

$$\mathcal{U}[\mathbf{X}] = \left(\mathbf{C}_{\mathrm{f}}^{\dagger}\mathbf{X}\right)^{*} \circ \left(\mathbf{C}_{\mathrm{f}}^{\dagger}\mathbf{X}\right) \tag{A.6}$$

For unitary $\mathbf{X}$, the matrix product $\mathbf{C}_{\mathrm{f}}^{\dagger}\mathbf{X}$ yields a transformation to $\mathbf{C}_{\mathrm{f}}$, which is orthonormal/unitary, the element-wise product with its complex conjugate gives a corresponding doubly-stochastic transformation matrix, that upon multiplication with a vector preserves its L1-norm and the sum of its elements. Hence, $\langle \mathbf{1}\,|\,\mathbf{v}\rangle = \left\langle \mathbf{1}\,\middle|\,\mathcal{U}[\mathbf{X}]\mathbf{v}\right\rangle$ and the transformation preserves the total energy. Inserting this into the previous expression (and transforming the result back to the eigenspace of the coupled system) gives

$$E_{\mathrm{MBD}}^{(\mathrm{int})} = \frac{1}{2}\left\{\langle \mathbf{1}\,|\,\boldsymbol{\omega}_{\mathrm{c}}\rangle - \langle \mathbf{1}\,|\,\boldsymbol{\omega}_{\mathrm{f}}\rangle - \left\langle \mathbf{1}\,\middle|\,\left(\boldsymbol{\eta}_{\mathrm{c}} - \boldsymbol{\eta}_{\mathrm{f}}\right)\right\rangle\right\} \tag{A.7}$$

$$= \frac{1}{2}\left\{\left\langle \mathbf{1}\,\middle|\,\mathcal{U}[\mathbf{C}_{\mathrm{c}}]\boldsymbol{\omega}_{\mathrm{c}}\right\rangle - \langle \mathbf{1}\,|\,\boldsymbol{\omega}_{\mathrm{f}}\rangle - \left\langle \mathbf{1}\,\middle|\,\mathcal{U}[\mathbb{1}]\left(\boldsymbol{\eta}_{\mathrm{c}} - \boldsymbol{\eta}_{\mathrm{f}}\right)\right\rangle\right\} \tag{A.8}$$

$$= \left\langle \mathbf{1}\,\middle|\,\frac{1}{2}\left\{\mathcal{U}[\mathbf{C}_{\mathrm{c}}]\boldsymbol{\omega}_{\mathrm{c}} - \boldsymbol{\omega}_{\mathrm{f}} - \mathcal{U}[\mathbb{1}]\left(\boldsymbol{\eta}_{\mathrm{c}} - \boldsymbol{\eta}_{\mathrm{f}}\right)\right\}\right\rangle \tag{A.9}$$

$$= \left\langle \mathbf{1}\,\middle|\,\frac{1}{2}\mathcal{U}^{\dagger}[\mathbf{C}_{\mathrm{c}}]\left\{\mathcal{U}[\mathbf{C}_{\mathrm{c}}]\boldsymbol{\omega}_{\mathrm{c}} - \boldsymbol{\omega}_{\mathrm{f}} - \mathcal{U}[\mathbb{1}]\left(\boldsymbol{\eta}_{\mathrm{c}} - \boldsymbol{\eta}_{\mathrm{f}}\right)\right\}\right\rangle \tag{A.10}$$

$$= \langle \mathbf{1}\,|\,\boldsymbol{\varepsilon}_{\mathrm{int}}\rangle = \sum_i \varepsilon_{\mathrm{int}}^{(i)} \tag{A.11}$$

With this we get the contribution $\varepsilon_{(\mathrm{int})}^{(i)}$ of the i^{th} collective electronic "eigenmode" to the interaction of the subsystems — in the present case the solute–solvent interaction.

A3 vdW Interaction for Improved Sampling of Disordered States

As detailed in Chapter 4, conventional MM force fields in conjunction with traditional water models such as the standard TIP3P, predict unfolded protein states that are too compact. This has largely been assigned to an unbalanced description of *intra*-protein and protein–water vdW interactions. Our present findings thereby provide a fundamental, quantum-mechanical explanation for this argument: Due to many-atom effects beyond the traditional pairwise formulation of vdW energetics, we observe a system- and conformation-dependent change in the relative magnitude of the individual vdW contributions. In the main part of Chapter 4, these conclusions have been based on structures obtained employing the above mentioned unbalanced description and thus featured unfolded states, which are too compact. As an ultimate confirmation of our results, we have performed our analysis based on a new sampling of the proteins' conformational space using the recent *a99SB-disp* force field together with the TIP4P-D water model as developed by Robustelli and coworkers. This setup has been designed and shown to provide an improved description of unfolded and disordered protein states avoiding spurious over-compaction.[228] To study potential effects of this over-compaction, we performed new molecular dynamics simulations of the Chignolin variant "cln025" in explicit water starting from an unfolded and folded state, respectively. Subsequently, we evaluate the vdW energetics in the same manner as before, where the only difference is the usage of a more correct and representative sampling of unfolded states as obtained from the new molecular dynamics simulations.

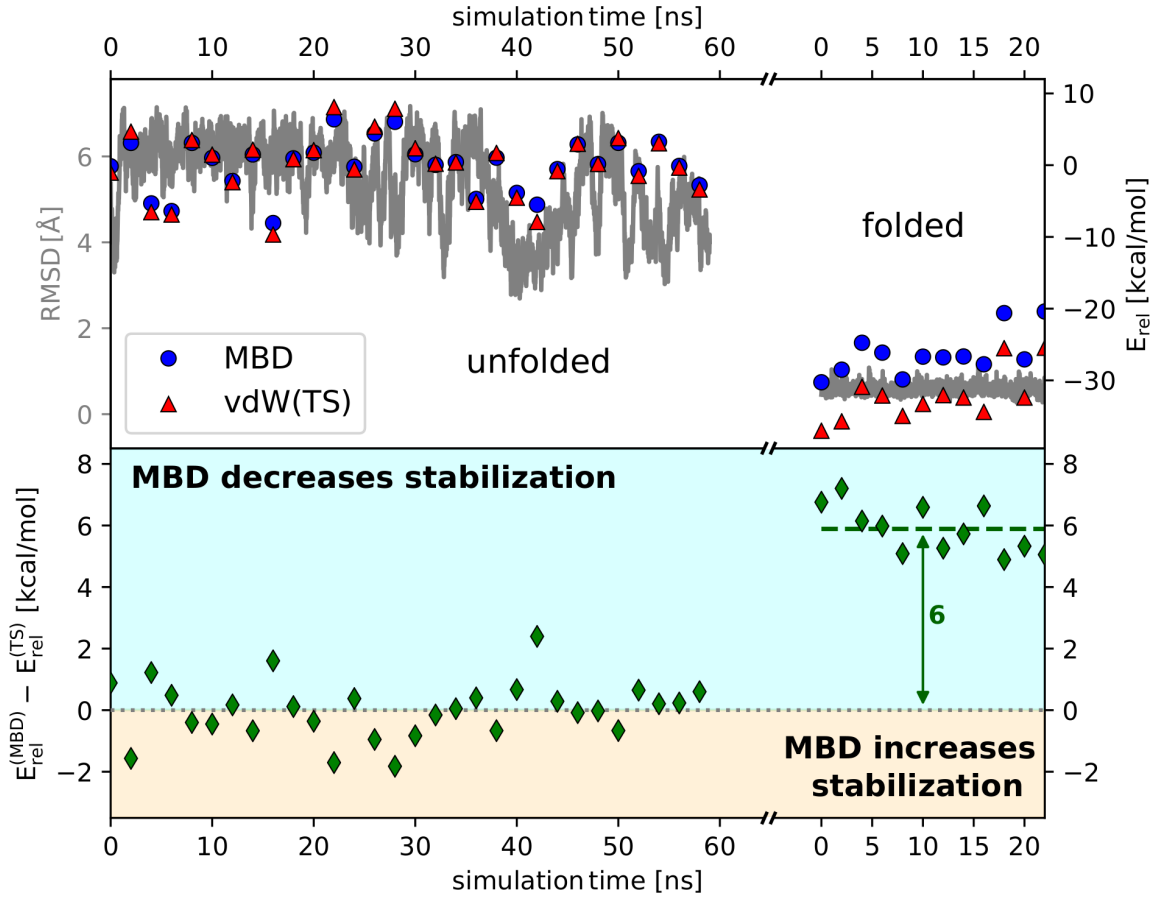

Fig. A.3: *Intra*-protein van der Waals energetics of Chignolin variant "cln025" as obtained from improved sampling of disordered states Top: relative *intra*-protein van der Waals interaction as obtained from pairwise description (vdW(TS)) and many-body formalism (MBD). Backbone RMSD (grey) taken with respect to native state. Bottom: many-body contributions as defined by the difference between MBD and vdW(TS).

Fig. A.3 shows the *intra*-protein vdW interaction energy based on this new sampling (left: unfolded state sampling, right: folded state sampling) in the top graph. The RMSD of both samplings is taken with respect to the native conformation. The bottom graph again depicts the difference between the pairwise description in vdW(TS) and the many-body energetics obtained from MBD. All in all, Fig. A.3 confirms our previous findings of the pairwise formalism overestimating the internal stabilization in the pairwise description. As can be seen from Fig. A.4, the improved sampling of unfolded conformations neither affects our conclusions for the vdW solvation energy: It still tracks with the (inverse) geometrical RMSD (top graph), and many-body effects increase the relative protein–water vdW interaction in the native state (bottom graph). In conclusion, the results obtained for the previous trajectories can clearly be assigned to a failure of the pairwise approximation and do not simply represent an artifact of the incorrect sampling of unfolded states.

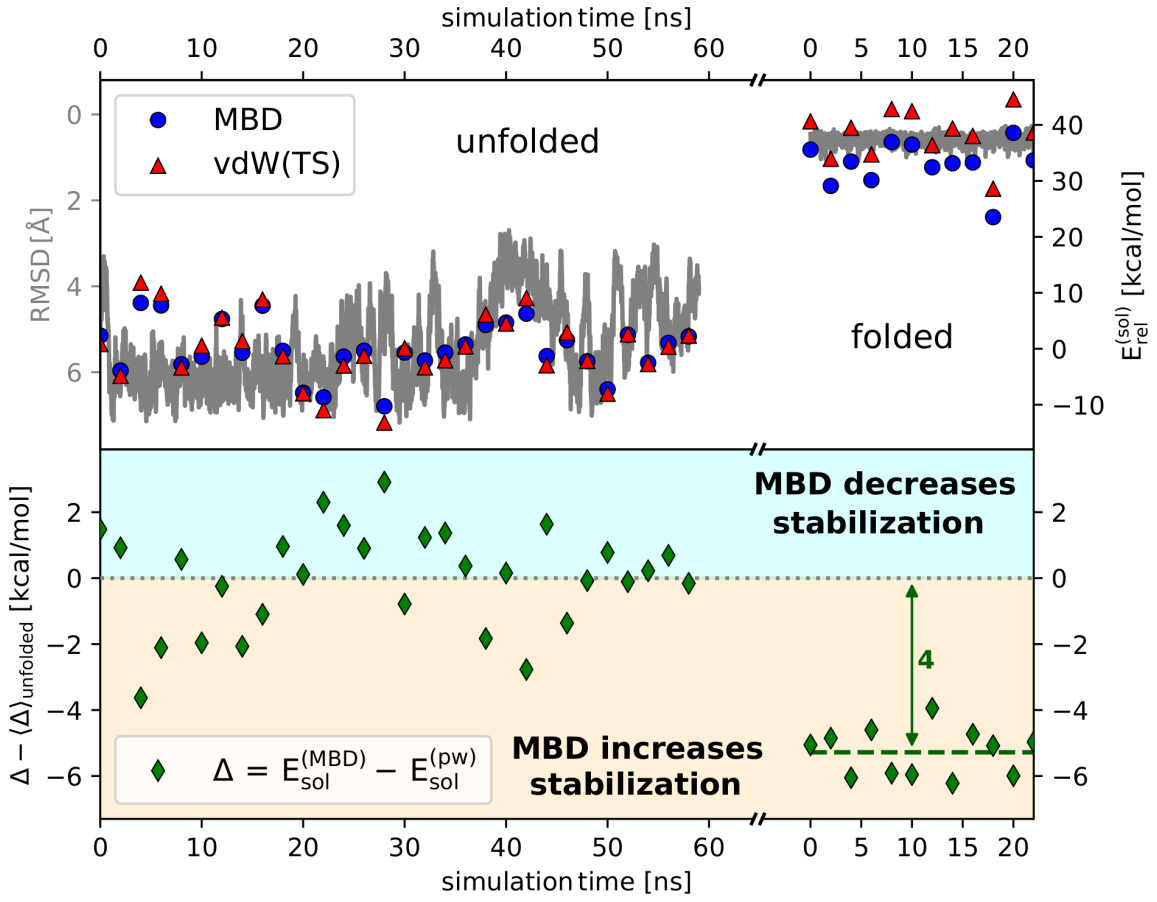

Fig. A.4: van der Waals solvation energy for Chignolin variant "cln025" as obtained from improved sampling of disordered states. Top: relative van der Waals solvation energy as obtained from pairwise description (vdW(TS)) and many-body formalism (MBD). Backbone RMSD (grey) taken with respect to native state. Bottom: beyond-pairwise contributions as defined by the difference between MBD and vdW(TS).

A4 Rescaling of Pairwise-Additive van der Waals Solvation Energies

As can be seen from Fig. 4.1 and Chapter 4, the overestimation of vdW energies increases with the absolute vdW interaction energy. Correspondingly, a simple rescaling of the pairwise approaches considerably improves the agreement with the many-body treatment. Fig. A.5 shows the correlation between such optimally rescaled vdW solvation energies and MBD. The obtained rescaling factors show that relying on electronic-structure-based C_6 interaction coefficients as done within vdW(TS) provides the best estimate for vdW solvation energies. This can mainly be attributed to the description of the pure solvent as the geometry-based D2 and D3 methods outperform vdW(TS) for gas-phase energetics (*vide supra*). Despite the overall improvement, the deviation between the optimally-rescaled pairwise approaches and MBD still regularly exceeds 4 kcal/mol. Furthermore, the optimal rescaling factors are highly system- and method-dependent and can only be obtained as an *a posteriori* correction.

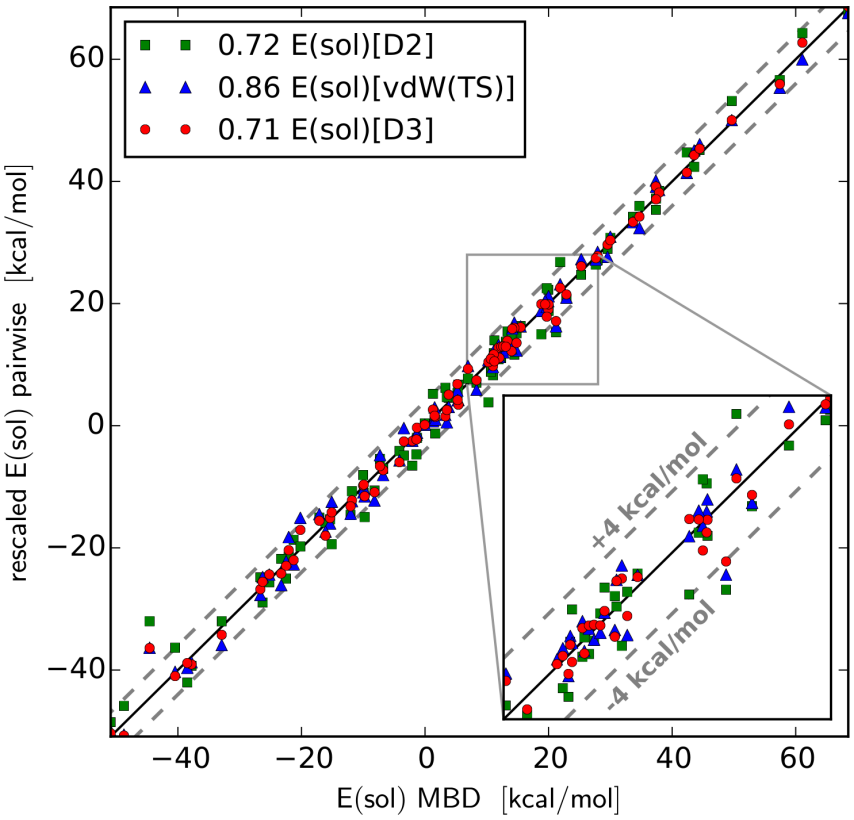

Fig. A.5: Correlation of rescaled relative van der Waals solvation energies as obtained from pairwise models in comparison to the results obtained from many-body treatment.

