Development and application of force fields for molecular simulations

Zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

von der KIT-Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

Dissertation

von

M.Sc. Manuel Konrad

Tag der mündlichen Prüfung: 23.04.2021

Referent Prof. Dr. Wolfgang Wenzel
Karlsruher Institut für Technologie

Korreferent Prof. Dr. Lars Pastewka
Albert-Ludwigs-Universität Freiburg

Abstract

Soft matter simulations include a wide range of applications, such as modeling biomolecules, polymers, and organic electronics materials. In order to achieve the length and time scales of relevant phenomena, the interactions in these systems are commonly calculated by computationally efficient analytical force fields. One part of this work describes an example application for force field based modeling of amorphous organic semiconductors. However, the conventional force field approach introduces parameters that have to be assigned from parameter sets suitable for the considered molecule. Mainly due to the simple function expressions for the non-covalent interactions, the fitting procedure for obtaining these parameter sets requires empirical target values, which are not always available. Bottom-up approaches, such as bottom-up force fields with fixed function expressions or neural network potentials, aim to replace the experimental data with results from ab initio calculations. For the application in large-scale molecular simulations, these methods still exhibit open challenges. Fixed function expressions suffer from limited flexibility to reproduce the ab initio potential energy surface and require manual type definitions to reduce the number of parameters. Neural network potentials improve both issues, but their high computational requirements limit the accessible length and time scales.

In this work, a novel bottom-up approach for modeling non-covalent interactions is presented, designed for large-scale simulations. The concept of efficient additive interactions is combined with the flexibility of artificial neural networks for the interpolation of different chemical configurations and geometric arrangements. The application of the model is demonstrated in molecular dynamics simulations, and the comparison of calculated thermodynamic properties of several small organic molecules with experimental data and conventional force fields reveals a promising predictive performance. Additionally, the model preserves the energy decomposition into physically motivated components provided by the symmetry-adapted perturbation theory used for the ab initio reference calculations. This separability and the independence from empirical data make this model potentially useful for future material design applications.

Zusammenfassung

Simulationen weicher Materie umfassen ein breites Spektrum von Anwendungen, wie z. B. die Modellierung von Biomolekülen, Polymeren und Materialien für die organische Elektronik. Um die Längen- und Zeitskalen relevanter Phänomene zu erreichen, werden die Wechselwirkungen in diesen Systemen üblicherweise durch recheneffiziente analytische Kraftfelder berechnet. Ein Teil dieser Arbeit beschreibt eine Beispielanwendung für die kraftfeldbasierte Modellierung von amorphen organischen Halbleitern. Der konventionelle Kraftfeldansatz führt jedoch Parameter ein, die aus für das betrachtete Molekül geeigneten Parametersätzen zugewiesen werden müssen. Vor allem aufgrund der einfachen Funktionsausdrücke für die nicht-kovalenten Wechselwirkungen erfordert das Verfahren zur Bestimmung dieser Parametersätze empirische Zielwerte, die nicht immer verfügbar sind. Bottom-up-Ansätze, wie z. B. Bottom-up-Kraftfelder mit festen Funktionsausdrücken oder Potentiale basierend auf neuronalen Netzen, zielen darauf ab, die experimentellen Daten durch Ergebnisse aus ab initio Rechnungen zu ersetzen. Für die Anwendung in umfangreichen Molekulardynamiksimulationen weisen diese Methoden noch offene Herausforderungen auf. Feste Funktionsausdrücke leiden unter einer begrenzten Flexibilität, die ab initio Potentialenergieoberfläche zu reproduzieren und erfordern manuelle Typdefinitionen, um die Anzahl der Parameter zu reduzieren. Potentiale, die auf neuronalen Netzen basieren, verbessern beide Aspekte, aber ihre hohen Rechenanforderungen begrenzen die zugänglichen Längen- und Zeitskalen.

In dieser Arbeit wird ein neuartiger Bottom-up-Ansatz zur Modellierung nicht-kovalenter Wechselwirkungen vorgestellt, der für großskalige Simulationen konzipiert ist. Das Konzept effizienter additiver Wechselwirkungen wird mit der Flexibilität künstlicher neuronaler Netze für die Interpolation verschiedener chemischer Zusammensetzungen und geometrischer Anordnungen kombiniert. Die Anwendung des Modells wird in Molekulardynamiksimulationen demonstriert, und der Vergleich der berechneten thermodynamischen Eigenschaften mehrerer kleiner organischer Moleküle mit experimentellen Daten und konventionellen Kraftfeldern zeigt eine vielversprechende Vorhersageleistung. Zusätzlich bewahrt das Modell die Energiezerlegung in physikalisch motivierte Komponenten, die von der symmetrieangepassten Störungstheorie, die für die ab initio Referenzrechnungen verwendet wird, bereitgestellt wird. Diese Trennbarkeit und die Unabhängigkeit von empirischen Daten machen dieses Modell potenziell nützlich für zukünftige Materialdesign-Anwendungen.

Contents

1 Introduction

Soft matter simulations include a wide range of applications such as protein folding [1], ligand docking [2], liquid crystal alignment [3], polymer self-assembly [4], and organic electronics [5]. In contrast to covalently or ionically bound hard matter, many key properties of soft matter systems are determined by non-covalent interactions (NCIs), causing the comparatively low cohesive energy density and eponymous soft character.

From an ab initio perspective, the two most popular concepts for the prediction of NCIs are the supermolecular approach, where the interaction between two molecules is deducted from the difference of dimer and monomer energies [6], and the symmetry-adapted perturbation theory (SAPT), where the NCI is obtained from a perturbation expansion [7]. The supermolecular approach is a very general concept applicable to many electronic structure methods that can model NCIs. In combination with coupled-cluster or Møller-Plesset theory [8–10], in particular CCSD(T) and MP2 [11–14], it is a popular choice for gold standard benchmark datasets [15, 16]. The SAPT method also offers a range of truncations, and the higher-order flavors can approach the accuracies of gold standard supermolecular calculations [17]. On top, it offers a natural decomposition of energies [7, 18]. The accessible length scales are limited for both approaches due to the unfavorable scaling of computational cost with system size [10, 17]. However, many phenomena in soft matter physics take place on large length and time scales where the application of ab initio methods is not feasible [19–25].

Alternatives to ab initio methods are molecular force fields which consist of analytical expressions for the inter- and intramolecular interactions. Traditionally, the intramolecular part is modeled by harmonic bond and angle terms and periodic torsional potentials, and the intermolecular part by pairwise additive Lennard-Jones potentials and partial charge interactions [26–28]. In combination with efficient simulation protocols such as Monte-Carlo sampling or molecular dynamics (MD) [29–32], the reduced computational effort extends the accessible length and time scales by several orders of magnitude compared to ab initio methods. However, this gain comes with the price of introducing parameters that have to be determined before the simulation. The intramolecular parameters can be chosen to reproduce vibrational and structural data, which can come from experiments or quantum mechanics [28, 33–35]. The partial charges can directly be fitted to the quantum mechanical electrostatic potential of the monomer geometry [36]. The simple Lennard-Jones potential requires the involvement of experimental target values in the parameter fitting procedure to achieve error cancellation [26, 27]. Especially for organic molecules, this top-down approach was successfully applied for the development of general force fields which provide transferable parameters for a wide range of functional groups [26–28, 37, 38]. In Chapter 3, I will present an application of a general force field in MD simulations for a series of organic semiconductors. The simulations are part of a collaborative study, where the MD trajectories are the starting point for a multiscale workflow to determine the material-specific molecular energy level distributions.

However, the top-down approach for the development of force fields also brings some drawbacks. The derived force fields are dependent on empirical data and, in general, provide an inaccurate description of physics at atomic scale [39]. This is the motivation for bottom-up force fields, where ab initio data for the NCIs is used for the parameter adjustment to, partially or completely, eliminate the empirical aspect from the model [40–52]. The design challenge is to find flexible and physically motivated function expressions that can reproduce the ab initio potential energy surface at an atomic scale [40]. Additionally, for the development of transferable force fields in general, a bookkeeping scheme is required to systematically classify atoms into a limited set of atom types to reduce the number of parameters and enable the application to new molecules not used in the parameter determination procedure [40].

An alternative bottom-up approach involves the use of artificial neural networks (ANNs) to predict interaction energies. In order to cope with high dimensional potential energy surfaces, Behler and Parrinello proposed a division of the model into submodels that compute atomic contributions to the total interaction [53]. Some of the first successful applications of these neural network potentials (NNPs) are models for predicting DFT level atomization energies of molecules containing a limited set of chemical elements [54]. For each atom, its contribution to the target property is computed by a separate ANN instance utilizing a symmetry function descriptor, a symmetrical representation of the neighboring atomic positions, whereby different instances for the same chemical element use identical ANN parameters [53, 54]. Compared to conventional bottom-up force fields, a manual definition of atom types or function expressions is not required, and the ANNs enable flexible regression capabilities. Furthermore, it was shown that the concept is also applicable with an atomic pairwise decomposition [55], which is the standard for the description of NCIs in conventional force field approaches. Recent studies tested both decomposition schemes to model NCIs of hydrogen-bonded dimer structures and for this task and the atomic pairwise NNP outperforms the model with atomic contributions [56, 57].

The NNPs mentioned above, with their main objective to predict energies across conformational and configurational space, represent successful examples of efficient surrogate models for quantum methods. When developing NNPs for dynamic soft matter simulations, there are further aspects to consider. For MD simulations, the relevant quantity is the force, which is related to the derivative of the energy. Models with high regression flexibility, such as NNPs with symmetry function descriptors, are prone to overfitting, which can lead to artifacts in the energy curves and, therefore, unstable forces [56, 57]. For large-scale simulations, another critical aspect is computational efficiency. Symmetry function descriptors have to be calculated on-the-fly at each step followed by an ANN inference, which is cheap compared to quantum methods but cannot compete with conventional force fields [58].

In Chapter 4, I will present the development and application of a neural network potential for non-covalent interactions, particularly designed to be applicable in large-scale MD simulations. The overall procedure can be described as a workflow consisting of several steps outlined in Section 4.1. The relevant fundamental concepts and theoretical methods are introduced in Chapter 2. The model is based on the Behler-Parrinello network architecture with an atomic pairwise decomposition [55, 57]. The submodels for the pairwise interactions are constructed to counteract overfitting and ensure smooth energy and force curves. As an alternative to symmetry functions, a pair fingerprint descriptor is introduced, which depends on equilibrium monomer properties and can therefore be precomputed for the application in MD, resulting in an efficient pairwise additive model. Similar to other bottom-up approaches, the model is trained to predict dimer interaction energies obtained by the SAPT method, whose inherent energy decomposition enables the separate training of an independent model for each component [40, 42, 56, 57]. The generation of the training samples, molecular dimers at different orientations and distances, is implemented in an automated procedure. The same applies to the fingerprint calculation and model training. Therefore, the overall model construction process requires no manual intervention.

In Section 4.2, the data efficiency of the model is examined. For a single molecule, the model performance is compared for different sized datasets. In Section 4.3, the developed workflow is deployed for the construction of a force field for several small organic molecules to test the ability of the model to interpolate between different dimer geometries. Furthermore, the model is applied in MD simulations to predict thermodynamic properties for all the molecules in the dataset. The predictions are compared to values from conventional force fields and experiments. In Section 4.4, a model is constructed for a set of hydrocarbons. In contrast to the previous application, the subsequent prediction of thermodynamic observables and comparison with experimental values is only performed for molecules that are not involved in the model training. With that, the model performance to interpolate in the space of pair descriptors is investigated and, therefore, its applicability for developing transferable force fields.

One key result of the present work is the development of a neural network potential for non-covalent interactions. Despite the lack of empirical target values in the training procedure, the applications show an outstanding agreement with experimental data. The resulting interactions are separable into physically motivated components and allow efficient large-scale MD simulations. The intervention-free training workflow enables a straightforward transfer to new scientific problems, making the approach potentially suitable for many applications and especially useful for material design of unknown molecular compounds.

2 Fundamental concepts and theoretical methods

In this chapter, the fundamental concepts and theoretical methods are introduced. First, an overview is given of the research field of multiscale modeling in the context of molecular simulation. The different scientific problems are classified into scale categories, and strategies are discussed on how to connect the scales. Then, for the subcategories electronic structure and molecular mechanics, the methods and concepts are introduced, which are relevant for this work. Finally, a brief introduction is given to machine learning with artificial neural networks and its application to develop neural network potentials.