A5 Coulomb Interaction Energy for Dipole-Coupled Oscillators

The derivations presented in this section have been carried out in similar form by Dr J. Hermann, see Ref. 363.

The expectation value of the Coulomb interaction energy for a set of dipole-coupled (DC) quantum harmonic oscillators (QHOs) is given by

$$E_{\mathrm{Coul}}[\Psi_{\mathrm{DC}}] = \left(\prod_{i=1}^{3N}\sqrt{\frac{\omega_i}{\pi}}\right)\int\cdots\int \sum_{A<B} g_{\mathrm{rs}}(R_{AB})\,\frac{q_A\,q_B}{\|\mathbf{r}_A-\mathbf{r}_B\|}\,e^{-\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\zeta}}\prod_{i=1}^{3N}d\xi_i\;. \tag{A.12}$$

We then separate the exponent into terms according to their dependency on the coordinates of the pseudo-particles of A and B (*i.e.*, on $\boldsymbol{\zeta}_{AB}=\boldsymbol{\zeta}_A\oplus\boldsymbol{\zeta}_B$):

$$\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\zeta} = \sum_{j,l}\boldsymbol{\Omega}_{jl}\zeta_j\zeta_l = \sum_{\substack{j\notin A,B\\ l\notin A,B}}\boldsymbol{\Omega}_{jl}\zeta_j\zeta_l \;+\; 2\sum_{\substack{m\in A,B\\ n\notin A,B}}\boldsymbol{\Omega}_{mn}\zeta_m\zeta_n \;+\; \sum_{\substack{p\in A,B\\ q\in A,B}}\boldsymbol{\Omega}_{pq}\zeta_p\zeta_q \tag{A.13}$$

$$= \boldsymbol{\zeta}'^{\mathrm{T}}_{AB}\boldsymbol{\Omega}''_{AB}\boldsymbol{\zeta}'_{AB} \;+\; 2\;\;\boldsymbol{\zeta}'^{\mathrm{T}}_{AB}\boldsymbol{\Omega}'_{AB}\boldsymbol{\zeta}_{AB} \;\;+\;\;\boldsymbol{\zeta}_{AB}\boldsymbol{\Omega}_{AB}\boldsymbol{\zeta}_{AB}\;. \tag{A.14}$$

By completing the square, in its generalized form for symmetric, invertible matrices $\mathbf{M}$,

$$\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}+\mathbf{x}^{\mathrm{T}}\mathbf{v}+d = \left(\mathbf{x}+{}^1\!/_2\mathbf{M}^{-1}\mathbf{v}\right)^{\mathrm{T}}\mathbf{M}\left(\mathbf{x}+{}^1\!/_2\mathbf{M}^{-1}\mathbf{v}\right)+d-{}^1\!/_4\mathbf{v}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{v}\;, \tag{A.15}$$

we complete the square with respect to $\boldsymbol{\zeta}_{AB}$:

$$\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\zeta} = \overbrace{\boldsymbol{\zeta}'^{\mathrm{T}}_{AB}}^{\mathbf{x}^{\mathrm{T}}}\overbrace{\boldsymbol{\Omega}''_{AB}}^{\mathbf{M}}\overbrace{\boldsymbol{\zeta}'_{AB}}^{\mathbf{x}} \;+\; \overbrace{\boldsymbol{\zeta}'^{\mathrm{T}}_{A}}^{\mathbf{x}^{\mathrm{T}}}\overbrace{2\boldsymbol{\Omega}'_{AB}\boldsymbol{\zeta}_{AB}}^{\mathbf{v}} \;+\; \overbrace{\boldsymbol{\zeta}_{AB}\boldsymbol{\Omega}_{AB}\boldsymbol{\zeta}_{AB}}^{d}$$

$$= \left(\boldsymbol{\zeta}'_{AB}+\boldsymbol{\Omega}''^{-1}_{AB}\boldsymbol{\Omega}'_{AB}\boldsymbol{\zeta}_{AB}\right)^{\mathrm{T}}\boldsymbol{\Omega}''_{AB}\left(\boldsymbol{\zeta}'_{AB}+\boldsymbol{\Omega}''^{-1}_{AB}\boldsymbol{\Omega}'_{AB}\boldsymbol{\zeta}_{AB}\right)+\boldsymbol{\zeta}^{\mathrm{T}}_{AB}\boldsymbol{\Omega}_{AB}\boldsymbol{\zeta}_{AB}-\boldsymbol{\zeta}^{\mathrm{T}}_{AB}\boldsymbol{\Omega}'^{\mathrm{T}}_{AB}\boldsymbol{\Omega}''^{-1}_{AB}\boldsymbol{\Omega}'_{AB}\boldsymbol{\zeta}_{AB}$$

$$= \boldsymbol{\chi}^{\mathrm{T}}_{AB}\;\boldsymbol{\Omega}''_{AB}\;\boldsymbol{\chi}_{AB}+\boldsymbol{\zeta}^{\mathrm{T}}_{AB}\;\boldsymbol{\Upsilon}_{AB}\;\boldsymbol{\zeta}_{AB}\;, \tag{A.16}$$

where $\boldsymbol{\Upsilon}_{AB}=\boldsymbol{\Omega}_{AB}-\boldsymbol{\Omega}'^{\mathrm{T}}_{AB}\boldsymbol{\Omega}''^{-1}_{AB}\boldsymbol{\Omega}'_{AB}$. Using the above notation, we re-order Eq. (A.12) such that

$$E_{\mathrm{Coul}}[\Psi_{\mathrm{DC}}] = \left(\prod_{i=1}^{3N}\sqrt{\frac{\omega_i}{\pi}}\right)\sum_{A<B}\iint g_{\mathrm{rs}}(R_{AB})\,\frac{q_A\,q_B}{\|\mathbf{r}_A-\mathbf{r}_B\|}\,e^{-\boldsymbol{\zeta}^{\mathrm{T}}_{AB}\boldsymbol{\Upsilon}_{AB}\boldsymbol{\zeta}_{AB}} \times\left(\int e^{-\boldsymbol{\chi}^{\mathrm{T}}_{AB}\boldsymbol{\Omega}''_{AB}\boldsymbol{\chi}_{AB}}\,d\boldsymbol{\zeta}'_{AB}\right)d\boldsymbol{\zeta}_{AB}\;, \tag{A.17}$$

Note that $\boldsymbol{\xi}\rightarrow\boldsymbol{\zeta}$ is a unitary transformation and hence we can substitute $\prod d\xi_i$ with $d\boldsymbol{\zeta}'_{AB}\,d\boldsymbol{\zeta}_{AB}$. For the integral in parenthesis, we change the variable of integration from $\boldsymbol{\zeta}'_{AB}$ to $\boldsymbol{\chi}_{AB}$ with $d\boldsymbol{\zeta}'_{AB}=d\boldsymbol{\chi}_{AB}$ and then substitute $\boldsymbol{\gamma}_{AB}=\mathbf{U}^{\mathrm{T}}\boldsymbol{\chi}_{AB}$, where $\mathbf{U}^{\mathrm{T}}\boldsymbol{\Omega}''_{AB}\mathbf{U}=\mathrm{diag}\{\lambda_{AB,j}\}$. Given that the Jacobian of this transformation is simply $\mathbf{U}^{\mathrm{T}}$, which is a unitary matrix, and that the (absolute of the) determinant of unitary matrices is 1, the integral in parentheses above can be simplified according to:

$$\int e^{-\boldsymbol{\chi}^{\mathrm{T}}_{AB}\boldsymbol{\Omega}''_{AB}\boldsymbol{\chi}_{AB}}\,d\boldsymbol{\zeta}'_{AB} = \int e^{-\boldsymbol{\chi}^{\mathrm{T}}_{AB}\boldsymbol{\Omega}''_{AB}\boldsymbol{\chi}_{AB}}\,d\boldsymbol{\chi}_{AB} = \int e^{-\boldsymbol{\gamma}^{\mathrm{T}}_{AB}\mathbf{U}^{\mathrm{T}}\boldsymbol{\Omega}''_{AB}\mathbf{U}\boldsymbol{\gamma}_{AB}}\,d\boldsymbol{\gamma}_{AB}$$

$$= \prod_{j=1}^{3N-6}\int e^{-\lambda_{AB,j}\gamma^2_{A,j}}\,d\gamma_{A,j} \;=\; \prod_{j=1}^{3N-6}\sqrt{\frac{\pi}{\lambda_{AB,j}}} = \frac{\pi^{-3}}{\sqrt{\det\{\boldsymbol{\Omega}''_{AB}\}}}\prod_{j=1}^{3N}\sqrt{\pi}\;, \tag{A.18}$$

where in the last step we have used the fact that the product of eigenvalues corresponds to the determinant. Using the same for the product of ω_i in the prefactor (*i.e.*, $\prod\omega_i = \det\{\boldsymbol{\Omega}\}$) and combining the occurrences of π, we can now write the Coulomb interaction energy as

$$E_{\mathrm{Coul}}[\Psi_{\mathrm{DC}}] = \pi^{-3}\sum_{A<B}\sqrt{\frac{\det\{\boldsymbol{\Omega}\}}{\det\{\boldsymbol{\Omega}''_{AB}\}}}\underbrace{\iint g_{\mathrm{rs}}(R_{AB})\,\frac{q_A\,q_B}{\|\mathbf{r}_A-\mathbf{r}_B\|}\,e^{-\boldsymbol{\zeta}_{AB}^{\mathrm{T}}\boldsymbol{\Upsilon}_{AB}\boldsymbol{\zeta}_{AB}}\,d\boldsymbol{\zeta}_{AB}}_{I_{\boldsymbol{\zeta}}}\,. \tag{A.19}$$

Let us now focus on the last (double) integral, $I_{\boldsymbol{\zeta}}$. First, we re-write the Coulomb potential,

$$\frac{1}{\|\mathbf{r}_A-\mathbf{r}_B\|} = \frac{2}{\sqrt{\pi}}\int_0^\infty e^{-\|\mathbf{r}_A-\mathbf{r}_B\|^2\,s^2}\,ds = \frac{2}{\sqrt{\pi}}\int_0^\infty e^{-\mathbf{r}_{AB}^{\mathrm{T}}\mathbf{S}_2\,\mathbf{r}_{AB}}\,ds\,, \tag{A.20}$$

where

$$\mathbf{r}_{AB} = \mathbf{r}_A\oplus\mathbf{r}_B \qquad\text{and}\qquad \mathbf{S}_2 = s^2\begin{pmatrix}1&0&0&-1&0&0\\0&1&0&0&-1&0\\0&0&1&0&0&-1\\-1&0&0&1&0&0\\0&-1&0&0&1&0\\0&0&-1&0&0&1\end{pmatrix} = s^2\begin{pmatrix}\mathbb{1}_3&-\mathbb{1}_3\\-\mathbb{1}_3&\mathbb{1}_3\end{pmatrix}. \tag{A.21}$$

We then insert $\boldsymbol{\zeta}_{AB} = \sqrt{m_A}(\mathbf{r}_A-\mathbf{R}_A)\oplus\sqrt{m_B}(\mathbf{r}_B-\mathbf{R}_B)$ and $d\boldsymbol{\zeta}_{AB} = (m_A\,m_B)^{3/2}\,d\mathbf{r}_{AB}$ and sort the exponents to arrive at

$$I_{\boldsymbol{\zeta}} = \frac{2\,q_A q_B}{\sqrt{\pi}}(m_A\,m_B)^{3/2}g_{\mathrm{rs}}(R_{AB})\int_0^\infty\underbrace{\iint e^{-\left[\mathbf{r}_{AB}^{\mathrm{T}}\left(\boldsymbol{\Gamma}_{AB}+\mathbf{S}_2\right)\mathbf{r}_{AB}\;-\;2\mathbf{r}_{AB}^{\mathrm{T}}\boldsymbol{\Gamma}_{AB}\mathbf{R}_{AB}\right]}\,d\mathbf{r}_{AB}}_{I_{\mathbf{r}}}\,ds\,, \tag{A.22}$$

where $\boldsymbol{\Gamma}_{AB} = \mathbf{D}_m\boldsymbol{\Upsilon}_{AB}\mathbf{D}_m$ with $\mathbf{D}_m = \mathrm{diag}\{m_A,m_A,m_A,m_B,m_B,m_B\}^{1/2} = \left(\sqrt{m_A}\,\mathbb{1}_3\right)\oplus\left(\sqrt{m_B}\,\mathbb{1}_3\right)$. For the integral $I_{\mathbf{r}}$, completing the square in the exponential according to Eq. (A.15) with respect to $\mathbf{r}_{AB}$ gives

$$\begin{aligned}I_{\mathbf{r}} &= \iint\exp\Big\{-\boldsymbol{\vartheta}_{AB}^{\mathrm{T}}\left(\boldsymbol{\Gamma}_{AB}+\mathbf{S}_2\right)\boldsymbol{\vartheta}_{AB}-\mathbf{R}_{AB}^{\mathrm{T}}\left[\boldsymbol{\Gamma}_{AB}-\boldsymbol{\Gamma}_{AB}^{\mathrm{T}}\left(\boldsymbol{\Gamma}_{AB}+\mathbf{S}_2\right)^{-1}\boldsymbol{\Gamma}_{AB}\right]\mathbf{R}_{AB}\Big\}\,d\mathbf{r}_{AB}\\ &= \exp\left(-\mathbf{R}_{AB}^{\mathrm{T}}\boldsymbol{\Theta}_{AB}\,\mathbf{R}_{AB}\right)\iint\exp\left(-\boldsymbol{\vartheta}_{AB}^{\mathrm{T}}\left(\boldsymbol{\Gamma}_{AB}+\mathbf{S}_2\right)\boldsymbol{\vartheta}_{AB}\right)d\mathbf{r}_{AB}\,,\end{aligned} \tag{A.23}$$

where we have introduced $\boldsymbol{\Theta}_{AB} = \boldsymbol{\Gamma}_{AB}-\boldsymbol{\Gamma}_{AB}^{\mathrm{T}}\left(\boldsymbol{\Gamma}_{AB}+\mathbf{S}_2\right)^{-1}\boldsymbol{\Gamma}_{AB}$ and $\boldsymbol{\vartheta}_{AB} = \mathbf{r}_{AB}-\left(\boldsymbol{\Gamma}_{AB}+\mathbf{S}_2\right)^{-1}\boldsymbol{\Gamma}_{AB}\,\mathbf{R}_{AB}$. The double integral in the above expression can be solved easily by diagonalization similar to Eq. (A.18). Taking into account that $\boldsymbol{\vartheta}_{AB}$ is six-dimensional, we obtain

$$\iint\exp\left(-\boldsymbol{\vartheta}_{AB}^{\mathrm{T}}\left(\boldsymbol{\Gamma}_{AB}+\mathbf{S}_2\right)\boldsymbol{\vartheta}_{AB}\right)d\mathbf{r}_{AB} = \frac{\pi^3}{\sqrt{\det\{\boldsymbol{\Gamma}_{AB}+\mathbf{S}_2\}}} \tag{A.24}$$

Combining the above with Eqs. (A.19), (A.22) and (A.23) finally yields

$$E_{\mathrm{Coul}}[\Psi_{\mathrm{DC}}] = \sum_{A<B}^{N}\frac{2\,q_A q_B}{\sqrt{\pi}}\left(m_A\,m_B\right)^{3/2}g_{\mathrm{rs}}(R_{AB})\sqrt{\frac{\det\{\boldsymbol{\Omega}\}}{\det\{\boldsymbol{\Omega}''_{AB}\}}}\int_0^\infty\frac{e^{-\mathbf{R}_{AB}^{\mathrm{T}}\boldsymbol{\Theta}_{AB}\,\mathbf{R}_{AB}}}{\sqrt{\det\{\boldsymbol{\Gamma}_{AB}+\mathbf{S}_2\}}}\,ds\,. \tag{A.25}$$

A6 Dipole Interaction Energy for Dipole-Coupled Oscillators

The dipolar interaction potential as used in the MBD formalism is given by

$$\hat{\mathcal{V}}_{\mathrm{dip}} = \frac{1}{2}\sum_{A,B} g_{\mathrm{rs}}(R_{AB})\,\eta_A\eta_B\sqrt{\alpha^{(\mathrm{sr})}_{A,0}\,\alpha^{(\mathrm{sr})}_{B,0}}\,\zeta_A^{\mathrm{T}}\tilde{\mathbf{T}}_{AB}\,\zeta_B = \frac{1}{2}\sum_{A,B}\zeta_A^{\mathrm{T}}\tilde{\mathbf{T}}^{(AB)}_{\zeta}\,\zeta_B = \frac{1}{2}\boldsymbol{\zeta}^{\mathrm{T}}\tilde{\mathbf{T}}_{\zeta}\boldsymbol{\zeta}\,, \tag{A.26}$$

where in the second step we have introduced $\tilde{\mathbf{T}}_{\zeta} = g_{\mathrm{rs}}(R_{AB})\,\eta_A\eta_B\sqrt{\alpha^{(\mathrm{sr})}_{A,0}\,\alpha^{(\mathrm{sr})}_{B,0}}\,\mathbf{T}_{AB}$ and in the last step we used the definition of the generalized $3N$-dimensional coordinate $\boldsymbol{\zeta}$ as used above. Using the definition $\boldsymbol{\xi} = \mathbf{C}^{\mathrm{T}}\boldsymbol{\zeta}$ and $\mathbf{C}\mathbf{C}^{\mathrm{T}} = \mathbb{1}_{3N}$, we can write

$$\hat{\mathcal{V}}_{\mathrm{dip}} = \frac{1}{2}\boldsymbol{\xi}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\tilde{\mathbf{T}}_{\zeta}\mathbf{C}\boldsymbol{\xi} = \frac{1}{2}\boldsymbol{\xi}^{\mathrm{T}}\tilde{\mathbf{T}}_{\xi}\,\boldsymbol{\xi} = \frac{1}{2}\sum_{i,j}\xi_i\,\xi_j\tilde{\mathbf{T}}^{(ij)}_{\xi}\left(= \frac{1}{2}\sum_{i,j}\xi_i\,\xi_j\sum_{k,l}\mathbf{C}_{ki}\mathbf{C}_{lj}\tilde{\mathbf{T}}^{(kl)}_{\zeta}\right), \tag{A.27}$$

which transforms the dipolar interaction potential to the collective variable $\boldsymbol{\xi}$. With this we can write the expectation value of the dipolar interaction potential as

$$\begin{aligned}
E_{\mathrm{dip}}[\Psi_{\mathrm{DC}}] &= \int\cdots\int \overbrace{\left[\prod_k \sqrt[4]{\frac{\omega_k}{\pi}}e^{-\frac{\omega_k}{2}\xi_k^2}\right]}^{\Psi_{\mathrm{MBD}}}\overbrace{\left(\frac{1}{2}\sum_{i,j}\xi_i\,\xi_j\tilde{\mathbf{T}}^{(ij)}_{\xi}\right)}^{\hat{\mathcal{V}}_{\mathrm{dip}}}\overbrace{\left[\prod_l \sqrt[4]{\frac{\omega_l}{\pi}}e^{-\frac{\omega_l}{2}\xi_l^2}\right]}^{\Psi_{\mathrm{MBD}}}\prod_n d\xi_n \\
&= \frac{1}{2}\sum_{i,j}\tilde{\mathbf{T}}^{(ij)}_{\xi}\int\cdots\int\xi_i\,\xi_j\left[\prod_k \sqrt[4]{\frac{\omega_k}{\pi}}e^{-\frac{\omega_k}{2}\xi_k^2}\right]\left[\prod_l \sqrt[4]{\frac{\omega_l}{\pi}}e^{-\frac{\omega_l}{2}\xi_l^2}\right]\prod_n d\xi_n \\
&= \frac{1}{2}\sum_{i,j}\tilde{\mathbf{T}}^{(ij)}_{\xi}\int\cdots\int\xi_i\,\xi_j\sqrt[4]{\frac{\omega_i}{\pi}}e^{-\frac{\omega_i}{2}\xi_i^2}\sqrt[4]{\frac{\omega_i}{\pi}}e^{-\frac{\omega_j}{2}\xi_j^2}\left[\prod_{k\neq i,j}\sqrt{\frac{\omega_k}{\pi}}e^{-\omega_k\xi_k^2}\right]\prod_n d\xi_n\,,
\end{aligned} \tag{A.28}$$

where the $3N$-dimensional integral can be factorized and simplified to give

$$\begin{aligned}
E_{\mathrm{dip}}[\Psi_{\mathrm{DC}}] &= \frac{1}{2}\sum_{i,j}\tilde{\mathbf{T}}^{(ij)}_{\xi}\iint\xi_i\,\xi_j\sqrt[4]{\frac{\omega_i}{\pi}}e^{-\frac{\omega_i}{2}\xi_i^2}\sqrt[4]{\frac{\omega_i}{\pi}}e^{-\frac{\omega_j}{2}\xi_j^2}\,d\xi_i d\xi_j\prod_{k\neq i,j}\int\sqrt{\frac{\omega_k}{\pi}}e^{-\omega_k\xi_k^2}\,d\xi_k \\
&= \frac{1}{2}\sum_{i,j}\tilde{\mathbf{T}}^{(ij)}_{\xi}\iint\xi_i\,\xi_j\sqrt[4]{\frac{\omega_i}{\pi}}e^{-\frac{\omega_i}{2}\xi_i^2}\sqrt[4]{\frac{\omega_i}{\pi}}e^{-\frac{\omega_j}{2}\xi_j^2}\,d\xi_i d\xi_j\,.
\end{aligned} \tag{A.29}$$

We now divide the sum into terms $i=j$ and $i\neq j$:

$$\begin{aligned}
E_{\mathrm{dip}}[\Psi_{\mathrm{DC}}] &= \frac{1}{2}\sum_i\tilde{\mathbf{T}}^{(ii)}_{\xi}\int\xi_i^2\sqrt{\frac{\omega_i}{\pi}}e^{-\omega_i\xi_i^2}\,d\xi_i + \frac{1}{2}\sum_{i\neq j}\tilde{\mathbf{T}}^{(ij)}_{\xi}\iint\xi_i\,\xi_j\sqrt[4]{\frac{\omega_i}{\pi}}e^{-\frac{\omega_i}{2}\xi_i^2}\sqrt[4]{\frac{\omega_i}{\pi}}e^{-\frac{\omega_j}{2}\xi_j^2}\,d\xi_i d\xi_j \\
&= \frac{1}{2}\sum_i\tilde{\mathbf{T}}^{(ii)}_{\xi}\int\xi_i^2\sqrt{\frac{\omega_i}{\pi}}e^{-\omega_i\xi_i^2}\,d\xi_i + \frac{1}{2}\sum_{i\neq j}\tilde{\mathbf{T}}^{(ij)}_{\xi}\underbrace{\int\xi_i\sqrt[4]{\frac{\omega_i}{\pi}}e^{-\frac{\omega_i}{2}\xi_i^2}\,d\xi_i}_{=0}\underbrace{\int\xi_j\sqrt[4]{\frac{\omega_i}{\pi}}e^{-\frac{\omega_j}{2}\xi_j^2}\,d\xi_j}_{=0} \\
&= \frac{1}{2}\sum_i\tilde{\mathbf{T}}^{(ii)}_{\xi}\sqrt{\frac{\omega_i}{\pi}}\int\xi_i^2e^{-\omega_i\xi_i^2}\,d\xi_i
\end{aligned} \tag{A.30}$$

where the last two integrals in the second line are zero by symmetry (odd functions). Solving the last integral, we get

$$E_{\mathrm{dip}}[\Psi_{\mathrm{DC}}] = \frac{1}{2}\sum_i\tilde{\mathbf{T}}^{(ii)}_{\xi}\sqrt{\frac{\omega_i}{\pi}}\frac{\sqrt{\pi}}{2\omega_i^{3/2}} = \sum_i\frac{\tilde{\mathbf{T}}^{(ii)}_{\xi}}{4\omega_i}\,. \tag{A.31}$$

So, the final expression of the expectation value of the dipolar interaction potential is simply,

$$E_{\mathrm{dip}}[\Psi_{\mathrm{DC}}] = \sum_i\frac{\tilde{\mathbf{T}}^{(ii)}_{\xi}}{4\omega_i} = \sum_i\frac{\left(\mathbf{C}^{\mathrm{T}}\tilde{\mathbf{T}}_{\zeta}\mathbf{C}\right)_{ii}}{4\omega_i} = \sum_{i,k,l}\frac{\mathbf{C}_{ki}\tilde{\mathbf{T}}^{(kl)}_{\zeta}\mathbf{C}_{li}}{4\omega_i}\,. \tag{A.32}$$

A7 Rank-Update Inversion *via* Woodbury Identity

A7.1 Submatrix Inversion as Inversion of Rank-Updated Matrix

For reformulating matrix operations on submatrices, let us first view the full matrix $\mathbf{\Omega}$ as an assembly of (3×3)-dimensional submatrices $\mathbf{\Omega}_{AB}$. For dealing with submatrices, that are obtained by "removing" the rows and columns corresponding to atoms A and B, we consider the corresponding *full* matrix, where the rows and columns (except the diagonal block) are replaced by zeros instead of removed. So, a matrix, where the (3×3) sub-block IJ is $(1-\delta_{IA}\delta_{JB})\mathbf{\Omega}_{IJ}$. This matrix can be obtained by a rank-update of the original matrix:

$$
\begin{pmatrix}
\mathbf{\Omega}_{11} & \cdots & \mathbf{\Omega}_{1(A-1)} & \mathbf{0} & \mathbf{\Omega}_{1(A+1)} & \cdots & \mathbf{\Omega}_{1(B-1)} & \mathbf{0} & \mathbf{\Omega}_{1(B+1)} & \cdots & \mathbf{\Omega}_{1N} \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\mathbf{\Omega}_{(A-1)1} & \cdots & \mathbf{\Omega}_{(A-1)(A-1)} & \mathbf{0} & \mathbf{\Omega}_{(A-1)(A+1)} & \cdots & \mathbf{\Omega}_{(A-1)(B-1)} & \mathbf{0} & \mathbf{\Omega}_{(A-1)(B+1)} & \cdots & \mathbf{\Omega}_{(A-1)N} \\
\mathbf{0} & \cdots & \mathbf{0} & \mathbf{\Omega}_{AA} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{\Omega}_{AB} & \mathbf{0} & \cdots & \mathbf{0} \\
\mathbf{\Omega}_{(A+1)1} & \cdots & \mathbf{\Omega}_{(A+1)(A-1)} & \mathbf{0} & \mathbf{\Omega}_{(A+1)(A+1)} & \cdots & \mathbf{\Omega}_{(A+1)(B-1)} & \mathbf{0} & \mathbf{\Omega}_{(A+1)(B+1)} & \cdots & \mathbf{\Omega}_{(A+1)N} \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\mathbf{\Omega}_{(B-1)1} & \cdots & \mathbf{\Omega}_{(B-1)(A-1)} & \mathbf{0} & \mathbf{\Omega}_{(B-1)(A+1)} & \cdots & \mathbf{\Omega}_{(B-1)(B-1)} & \mathbf{0} & \mathbf{\Omega}_{(B-1)(B+1)} & \cdots & \mathbf{\Omega}_{(B-1)N} \\
\mathbf{0} & \cdots & \mathbf{0} & \mathbf{\Omega}_{BA} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{\Omega}_{BB} & \mathbf{0} & \cdots & \mathbf{0} \\
\mathbf{\Omega}_{(B+1)1} & \cdots & \mathbf{\Omega}_{(B+1)(A-1)} & \mathbf{0} & \mathbf{\Omega}_{(B+1)(A+1)} & \cdots & \mathbf{\Omega}_{(B+1)(B-1)} & \mathbf{0} & \mathbf{\Omega}_{(B+1)(B+1)} & \cdots & \mathbf{\Omega}_{(B+1)N} \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\mathbf{\Omega}_{N1} & \cdots & \mathbf{\Omega}_{N(A-1)} & \mathbf{0} & \mathbf{\Omega}_{N(A+1)} & \cdots & \mathbf{\Omega}_{N(B-1)} & \mathbf{0} & \mathbf{\Omega}_{N(B+1)} & \cdots & \mathbf{\Omega}_{NN}
\end{pmatrix}
$$

$$
= \mathbf{\Omega} + \mathbf{U}_{AB}\mathbf{V}_{AB}
$$

where the matrix perturbation can be written in terms of the reduced-rank matrices $\mathbf{U}_{AB}$ and $\mathbf{V}_{AB}$:

$$
\mathbf{U}_{AB} = \begin{pmatrix}
\mathbf{0} & \cdots & \mathbf{0} & \mathbb{1}_3 & \mathbf{0} & \cdots & \mathbf{0} & \mathbb{1}_3 & \mathbf{0} & \cdots & \mathbf{0} \\
-\mathbf{\Omega}_{1A} & \cdots & -\mathbf{\Omega}_{(A-1)A} & \mathbf{0} & -\mathbf{\Omega}_{(A+1)A} & \cdots & -\mathbf{\Omega}_{(B-1)A} & \mathbf{0} & -\mathbf{\Omega}_{(B+1)A} & \cdots & -\mathbf{\Omega}_{NA} \\
-\mathbf{\Omega}_{1B} & \cdots & -\mathbf{\Omega}_{(A-1)B} & \mathbf{0} & -\mathbf{\Omega}_{(A+1)B} & \cdots & -\mathbf{\Omega}_{(B-1)B} & \mathbf{0} & -\mathbf{\Omega}_{(B+1)B} & \cdots & -\mathbf{\Omega}_{NB}
\end{pmatrix}^{\mathrm{T}}
$$

$$
\mathbf{V}_{AB} = \begin{pmatrix}
-\mathbf{\Omega}_{B1} & \cdots & -\mathbf{\Omega}_{B(A-1)} & \mathbf{0} & -\mathbf{\Omega}_{B(A+1)} & \cdots & -\mathbf{\Omega}_{B(B-1)} & \mathbf{0} & -\mathbf{\Omega}_{B(B+1)} & \cdots & -\mathbf{\Omega}_{BN} \\
-\mathbf{\Omega}_{A1} & \cdots & -\mathbf{\Omega}_{A(A-1)} & \mathbf{0} & -\mathbf{\Omega}_{A(A+1)} & \cdots & -\mathbf{\Omega}_{A(B-1)} & \mathbf{0} & -\mathbf{\Omega}_{A(B+1)} & \cdots & -\mathbf{\Omega}_{AN} \\
\mathbf{0} & \cdots & \mathbf{0} & \mathbb{1}_3 & \mathbf{0} & \cdots & \mathbf{0} & \mathbb{1}_3 & \mathbf{0} & \cdots & \mathbf{0}
\end{pmatrix}
$$

The subscript to $\mathbf{U}$ and $\mathbf{V}$ thereby dictates the position of the two (3×3) identity matrices (and the $\mathbf{0}$-entries, correspondingly). With the help of permutation matrices, $\mathbf{M}_A$ and $\mathbf{M}_B$, we can re-order columns and rows such that the zeros are in the first rows and columns, which gives

$$
\tilde{\mathbf{\Omega}}_{AB} = \mathbf{M}_B^{\mathrm{T}}\mathbf{M}_A^{\mathrm{T}}\left(\mathbf{\Omega} + \mathbf{U}_{AB}\mathbf{V}_{AB}\right)\mathbf{M}_A\mathbf{M}_B = \begin{pmatrix}
\mathbf{\Omega}_{AA} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\
\mathbf{0} & \mathbf{\Omega}_{BB} & \mathbf{0} & \cdots & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & & & \\
\vdots & \vdots & & \mathbf{\Omega}''_{AB} & \\
\mathbf{0} & \mathbf{0} & & &
\end{pmatrix}. \tag{A.33}
$$

Using $(\mathbf{A_1}\mathbf{A_2}\mathbf{A_3})^{-1} = \mathbf{A}_3^{-1}\mathbf{A}_2^{-1}\mathbf{A}_1^{-1}$, the property of (single) permutation matrices $\mathbf{M}^{\mathrm{T}} = \mathbf{M} = \mathbf{M}^{-1}$ and the rules for inverting block-diagonal matrices, we can write the inverse of this matrix as

$$
\tilde{\mathbf{\Omega}}_{AB}^{-1} = \mathbf{M}_B^{\mathrm{T}}\mathbf{M}_A^{\mathrm{T}}\left(\mathbf{\Omega} + \mathbf{U}_{AB}\mathbf{V}_{AB}\right)^{-1}\mathbf{M}_A\mathbf{M}_B = \begin{pmatrix}
\mathbf{\Omega}_{AA}^{-1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\
\mathbf{0} & \mathbf{\Omega}_{BB}^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & & & \\
\vdots & \vdots & & \left(\mathbf{\Omega}''_{AB}\right)^{-1} & \\
\mathbf{0} & \mathbf{0} & & &
\end{pmatrix}. \tag{A.34}
$$

Thus, we can obtain the inverse of $\mathbf{\Omega}''_{AB}$ by removing the first six rows and columns of the inverse of $\tilde{\mathbf{\Omega}}_{AB}$. Alternatively, we can restore the previous order by applying the opposite re-shuffling of rows and columns and then remove the A^{th} and B^{th} rows and columns. Restoring the previous ordering of rows and columns is achieved by applying the inverse of $\mathbf{M}_A$ and $\mathbf{M}_B$ to the obtained result. With Eq. (A.34) and (single) permutation matrices being unitary, this gives

$$\mathbf{M}_A\mathbf{M}_B\tilde{\mathbf{\Omega}}^{-1}_{AB}\mathbf{M}_B^{\mathrm{T}}\mathbf{M}_A^{\mathrm{T}} = \left(\mathbf{\Omega}+\mathbf{U}_{AB}\mathbf{V}_{AB}\right)^{-1} \tag{A.35}$$

So, the inverse $\mathbf{\Omega}''^{-1}_{AB}$ is given by removing the A^{th} and B^{th} rows and columns from $\left(\mathbf{\Omega}+\mathbf{U}_{AB}\mathbf{V}_{AB}\right)^{-1}$. With this, we can obtain the inverse of $\mathbf{\Omega}''_{AB}$ from the inverse of the corresponding rank-updated matrix $\mathbf{\Omega}+\mathbf{U}_{AB}\mathbf{V}_{AB}$.