2.1 Multiscale modeling of molecular systems

Multiscale modeling combines different computational methods to describe phenomena which are connected to multiple scales. The scales and related method categories relevant for molecular simulation are shown in Fig. 2.1.

On a **quantum mechanics** (QM) level, the goal is to obtain knowledge about the electronic structure of a system, i.e., the electronic wavefunction or density. QM methods enable investigations of various aspects, such as the potential energy surface for an arrangement of atoms [59], electronic transport properties [60], and reaction barriers [61]. Some results can directly be related to macroscopic quantities, such as band gaps [62], ionization potentials [63], and absorption spectra [64]. Several methods exist that range from parameter-free ab initio methods [65] to semi-empirical approaches [66], which vary in predictive power and computational cost.

In **molecular mechanics** (MM), the main idea is to replace the quantum methods with approximate analytical expressions, so-called force fields, that describe the potential energy surface, enabling a massive extension of length

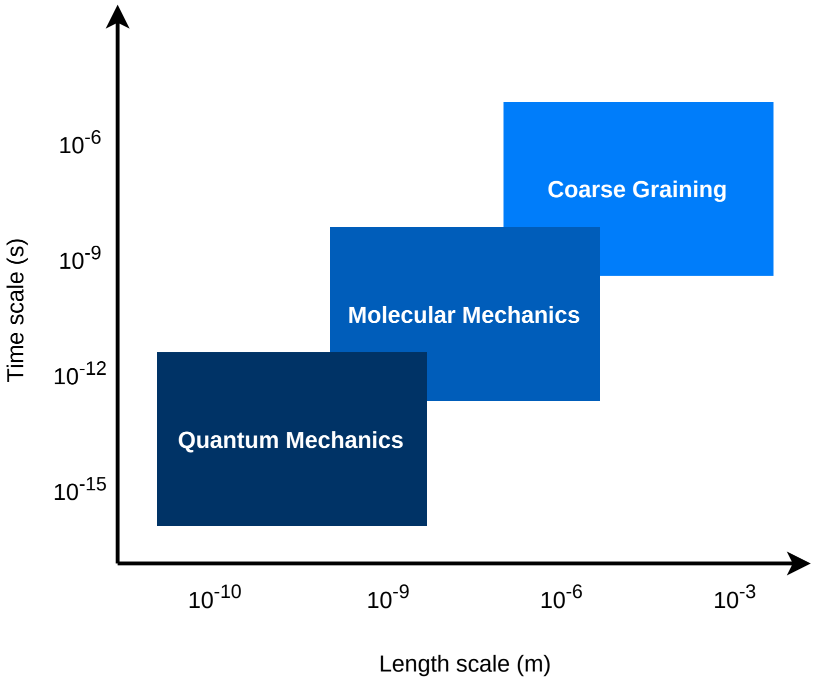


Fig. 2.1.: Multiscale modeling categories related to molecular simulation and their typical length and time scales.

and time scales while losing the electronic degrees of freedom. The force fields can be used for Monte Carlo sampling or to integrate the equations of motions in molecular dynamics simulations. Applications are numerous and include biomolecular simulations [67], morphology evolution [68], and reaction dynamics [69].

Coarse graining (CG) is a technique to further extend the time and length scales of molecular simulations by introducing super-atoms (beads) which are reduced representations of several atoms. CG models exist at different levels of detail where beads can represent atoms with implicit hydrogen atoms (united atoms) [70], functional groups [71] or larger fragments [72]. Similar to all-atomistic models, they can be applied with various sampling algorithms such as molecular dynamics [73], Monte Carlo [74] or dissipative particle dynamics [75]. Besides the reduced degrees of freedom, the elimination of fast bond vibrations and the resulting large timestep enables extended time scales for different applications in biomolecular and soft matter modeling [76, 77].

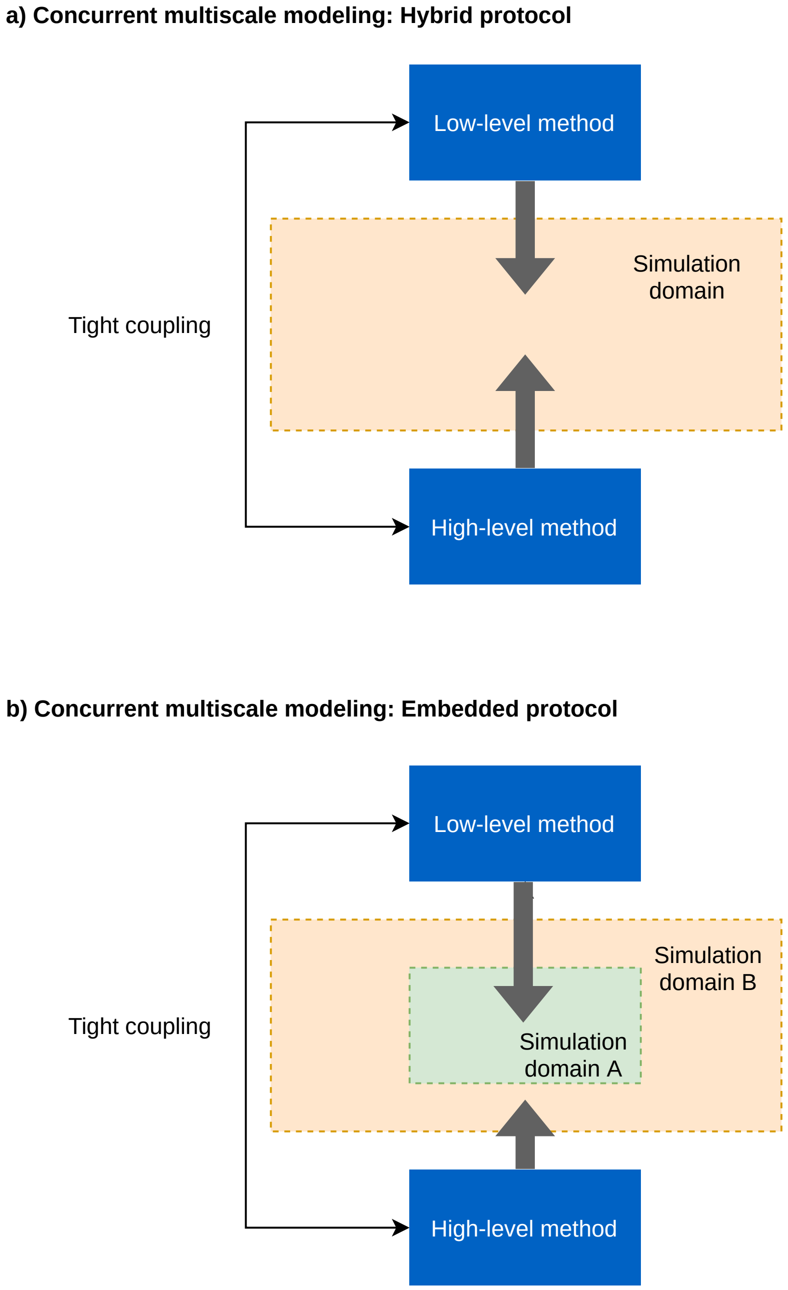
2.1.1 Bridging the scales

In multiscale modeling, there are different ways to connect the scales. In this section, some of the approaches are introduced.

Concurrent multiscale modeling describes tightly coupled simulation methods. In hybrid approaches, methods of different scales are combined in one simulation protocol (Fig. 2.2 a)). For example, in ab initio molecular dynamics, at each time step, forces are calculated by a QM method and used to integrate the classical equations of motions of the atoms [78]. Embedded approaches spatially divide the total system into regions of different levels of theory, which are connected by an interface region (Fig. 2.2 b)). In QM/MM approaches for example, a subsystem is treated at QM level and interfaces with a region treated by an analytical MM model [79]. For both approaches, the computational cost of the expensive method is a potential bottleneck for the accessible time scales.

In **sequential** multiscale modeling, the methods of the different scales are decoupled and connected via parameter passing (Fig. 2.3 a)). One example is force field development, where force constants or partial charges are calculated at the QM level and passed to the molecular mechanics level [80]. Another example is the modeling of charge transport in amorphous organic semiconductors, where energy levels and coupling parameters are computed by QM methods that enable kinetic Monte Carlo simulations of hopping transport on a lattice [81].

Mixed approaches extend sequential multiscale modeling by a feedback loop which introduces a loose coupling (Fig. 2.3 b)). For instance, if a more expensive low-level method passes parameters to a high-level method, the calculation of the parameters can be dependent on the system state exploration. Therefore, in order to account for this dependency, the parameters need to be updated by passing back the system state to the low-level method. In general, several simulation steps at the high-level method can be performed between the feedback calls, and the interval can increase with higher system state exploration. An application example is on-the-fly machine learning of force fields. [82]



a) Concurrent multiscale modeling: Hybrid protocol

b) Concurrent multiscale modeling: Embedded protocol

Fig. 2.2.: Two types of concurrent multiscale modeling approaches. a) Hybrid protocol: At each simulation step, the expensive low-level and cheap high-level methods are employed in the whole simulation domain. b) Embedded protocol: At each simulation step, the low- and high-level methods are employed in spatially separate domains of the simulation.

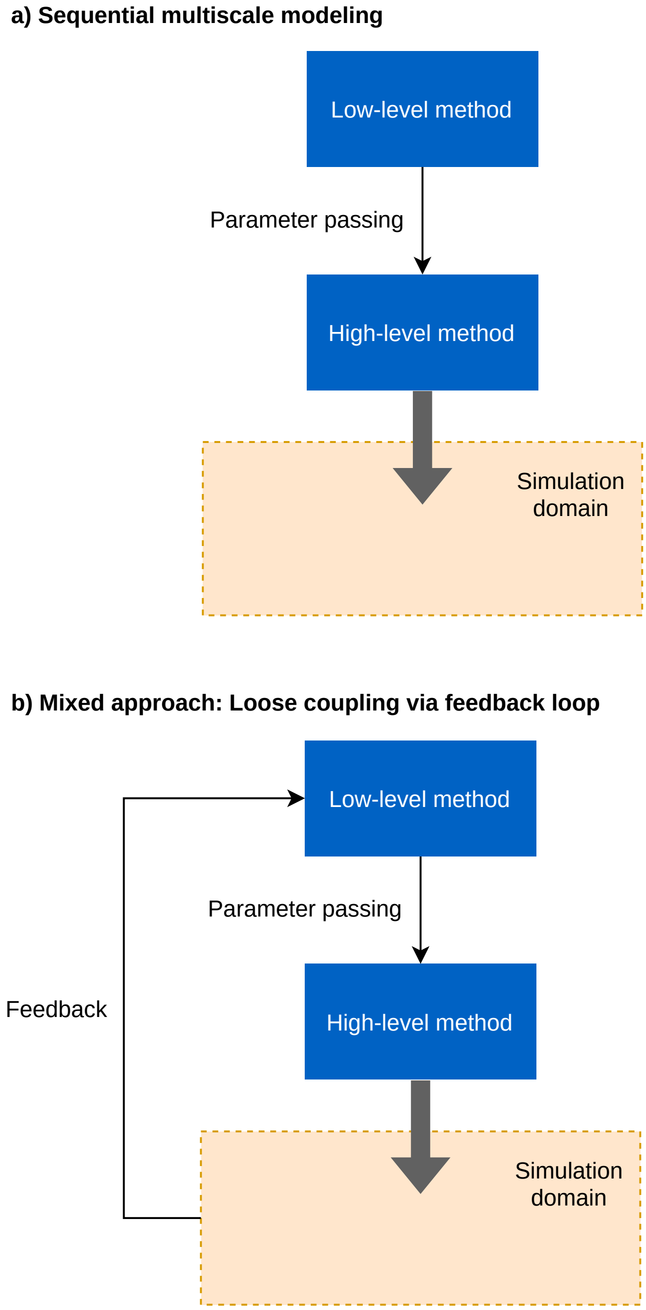


Fig. 2.3.: a) Sequential multiscale modeling approach: Parameters are computed by a low-level method and passed to a high-level method which is employed in the simulation domain. b) Loose coupling: Similar to the sequential approach, however, if required, a feedback is triggered to update the parameters on the basis of the system state.