A7.2 Inversion of Rank-Updated Matrix as Rank-Update of Inverse

For finding the inverse of a rank-updated matrix, we can make use of the *Woodbury Identity*. In its most general form,

$$\left(\mathbf{A}+\mathbf{DGF}\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{D}\left(\mathbf{G}^{-1}+\mathbf{FA}^{-1}\mathbf{D}\right)^{-1}\mathbf{FA}^{-1}\ . \tag{A.36}$$

So, instead of finding the inverse of the rank-updated matrix, we can simply rank-update the inverse of the original matrix. In the present case, $\mathbf{A}=\mathbf{\Omega}$, $\mathbf{D}=\mathbf{U}_{AB}$, $\mathbf{G}=\mathbb{1}_9$, and $\mathbf{F}=\mathbf{V}_{AB}$. Hence,

$$\left(\mathbf{\Omega}+\mathbf{U}_{AB}\mathbf{V}_{AB}\right)^{-1} = \mathbf{\Omega}^{-1} - \mathbf{\Omega}^{-1}\mathbf{U}_{AB}\left(\mathbb{1}_9+\mathbf{V}_{AB}\mathbf{\Omega}^{-1}\mathbf{U}_{AB}\right)^{-1}\mathbf{V}_{AB}\mathbf{\Omega}^{-1}\ . \tag{A.37}$$

The great advantage of this reformulation comes from the fact that $\mathbf{V}_{AB}\mathbf{\Omega}^{-1}\mathbf{U}_{AB}$ gives a (9×9)-dimensional matrix. Thus, we only have to find the inverse of the 9-by-9 matrix $(\mathbb{1}_9+\mathbf{V}_{AB}\mathbf{\Omega}^{-1}\mathbf{U}_{AB})$ and perform basic matrix operations including only addition and multiplication, while the *full* matrix $\mathbf{\Omega}$ needs to be inverted only once at the beginning. In summary, instead of performing N^2 inversions of a $(3N–6\times 3N–6)$-dimensional matrix, we perform N^2 inversions of an only (9×9)-dimensional matrix and a few matrix-matrix multiplications to obtain

$$\mathbf{\Omega}''^{-1}_{AB} = \left[\mathbf{\Omega}^{-1} - \mathbf{\Omega}^{-1}\mathbf{U}_{AB}\left(\mathbb{1}_9+\mathbf{V}_{AB}\mathbf{\Omega}^{-1}\mathbf{U}_{AB}\right)^{-1}\mathbf{V}_{AB}\mathbf{\Omega}^{-1}\right]_{ij}\ \forall\,(i,j)\in(A,B)\ , \tag{A.38}$$

where $[\mathbf{A}]_{ij}$ denotes the matrix obtained after removing the i^{th} column and j^{th} row from matrix $\mathbf{A}$. For large N, the computational cost of inverting $(3N\times 3N)$- or $(3N–6\times 3N–6)$-dimensional matrices is nearly the same. So, trading N^2 inversions of $(3N–6\times 3N–6)$-dimensional matrices for a single inversion of a $(3N\times 3N)$-dimensional matrix offers a substantial reduction of computational costs.

A8 Rank-Update Determinant *via* Matrix Determinant Lemma

Considering that

$$\det\left\{\begin{pmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{B}\end{pmatrix}\right\} = \det\{\mathbf{A}\}\cdot\det\{\mathbf{B}\}\ , \tag{A.39}$$

we first write the matrix

$$\tilde{\mathbf{\Omega}}_{AB} = \mathbf{M}_B^{\mathrm{T}}\mathbf{M}_A^{\mathrm{T}}\left(\mathbf{\Omega}+\mathbf{U}_{AB}\mathbf{V}_{AB}\right)\mathbf{M}_A\mathbf{M}_B = \begin{pmatrix}\mathbf{\Omega}_{AA} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{\Omega}_{BB} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & & & \\ \vdots & \vdots & & \mathbf{\Omega}''_{AB} & \\ \mathbf{0} & \mathbf{0} & & & \end{pmatrix} \tag{A.40}$$

as introduced in Appendix A7. According to $\det\{\mathbf{AB}\} = \det\{\mathbf{BA}\}$, the unitarity of (single) permutation matrices and Eq. (A.39), the determinant of this matrix is given by

$$\det\{\tilde{\mathbf{\Omega}}_{AB}\} = \det\{\mathbf{\Omega}+\mathbf{U}_{AB}\mathbf{V}_{AB}\} = \det\{\mathbf{\Omega}_{AA}\}\cdot\det\{\mathbf{\Omega}_{BB}\}\cdot\det\{\mathbf{\Omega}''_{AB}\}\ . \tag{A.41}$$

Using the *Matrix Determinant Lemma*, in its most general form,

$$\det\{\mathbf{A}+\mathbf{DGF}\} = \det\{\mathbf{G}^{-1}+\mathbf{FA}^{-1}\mathbf{D}\}\cdot\det\{\mathbf{G}\}\cdot\det\{\mathbf{A}\}\ , \tag{A.42}$$

we can also write the determinant in Eq. (A.41) as

$$\begin{aligned}\det\{\mathbf{\Omega}+\mathbf{U}_{AB}\mathbf{V}_{AB}\} &= \det\{\mathbb{1}_9+\mathbf{V}_{AB}\mathbf{\Omega}^{-1}\mathbf{U}_{AB}\}\cdot\det\{\mathbb{1}_9\}\cdot\det\{\mathbf{\Omega}\}\\ &= \det\{\mathbb{1}_9+\mathbf{V}_{AB}\mathbf{\Omega}^{-1}\mathbf{U}_{AB}\}\cdot\det\{\mathbf{\Omega}\} = \det\{\mathbf{\Omega}_{AA}\}\cdot\det\{\mathbf{\Omega}_{BB}\}\cdot\det\{\mathbf{\Omega}''_{AB}\}\ .\end{aligned} \tag{A.43}$$

Hence,

$$\det\{\mathbf{\Omega}''_{AB}\} = \det\{\mathbb{1}_9+\mathbf{V}_{AB}\mathbf{\Omega}^{-1}\mathbf{U}_{AB}\}\cdot\det\{\mathbf{\Omega}\}\cdot\det\{\mathbf{\Omega}_{AA}\}^{-1}\cdot\det\{\mathbf{\Omega}_{BB}\}^{-1}\ . \tag{A.44}$$

Even better, the expression for the Coulomb interaction energy of a set of quantum harmonic oscillators as given in Eq. (5.12) contains the prefactor $\det\{\mathbf{\Omega}\}/\det\{\mathbf{\Omega}''_{AB}\}$, which becomes

$$\frac{\det\{\mathbf{\Omega}\}}{\det\{\mathbf{\Omega}''_{AB}\}} \overset{(A.44)}{=} \frac{\det\{\mathbf{\Omega}_{AA}\}\cdot\det\{\mathbf{\Omega}_{BB}\}}{\det\{\mathbb{1}_9+\mathbf{V}_{AB}\mathbf{\Omega}^{-1}\mathbf{U}_{AB}\}}\ . \tag{A.45}$$

So, we only need to evaluate low-dimensional determinants (3×3 and 9×9) plus lower-cost matrix operations involving the inverse of the global matrix $\mathbf{\Omega}$, which we need to obtain only once in the beginning.

A9 Qualitative Descriptors of Dipole-Correlated Coulomb Singles

The Dipole-Correlated Coulomb Singles (DCS) energy can be separated into charge polarization effects arising from long-range electron correlation and beyond-dipolar vdW interactions. Correspondingly, the displaced charge due to vdW dispersion as obtained within the MBD formalism resembles the relative trend in DCS interaction energies. Fig. A.6A shows a correlation plot of the displaced charge with the DCS contribution to the binding energy of a fullerene to various host molecules.

To further analyze geometric descriptors for spatial proximity, we tested a variety of power law summations of the atom-pairwise distances. As can be seen from Fig. A.6B, the sum of inverse distances to the power eight also provides a qualitative descriptor for E_{DCS}. Within the atom-pairwise formulation of vdW interactions, such a sum also represents dipole–quadrupole vdW contributions. This connection is merely coincidental, however. DCS reduce the binding energies in all cases, whereas extending MBD with atom-pairwise dipole–quadrupole vdW interactions ($-C_8/R^8$-potentials), would add additional, *attractive* interaction.

The best correlation (*i.e.*, the least spread from linear behavior) between E_{DCS} and power law summations can be found for the fifth power of inverse atom-pairwise distances (*cf.* Fig. A.6B). This can be rationalized by the fact that, as a leading-order term, dipolar quantum fluctuations induce static quadrupole moments, whose interaction decays as R^{-5}. Further investigation of the general validity and origin of the observed correlations is subject to future studies.

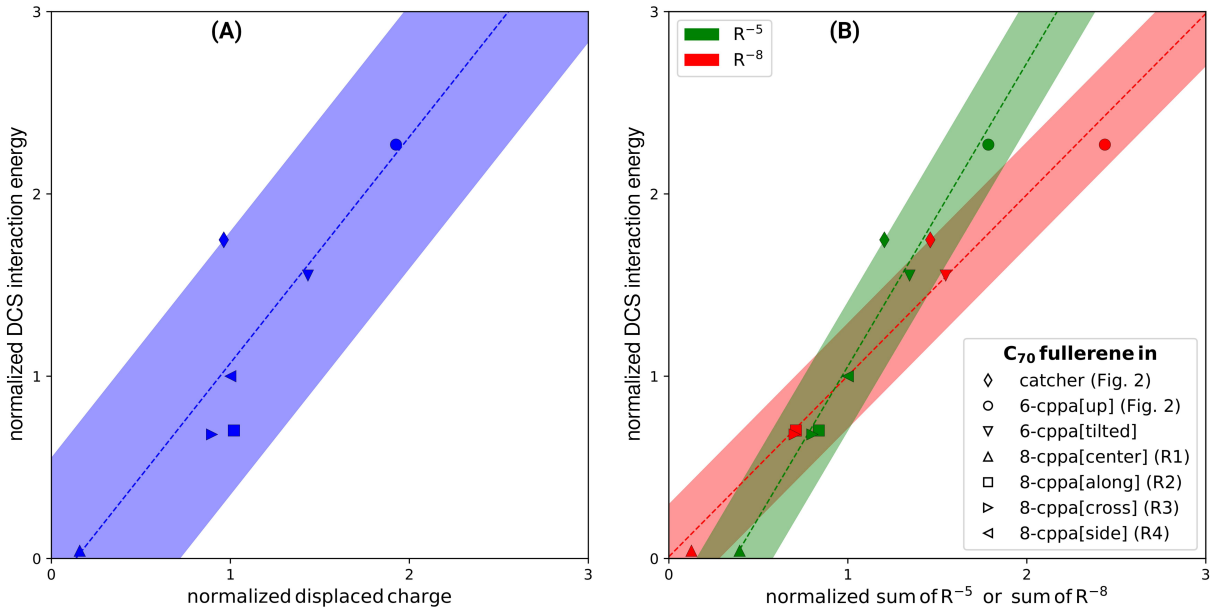

Fig. A.6: Correlation of displaced charge (A) and the sum of R^{-5}- or R^{-8}-terms (B) with *Dipole-Correlated Coulomb Singles* interaction energies for the C_{70}-fullerene in various host molecules.

A10 Summary of Machine-Learned Repulsive Potentials for DFTB

The SchNetPack toolbox allows to train deep tensor neural networks (DTNNs) based on energies as well as on forces. We performed our tests of different sampling algorithms for the training set selection for DTNN-models trained on repulsive energies. The corresponding learning curves for random sampling, our eq+X approach and farthest point sampling (FPS, non-monotonic behavior due to stochastic nature of the algorithm) are shown in Fig. A.7. Our NN_{rep} model can thereby be combined with arbitrary electronic parameterizations and different levels of DFTB. This is exemplified for DFTB2 with *mio* parameterization in Fig. A.7.

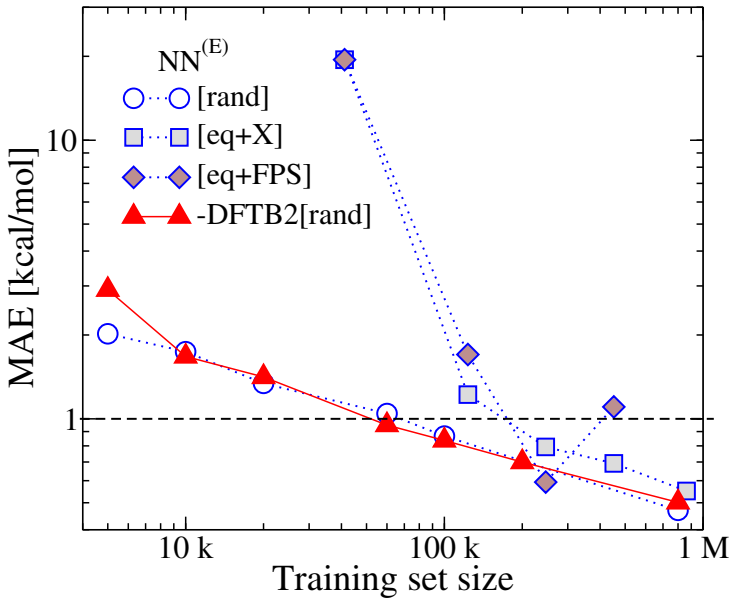

Fig. A.7: Learning curves of NN_{rep} for DFTB3 using different algorithms of training set selection and for DFTB2. Blue circles: Random selection of training points, blue squares: eq+X selection, blue diamonds: eq+FPS selection, and red triangles: NN_{rep} model of repulsive energy for DFTB2. Reproduced with permission from Ref. 4. ©2020 American Chemical Society.

The repulsive energy can be also learned by using kernel ridge regression (KRR) as implemented in the QML package, [364,365] for instance. Fig. A.8 shows the learning curves (a) and error distribution (b) for several molecular descriptors. As expected due to the nature of the property target, local descriptors (labeled with the prefix "at.") provide a better performance than global ones.