2.2 Electronic structure methods

The time-independent non-relativistic Schrödinger equation for a system of n electrons and K nuclei is given by

$$\hat{H}(\vec{r},\vec{R})\Psi(\vec{r},\vec{R}) = E\Psi(\vec{r},\vec{R}) \tag{2.1}$$

where E is the eigenvalue of the molecular wavefunction Ψ, $\vec{r}$ are the coordinates of the electrons and $\vec{R}$ of the nuclei (the spin degrees of freedom are omitted). The non-relativistic Hamilton operator is given by

$$\hat{H} = \hat{T}_\mathrm{e} + \hat{T}_\mathrm{N} + \hat{V}_\mathrm{ee} + \hat{V}_\mathrm{NN} + \hat{V}_\mathrm{eN} \tag{2.2}$$

with the operators for the kinetic energies of the electrons

$$\hat{T}_\mathrm{e} = -\sum_{i=1}^{n} \frac{\hbar^2}{2m_\mathrm{e}}\nabla_i^2 \tag{2.3}$$

the kineteic energy of the nuclei

$$\hat{T}_\mathrm{N} = -\sum_{k=1}^{K} \frac{\hbar^2}{2M_k}\nabla_k^2 \tag{2.4}$$

the interaction between electrons

$$\hat{V}_\mathrm{ee} = \frac{q_\mathrm{e}^2}{4\pi\varepsilon_0}\sum_{i<j}\frac{1}{|\vec{r}_i - \vec{r}_j|} \tag{2.5}$$

the interaction between nuclei

$$\hat{V}_\mathrm{NN} = \frac{q_\mathrm{e}^2}{4\pi\varepsilon_0}\sum_{k<l}\frac{Z_k Z_l}{|\vec{R}_k - \vec{R}_l|} \tag{2.6}$$

and the interaction between electrons and nuclei

$$\hat{V}_\mathrm{eN} = -\frac{q_\mathrm{e}^2}{4\pi\varepsilon_0}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{Z_k}{|\vec{r}_i - \vec{R}_k|} \tag{2.7}$$

where q_e is the elementary charge, m_e is the electron mass, M_k and Z_k are the mass and atomic number of nucleus k, and ε_0 is the vacuum permittivity. Due to the high dimensionality of the problem an analytical or numerical solution for the full system is only possible for very few simple systems.

2.2.1 Born-Oppenheimer approximation

For most quantum chemistry applications, the standard approach to simplify Eq. 2.1 is the Born-Oppenheimer approximation. It is motivated by the weight difference of electrons and nuclei, enabling an adiabatic adjustment of the fast electron cloud to the slow nuclei. Therefore, it is reasonable to decouple the electronic and ionic part and to factorize the wavefunction:

$$\Psi(\vec{r}, \vec{R}) = \Psi_{\mathrm{N}}(\vec{R})\Psi_{\mathrm{e}}(\vec{r}, \vec{R}) \tag{2.8}$$

For the electronic part, the electrons are assumed to move in the fixed ionic potential. Without the kinetic energy terms of the nuclei, the resulting electronic Schrödinger equation only contains a parametric dependence on nuclei positions:

$$\hat{H}_{\mathrm{e}}(\vec{r}, \vec{R})\Psi_{\mathrm{e}}(\vec{r}, \vec{R}) = E_{\mathrm{e}}(\vec{R})\Psi_{\mathrm{e}}(\vec{r}, \vec{R}) \tag{2.9}$$

with the electronic Hamiltonian

$$\hat{H}_{\mathrm{e}} = \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{ee}} + \hat{V}_{\mathrm{eN}} \tag{2.10}$$

2.2.2 Density functional theory

Several methods have been developed to solve Eq. 2.9, such as Hartree-Fock and density functional theory (DFT). Here, a brief overview of DFT is given, which is a popular choice for a broad range of applications. DFT is based on the Hohenberg-Kohn theorem [83], which states that the external potential V_{ext} can be determined uniquely (except for constant) from the the ground state density ρ_0.

$$\rho_0 \overset{\mathbf{HK}}{\Longrightarrow} V_{\mathrm{ext}} \tag{2.11}$$

Since V_{ext} fully defines the Hamiltonian, the ground state density also defines all system properties. Therefore, the ground state energy E_0 is a universal functional of the ground state density:

$$E[\rho_0] = E_0 \tag{2.12}$$

Furthermore, the second part of the Hohenberg-Kohn theorem states that the true ground state density ρ_0 gives the lowest energy:

$$E[\rho_0] = E_0 \leq E[\rho] \tag{2.13}$$

where ρ is an arbitrary trial density. Unfortunately, the real functional is unknown. Therefore, Kohn and Sham reformulated the problem and paved the way for approximate functionals [84]. The Kohn-Sham approach introduces a fictitious auxiliary system of non-interacting particles with the same ground state density as the interacting system. The one-electron Kohn-Sham orbitals φ_i are introduced, which are related to the density of the n-particle system by

$$\rho = \sum_i^n |\varphi_i(\vec{r})|^2 \tag{2.14}$$

The functional of the total energy of the real system can now be regrouped into known and unknown parts:

$$E[\rho] = T_s[\rho] + \int \mathrm{d}^3r \; V_{\text{ext}}(\vec{r})\rho(\vec{r}) + J[\rho] + E_{\text{xc}}[\rho] \tag{2.15}$$

with the Kohn-Sham kinetic energy of n independent particles

$$T_s[\rho] = \sum_i^n \int \mathrm{d}^3r \; \varphi^*(\vec{r}) \left(-\frac{\hbar^2}{2m_{\text{e}}}\nabla^2\right) \varphi(\vec{r}) \tag{2.16}$$

and the Coulomb energy

$$J[\rho] = \frac{q_{\text{e}}^2}{8\pi\varepsilon_0} \int \mathrm{d}^3r \int \mathrm{d}^3r\,' \; \frac{\rho(\vec{r})\rho(\vec{r}\,')}{|\vec{r} - \vec{r}\,'|} \tag{2.17}$$

with the elementary charge q_{e}, the electron mass m_{e}, and the vacuum permittivity ε_0.

The exchange-correlation functional E_{xc} includes all the deviations from the known terms due to many-body effects:

$$E_{\mathrm{xc}}[\rho] = (T[\rho] - T_s[\rho]) + (E_{\mathrm{ee}}[\rho] - J[\rho]) \tag{2.18}$$

where $T[\rho]$ is the kinetic energy and $E_{\mathrm{ee}}[\rho]$ the electronic interaction of the real system. The Kohn-Sham orbitals are computed by the solving the Kohn-Sham equations:

$$\left(-\frac{\hbar^2}{2m_{\mathrm{e}}}\nabla^2 + V_{\mathrm{eff}}(\vec{r})\right)\varphi_i(\vec{r}) = \epsilon_i\varphi_i(\vec{r}) \tag{2.19}$$

where ϵ_i is the orbital energy of the Kohn-Sham orbital φ_i and the effective potential V_{eff} is dependent on the density itself:

$$V_{\mathrm{eff}}(\vec{r}) = V_{\mathrm{ext}}(\vec{r}) + \frac{q_{\mathrm{e}}^2}{4\pi\varepsilon_0}\int \mathrm{d}^3r'\,\frac{\rho(\vec{r}\,')}{|\vec{r}-\vec{r}\,'|} + \frac{\delta E_{\mathrm{xc}}[\rho]}{\delta\rho(\vec{r})} \tag{2.20}$$

Therefore, the solution is performed in a self-consistent loop starting from an initial guess for the Kohn-Sham orbitals (Fig. 2.4). The theoretical challenge is to find good approximations for the exchange-correlation functional E_{xc}. The simplest approach is the local density approximation which describes a dependence of the functional on the local density [83]. Many more extensions were developed, such as the generalized gradient approximation [85], which includes the dependence on the density gradient, or hybrid approaches that combine different approximations and energy terms from Hartree-Fock theory [86–88]. Different basis sets for the Kohn-Sham orbitals provide further options to choose a balance between accuracy and computational cost. Besides energies and densities, DFT also enables the evaluation of forces that can be used for geometry relaxations and dynamical simulations.

2.2.3 Partial charge fit and population analysis

Especially for molecular systems, several schemes have been developed to define atomic partial charges from QM results. A fundamental approach is to optimize point charges on the atomic positions to mimic the QM-derived electrostatic potential. For this, some methods use sample points which are placed in shells around molecules to compare the potential [36, 89, 90] (Fig. 2.5), others use a volume integral of the potential difference as cost function

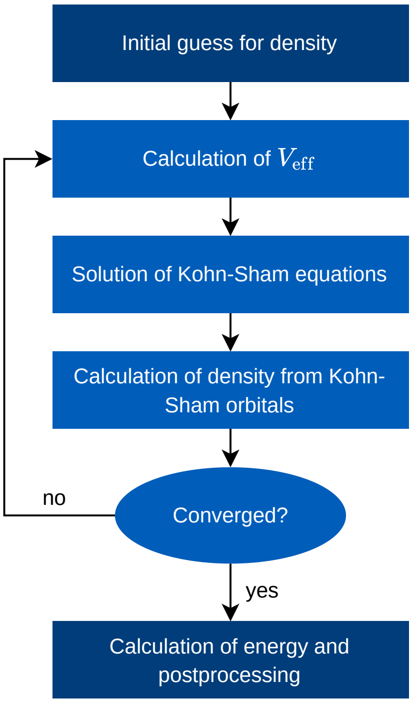

Initial guess for density
Calculation of V_{eff}
Solution of Kohn-Sham equations
Calculation of density from Kohn-Sham orbitals
Converged?
no
yes
Calculation of energy and postprocessing

Fig. 2.4.: Outline of the Self-Consistent Field approach for the iterative solution of the Kohn-Sham equations [84].

[91, 92]. The partial charges from these methods are often used to approximate long-range electrostatic interactions.

Other methods are based on the partitioning of the electronic density. One example is the Hirshfeld population analysis [94]. For the assignment of the molecular density to individual atoms, a promolecule is defined as the sum of all spherically averaged densities ρ_j^0 of the isolated atoms in the molecule:

$$\rho^{\text{pro}} = \sum_j \rho_j^0(\vec{r}) \tag{2.21}$$

With that, the Hirshfeld charge of atom i is defined by

$$q_i = Z_i - \int \mathrm{d}^3 r \, \frac{\rho_i^0(\vec{r})}{\rho^{\text{pro}}} (\rho(\vec{r}) - \rho^{\text{pro}}) \tag{2.22}$$

where Z_i is the nuclear charge of atom i, ρ_i^0 is the spherically averaged density of isolated atom i, and ρ is the molecular density as calculated from QM.

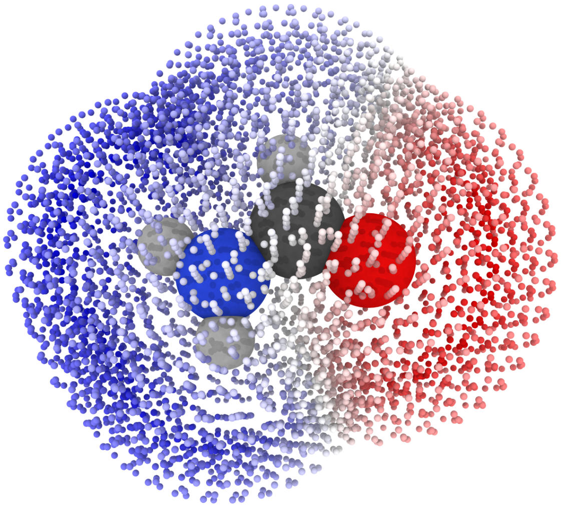


Fig. 2.5.: Shells of probing points used in the Merz-Singh-Kollmann ESP fitting scheme [89], shown here for the formamide molecule. Red probing points indicate a negative and blue a positive electrostatic potential. Visualization software: OVITO [93].

Hirshfeld charges are a measure for the charge reorganization due to bond formation [94].

Another way to extract information from QM results is to calculate bond orders which characterize the nature and strength of covalent bonds between the atoms in a molecule. In the Mayer bond order analysis [95], this information is extracted from the wave function. For the restricted closed shell case (no unpaired electrons), the bond order between atoms A and B is defined by

$$B_{AB}^{\mathrm{Mayer}} = \sum_{a \in A} \sum_{b \in B} (PS)_{ab} (PS)_{ba} \tag{2.23}$$

with the density matrix P and the overlap matrix S given by

$$\begin{aligned} P_{ab} &= 2 \sum_{i=1} c_{ai} c_{bi}^{*} && (2.24) \\ S_{ab} &= \int \mathrm{d}\vec{r} \chi_a(\vec{r}) \chi_b(\vec{r}) && (2.25) \end{aligned}$$

where the summation is performed over doubly occupied molecular orbitals i and c_{ai} is the coefficient of the basis function χ_a.

2.2.4 Non-covalent interactions and energy decomposition

Non-covalent interactions (NCIs) act between molecules or atoms, and as opposed to covalent interactions, they do not involve electron sharing. NCIs can also occur between fragments of the same molecule. However, in the following, the methods are described in the picture of two separate monomers. One general procedure to calculate NCIs from first principles is the supermolecular approach. Here, the non-covalent interaction energy between two fragments, such as molecules, is computed via the difference of the total energy E_{total} of the complex and the energies E_1 and E_2 of the two individual fragments:

$$E_{\text{NCI}} = E_{\text{total}} - E_1 - E_2 \tag{2.26}$$

In principle, any electronic structure method can be used for this approach. However, most DFT functionals do not model electron correlation adequately. Therefore, DFT calculations of NCIs require analytical dispersion energy corrections or non-local van-der-Waals functionals [96–99]. Alternatives are methods with an explicit treatment of electron correlation, such as the post-Hartree-Fock methods coupled-cluster and Møller-Plesset theory [8–10]. Both methods are often based on wavefunctions resulting from the Hartree-Fock method, an alternative to DFT, which treats electron exchange but neglects correlation. Furthermore, the supermolecular approach has an inherent inaccuracy for finite basis sets, the basis set superposition error, which describes the error that arises from the different number of basis functions in the monomer and dimer calculations. In the dimer calculation, each monomer can borrow basis functions from the other monomer, which improves the description of the wavefunction and lowers the energy. In the monomer calculations, these basis functions are missing. Therefore, the interaction energy is overestimated by Eq. 2.26. A simple approximate fix is the counterpoise correction. Additional basis functions are included in the monomer calculations, placed in each case at the atomic positions of the other monomer as ghost atoms.

A different approach to calculate NCIs is provided by symmetry-adapted perturbation theory (SAPT) [7]. The method is based on a perturbation expansion of the intermolecular energy, which directly computes the NCI and is free from basis set superposition errors. An important feature of the method is its intrinsic decomposition into separate energy contributions, which allows a physically motivated grouping into dispersion, exchange, electrostatic, and induction components:

$$E_{\text{inter}} = E_{\text{disp}} + E_{\text{exch}} + E_{\text{elec}} + E_{\text{ind}} \tag{2.27}$$

The attractive dispersion energy describes the interaction between correlated and instantaneous density fluctuations on the monomers. The repulsive exchange energy arises from the Pauli exclusion principle due to overlapping electron densities at short distances. The electrostatic energy describes the interaction between the charge densities of the monomers. At large distances, it is mainly determined by the interaction of the permanent multipoles of the monomers. At short distances, charge penetration effects due to density overlap can have a stabilizing effect. The induction component summarizes the contributions from the mutual polarization of the monomers.

There are several SAPT approaches available that truncate the expansion at different orders. Furthermore, a distinction is made between methods that describe the monomers through Hartree-Fock wavefunctions and Kohn-Sham orbitals of density functional theory. An overview of benchmarks for different variations and basis sets can be found in Ref. [17].

2.3 Classical molecular simulation

This section provides an introduction to classical simulation methods with an emphasis on dynamic simulations of molecules. However, the basic concepts are valid in general. A comprehensive overview of the topic can be found in Ref. [100].

2.3.1 Molecular force fields

The idea in molecular mechanics is to find analytical expressions for the quantum energy and forces, which drastically extends the accessible time and length scales. For molecules with a fixed configuration, it is convenient to divide the potential energy into intra- and intermolecular contributions.

$$E_{\text{pot}} = E_{\text{intra}} + E_{\text{inter}} \tag{2.28}$$

The intermolcular interaction is often computed by a sum over the pairwise Lennard-Jones and Coulomb contributions of atoms i and j:

$$E_{\text{inter}} = \sum_{\text{pairs}} \left[E_{ij}^{\text{LJ}} + E_{ij}^{\text{Coul}} \right] \tag{2.29}$$

The Coulomb energy is an estimate based on the interaction of partial charges or multipoles. For partial charges centered on the atoms, the functional form is (Fig. 2.6 a)):

$$E_{ij}^{\text{Coul}} = \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}} \tag{2.30}$$

where q_i and q_j are the partial charges on the atoms i and j, ε_0 is the vacuum permittivity, and r_{ij} is the distance between the atoms. The Lennard-Jones potential combines all the remaining attractive and repulsive contributions. A common choice is the 12/6-Lennard-Jones potential (Fig. 2.6 b)):

$$E_{ij}^{\text{LJ}} = 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) \tag{2.31}$$

where ϵ_{ij} and σ_{ij} are parameters which correspond to the potential well depth and the zero crossing. The exponent of the attractive term has its physical origin in the leading distance dependence of dispersion interactions. The main reason for the choice of the repulsive r^{-12} dependency is computational efficiency, since intermediate results can be reused during force field evaluation. However, other variations exist such as r^{-9} or exponential repulsion terms. The intramolecular interactions can be divided into several contributions:

$$\begin{aligned} E_{\text{intra}} = & \sum_{\text{bonds}} E_i^{\text{bond}} + \sum_{\text{angles}} E_i^{\text{angle}} \\ & + \sum_{\text{dihedrals}} E_i^{\text{dihedral}} + \sum_{\text{impropers}} E_i^{\text{improper}} \end{aligned} \tag{2.32}$$

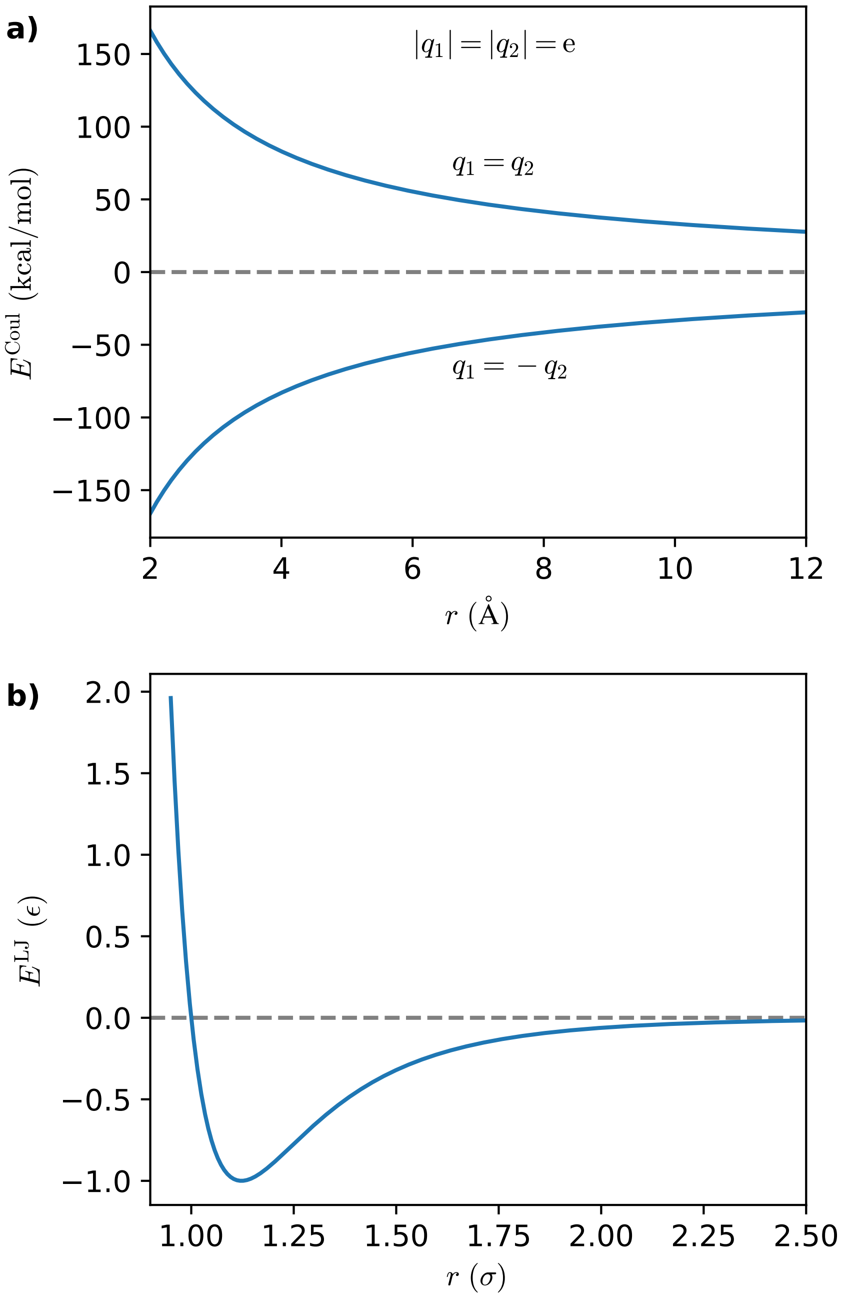


Fig. 2.6.: Examples for non-covalent force field contributions. a) Coulomb potential between two point charges q_1 and q_2. b) Lennard-Jones potential (in units of its parameters σ and ϵ).

In Fig. 2.7, the definitions of these two-, three-, and four-body terms are illustrated. For the bond, angle and improper terms, a popular choice is the harmonic potential:

$$E_i^{\text{bond}} = \frac{1}{2}k_{\text{b}}(r - r_0)^2 \tag{2.33}$$

with the bond length r, the equilibrium bond length r_0 and the force constant k_{b} (Fig. 2.7 a)).

$$E_i^{\text{angle}} = \frac{1}{2}k_{\theta}(\theta - \theta_0)^2 \tag{2.34}$$

with the angle θ, the equilibrium angle θ_0 and the force constant k_θ (Fig. 2.7 b)).

$$E_i^{\text{improper}} = \frac{1}{2}k_{\vartheta}(\vartheta - \vartheta_0)^2 \tag{2.35}$$

with the improper angle ϑ, the equilibrium angle ϑ_0 and the force constant k_ϑ (Fig. 2.7 c)). The dihedral potential is usually expressed by a periodic function. As an example, the GAFF force field uses [28]:

$$E_i^{\text{dihedral}} = \frac{1}{2}k_{\phi}[1 + \cos(n\phi - \gamma)] \tag{2.36}$$

where ϕ is the dihedral angle, n is the multiplicity, γ is the phase angle, and k_ϕ the force constant (Fig. 2.7 d)). Some force fields combine several terms with different multiplicities [27, 28].

An essential concept for the simulation of bulk systems is the introduction of periodic boundary conditions. In order to avoid a vacuum interface at the edge of the simulation box, the system is extended by copies of itself (periodic images) translated in all directions as illustrated in Fig. 2.8. Therefore, a finite unit cell can approximate a quasi-infinite system.

The calculation of the pairwise intermolecular interactions in Eq. 2.29 is a sum over atom pairs. The number of all possible pairs in a system of N particles is $\mathcal{O}(N^2)$. To improve this scaling issue, an approach is to only take into account pair interactions within a finite spherical cutoff (Fig. 2.8). For short-range contributions which converge to zero quickly, this is a good approximation. For long-range interactions, such as electrostatics, a special treatment is necessary. One example is the Ewald summation where the distance dependency of the Coulomb interaction is decomposed into two parts:

$$\frac{1}{r} = \frac{\text{erfc}(\alpha \text{r})}{r} + \frac{\text{erf}(\alpha r)}{r} \tag{2.37}$$

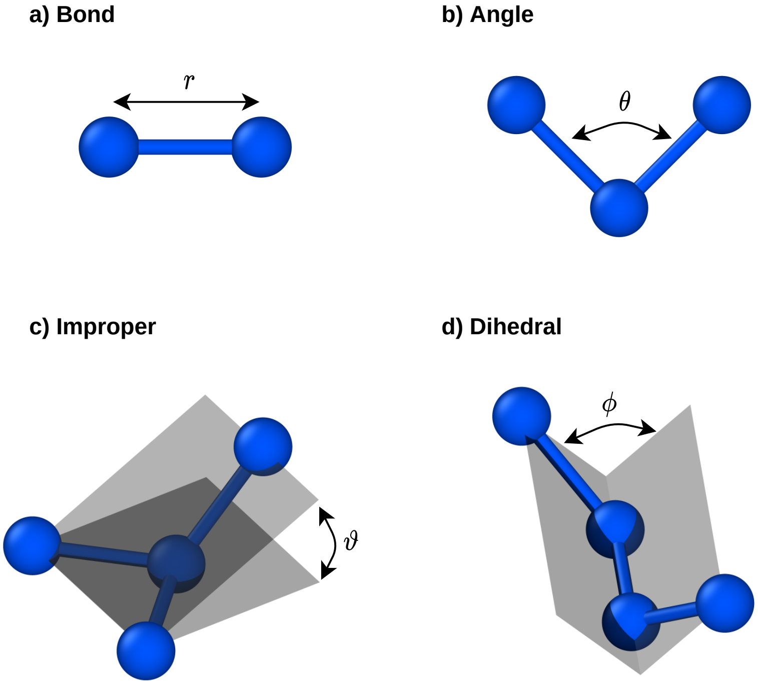


Fig. 2.7.: Definitions of different intramolecular force field contributions.

where α is the Ewald splitting parameter. The error function erf and complimentary error function erfc are defined by

$$\mathrm{erf}(x) = 1 - \mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\tau^2} \mathrm{d}\tau \tag{2.38}$$

The first term in Eq. 2.37 is convergent in real space and is computed by a direct summation, and the second part is a convergent sum in reciprocal space.