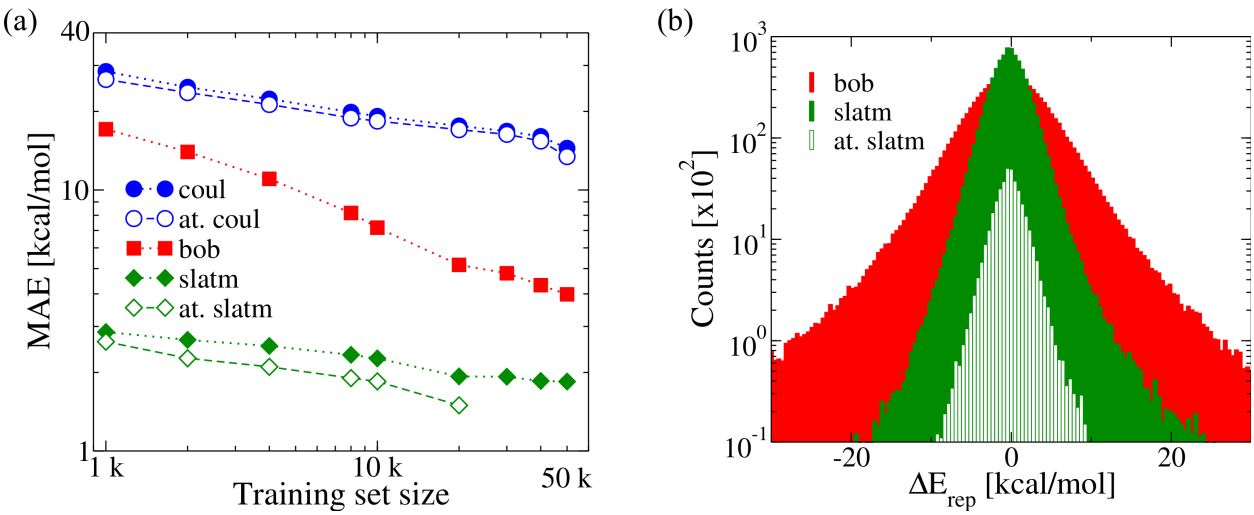

Fig. A.8: Learning curves for Kernel Ridge Regression of the repulsive energy using different global and local atomic representations. (a) Learning curves and (b) distribution of errors. Reproduced with permission from Ref. 4. ©2020 American Chemical Society.

A11 Dimensionality & Complexity of Global vs Local Chemical Space

The complexity of the global chemical space spanned by QM7-X can be estimated by the number of distinct configuration times the conformational degrees of freedom (DOF). We can further approximate the required training set size or complexity of the set, $\mathcal{C}$, by multiplying the resulting number of DOF by an estimate for the number of samples per DOF. Given ~42k unique equilibrium structures, an average of ~16 atoms per molecules and approximately 5 samples per DOF, we get

$$N_{\mathrm{DOF}}^{(\mathrm{glob})} = N_{\mathrm{mols}} \cdot (3\,\bar{N}_{\mathrm{atoms}} - 6) \approx 2 \cdot 10^6$$

$$\text{and} \quad \mathcal{C}_{\mathrm{glob}} \lesssim 9 \cdot 10^6$$

We note that thanks to the extensive sampling of configurational space by choosing all equilibrium structures, the $(3\bar{N}_{\mathrm{atoms}} - 6)$ conformational DOF can be taken in eigenmode representation, which justifies the rather low choice of 5 samples per DOF. Considering that many DOF share common characteristics, one may see the above as an upper bound.

One can alternatively consider the average number of non-hydrogen elements per molecule being 6.9. Limiting the minimum content of carbon atoms to 50 %. This leaves an average of 3.45 carbon atoms per molecule, while 3.45 atoms can be chosen from {C, N, O}, which gives a compositional diversity of $3^{3.45}$. In order to account for configurational and conformational diversity, we may now place each composition in an appropriately-sized volume. From the conformations of benzene as contained in QM7-X, we can estimate this to be $V \approx 6 \times 6 \times 3$ Å. All possible configurations and conformations can then be represented by covering the $(3\bar{N}_{\mathrm{atoms}} - 6)$ displacements of each atom of the molecule (including previously neglected hydrogens) within our estimated "molecular volume". In this case, the $(3\bar{N}_{\mathrm{atoms}} - 6)$ displacements have to cover both configurational and conformational space and thus sample the full "molecular volume". A reasonable discretization step for this sampling, that accounts for the fact that we cannot place atoms at random, is $\Delta V \approx 0.5 \times 0.5 \times 0.5$ Å. Together with $\bar{N}_{\mathrm{atoms}} = 16$, we arrive at

$$\mathcal{C}_{\mathrm{glob}} \gtrsim 3^{3.45} \cdot (3\bar{N}_{\mathrm{atoms}} - 6) \cdot V / \Delta V = 10^6$$

This measure does not account for the compositional diversity of different hydrogen contents at the same non-hydrogen composition (*i.e.*, ethane vs ethylene vs acetylene, for example). The above number thus represents a lower bound to the actual complexity and number of samples needed to fully describe QM7-X. Combining both approaches, we get

$$10^6 \lesssim \mathcal{C}_{\mathrm{glob}} \lesssim 9 \cdot 10^6$$

and can assume

$$\boxed{N_{\mathrm{DOF}}^{(\mathrm{glob})} \approx 2 \cdot 10^6 \quad \text{and} \quad \mathcal{C}_{\mathrm{glob}} \approx 5 \cdot 10^6}$$

We can thus expect a purely data-driven, global model to require about five million training instances in order to be applicable to the whole QM7-X dataset.

For the space of local (chemical) environments, we may expand the structure to property mapping into local, force-field-like descriptors (*i.e.*, bonds, angles, dihedrals). The total complexity of the local chemical space can then be defined as the number of possible two-, three-, and four-body terms times an estimate for how many distinct samples one needs in order to describe the corresponding term. The individual n-body terms, we get for the set {H,C,N,O} are summarized in Tab. A.2.

Tab. A.2: Summary of two-, three-, and four-body terms defining the space of local chemical environments for molecules containing {H,C,N,O}.

n				**Terms**
2	**Pairs**	$A = B$	$N_{AA} = 4$	
	A–B	$A \neq B$	$N_{AB} = \binom{4}{2} = 6$	
		Exclude H–H (won't be in larger molecules)	$N_b = N_{AA} + N_{AB} - 1$	**9**
3	**Angles**	$B \in$ {C,N,O}, giving 3 options to insert an element into each possible pair AB.	$N_{ABC} = 3 \cdot (N_{AA} + N_{AB})$	
	A–B–C	We neglect H–O–H.	$N_a = N_{ABC} - 1$	**29**
4	**Dihedrals**	Let's first take the case $A = D \wedge B \neq C$ and remember A–B–C–$A = A$–C–B–A.	$N_{ABCA} = N_{AA} \cdot 3 = 12$	
	A–B–C–D	Next, $A = D$ and $B = C$.	$N_{ABBA} = N_{AA} \cdot 3 = 12$	
	$B, C \neq$ H	For $A \neq D$, we have $B, C \in$ {C,N,O}.	$N_{ABCD} = N_{AB} \cdot 3 \cdot 3 = 54$	
		We neglect H–O–O–H and O–O–O–O, which won't be part of larger molecules.	$N_d = N_{ABCA} + N_{ABBA} + N_{ABCD} - 2$	**76**

As a result we obtain a total of 114 DOF. In terms of the number of required samples, we can estimate for angles (a) and dihedrals (d) $\Delta a = \Delta d \approx 5^\circ \Rightarrow n_s(a) = n_s(d) \approx 36$. Assuming the space of bonds (b) to be within a range of 0.5 to 4 Å, we can say $\Delta b \approx 0.1$ Å $\Rightarrow n_s(b) \approx 36$. So, the number of DOF and the complexity of or the number of samples needed to sample the local chemical space are

$$N_{\mathrm{DOF}}^{(\mathrm{loc})} = N_b + N_a + N_d = 114 \qquad \text{and} \qquad \mathcal{C}_{\mathrm{loc}} = N_b \cdot n_s(b) + N_a \cdot n_s(a) + N_d \cdot n_s(d) \approx 4 \cdot 10^3$$

The required training set for a global model scales with the number of distinct molecules and ranges at $\sim 10^6$ in the present case of the QM7-X dataset. The necessary number of structures to sample all possible local environments as detailed above is thus **3 orders of magnitude less than for the global chemical space**.

Now, this is for arbitrary conformations of points in space. In nature, the individual terms are strongly limited, however. Angles and dihedrals preferably take a value from a very limited set of the total 180° span assumed above plus a considerable number of combinations is prohibited due to the corresponding three-dimensional arrangements potentially resulting in a fusing of atoms. For the sake of simplicity and in order to stick to a complete description of local environments including potential rare transition states, we do not include these aspects in the current analysis.

Note that, contrary to the global complexity, the dimensionality and complexity of the local chemical space is *independent* of the total size of the molecule and thus independent of the configurational/conformational dimensionality. For local environments considering four elements, the dimensionality is fixed at 114 and the required number of samples fixed at the above approximate value. When extending beyond four elements, the complexity grows solely due to combinatorics. In the case of the global complexity, one additionally has to consider the rapidly exploding configurational and conformational part.

A12 Analytical Repulsive Potentials in Two-Center Approximation

In order to evaluate the individual terms for the pair-integrals in Eq. (6.15), we will rely on the following reformulation of the Coulomb potential between points $\mathbf{r}$ and $\mathbf{r}'$:

$$\frac{1}{\|\mathbf{r}-\mathbf{r}'\|} = \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-s^2 \|\mathbf{r}-\mathbf{r}'\|^2} \, ds \,. \tag{A.46}$$

Furthermore, we will employ several Gaussian integrals, in a general form,

$$\int_{\mathbb{R}} e^{-ax^2+bx-c} \, dx = \sqrt{\frac{\pi}{a}} \, e^{\frac{b^2}{4a}-c} \,. \tag{A.47}$$

Hartree Energy Contributions

When expressing the electron density in the form of simple Gaussians as proposed in Eq. (6.16) (or linear combinations thereof), the contributions from the Hartree energy terms in Eq. (6.15) can be written in the form of a general double integral (or linear combinations thereof). Setting $\mathbf{R}_A = \mathbf{0}$, $\mathbf{R} = (R_x, R_y, R_z)^{\mathrm{T}} \equiv \mathbf{R}_B - \mathbf{R}_A = \mathbf{R}_B$, we can write the general form

$$\Omega_{AB}^{(\mathrm{H})} \equiv \iint \frac{\rho_A \rho'_B}{\|\mathbf{r}-\mathbf{r}'\|} \, d\mathbf{r} d\mathbf{r}' = n_A n_B \iint \frac{e^{-\frac{\|\mathbf{r}\|^2}{\sigma_A}} \, e^{-\frac{\|\mathbf{r}'-\mathbf{R}\|^2}{\sigma_B}}}{\|\mathbf{r}-\mathbf{r}'\|} \, d\mathbf{r} d\mathbf{r}' \,. \tag{A.48}$$

With identity (A.46), we can expand the norm in the exponentials into its x, y, z components, such that

$$\Omega_{AB}^{(\mathrm{H})} = \frac{2\, n_A n_B}{\sqrt{\pi}} \int_0^\infty \int_{\mathbb{R}} I_x \, e^{-\frac{(r'_x-R_x)^2}{\sigma_B}} \, dr'_x \int_{\mathbb{R}} I_y \, e^{-\frac{(r'_y-R_y)^2}{\sigma_B}} \, dr'_y \int_{\mathbb{R}} I_z \, e^{-\frac{(r'_z-R_z)^2}{\sigma_B}} \, dr'_z \, ds \,, \tag{A.49}$$

where, with the help of Eq. (A.47), we can obtain

$$I_k = \int_{\mathbb{R}} e^{-s^2\left(r_k-r'_k\right)^2 - \frac{r_k^2}{\sigma_A}} \, dr_k = \int_{\mathbb{R}} e^{-\left(s^2+\sigma_A^{-1}\right) r_k^2 \, + \, 2\, s^2 \, r_k \, r'_k \, - \, s^2 r'^2_k} \, dr_k \;=\; \sqrt{\frac{\pi}{s^2+\sigma_A^{-1}}} \, e^{-\frac{s^2\sigma_A^{-1} \, r'^2_k}{s^2+\sigma_A^{-1}}} \,. \tag{A.50}$$

Thus,

$$\Omega_{AB}^{(\mathrm{H})} = \frac{2\, n_A n_B}{\sqrt{\pi}} \int_0^\infty \left(\frac{\pi}{s^2+\sigma_A^{-1}} \right)^{3/2} \cdot \Gamma_x \cdot \Gamma_y \cdot \Gamma_z \, ds \tag{A.51}$$

$$\text{where} \qquad \Gamma_k = \int_{\mathbb{R}} e^{-\frac{s^2\sigma_A^{-1} \, r'^2_k}{s^2+\sigma_A^{-1}} - \frac{\left(r'_k-R_k\right)^2}{\sigma_B}} \, dr'_k \overset{(\mathrm{A.47})}{=} \sqrt{\frac{\pi}{\frac{s^2\sigma_A^{-1}}{s^2+\sigma_A^{-1}} + \sigma_B^{-1}}} \exp\left(\frac{\frac{s^2\sigma_A^{-1}}{s^2+\sigma_A^{-1}} \sigma_B^{-1} \, R_k^2}{\frac{s^2\sigma_A^{-1}}{s^2+\sigma_A^{-1}} + \sigma_B^{-1}} \right) . \tag{A.52}$$

So,

$$\Omega_{AB}^{(\mathrm{H})} = \frac{2\, n_A n_B}{\sqrt{\pi}} \int_0^\infty \left(\frac{\pi}{s^2+\sigma_A^{-1}} \right)^{3/2} \left(\frac{\pi}{\frac{s^2\sigma_A^{-1}}{s^2+\sigma_A^{-1}} + \sigma_B^{-1}} \right)^{3/2} \exp\left(\frac{\frac{s^2\sigma_A^{-1}}{s^2+\sigma_A^{-1}} \sigma_B^{-1} \, \|\mathbf{R}\|^2}{\frac{s^2\sigma_A^{-1}}{s^2+\sigma_A^{-1}} + \sigma_B^{-1}} \right) ds \tag{A.53}$$

$$= \frac{2\, n_A n_B}{\sqrt{\pi}} \, \pi^3 \int_0^\infty \left(\frac{\sigma_A \sigma_B}{(\sigma_A+\sigma_B)\, s^2 + 1} \right)^{3/2} e^{-\frac{s^2}{(\sigma_A+\sigma_B) s^2+1} \cdot \|\mathbf{R}\|^2} \, ds \,. \tag{A.54}$$

Substituting $t^2 = \|\mathbf{R}\|^2\, s^2 \cdot \left[(\sigma_A+\sigma_B)\, s^2+1\right]^{-1}$, which includes

$$ds = \frac{\left[(\sigma_A+\sigma_B)\, s^2+1\right]^{3/2}}{\|\mathbf{R}\|}\, dt \qquad \text{and} \qquad \lim_{s\to\infty} t = \frac{\|\mathbf{R}\|}{\sqrt{\sigma_A+\sigma_B}}\,, \tag{A.55}$$

we get

$$\Omega_{AB}^{(\mathrm{H})} = n_A n_B \frac{\pi^3\,(\sigma_A\sigma_B)^{3/2}}{\|\mathbf{R}\|} \frac{2}{\sqrt{\pi}} \int_0^{\frac{\|\mathbf{R}\|}{\sqrt{\sigma_A+\sigma_B}}} e^{-t^2}\, dt \;=\; n_A n_B \frac{\pi^3\,(\sigma_A\sigma_B)^{3/2}}{\|\mathbf{R}\|}\, \mathrm{erf}\!\left(\frac{\|\mathbf{R}\|}{\sqrt{\sigma_A+\sigma_B}}\right). \tag{A.56}$$

When using an "outer-shell" or p-orbital-like ansatz for the valence–valence Hartree contribution as given in Eq. (6.16), the linear combination of $\Omega_{AB}^{(\mathrm{H})}$ terms can involve positive as well as negative contributions.