2.3.2 Force field parameters

The function expressions described in the previous section have several material-specific parameters that have to be determined before the simulation. As discussed in the introduction, the parameter values of a top-down force field can in part come from ab initio calculations, such as partial charges or covalent

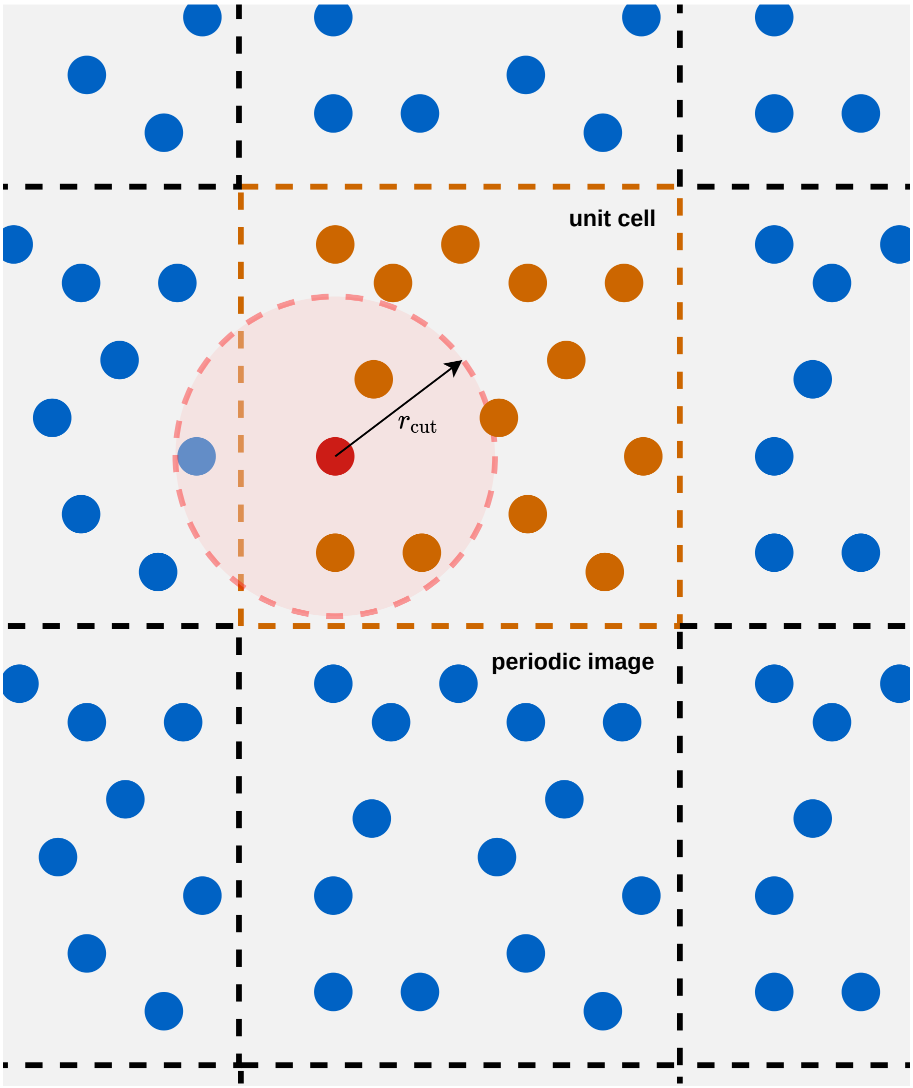


Fig. 2.8.: Periodic boundary conditions and the cutoff for pair interactions, illustrated for two dimensions.

interaction terms. However, especially for simple function expressions, such as the Lennard-Jones potential, empirical target values are required to achieve error cancellation, e.g., the enthalpy of vaporization and mass density [27, 28]. For bottom-up approaches, where also the non-covalent part is fitted to ab initio data, more flexible function expressions are required [40].

In general, a distinction can be made between custom and transferable force fields. While the function expressions can be the same, the main difference is the transferability of the parameters. For custom force fields, the parameters are fitted for specific molecular species and conditions. For transferable force fields, the idea is to create a set of parameters, which can also be used for new molecules that were not involved in the fitting process. The basic assumption is that an atom is mainly characterized by the functional group and not the entire molecule. This concept enables the definition of a set of parameters for a class of molecules that share characteristic functional groups such as organic molecules. In order to establish transferable parameters, the definition of appropriate atom types for the assignment of Lennard-Jones parameters is required. In Fig. 2.9, such a classification scheme is illustrated for the basic types of carbon and hydrogen in the GAFF force field [28]. As it can be seen, the same chemical element can be represented by various atom types, which are chosen depending on the topology of the molecule. The scheme varies for different force field families, but usually, mixing rules are applied to generate Lennard-Jones parameters for pairs of dissimilar atom types, which drastically reduces the number of parameters. Common choices are:

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \tag{2.39}$$

$$\sigma_{ij} = \sqrt{\sigma_{ii}\sigma_{jj}} \quad \text{or} \quad \sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2} \tag{2.40}$$

Similarly, bond, angle, improper, and dihedral types are defined, which can be classified depending on the involved atom types. The total number of types determines the number of free parameters of the force field.

2.3.3 Molecular dynamics

A simulation domain at time t can include several atoms with indices i at positions $r_i(t)$ with velocities $v_i(t)$ and forces $f_i(t)$ defined by a force field. In

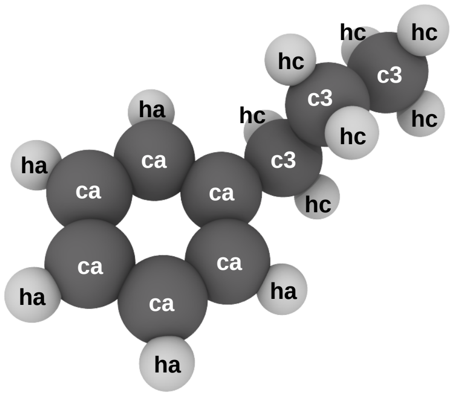


Fig. 2.9.: Assignment of GAFF atom types for the organic molecule propylbenzene (via the AmberTools software [101]). The molecule is described by the aromatic types "ca" (carbon) and "ha" (hydrogen) and aliphatic types "c3" and "hc". Visualization software: OVITO [93].

order to propagate the system state in time, one has to solve the equations of motion, a set of coupled differential equations:

$$m_i \vec{a}_i(t) = \vec{f}_i(t) \tag{2.41}$$

where m_i is the mass and $\vec{a}_i(t)$ the acceleration of atom i. The standard approach for this problem is an iterative solution, which propagates the system in discrete time steps. There are several algorithms available, here the common Velocity-Verlet method is briefly introduced. For each time t, the positions $\vec{r}(t)$ and velocities $\vec{v}(t)$ of the particles are saved. To propagate both vectors by the time step Δt, the following algorithm is applied:

$$\vec{r}_i(t+\Delta t) = \vec{r}_i(t) + \vec{r}_i(t)\Delta t + \frac{1}{2}\vec{a}_i(t)\Delta t^2 \tag{2.42}$$

$$\vec{v}_i(t+\Delta t) = \vec{v}_i(t) + \frac{1}{2}[\vec{a}_i(t) + \vec{a}_i(t+\Delta t)]\Delta t \tag{2.43}$$

The accelerations $\vec{a}_i(t)$ and $\vec{a}_i(t+\Delta t)$ are computed from the force field which also involves information about all interacting neighbors of atom i. The choice of the time step is a compromise between computational efficiency and integration accuracy. A time step too small limits the accessible time scale, and a time step too big can lead to energy drift and unstable simulations. In principle, the force field and the integration algorithm are the basic ingredients for conducting molecular dynamics simulations in the *NVE* ensemble. Often, however, it is of interest to run simulations at a specific temperature. The simulation temperature T is defined by the kinetic energy E_{kin}:

$$T(t) = \frac{2}{k_{\text{B}}(3N - N_c)} E_{\text{kin}} = \frac{2}{k_{\text{B}}(3N - N_c)} \sum_{i=1}^{N} \frac{m_i |\vec{v}_i(t)|^2}{2} \tag{2.44}$$

where N is the number of atoms, N_c the number of constraints, and k_{B} the Boltzmann constant. The equilibration of the system at a specific target temperature is realized by the application of a thermostat, such as the Berendsen method. At each step the velocities of the atoms are scaled by a factor $\lambda(t)$, which is defined to generate a temperature change ΔT that is proportional to the deviation from the target temperature T_0:

$$\Delta T = \frac{\Delta t}{\tau_T}(T_0 - T(t)) \overset{\text{Eq. 2.44}}{=} (\lambda(t)^2 - 1)T(t) \tag{2.45}$$

$$\lambda_T(t) = \sqrt{1 + \frac{\Delta t}{\tau_T}\left(\frac{T_0}{T(t)} - 1\right)} \tag{2.46}$$

where τ_T is the coupling parameter which is small for strong and high for weak coupling. The approach can also be extended to function as a barostat to control the simulation pressure by adjusting the simulation box volume V. The scaling factor $\lambda_P(t)$ is then applied to the coordinates and simulation cell vectors:

$$\lambda_P(t) = \left[1 - \frac{\beta \Delta t}{\tau_P}(P_0 - P(t))\right]^{\frac{1}{3}} \tag{2.47}$$

where P_0 is the target pressure, β the isothermal compressibility and τ_P the coupling parameter. Since β enters only via the ratio β/τ_P, the knowledge of an exact value is not crucial. The instantaneous pressure $P(t)$ is defined as:

$$P(t) = \frac{2E_{\text{kin}}}{3V} - \frac{1}{3V}\sum_{i<j} \vec{r}_{ij}(t) \cdot \vec{f}_{ij}(t) \tag{2.48}$$

The scaling factors τ_T and τ_P control how fast the deviation decays. In the limit of $\tau = \Delta t$, the observables are scaled to the target values at each time step, and for $\tau \to \infty$, the scaling is not active. The method effectively equilibrates a system to the desired temperature and pressure. However, it does not result in a correct *NVT* or *NPT* ensemble. Therefore, the Berendsen approach is mainly applied during system preparation. For production runs, where a consistent ensemble is essential, the Nosé-Hoover thermostat is a common choice [102, 103]. Here, an extended system is defined, which has an additional degree of freedom representing a heat bath and is constructed to sample a microcanonical ensemble with constant energy. Due to the equipartition theorem, the heat bath preserves a canonical ensemble at temperature T_0 in the real system. If the real system fluctuates from the target temperature, a heat flux from or to the extended degree of freedom counteracts as a friction term $\zeta(t)$ in the equation of motions [103]:

$$\begin{aligned} \vec{a}(t) &= \frac{\vec{f}(t)}{m} - \zeta(t)\vec{v}(t) && (2.49) \\ \frac{\mathrm{d}\zeta(t)}{\mathrm{d}t} &= \frac{1}{Q}\left(\sum_{i=1}^{N} m_i\vec{v}(t)^2 - gk_\mathrm{B}T_0\right) && (2.50) \end{aligned}$$

where g corresponds to the degrees of freedom in the real system. The parameter Q determines the coupling strength of the real system to the heat bath and has to be chosen carefully. The general approach of using an extended system can also be used for the development of a barostat to realize the *NPT* ensemble [104, 105].

Another approach to control the temperature is Langevin dynamics, where the equations of motion are supplemented by a friction and a stochastic term which act as a thermostat :

$$m_i\vec{a}_i(t) = \vec{f}_i(t) - \gamma\vec{v}_i(t) + \vec{\eta}_i(t) \qquad (2.51)$$

with the damping coefficient γ and a random force $\vec{\eta}_i(t)$ that fulfills:

$$\begin{aligned} \langle\vec{\eta}_i(t)\rangle &= 0 && (2.52) \\ \langle\vec{\eta}_i(t)\vec{\eta}_j(t')\rangle &= C\gamma k_\mathrm{B}T\delta_{ij}\delta(t-t') && (2.53) \end{aligned}$$

where C is a normalization factor depending on the distribution used to generate the random directions. Langevin dynamics is a popular choice for vacuum sampling simulations since it does not lead to the accumulation of errors in the total rotational and translational degrees of freedom of the system [106].