External Potential Contributions

Using the same steps as for the Hartree potential energy contributions above, we can also derive the interaction of nucleus A and the (core) electrons on atom B. In this case, we set $\mathbf{R}_B = \mathbf{0}$ and $\mathbf{R}_A = \mathbf{R}_A - \mathbf{R}_B = \mathbf{R}$. With identity (A.46), we can write the general term

$$\Omega_{AB}^{(\mathrm{ext})} = \int \frac{Z_A \varrho_B}{\|\mathbf{r}-\mathbf{R}_A\|}\, d\mathbf{r} = \frac{2\,Z_A\, n_B}{\sqrt{\pi}} \int_0^\infty \int_{\mathbb{R}^3} e^{-s^2\|\mathbf{r}-\mathbf{R}\|^2}\, e^{-\frac{1}{\varsigma_B}\|\mathbf{r}\|^2}\, d\mathbf{r}ds \tag{A.57}$$

$$= \frac{2\,Z_A\, n_B}{\sqrt{\pi}} \int_0^\infty \prod_k^{\{x,y,z\}} \int_{\mathbb{R}} e^{-s^2(r_k-R_k)^2 \,-\, \frac{1}{\varsigma_B} r_k^2}\, dr_k ds \tag{A.58}$$

$$= \frac{2\,Z_A\, n_B}{\sqrt{\pi}} \int_0^\infty \prod_k^{\{x,y,z\}} \sqrt{\frac{\varsigma_B\,\pi}{\varsigma_B\, s^2+1}}\, e^{-\frac{s^2}{\varsigma_B\, s^2+1} R_k^2}\, ds \tag{A.59}$$

$$= \frac{2\,Z_A\, n_B}{\sqrt{\pi}} \int_0^\infty \left(\frac{\varsigma_B\,\pi}{\varsigma_B\, s^2+1}\right)^{3/2} e^{-\frac{s^2}{\varsigma_B\, s^2+1}\|\mathbf{R}\|^2}\,. \tag{A.60}$$

Next, we substitute

$$t^2 = \frac{s^2\,\|\mathbf{R}\|^2}{\varsigma_B\, s^2+1} \qquad \text{with} \qquad ds = \frac{\left(\varsigma_B\, s^2+1\right)^{3/2}}{\|\mathbf{R}\|}\, dt \qquad \text{and} \qquad \lim_{s\to\infty} t = \frac{\|\mathbf{R}\|}{\sqrt{\varsigma_B}} \tag{A.61}$$

to arrive at the final expression

$$\Omega_{AB}^{(\mathrm{ext})} = \frac{Z_A\, n_B\,(\alpha\,\pi)^{3/2}}{\|\mathbf{R}\|} \frac{2}{\sqrt{\pi}} \int_0^{\|\mathbf{R}\|/\sqrt{\varsigma_B}} e^{-t^2}\, dt \;=\; \frac{Z_A\, n_B\,(\varsigma_B\,\pi)^{3/2}}{\|\mathbf{R}\|}\, \mathrm{erf}\!\left(\frac{\|\mathbf{R}\|}{\sqrt{\varsigma_B}}\right). \tag{A.62}$$

Total Two-Center-Approximated Repulsive Potential

Together with the classical nuclear repulsion energy and in two-center approximation, the Gaussian expansion of electron densities leads to a repulsive potential between atoms A and B, which can in general form be written as

$$\mathcal{V}_{\mathrm{tcr}}^{(AB)}\left(R_{AB}\right) = \frac{Z_A\, Z_B}{R_{AB}} + \sum_{i,j} \Omega_{A_iB_j}^{(\mathrm{H})} + \Omega_{A_jB_j}^{(\mathrm{ext})} = \frac{Z_A\, Z_B}{R_{AB}} + \sum_{i,j} p_{A_i}\, p_{B_j}\, \mathrm{erf}\!\left(\frac{R_{AB}}{\sqrt{q_{A_i}+q_{B_j}}}\right) R_{AB}^{-1}\,, \tag{A.63}$$

where the prefactors incorporate the sign of the individual contributions entering Eq. (6.15).

Publication Record

Publications during the course of the doctorate:

- M. Stöhr, M. Sadhukhan, Y. S. Al-Hamdani, J. Hermann, and A. Tkatchenko "Coulomb Interactions between Dipolar Quantum Fluctuations in van der Waals Bound Molecules and Materials" *Nat. Commun.* accepted (2020). Preprint available at https://arxiv.org/abs/2007.12505.
- M. Stöhr, L. Medrano Sandonas, and A. Tkatchenko "Accurate Many-Body Repulsive Potentials for Density-Functional Tight Binding from Deep Tensor Neural Networks" *J. Phys. Chem. Lett.* **11**, 6835 (2020).
- B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshaye, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. van der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W.-z. Yu, and T. Frauenheim "DFTB+, a software package for efficient approximate density functional theory based atomistic simulations" *J. Chem. Phys.* **152**, 124101 (2020).
- M. Stöhr and A. Tkatchenko "Quantum mechanics of proteins in explicit water: The role of plasmon-like solute–solvent interactions" *Sci. Adv.* **5**, eaax0024 (2019).
- M. Stöhr, T. Van Voorhis, and A. Tkatchenko "Theory and practice of modeling van der Waals interactions in electronic-structure calculations" *Chem. Soc. Rev.* **48**, 4118 (2019).
- D. V. Fedorov, M. Sadhukhan, M. Stöhr, and A. Tkatchenko "Quantum-Mechanical Relation between Atomic Dipole Polarizability and the van der Waals Radius" *Phys. Rev. Lett.* **121**, 183401 (2018).

Previous publications:

- M. Stöhr, G. S. Michelitsch, J. C. Tully, K. Reuter, and R. J. Maurer "Communication: Charge-population based dispersion interactions for molecules and materials" *J. Chem. Phys.* **144**, 151101 (2016).
- J. A. Lloyd, A. C. Papageorgiou, S. Fischer, S. C. Oh, Ö. Sağlam, K. Diller, D. A. Duncan, F. Allegretti, F. Klappenberger, M. Stöhr, R. J. Maurer, K. Reuter, J. Reichert, and J. V. Barth "Dynamics of Spatially Confined Bisphenol A Trimers in a Unimolecular Network on Ag(111)" *Nano Lett.* **16**, 1884 (2016).

Software (co-)development during the doctorate:

- DFTB+ (https://github.com/dftbplus/dftbplus)
  Software package for density-functional tight-binding calculations and extended methods.
  Contribution: MBD and vdW(TS) dispersion models.
- libMBD (https://github.com/jhrmnn/libmbd)
  Library for many-body dispersion calculations and beyond with various interfaces.
  Contribution: Original ScaLAPACK implementation of MBD formalism.
- Hotbit (https://github.com/pekkosk/hotbit)
  Open-source ASE density-functional tight-binding calculator and companion for DFT.
  Contribution: GGA exchange-correlation functionals for atomic DFT calculations.

Bibliography

[1] M. Stöhr, T. Van Voorhis, and A. Tkatchenko, *Chem. Soc. Rev.* **48**, 4118 (2019).

[2] M. Stöhr and A. Tkatchenko, *Sci. Adv.* **5**, eaax0024 (2019).

[3] M. Stöhr, M. Sadhukhan, Y. S. Al-Hamdani, J. Hermann, and A. Tkatchenko, *Nat. Commun.* , accepted (2020).

[4] M. Stöhr, L. Medrano Sandonas, and A. Tkatchenko, *J. Phys. Chem. Lett.* **11**, 6835 (2020).

[5] P. A. M. Dirac, *Proc. R. Soc. Lond. A* **123**, 714 (1929).

[6] M. Levitt and A. Warshel, *Nature* **253**, 694 (1975).

[7] J. A. McCammon, B. R. Gelin, and M. Karplus, *Nature* **267**, 585 (1977).

[8] S. Piana, J. L. Klepeis, and D. E. Shaw, *Curr. Opin. Struct. Biol.* **24**, 98 (2014).

[9] C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele, *Nature* **420**, 102 (2002).

[10] M. B. Prigozhin and M. Gruebele, *Phys. Chem. Chem. Phys.* **15**, 3372 (2013).

[11] P. van der Ploeg and H. J. C. Berendsen, *J. Chem. Phys.* **76**, 3271 (1982).

[12] S. J. Marrink *et al.*, *Chem. Rev.* **119**, 6184 (2019).

[13] D. L. Ensign, P. M. Kasson, and V. S. Pande, *J. Mol. Biol.* **374**, 806 (2007).

[14] D. E. Shaw *et al.*, *Science* **330**, 341 (2010).

[15] T. R. Sosnick and J. R. Hinshaw, *Science* **334**, 464 (2011).

[16] J. Ma and M. Karplus, *Proc. Natl. Acad. Sci. U. S. A.* **94**, 11905 (1997).

[17] M. Lei *et al.*, *J. Mol. Biol.* **392**, 823 (2009).

[18] A. Warshel, *Proc. Natl. Acad. Sci. U. S. A.* **75**, 5250 (1978).

[19] D. K. Chakravorty and S. Hammes-Schiffer, *J. Am. Chem. Soc.* **132**, 7549 (2010).

[20] C. Maffeo, S. Bhattacharya, J. Yoo, D. Wells, and A. Aksimentiev, *Chem. Rev.* **112**, 6250 (2012).

[21] C. Kutzner *et al.*, *Biochim. Biophys. Acta - Biomembr.* **1858**, 1741 (2016).

[22] W. E. Miranda, V. A. Ngo, L. L. Perissinotti, and S. Y. Noskov, *Biochim. Biophys. Acta - Proteins Proteomics* **1865**, 1643 (2017).

[23] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, *Biophys. J.* **100**, L47 (2011).

[24] S. Rauscher *et al.*, *J. Chem. Theory Comput.* **11**, 5513 (2015).

[25] A. J. Stone, *The Theory of Intermolecular Forces*, Oxford University Press, Oxford, 2nd edition, 2016.

[26] W. Heisenberg, *Z. Phys.* **43**, 172 (1927).

[27] E. H. Kennard, *Z. Phys.* **44**, 326 (1927).

[28] W. Pauli, *Z. Phys.* **31**, 765 (1925).

[29] A. M. Reilly and A. Tkatchenko, *J. Chem. Phys.* **139**, 024705 (2013).

[30] N. Marom *et al.*, *Angew. Chemie Int. Ed.* **52**, 6629 (2013).

[31] A. M. Reilly and A. Tkatchenko, *Phys. Rev. Lett.* **113**, 055701 (2014).

[32] L. Kronik and A. Tkatchenko, *Acc. Chem. Res.* **47**, 3208 (2014).

[33] A. Ambrosetti, D. Alfè, R. A. DiStasio Jr., and A. Tkatchenko, *J. Phys. Chem. Lett.* **5**, 849 (2014).

[34] J. Hoja, A. M. Reilly, and A. Tkatchenko, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **7**, e1294 (2017).

[35] J. Hermann, D. Alfè, and A. Tkatchenko, *Nat. Commun.* **8**, 14052 (2017).

[36] P. Atkins and R. Friedman, *Molecular Quantum Mechanics*, Oxford University Press, Oxford, 4th edition, 2005.

[37] I. N. Levine, *Quantum Chemistry*, Pearson Advanced Chemistry Series, Pearson Education, Upper Saddle River, NJ, 7th edition, 2013.

[38] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry*, McGraw-Hill, New York, 1989.

[39] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, John Wiley & Sons, Chichester, 2nd edition, 2004.

[40] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, International Series of Monographs on Chemistry, Oxford University Press USA, New York, 1994.

[41] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, *Phys. Rev. B* **51**, 12947 (1995).

[42] M. Elstner *et al.*, *Phys. Rev. B* **58**, 7260 (1998).

[43] M. Gaus, Q. Cui, and M. Elstner, *J. Chem. Theory Comput.* **7**, 931 (2011).

[44] P. Koskinen and V. Mäkinen, *Comput. Mater. Sci.* **47**, 237 (2009).

[45] A. F. Oliveira, G. Seifert, T. Heine, and H. A. Duarte, *J. Braz. Chem. Soc.* **20**, 1193 (2009).

[46] M. Elstner and G. Seifert, *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **372**, 20120483 (2014).

[47] B. Hourahine *et al.*, *J. Chem. Phys.* **152**, 124101 (2020).

[48] A. Tkatchenko and M. Scheffler, *Phys. Rev. Lett.* **102**, 073005 (2009).

[49] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, *J. Chem. Phys.* **132**, 154104 (2010).

[50] A. Tkatchenko, R. A. DiStasio Jr., R. Car, and M. Scheffler, *Phys. Rev. Lett.* **108**, 236402 (2012).

[51] A. Ambrosetti, A. M. Reilly, R. A. DiStasio Jr., and A. Tkatchenko, *J. Chem. Phys.* **140**, 18A508 (2014).

[52] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics: Non-relativistic Theory*, Vol. **3** of *Course of Theoretical Physics*, Pergamon Press, Oxford, 2nd edition, 1965.

[53] E. Schrödinger, *Ann. Phys.* **384**, 361 (1926).

[54] E. Schrödinger, *Ann. Phys.* **384**, 489 (1926).

[55] E. Schrödinger, *Ann. Phys.* **385**, 437 (1926).

[56] E. Schrödinger, *Ann. Phys.* **386**, 109 (1926).

[57] M. Born and R. Oppenheimer, *Ann. Phys.* **389**, 457 (1927).

[58] P.-O. Löwdin, *Phys. Rev.* **97**, 1474 (1955).

[59] J. Čížek, *J. Chem. Phys.* **45**, 4256 (1966).

[60] C. Møller and M. S. Plesset, *Phys. Rev.* **46**, 618 (1934).

[61] P. Hohenberg and W. Kohn, *Phys. Rev.* **136**, B864 (1964).

[62] W. Kohn and L. J. Sham, *Phys. Rev.* **140**, A1133 (1965).

[63] C. Adamo and V. Barone, *J. Chem. Phys.* **110**, 6158 (1999).

[64] B. Santra, A. Michaelides, and M. Scheffler, *J. Chem. Phys.* **127**, 184104 (2007).

[65] N. Mardirossian and M. Head-Gordon, *Mol. Phys.* **115**, 2315 (2017).

[66] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, *J. Am. Chem. Soc.* **107**, 3902 (1985).

[67] O. F. Sankey and D. J. Niklewski, *Phys. Rev. B* **40**, 3979 (1989).

[68] J. P. Lewis *et al.*, *Phys. Rev. B* **64**, 195103 (2001).

[69] J. J. P. Stewart, *J. Mol. Model.* **13**, 1173 (2007).

[70] W. Thiel, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 145 (2014).

[71] C. Bannwarth, S. Ehlert, and S. Grimme, *J. Chem. Theory Comput.* **15**, 1652 (2019).

[72] H. A. Witek, C. Köhler, T. Frauenheim, K. Morokuma, and M. Elstner, *J. Phys. Chem. A* **111**, 5712 (2007).

[73] J. C. Slater and G. F. Koster, *Phys. Rev.* **94**, 1498 (1954).

[74] R. S. Mulliken, *J. Chem. Phys.* **23**, 1833 (1955).

[75] J. F. Janak, *Phys. Rev. B* **18**, 7165 (1978).

[76] M. Elstner, *J. Phys. Chem. A* **111**, 5614 (2007).

[77] C. Köhler *et al.*, *Phys. Chem. Chem. Phys.* **3**, 5109 (2001).

[78] C. Köhler, T. Frauenheim, B. Hourahine, G. Seifert, and M. Sternberg, *J. Phys. Chem. A* **111**, 5622 (2007).

[79] M. Gaus, A. Goez, and M. Elstner, *J. Chem. Theory Comput.* **9**, 338 (2013).

[80] M. Gaus, X. Lu, M. Elstner, and Q. Cui, *J. Chem. Theory Comput.* **10**, 1518 (2014).

[81] M. Kubillus, T. Kubař, M. Gaus, J. Řezáč, and M. Elstner, *J. Chem. Theory Comput.* **11**, 332 (2015).

[82] A. S. Christensen, T. Kubař, Q. Cui, and M. Elstner, *Chem. Rev.* **116**, 5301 (2016).

[83] S. Sanna, B. Hourahine, T. Frauenheim, and U. Gerstmann, *Phys. status solidi* **5**, 2358 (2008).

[84] T. A. Niehaus and F. Della Sala, *Phys. status solidi* **249**, 237 (2012).

[85] V. Lutsker, B. Aradi, and T. A. Niehaus, *J. Chem. Phys.* **143**, 184107 (2015).

[86] T. A. Niehaus, *J. Mol. Struct. THEOCHEM* **914**, 38 (2009).

[87] J. J. Kranz *et al.*, *J. Chem. Theory Comput.* **13**, 1737 (2017).

[88] T. Kowalczyk, K. Le, and S. Irle, *J. Chem. Theory Comput.* **12**, 313 (2016).

[89] J. R. Reimers *et al.*, *J. Phys. Chem. A* **111**, 5692 (2007).

[90] A. Pecchia, G. Penazzi, L. Salvucci, and A. Di Carlo, *New J. Phys.* **10**, 065022 (2008).

[91] A. Otero-de-la Roza and E. R. Johnson, *J. Chem. Phys.* **136**, 174109 (2012).

[92] V. V. Gobre and A. Tkatchenko, *Nat. Commun.* **4**, 2341 (2013).

[93] A. M. Reilly and A. Tkatchenko, *Chem. Sci.* **6**, 3289 (2015).

[94] S. Grimme, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 211 (2011).

[95] J. Klimeš and A. Michaelides, *J. Chem. Phys.* **137**, 120901 (2012).

[96] T. Björkman, A. Gulans, A. V. Krasheninnikov, and R. M. Nieminen, *Phys. Rev. Lett.* **108**, 235502 (2012).

[97] D. A. Egger and L. Kronik, *J. Phys. Chem. Lett.* **5**, 2728 (2014).

[98] A. Tkatchenko, *Adv. Funct. Mater.* **25**, 2054 (2015).