2.4 Machine learning of potential energy surfaces

Machine learning (ML) is the umbrella term for computer algorithms, which learn to perform specific tasks through a training procedure. The subject area can be divided into several approaches, including supervised learning, where the algorithm learns from known input/output value pairs, unsupervised learning, such as clustering analysis, and reinforcement learning, where an agent in an environment learns a behavior that maximizes the cumulative reward for its actions. Supervised learning algorithms have gained tremendous popularity over the last years, especially due to their advances in image classification. However, the concept is also applicable for regression tasks. The following subsections are a brief overview of supervised machine learning in the context of artificial neural networks for regression. Furthermore, the concept of neural network potentials is introduced.

2.4.1 Artificial neural networks

The basic building block of artificial neural networks (ANN) is the artificial neuron (Fig. 2.10 a)). It consists of a weighted sum T of its input features x_i, which is transformed by an activation function φ. The output y of an artificial neuron is given by

$$T = \sum_i w_i x_i \tag{2.54}$$

$$y = \varphi(T) \tag{2.55}$$

where the weights w_i are the parameters of the model. Artificial neurons can be used to construct many different network architectures. In the following, the feed-forward artificial neural network with fully-connected layers is discussed (Fig. 2.10 b)). It is composed of one or more consecutive layers that each consists of several neurons. The last layer is identified as the output layer, and if there are more layers, they are called hidden layers. Each layer uses either the input features or the output of the preceding layer to generate a new representation. As a generalization, the output of a neuron k inside an ANN can be defined as:

$$T_k = \sum_i w_{ik} y_i \tag{2.56}$$

$$y_k = \varphi(T_k) \tag{2.57}$$

where y_i is the ith output of the preceding layer, or if neuron k is in the first layer, y_i corresponds to the model input feature x_i. The number of neurons in a layer and the number of layers are hyperparameters of the model, which are the model properties that are not optimized directly during training. The term deep learning is usually used for ML approaches that include at least one hidden layer. Another hyperparameter is the choice of the activation function. In Fig. 2.11, several activation functions are shown. In order to model non-linear relationships between input and output, a non-linear activation function has to be chosen.

2.4.2 Backpropagation and model optimization

The goal of machine learning is to maximize the model performance by optimizing its parameters, the weights and biases. Therefore, it is first required to define a metric that measures the model error, which is also called loss function. For regression applications, a popular choice is the squared error loss function L^{SE}. The loss for a prediction/target value pair of an output neuron is defined as

$$L^{\mathrm{SE}} = (y^{\mathrm{target}} - y^{\mathrm{out}})^2 \tag{2.58}$$

where y^{target} is the true value and y^{out} the prediction of the output neuron. Effective optimization methods for finding the parameter values which minimize the loss function require the computation of a gradient. The partial derivative of the loss with respect to the weights is resulting from chain rule:

$$\frac{\partial L^{\mathrm{SE}}}{\partial w_{ik}} = \frac{\partial L^{\mathrm{SE}}}{\partial y_k} \frac{\partial y_k}{\partial T_k} \frac{\partial T_k}{\partial w_{ik}} = \delta_k y_i \tag{2.59}$$

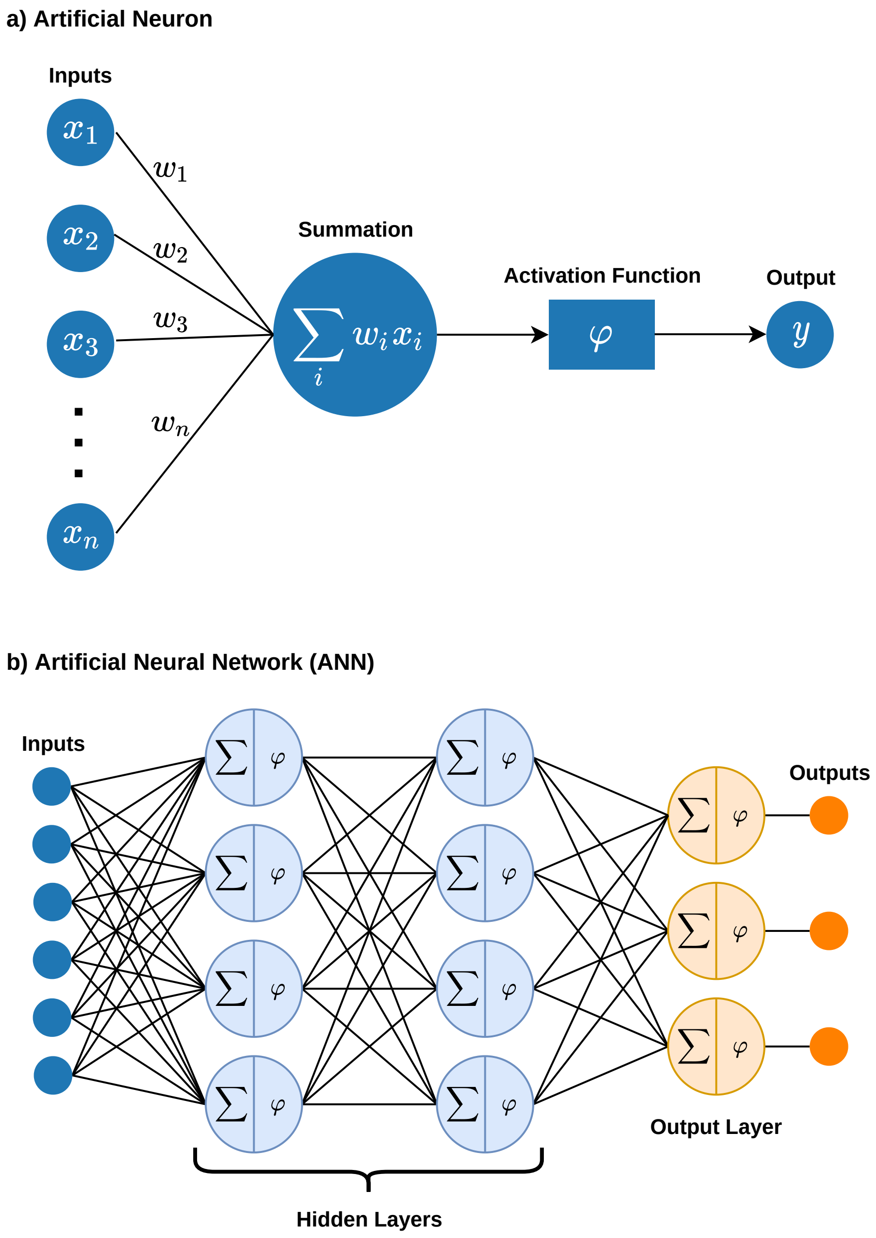


Fig. 2.10.: a) Artificial neuron: The output results from a weighted sum of all input values, which is transformed by an activation function. b) Fully-connected artificial neural network (ANN): Consists of several artificial neurons arranged in multiple layers. The number of layers and neurons in each layer are hyperparameters of the model.

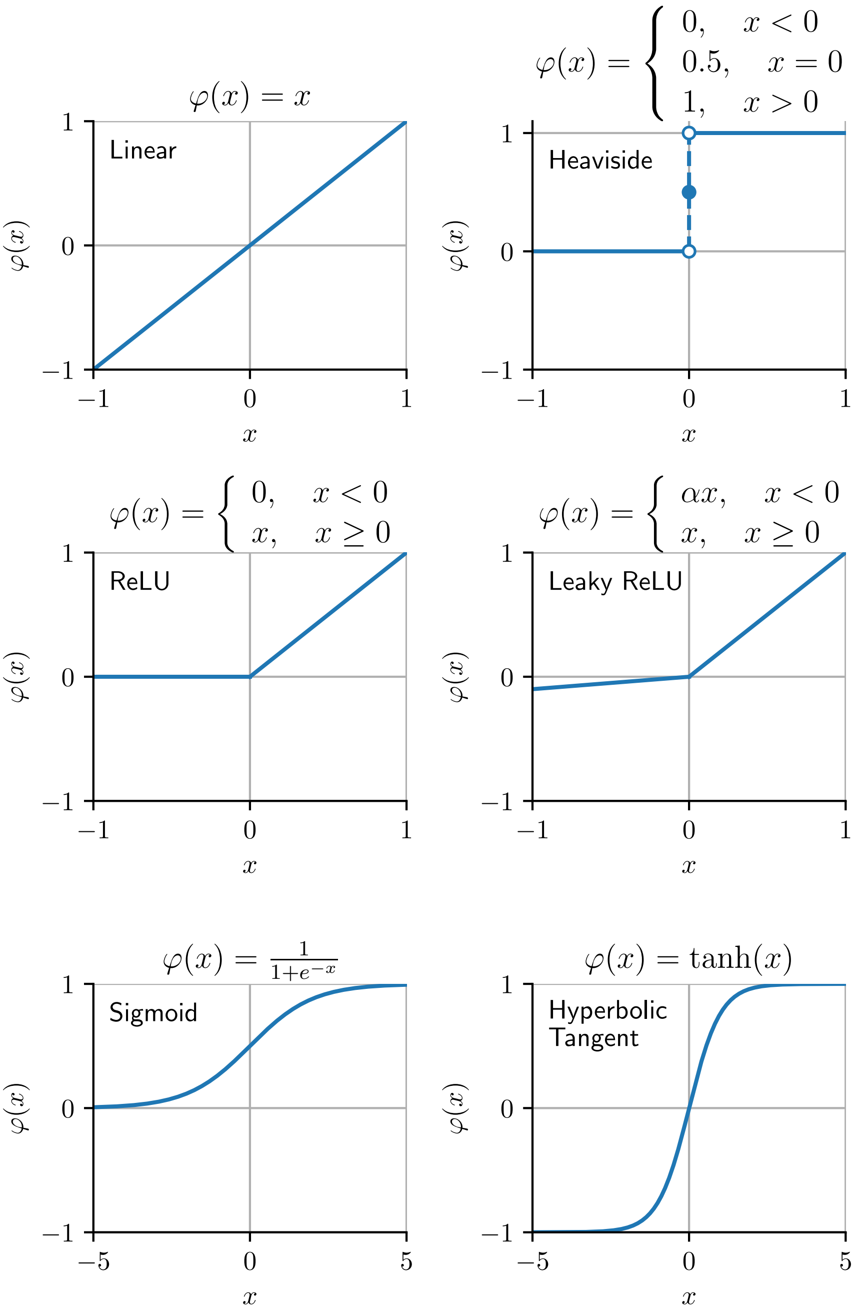


Fig. 2.11.: Expressions and plots of different activation functions.

The expression for δ_k depends on the position of the neuron k in the network, either in the output layer or in one of the hidden layers. For neurons in the output layer, y_k corresponds to y^{out}, and δ_k is given by:

$$\delta_k = 2\frac{\partial\varphi(T_k)}{\partial T_k}(y_k - y^{\text{target}}) \tag{2.60}$$

For the neurons in a hidden layer, δ_k is given by a recursive expression:

$$\delta_k = \frac{\partial\varphi(T_k)}{\partial T_k}\sum_j w_{kj}\delta_j \tag{2.61}$$

where the summation is performed over all neurons j of the next layer which are connected to neuron k. With that, the partial derivatives of the loss with respect to all model parameters are defined, which is the basis for the backpropagation algorithm. It includes the following steps:

- Forward propagation: The input features of a sample are propagated through the network to compute the neuron values T and the model output. With the model output and the target values of the sample, the loss is calculated.

- Backpropagation: Starting from the loss function, the gradients are calculated layer by layer using multiple iterations of the chain rule.