[99] Y. V. Shtogun and L. M. Woods, *J. Phys. Chem. Lett.* **1**, 1356 (2010).

[100] W. Gao and A. Tkatchenko, *Phys. Rev. Lett.* **114**, 096101 (2015).

[101] R. J. Maurer *et al.*, *Phys. Rev. Lett.* **116**, 146101 (2016).

[102] W. Gao and A. Tkatchenko, *Phys. Rev. Lett.* **111**, 045501 (2013).

[103] D. A. Egger, L. Kronik, and A. M. Rappe, *Angew. Chemie Int. Ed.* **54**, 12437 (2015).

[104] A. Fabrizio and C. Corminboeuf, *J. Phys. Chem. Lett.* **9**, 464 (2018).

[105] É. Brémond, N. Golubev, S. N. Steinmann, and C. Corminboeuf, *J. Chem. Phys.* **140**, 18A516 (2014).

[106] N. Ferri, R. A. DiStasio Jr., A. Ambrosetti, R. Car, and A. Tkatchenko, *Phys. Rev. Lett.* **114**, 176802 (2015).

[107] O. Sinanoğlu, Many-Electron Theory of Atoms, Molecules and Their Interactions, in *Adv. Chem. Phys.*, edited by I. Prigogine, pages 315–412, John Wiley & Sons, London, 1964.

[108] S. Wilson, *Electron Correlation in Molecules*, Dover Publications, New York, dover edition, 1984.

[109] K. Hirao, *Recent Advances in Multireference Methods*, Vol. **4** of *Recent Advances in Computational Chemistry*, World Scientific, Singapore, 1999.

[110] J. F. Dobson, A. White, and A. Rubio, *Phys. Rev. Lett.* **96**, 073201 (2006).

[111] J. Chen, A. Zen, J. G. Brandenburg, D. Alfè, and A. Michaelides, *Phys. Rev. B* **94**, 220102 (2016).

[112] L. M. Woods *et al.*, *Rev. Mod. Phys.* **88**, 045003 (2016).

[113] S. Y. Buhmann, *Dispersion Forces I*, Vol. **247** of *Springer Tracts in Modern Physics*, Springer-Verlag, Berlin/Heidelberg, 2012.

[114] J. Hermann, R. A. DiStasio Jr., and A. Tkatchenko, *Chem. Rev.* **117**, 4714 (2017).

[115] L. D. Landau and E. M. Lifshitz, *Statistical Physics*, Vol. **5** of *Course of Theoretical Physics*, Pergamon Press, Oxford, 2nd edition, 1970.

[116] M. Born and V. Fock, *Z. Phys.* **51**, 165 (1928).

[117] T. Kato, *J. Phys. Soc. Japan* **5**, 435 (1950).

[118] S. L. Adler, *Phys. Rev.* **126**, 413 (1962).

[119] N. Wiser, *Phys. Rev.* **129**, 62 (1963).

[120] J. Toulouse, F. Colonna, and A. Savin, *Phys. Rev. A* **70**, 062505 (2004).

[121] H. B. G. Casimir and D. Polder, *Phys. Rev.* **73**, 360 (1948).

[122] H. Hamaker, *Physica* **4**, 1058 (1937).

[123] F. London, *Z. Phys. Chem. Abt. B* **11**, 222 (1930).

[124] S. Grimme, *J. Comput. Chem.* **25**, 1463 (2004).

[125] A. D. Becke and E. R. Johnson, *J. Chem. Phys.* **122**, 154104 (2005).

[126] E. R. Johnson and A. D. Becke, *J. Chem. Phys.* **123**, 024101 (2005).

[127] S. Grimme, *J. Comput. Chem.* **27**, 1787 (2006).

[128] T. Sato and H. Nakai, *J. Chem. Phys.* **131**, 224104 (2009).

[129] S. N. Steinmann and C. Corminboeuf, *J. Chem. Theory Comput.* **7**, 3567 (2011).

[130] S. Grimme, S. Ehrlich, and L. Goerigk, *J. Comput. Chem.* **32**, 1456 (2011).

[131] A. Tkatchenko, A. Ambrosetti, and R. A. DiStasio Jr., *J. Chem. Phys.* **138**, 074106 (2013).

[132] Q. Wu and W. Yang, *J. Chem. Phys.* **116**, 515 (2002).

[133] J. Tao, J. P. Perdew, and A. Ruzsinszky, *Phys. Rev. B* **81**, 233102 (2010).

[134] J. Tao, J. P. Perdew, and A. Ruzsinszky, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 18 (2012).

[135] P. L. Silvestrelli, *Phys. Rev. Lett.* **100**, 053002 (2008).

[136] P. L. Silvestrelli, K. Benyahia, S. Grubisiĉ, F. Ancilotto, and F. Toigo, *J. Chem. Phys.* **130**, 074702 (2009).

[137] K. T. Tang and M. Karplus, *Phys. Rev.* **171**, 70 (1968).

[138] K. T. Tang, *Phys. Rev.* **177**, 108 (1969).

[139] T. Brinck, J. S. Murray, and P. Politzer, *J. Chem. Phys.* **98**, 4305 (1993).

[140] F. L. Hirshfeld, *Theor. Chim. Acta* **44**, 129 (1977).

[141] D. V. Fedorov, M. Sadhukhan, M. Stöhr, and A. Tkatchenko, *Phys. Rev. Lett.* **121**, 183401 (2018).

[142] M. Stöhr, G. S. Michelitsch, J. C. Tully, K. Reuter, and R. J. Maurer, *J. Chem. Phys.* **144**, 151101 (2016).

[143] R. Petraglia, S. N. Steinmann, and C. Corminboeuf, *Int. J. Quantum Chem.* **115**, 1265 (2015).

[144] P. W. Ayers, *J. Chem. Phys.* **113**, 10886 (2000).

[145] P. Bultinck, C. Van Alsenoy, P. W. Ayers, and R. Carbó-Dorca, *J. Chem. Phys.* **126**, 144111 (2007).

[146] T. Bučko, S. Lebègue, J. Hafner, and J. G. Ángyán, *J. Chem. Theory Comput.* **9**, 4293 (2013).

[147] T. Bučko, S. Lebègue, J. G. Ángyán, and J. Hafner, *J. Chem. Phys.* **141**, 034114 (2014).

[148] T. Gould, S. Lebègue, J. G. Ángyán, and T. Bučko, *J. Chem. Theory Comput.* **12**, 5920 (2016).

[149] V. G. Ruiz, W. Liu, E. Zojer, M. Scheffler, and A. Tkatchenko, *Phys. Rev. Lett.* **108**, 146103 (2012).

[150] E. M. Lifshitz, *Sov. Phys. – JETP* **2**, 73 (1956).

[151] E. Zaremba and W. Kohn, *Phys. Rev. B* **13**, 2270 (1976).

[152] J. Carrasco, W. Liu, A. Michaelides, and A. Tkatchenko, *J. Chem. Phys.* **140**, 084704 (2014).

[153] W. Liu *et al.*, *Phys. Rev. Lett.* **115**, 036104 (2015).

[154] R. J. Maurer *et al.*, *Prog. Surf. Sci.* **91**, 72 (2016).

[155] V. G. Ruiz, W. Liu, and A. Tkatchenko, *Phys. Rev. B* **93**, 035118 (2016).

[156] J. G. Brandenburg and S. Grimme, *J. Phys. Chem. Lett.* **5**, 1785 (2014).

[157] S. Ehrlich, J. Moellmann, W. Reckien, T. Bredow, and S. Grimme, *ChemPhysChem* **12**, 3414 (2011).

[158] E. Caldeweyher, C. Bannwarth, and S. Grimme, *J. Chem. Phys.* **147**, 034112 (2017).

[159] G. Starkschall and R. G. Gordon, *J. Chem. Phys.* **56**, 2801 (1972).

[160] K. T. Tang and J. P. Toennies, *J. Chem. Phys.* **80**, 3726 (1984).

[161] A. J. Thakkar, H. Hettema, and P. E. S. Wormer, *J. Chem. Phys.* **97**, 3252 (1992).

[162] J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.* **10**, 6615 (2008).

[163] B. M. Axilrod and E. Teller, *J. Chem. Phys.* **11**, 299 (1943).

[164] Y. Muto, *J. Phys.-Math. Soc. Japan* **17**, 629 (1943).

[165] M. T. Cvitaš, P. Soldán, and J. M. Hutson, *Mol. Phys.* **104**, 23 (2006).

[166] W. Cencek, M. Jeziorska, O. Akin-Ojo, and K. Szalewicz, *J. Phys. Chem. A* **111**, 11311 (2007).

[167] O. Anatole von Lilienfeld and A. Tkatchenko, *J. Chem. Phys.* **132**, 234109 (2010).

[168] J. Řezáč, Y. Huang, P. Hobza, and G. J. O. Beran, *J. Chem. Theory Comput.* **11**, 3065 (2015).

[169] G. J. O. Beran, *Chem. Rev.* **116**, 5567 (2016).

[170] J. F. Dobson, *Int. J. Quantum Chem.* **114**, 1157 (2014).

[171] V. A. Parsegian, *Van der Waals Forces*, Cambridge University Press, Cambridge, 2005.

[172] A. Ambrosetti, N. Ferri, R. A. DiStasio Jr., and A. Tkatchenko, *Science* **351**, 1171 (2016).

[173] C. Wagner *et al.*, *Nat. Commun.* **5**, 5568 (2014).

[174] C. Wagner *et al.*, *Phys. Rev. B* **81**, 035423 (2010).

[175] C. Kleimann, B. Stadtmüller, S. Schröder, and C. Kumpf, *J. Phys. Chem. C* **118**, 1652 (2014).

[176] X. Liu, Y. Wei, J. E. Reutt-Robey, and S. W. Robey, *J. Phys. Chem. C* **118**, 3523 (2014).

[177] G. A. Rance, D. H. Marsh, S. J. Bourne, T. J. Reade, and A. N. Khlobystov, *ACS Nano* **4**, 4920 (2010).

[178] C. A. S. Batista, R. G. Larson, and N. A. Kotov, *Science* **350**, 1242477 (2015).

[179] P. Loskill *et al.*, *Adv. Colloid Interface Sci.* **179-182**, 107 (2012).

[180] P. Loskill *et al.*, *J. R. Soc. Interface* **10**, 20120587 (2013).

[181] S. Tsoi *et al.*, *ACS Nano* **8**, 12410 (2014).

[182] A. Ambrosetti and P. L. Silvestrelli, *Carbon* **139**, 486 (2018).

[183] W. L. Bade, *J. Chem. Phys.* **27**, 1280 (1957).

[184] F. Wang and K. D. Jordan, *J. Chem. Phys.* **114**, 10717 (2001).

[185] A. P. Jones, J. Crain, V. P. Sokhan, T. W. Whitfield, and G. J. Martyna, *Phys. Rev. B* **87**, 144103 (2013).

[186] V. P. Sokhan, A. P. Jones, F. S. Cipcigan, J. Crain, and G. J. Martyna, *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6341 (2015).

[187] B. Thole, *Chem. Phys.* **59**, 341 (1981).

[188] R. A. DiStasio Jr., V. V. Gobre, and A. Tkatchenko, *J. Phys. Condens. Matter* **26**, 213202 (2014).

[189] M. A. Blood-Forsythe, T. Markovich, R. A. DiStasio Jr., R. Car, and A. Aspuru-Guzik, *Chem. Sci.* **7**, 1712 (2016).

[190] W. L. Bade and J. G. Kirkwood, *J. Chem. Phys.* **27**, 1284 (1957).

[191] G. D. Mahan, *J. Chem. Phys.* **43**, 1569 (1965).

[192] A. A. Lucas, *Physica* **35**, 353 (1967).

[193] B. Nijboer and M. Renne, *Chem. Phys. Lett.* **2**, 35 (1968).

[194] A. G. Donchev, *J. Chem. Phys.* **125**, 074713 (2006).

[195] H.-Y. Kim, J. O. Sofo, D. Velegol, M. W. Cole, and A. A. Lucas, *J. Chem. Phys.* **124**, 074504 (2006).

[196] M. W. Cole, D. Velegol, H.-Y. Kim, and A. A. Lucas, *Mol. Simul.* **35**, 849 (2009).

[197] J. Hermann and A. Tkatchenko, *J. Chem. Theory Comput.* **14**, 1361 (2018).

[198] T. Bučko, S. Lebègue, T. Gould, and J. G. Ángyán, *J. Phys. Condens. Matter* **28**, 045201 (2016).

[199] F. London, *Trans. Faraday Soc.* **33**, 8b (1937).

[200] F. London, *Z. Phys.* **63**, 245 (1930).

[201] T. W. Whitfield and G. J. Martyna, *Chem. Phys. Lett.* **424**, 409 (2006).

[202] A. P. Jones, A. Thompson, J. Crain, M. H. Müser, and G. J. Martyna, *Phys. Rev. B* **79**, 144119 (2009).

[203] A. P. Jones, F. S. Cipcigan, V. P. Sokhan, J. Crain, and G. J. Martyna, *Phys. Rev. Lett.* **110**, 227801 (2013).

[204] W. Kauzmann, Some Factors in the Interpretation of Protein Denaturation, Vol. **14** of *Advances in Protein Chemistry*, pages 1–63, Elsevier, 1959.

[205] D. Chandler, *Nature* **437**, 640 (2005).

[206] M.-C. Bellissent-Funel *et al.*, *Chem. Rev.* **116**, 7673 (2016).

[207] F. Persson, P. Söderhjelm, and B. Halle, *J. Chem. Phys.* **148**, 215104 (2018).

[208] S. Ebbinghaus *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **104**, 20749 (2007).

[209] K. Meister *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **110**, 1617 (2013).

[210] Y. Xu and M. Havenith, *J. Chem. Phys.* **143**, 170901 (2015).

[211] J. Duboisset and P.-F. Brevet, *Phys. Rev. Lett.* **120**, 263001 (2018).

[212] K. L. C. Hunt, *J. Chem. Phys.* **78**, 6149 (1983).

[213] P. Hauseux *et al.*, *Nat. Commun.* **11**, 1651 (2020).

[214] S. T. John and G. Csányi, *J. Phys. Chem. B* **121**, 10934 (2017).

[215] J. Wang *et al.*, *ACS Cent. Sci.* **5**, 755 (2019).

[216] H. DeVoe, *J. Chem. Phys.* **41**, 393 (1964).

[217] H. DeVoe, *J. Chem. Phys.* **43**, 3199 (1965).

[218] H. DeVoe, *J. Phys. Chem.* **75**, 1509 (1971).

[219] B. Self and D. Moore, *Biophys. J.* **73**, 339 (1997).

[220] S. Superchi, E. Giorgio, and C. Rosini, *Chirality* **16**, 422 (2004).

[221] F. Xu and T. A. Cross, *Proc. Natl. Acad. Sci. U. S. A.* **96**, 9057 (1999).

[222] A. Patriksson, C. M. Adams, F. Kjeldsen, R. A. Zubarev, and D. van der Spoel, *J. Phys. Chem. B* **111**, 13147 (2007).

[223] J. J. Dannenberg, The Importance of Cooperative Interactions and a Solid-State Paradigm to Proteins: What Peptide Chemists Can Learn from Molecular Crystals, in *Peptide Solvation and H-Bonds*, Vol. **72** of *Advances in Protein Chemistry*, pages 227–273, Academic Press, 2005.

[224] I. Shvab and R. J. Sadus, *J. Chem. Phys.* **139**, 194505 (2013).

[225] M. Rossi, W. Fang, and A. Michaelides, *J. Phys. Chem. Lett.* **6**, 4233 (2015).

[226] R. B. Best, W. Zheng, and J. Mittal, *J. Chem. Theory Comput.* **10**, 5113 (2014).

[227] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw, *J. Phys. Chem. B* **119**, 5113 (2015).

[228] P. Robustelli, S. Piana, and D. E. Shaw, *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4758 (2018).

[229] C. Baldauf and M. Rossi, *J. Phys. Condens. Matter* **27**, 493002 (2015).

[230] F. Schubert *et al.*, *Phys. Chem. Chem. Phys.* **17**, 7373 (2015).

[231] P. S. Nerenberg, B. Jo, C. So, A. Tripathy, and T. Head-Gordon, *J. Phys. Chem. B* **116**, 4524 (2012).

[232] J. Hoja *et al.*, *Sci. Adv.* **5**, eaau3338 (2019).

[233] W. Gao, Y. Chen, and Q. Jiang, *Phys. Rev. Lett.* **117**, 246101 (2016).

[234] H. Yada, M. Nagai, and K. Tanaka, *Chem. Phys. Lett.* **473**, 279 (2009).