- Weight update: The calculated gradients are used to update the weights according to a predefined rule. An example for a simple update rule for the weight w_{ik} is the gradient descent method with a fixed learning rate α:

$$\Delta w_{ik} = w_{ik}^{\text{new}} - w_{ik}^{\text{old}} = -\alpha\frac{\partial L^{\text{SE}}}{\partial w_{ik}} \tag{2.62}$$

As seen in Eq. 2.60 and 2.61, the algorithm contains a derivative of the activation function, therefore, the chosen function should at least partially have a non-zero derivative. Furthermore, the basic gradient descent method (Eq. 2.62) represents one of the simplest approaches to update the weights. Other more sophisticated parameter update algorithms exist such as the adaptive moment estimation optimizer (Adam) [107]. Finally, in practice, a bias parameter b_k is added to the summation in T_k:

$$T_k = b_k + \sum_i w_{ik} y_i \tag{2.63}$$

2.4.3 Training procedure

In general, the training of an ANN requires several training samples and multiple iterations of the backpropagation algorithm described in the previous section. For computational efficiency reasons, it is common to divide the training dataset into batches, for which the backpropagation is performed simultaneously. The gradient is then averaged over the samples of the batch. A training cycle where all batches are presented to the model once is called an epoch. Usually, multiple epochs are required to complete the model training. The total number of epochs needs to be chosen with care. If the training is only stopped upon the convergence of the training loss, high dimensional models such as ANNs can already be in an overfitting regime. Although the model parameters minimize the training loss, the performance for new samples not involved in the training is not optimal. In Fig. 2.12 a), the overfitting issue is illustrated for a simple example.

A simple approach to improve the generalization capabilities is the early stopping method (Fig. 2.12 b)). Before the training, the dataset is split into a training set used for parameter optimization and a validation set representing new samples unknown to the model. Then, in regular intervals during the training, the loss of the validation set is checked. If the validation loss is not improving for a predefined number of epochs, the training is stopped. Finally, the generalization performance of the trained model can be computed on an independent test set not involved in training or validation. There are also other techniques to improve the generalization of the model, such as regularization methods, e.g., the L1 regularization, where an additional loss term penalizes the weight parameters:

$$L^{\mathrm{L1}} = \lambda \sum_{m \in M} |w_m| \tag{2.64}$$

$$L^{\mathrm{total}} = L^{\mathrm{SE}} + L^{\mathrm{L1}} \tag{2.65}$$

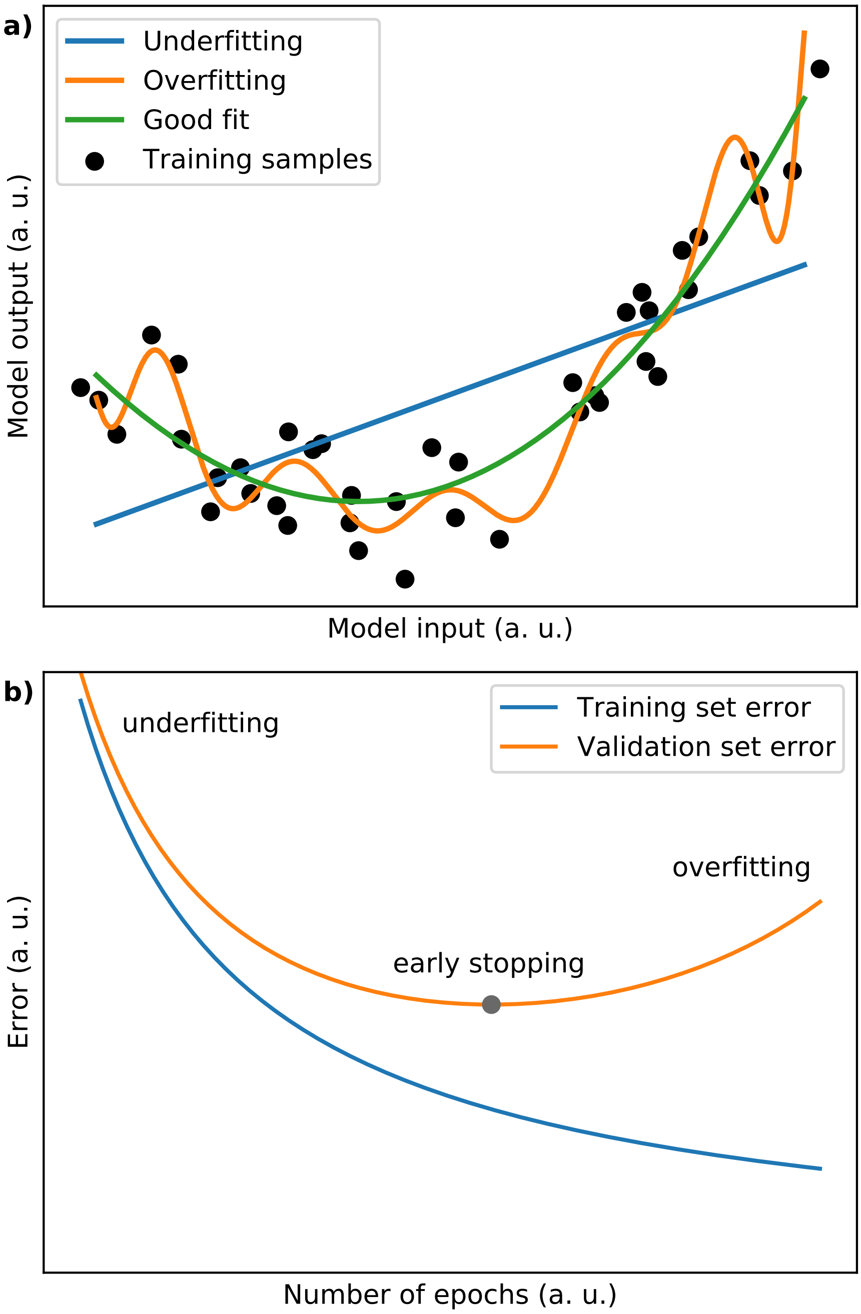


Fig. 2.12.: a) Illustration of under- and overfitting. b) Early stopping method: When the validation set error is not decreasing any more the training is stopped to avoid overfitting.

where λ is the penalty strength, and the sum is performed over the weights of choice M, which can be a subset of layers. The L^{L1} term competes with the model loss L^{SE} and has the effect that weights that do not improve L^{SE} significantly are driven towards zero. When applied to the input layer, this method effectively acts as an input feature selection.

2.4.4 Neural network potentials

The motivation for developing a neural network potential is to use the regression capabilities of an ANN to find a relationship between the geometric structure of an arrangement of atoms and some property, such as the potential energy surface, to realize cheap surrogate models for expensive quantum chemical methods. The methodological challenge is to find an appropriate geometry representation and network structure. In the following, the discussion is continued in the context of atomization energy predictions for covalent complexes. However, some aspects also apply to other properties.

If one considers directly using the Cartesian coordinates of the arrangement as input for a regular ANN, several problems become obvious. First, the Cartesian coordinates are not invariant to symmetry transformations, which means that energetically equivalent structures can have completely different numerical input features. Another issue, also related to the network structure, is the missing invariance for index exchanges. Even when only considering one chemical element, the ANN output is dependent on the sorting of the input features. The fixed network structure makes it also difficult to develop models for a varying number of atoms. Therefore, many alternative representations have been developed to eliminate the discussed issues [53, 108, 109]. One of them has been developed by Behler and Parrinello, which is briefly outlined in the following [53].

The main concept of the Behler-Parrinello approach is the reduction of dimensionality by dividing the model into submodels, which each compute a partial contribution of the total quantity (Fig. 2.13). In this example, the submodels compute atomic contributions to the atomization energy of a molecule. As input features for the submodels, the symmetry function descriptor is used, which describes the local environment of the corresponding atom and obeys the required symmetries. The radial symmetry functions are defined by a

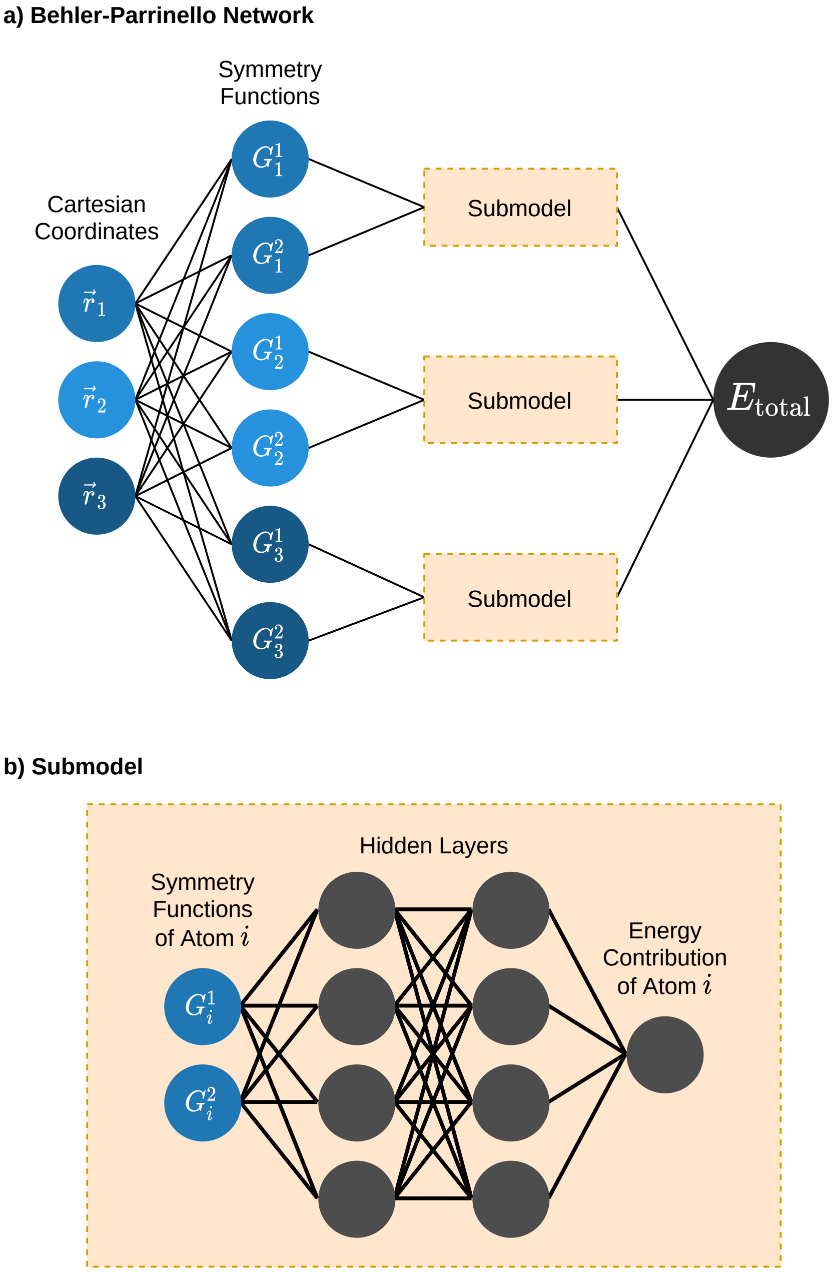


Fig. 2.13.: a) Behler-Parrinello network: The Cartesian coordinates are transformed to multiple symmetry function descriptors for each atom. The descriptors are the input for a submodel instance which computes the atomic contribution to the total energy. b) The submodel is represented by a fully-connected ANN with a single output node. Submodel instances corresponding to the same chemical element use identical parameters. In actual applications, usually more symmetry functions and nodes are used than shown here. Further details can be found in Ref. [53].

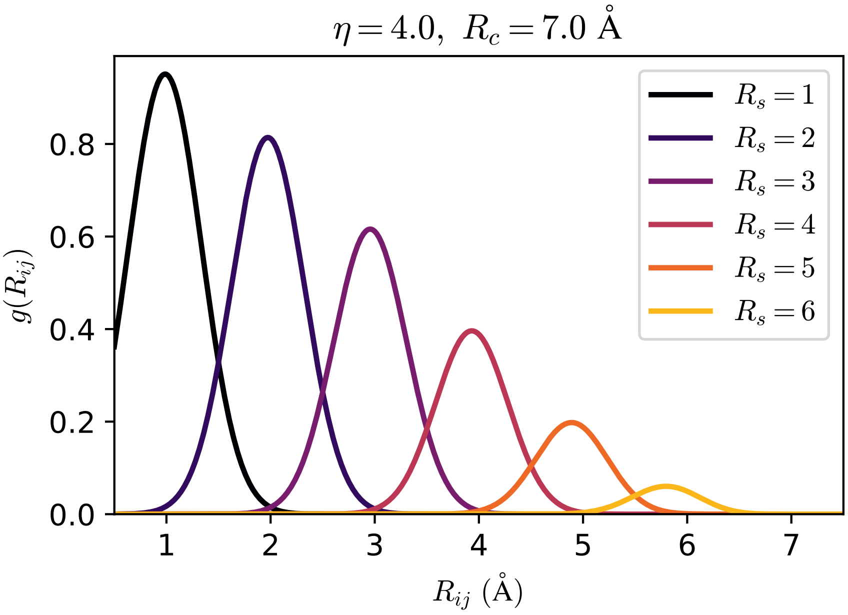


Fig. 2.14.: Plots of the summand used in the radial symmetry function descriptor as defined by Behler and Parrinello [53] for a series of R_s values.

summation of Gaussian functions over neighboring atom distances, scaled by cutoff function f_c [53]:

$$G_i(\eta, R_s) = \sum_{j \neq i} \underbrace{e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij})}_{g(R_{ij})} \tag{2.66}$$

$$f_c(R_{ij}) = \begin{cases} 0.5 \left[\cos\left(\frac{\pi R_{ij}}{R_c}\right) + 1\right], & R_{ij} \leq R_c \\ 0, & R_{ij} > R_c \end{cases} \tag{2.67}$$

where R_{ij} is the neighbor distance, R_c is the cutoff distance, and η and R_s are parameters of the Gaussian function. The total descriptor for each atom i can contain multiple symmetry functions G_i with different η and R_s parameters. In Fig. 2.14, the summands $g(R_{ij})$ of Eq. 2.66 are plotted for a series of R_s values. A similar descriptor is also defined for three body angles the atom i is involved in.