[235] G. Folpini *et al.*, *Phys. Rev. Lett.* **119**, 097404 (2017).

[236] H. Fröhlich, *Int. J. Quantum Chem.* **2**, 641 (1968).

[237] G. Acbas, K. A. Niessen, E. H. Snell, and A. Markelz, *Nat. Commun.* **5**, 3076 (2014).

[238] J. Preto, M. Pettini, and J. A. Tuszynski, *Phys. Rev. E* **91**, 52710 (2015).

[239] I. Nardecchia *et al.*, *Phys. Rev. X* **8**, 031061 (2018).

[240] P. Kurian, G. Dunston, and J. Lindesay, *J. Theor. Biol.* **391**, 102 (2016).

[241] T. J. Craddock *et al.*, *Sci. Rep.* **7**, 9877 (2017).

[242] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, *Science* **334**, 517 (2011).

[243] J. Hermann, Libmbd software library, https://github.com/jhrmnn/libmbd/.

[244] S. Yang, Y. Jiang, S. Li, and W. Liu, *Carbon* **111**, 513 (2017).

[245] M. Rossi, S. Chutia, M. Scheffler, and V. Blum, *J. Phys. Chem. A* **118**, 7349 (2014).

[246] A. Tkatchenko, M. Rossi, V. Blum, J. Ireta, and M. Scheffler, *Phys. Rev. Lett.* **106**, 118102 (2011).

[247] M. Rossi, M. Scheffler, and V. Blum, *J. Phys. Chem. B* **117**, 5574 (2013).

[248] D. F. Plusquellic, K. Siegrist, E. J. Heilweil, and O. Esenturk, *ChemPhysChem* **8**, 2412 (2007).

[249] R. M. Levy, D. Perahia, and M. Karplus, *Proc. Natl. Acad. Sci. U. S. A.* **79**, 1346 (1982).

[250] E. Fischer, *Ber. Dtsch. Chem. Ges.* **27**, 2985 (1894).

[251] A. Cooper and D. T. F. Dryden, *Eur. Biophys. J.* **11**, 103 (1984).

[252] D. Kern and E. R. Zuiderweg, *Curr. Opin. Struct. Biol.* **13**, 748 (2003).

[253] S. J. Wodak *et al.*, *Structure* **27**, 566 (2019).

[254] M. F. Perutz, A. J. Wilkinson, M. Paoli, and G. G. Dodson, *Annu. Rev. Biophys. Biomol. Struct.* **27**, 1 (1998).

[255] Y. Yuan, M. F. Tam, V. Simplaceanu, and C. Ho, *Chem. Rev.* **115**, 1702 (2015).

[256] H. N. Motlagh, J. O. Wrabl, J. Li, and V. J. Hilser, *Nature* **508**, 331 (2014).

[257] P. Pagano, Q. Guo, A. Kohen, and C. M. Cheatum, *J. Phys. Chem. Lett.* **7**, 2507 (2016).

[258] J. C. Brookes, *Proc. R. Soc. A Math. Phys. Eng. Sci.* **473**, 20160822 (2017).

[259] M. Kržan *et al.*, *PLoS One* **11**, e0154002 (2016).

[260] D. Havelka, M. A. Deriu, M. Cifra, and O. Kučera, *Sci. Rep.* **7**, 4227 (2017).

[261] M. Heyden *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12068 (2010).

[262] R. D. Hoehn, D. E. Nichols, J. D. McCorvy, H. Neven, and S. Kais, *Proc. Natl. Acad. Sci. U. S. A.* **114**, 5595 (2017).

[263] T. A. Bickle and D. H. Krüger, *Microbiol. Rev.* **57**, 434 (1993).

[264] A. Jeltsch, *Gene* **317**, 13 (2003).

[265] A. M. Pingoud, M. Fuxreiter, V. Pingoud, and W. Wende, *Cell. Mol. Life Sci.* **62**, 685 (2005).

[266] R. Podgornik, R. H. French, and V. A. Parsegian, *J. Chem. Phys.* **124**, 044709 (2006).

[267] A. Grigorescu, M. Horvath, P. A. Wilkosz, K. Chandrasekhar, and J. M. Rosenberg, The Integration of Recognition and Cleavage: X-Ray Structures of Pre-Transition State Complex, Post-Reactive Complex, and the DNA-Free Endonuclease, in *Restriction Endonucleases*, edited by A. M. Pingoud, pages 137–177, Springer, Berlin/Heidelberg, 2004.

[268] F. Stahl, W. Wende, A. Jeltsch, and A. M. Pingoud, *Proc. Natl. Acad. Sci. U. S. A.* **93**, 6175 (1996).

[269] M. Sadhukhan and A. Tkatchenko, *Phys. Rev. Lett.* **118**, 210402 (2017).

[270] R. Pollice, M. Bot, I. J. Kobylianskii, I. Shenderovich, and P. Chen, *J. Am. Chem. Soc.* **139**, 13126 (2017).

[271] E. Secchi *et al.*, *Nature* **537**, 210 (2016).

[272] S. K. Kannam, B. D. Todd, J. S. Hansen, and P. J. Daivis, *J. Chem. Phys.* **138**, 094701 (2013).

[273] D. Mattia, H. Leese, and K. P. Lee, *J. Memb. Sci.* **475**, 266 (2015).

[274] A. Striolo, A. Michaelides, and L. Joly, *Annu. Rev. Chem. Biomol. Eng.* **7**, 533 (2016).

[275] S. Thussing and P. Jakob, *J. Phys. Chem. C* **120**, 9904 (2016).

[276] S. Granick, *Science* **253**, 1374 (1991).

[277] U. Raviv, P. Laurat, and J. Klein, *Nature* **413**, 51 (2001).

[278] J. Baugh, A. Kleinhammes, D. Han, Q. Wang, and Y. Wu, *Science* **294**, 1505 (2001).

[279] G. Algara-Siller *et al.*, *Nature* **519**, 443 (2015).

[280] L. Fumagalli *et al.*, *Science* **360**, 1339 (2018).

[281] T. N. Bhat *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **91**, 1089 (1994).

[282] U. Langhorst, J. Backmann, R. Loris, and J. Steyaert, *Biochemistry* **39**, 6586 (2000).

[283] B. Jayaram and T. Jain, *Annu. Rev. Biophys. Biomol. Struct.* **33**, 343 (2004).

[284] Y. Levy and J. N. Onuchic, *Annu. Rev. Biophys. Biomol. Struct.* **35**, 389 (2006).

[285] M. Ahmad, W. Gu, T. Geyer, and V. Helms, *Nat. Commun.* **2**, 261 (2011).

[286] P. Dutta, M. Botlani, and S. Varma, *J. Phys. Chem. B* **118**, 14795 (2014).

[287] G. Jiménez-Cadena, J. Riu, and F. X. Rius, *Analyst* **132**, 1083 (2007).

[288] Y.-G. Guo, J.-S. Hu, and L.-J. Wan, *Adv. Mater.* **20**, 2878 (2008).

[289] F. Zaera, *Chem. Soc. Rev.* **42**, 2746 (2013).

[290] M. E. Casco *et al.*, *Nat. Commun.* **6**, 6432 (2015).

[291] K. S. Vasu *et al.*, *Nat. Commun.* **7**, 12168 (2016).

[292] S. M. J. Rogge *et al.*, *Chem. Soc. Rev.* **46**, 3134 (2017).

[293] H. Li, J. Xiao, Q. Fu, and X. Bao, *Proc. Natl. Acad. Sci. U. S. A.* **114**, 5930 (2017).

[294] F. Giorgi *et al.*, *Sci. Rep.* **9**, 12689 (2019).

[295] M. Sadhukhan and F. R. Manby, *Phys. Rev. B* **94**, 115106 (2016).

[296] G. E. Scuseria, T. M. Henderson, and D. C. Sorensen, *J. Chem. Phys.* **129**, 231101 (2008).

[297] X. Ren, A. Tkatchenko, P. Rinke, and M. Scheffler, *Phys. Rev. Lett.* **106**, 153003 (2011).

[298] E. R. Johnson and A. D. Becke, *J. Chem. Phys.* **124**, 174104 (2006).

[299] J. G. Ángyán, *J. Chem. Phys.* **127**, 024108 (2007).

[300] P. W. Ayers, *J. Math. Chem.* **46**, 86 (2009).

[301] A. Heßelmann, *J. Chem. Phys.* **130**, 084104 (2009).

[302] V. Blum *et al.*, *Comput. Phys. Commun.* **180**, 2175 (2009).

[303] J. Řezáč, K. E. Riley, and P. Hobza, *J. Chem. Theory Comput.* **7**, 2427 (2011).

[304] T. Risthaus and S. Grimme, *J. Chem. Theory Comput.* **9**, 1580 (2013).

[305] J. F. Dobson, T. Gould, and G. Vignale, *Phys. Rev. X* **4**, 021040 (2014).

[306] T. Bereau and O. A. von Lilienfeld, *J. Chem. Phys.* **141**, 034101 (2014).

[307] C. Panosetti, A. Engelmann, L. Nemec, K. Reuter, and J. T. Margraf, *J. Chem. Theory Comput.* **16**, 2181 (2020).

[308] J. M. Knaup, B. Hourahine, and T. Frauenheim, *J. Phys. Chem. A* **111**, 5637 (2007).

[309] M. Gaus, C. P. Chou, H. A. Witek, and M. Elstner, *J. Phys. Chem. A* **113**, 11866 (2009).

[310] Z. Bodrog, B. Aradi, and T. Frauenheim, *J. Chem. Theory Comput.* **7**, 2654 (2011).

[311] M. Doemer, E. Liberatore, J. M. Knaup, I. Tavernelli, and U. Rothlisberger, *Mol. Phys.* **111**, 3595 (2013).

[312] A. F. Oliveira, P. Philipsen, and T. Heine, *J. Chem. Theory Comput.* **11**, 5209 (2015).

[313] C.-P. Chou *et al.*, *J. Chem. Theory Comput.* **12**, 53 (2016).

[314] A. Krishnapriyan, P. Yang, A. M. N. Niklasson, and M. J. Cawkwell, *J. Chem. Theory Comput.* **13**, 6191 (2017).

[315] V. L. Deringer and G. Csányi, *Phys. Rev. B* **95**, 094203 (2017).

[316] T. A. Niehaus, S. T. A. G. Melissen, B. Aradi, and S. M. Vaez Allaei, *J. Phys. Condens. Matter* **31**, 395901 (2019).

[317] T. Kamencek *et al.*, *J. Chem. Theory Comput.* **16**, 2716 (2020).

[318] M. Mortazavi, J. G. Brandenburg, R. J. Maurer, and A. Tkatchenko, *J. Phys. Chem. Lett.* **9**, 399 (2018).

[319] P. Goyal, M. Elstner, and Q. Cui, *J. Phys. Chem. B* **115**, 6790 (2011).

[320] M. Hellström *et al.*, *J. Phys. Chem. C* **117**, 17004 (2013).

[321] E. Małolepsza, H. A. Witek, and K. Morokuma, *Chem. Phys. Lett.* **412**, 237 (2005).

[322] J. J. Kranz, M. Kubillus, R. Ramakrishnan, O. A. von Lilienfeld, and M. Elstner, *J. Chem. Theory Comput.* **14**, 2341 (2018).

[323] K. Rajan, *Mater. Today* **8**, 38 (2005).

[324] G. Montavon *et al.*, *New J. Phys.* **15**, 095003 (2013).

[325] K. T. Schütt, F. Arbabzadah, S. Chmiela, K.-R. Müller, and A. Tkatchenko, *Nat. Commun.* **8**, 13890 (2017).

[326] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, *Nature* **559**, 547 (2018).

[327] O. A. von Lilienfeld, *Angew. Chemie Int. Ed.* **57**, 4164 (2018).

[328] J. Behler and M. Parrinello, *Phys. Rev. Lett.* **98**, 146401 (2007).

[329] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, *Phys. Rev. Lett.* **104**, 136403 (2010).

[330] K. Hansen *et al.*, *J. Phys. Chem. Lett.* **6**, 2326 (2015).

[331] S. Chmiela *et al.*, *Sci. Adv.* **3**, e1603015 (2017).

[332] W. Pronobis, A. Tkatchenko, and K.-R. Müller, *J. Chem. Theory Comput.* **14**, 2991 (2018).

[333] K. T. Schütt *et al.*, *J. Chem. Theory Comput.* **15**, 448 (2019).

[334] F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld, *J. Chem. Phys.* **148**, 241717 (2018).

[335] A. Grisafi and M. Ceriotti, *J. Chem. Phys.* **151**, 204105 (2019).

[336] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, *J. Chem. Theory Comput.* **11**, 2087 (2015).

[337] J. Zhu, V. Q. Vuong, B. G. Sumpter, and S. Irle, *MRS Commun.* **9**, 867 (2019).

[338] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, *J. Chem. Phys.* **148**, 241722 (2018).

[339] J. Hoja *et al.*, *QM7-X: A comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules* (2020).

[340] J. Hoja *et al.*, *arXiv* , 2006.15139 (2020).

[341] S. R. Bahn and K. W. Jacobsen, *Comput. Sci. Eng.* **4**, 56 (2002).

[342] National Institute of Standards and Technology, NIST Standard Reference Database Number 69, in *NIST Chem. Webb.*, edited by P.J. Linstrom and W.G. Mallard, Gaithersburg MD, 2018.

[343] L. Medrano Sandonas, *HC160: A dataset of linear, cyclic, and aromatic hydrocarbons* (2020).

[344] N. F. Aguirre, A. Morgenstern, M. J. Cawkwell, E. R. Batista, and P. Yang, *J. Chem. Theory Comput.* **16**, 1469 (2020).

[345] M. P. Repasky, J. Chandrasekhar, and W. L. Jorgensen, *J. Comput. Chem.* **23**, 1601 (2002).

[346] K. W. Sattelmeyer, J. Tirado-Rives, and W. L. Jorgensen, *J. Phys. Chem. A* **110**, 13551 (2006).

[347] S. Grimme, M. Steinmetz, and M. Korth, *J. Org. Chem.* **72**, 2118 (2007).

[348] M. Gruden *et al.*, *J. Comput. Chem.* **38**, 2171 (2017).

[349] S. Kullback and R. A. Leibler, *Ann. Math. Stat.* **22**, 79 (1951).

[350] M. Rossi *et al.*, *J. Phys. Chem. Lett.* **1**, 3465 (2010).

[351] F. Schubert *et al.*, *Phys. Chem. Chem. Phys.* **17**, 5376 (2015).

[352] J. Schmidt *et al.*, *J. Phys. Chem. B* **113**, 11959 (2009).

[353] A. Møgelhøj *et al.*, *J. Phys. Chem. B* **115**, 14149 (2011).

[354] R. Jonchiere, A. P. Seitsonen, G. Ferlat, A. M. Saitta, and R. Vuilleumier, *J. Chem. Phys.* **135**, 154503 (2011).

[355] J. Wang, G. Roman-Perez, J. M. Soler, E. Artacho, and M.-V. Fernández-Serra, *J. Chem. Phys.* **134**, 024516 (2011).

[356] R. A. DiStasio Jr., B. Santra, Z. Li, X. Wu, and R. Car, *J. Chem. Phys.* **141**, 084502 (2014).

[357] G. Tocci, L. Joly, and A. Michaelides, *Nano Lett.* **14**, 6872 (2014).

[358] P. Debye, *Ann. Phys.* **344**, 789 (1912).

[359] P. S. Venkataram, J. Hermann, A. Tkatchenko, and A. W. Rodriguez, *Phys. Rev. Lett.* **118**, 1 (2017).

[360] A. Kleshchonok and A. Tkatchenko, *Nat. Commun.* **9**, 3017 (2018).

[361] R. Podeszwa, W. Jankiewicz, M. Krzuś, and H. A. Witek, *J. Chem. Phys.* **150**, 234110 (2019).

[362] J. Řezáč, K. E. Riley, and P. Hobza, *J. Chem. Theory Comput.* **7**, 3466 (2011).

[363] J. Hermann, *Towards unified density-functional model of van der Waals interactions*, Dissertation, Humboldt-Universität zu Berlin, 2017.

[364] B. Huang, N. O. Symonds, and O. A. von Lilienfeld, *Quantum Machine Learning in Chemistry and Materials*, pages 1–27, Handbook of Materials Modeling, Springer International Publishing, Cham, 2nd edition, 2018.

[365] A. S. Christensen *et al.*, QML: A Python Toolkit for Quantum Machine Learning, https://github.com/qmlcode/qml, 2017.