A key feature of the Behler-Parrinello approach is the sharing of weights between submodels. Therefore, the energy prediction is inherently invariant to index exchanges. Furthermore, the extensible modular approach allows the modeling of a variable number of atoms while keeping the number of parameters fixed. The original work shows an example of a specific model for homogeneous systems containing only one element. Later studies, such as the ANI-1 model, further improved the descriptors and demonstrated the development of transferable models for a small set of chemical elements [54]. Independent submodels are then introduced for each element, and the symmetry functions are calculated separately for each neighboring species.

3 Application of molecular dynamics for organic semiconductors

In this chapter, an application of molecular dynamics (MD) is presented in the context of a multiscale workflow to determine the static and dynamic conformational disorder of amorphous organic semiconductors. First, in Section 3.1, an introduction to the field of organic semiconductors is given, followed by the description of the workflow in Section 3.2.

This chapter is based on Reiser et al. [110]. My main contribution is presented in Section 3.2.1. Further results of the study are briefly summarized in Section 3.2.2.

3.1 Organic semiconductors

Since the invention of organic light-emitting diodes (OLEDs) [111, 112], organic semiconductors were continuously developed and put forth further promising applications, such as organic photovoltaics (OPVs) [113] and organic field-effect transistors (OFETs) [114], and already have made their way into many commercial applications including displays for consumer electronics and solar cells for photovoltaic systems. Devices typically consist of multiple stacked thin layers of different amorphous or semi-crystalline organic materials whose unique properties create new design possibilities such as flexible and semi-transparent devices [115]. However, one main disadvantage of organic semiconductors compared to inorganic alternatives is their poor charge carrier mobility.

The charge transport in these materials is often described by charge carriers which are localized on molecules and propagate through the system via ther-

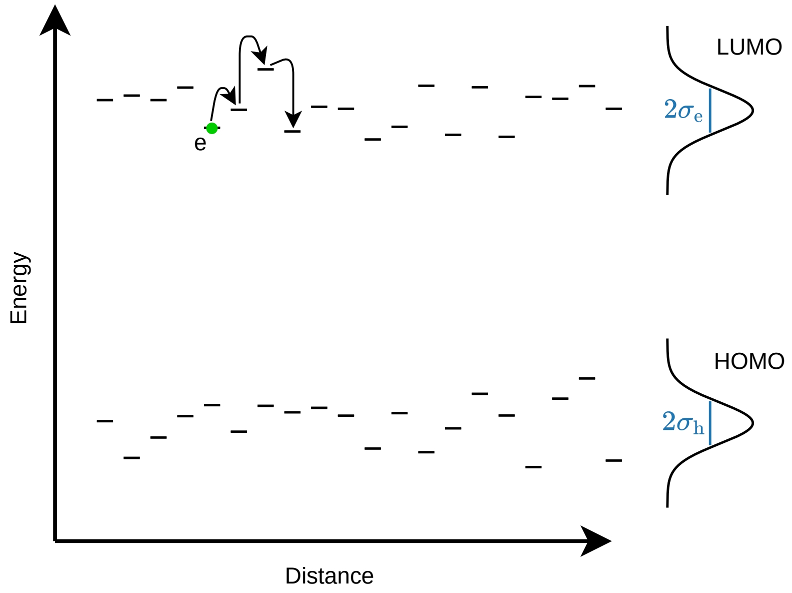


Fig. 3.1.: Simplified illustration of hopping transport in amorphous organic semiconductors. Each short horizontal line represents a localized energy level on a molecule. An electron (green) moves through the system by subsequent hopping to nearby molecules. Some processes involve an energy barrier and are thermally activated. The widths σ_{e} and σ_{h} of the distributions of the lowest unoccupied molecular orbitals (LUMOs), relevant for electron transport, and the highest occupied molecular orbitals (HOMOs), relevant for hole transport, are related to the average barrier heights and therefore also to the mobilities. In reality, the charge carriers take a complex three-dimensional path through the morphology and are often driven by an external electric field.

mally assisted hopping processes, as illustrated in Fig. 3.1 [116]. The hopping rates, which influence the mobility, depend on several quantities such as the coupling between molecules, the reorganization energy, and the molecular energy levels [117, 118]. Bässler et al. could show that for typical OLED materials, the energy distribution can be approximated by a Gaussian distribution and that the charge carrier mobility μ is related to the material-specific width of this distribution, the so-called energy disorder σ, via [119–121]

$$\mu \propto \exp\left[-C\left(\frac{\sigma}{k_{\mathrm{B}}T}\right)^2\right] \tag{3.1}$$

where k_{B} is the Boltzmann constant, T the temperature, and C a fitting constant. Larger disorders correspond to more and deeper tail states which act as charge carrier traps and lower the mobility. In order to gain more insight about the charge transport in organic materials, several multiscale modeling approaches have been developed [5, 81, 122–128]. They apply molecular dynamics (MD) or Monte Carlo simulations to create a bulk structure of the material, which is used to calculate hopping rates with density functional theory (DFT) and to define the hopping network, the spatial arrangement of molecular energy levels [122, 129, 130]. The results can then be used in kinetic Monte Carlo simulations [131], master equation approaches [132–134], or mean field models [135] to obtain a prediction for the mobility of the material. Due to the high computational cost of the DFT calculations, these approaches consider only individual snapshots of the molecular morphology and assume a static spatial arrangement of time-independent molecular energy levels [122].

In general however, the total disorder consists of static and dynamic contributions [136–139]. In amorphous materials, the static part results from the variation in the conformations and environments of the individual molecules, which causes a broadening of the distribution of time-averaged energies [136]. The dynamic part describes the energy fluctuations due to the thermal vibration and movement of the molecule in its confining environment [137–139].

3.2 Modeling of dynamic and static disorder

This work is based on a recent study by Friederich et al., where the application of machine learning models to determine static and dynamic disorder contributions was demonstrated for the material PEDOT:PSS [140]. In this section, a comprehensive study is presented, including a wide selection of materials relevant to OLED and OPV research. Furthermore, the workflow is limited to the calculation of the conformational disorder, and the calculation of molecular energy levels does not take into account the environment of neighboring molecules and charge carriers. Therefore, the electrostatic disorder is neglected in this study. However, it also has a relevant influence on the mobility and is investigated in other studies [5, 122].

The workflow consists of several steps which use different computational methods. First, in MD simulations, trajectories of bulk morphologies are generated for each considered molecule. For a small part of the monomer structures from the trajectories, the molecular energy levels are obtained from single-point DFT calculations in vacuum. The results represent the dataset used to perform machine learning (ML) of a regression model for the inexpensive prediction of molecular energy levels. The ML model is then applied to calculate the energy levels of the remaining structures from the trajectories, which allows the subsequent analysis of the static and dynamic disorder of the material.

The geometries of the materials considered in this study are shown in Fig. 3.2 and 3.3. They result from a minimum conformer search using the CREST module of the XTB program package [141–143]. A distinction is made between n-type materials, suitable for electron transport, p-type materials, suitable for hole transport, and materials suitable for both charge carriers.

3.2.1 Molecular dynamics of organic semiconductors

The bulk morphology simulations are conducted with the MD package LAMMPS [31] using the GROMOS force field [38]. For all simulations, a timestep of 1 fs and a short-range cutoff of 14 Å is applied. Furthermore, for long-range electrostatics, the particle-particle particle-mesh (PPPM) solver and for thermostatting and barostatting, the Nosé-Hoover approach is used, both as implemented in LAMMPS.

Initially, the optimized geometries are submitted to the web-accessible tool ATB [144, 145], which provides force field parameters and partial charges. For the initialization of the simulation, 512 molecules are placed on a cubic grid, each randomly oriented and with enough space between them to avoid overlaps (except for mCP, where 1000 molecules are simulated). To bring the

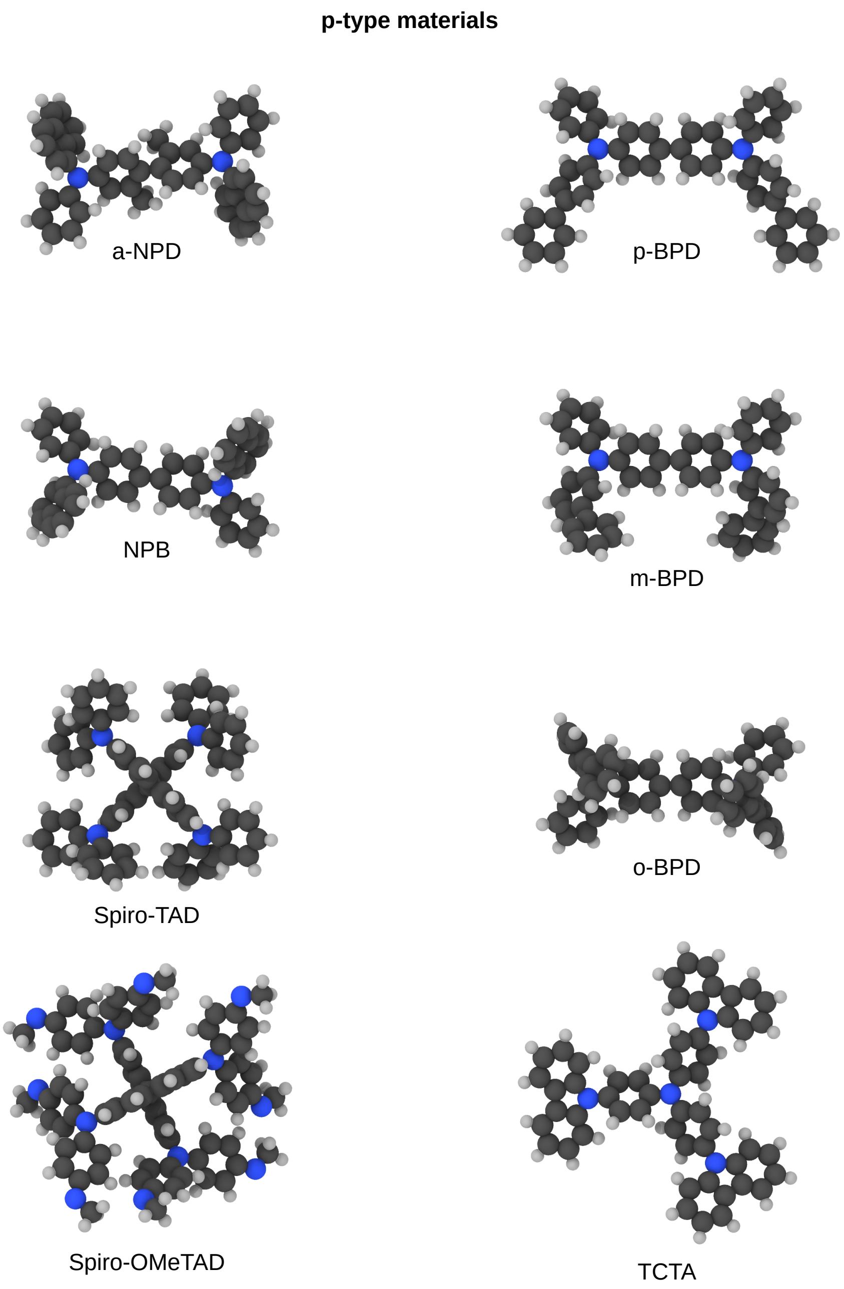


Fig. 3.2.: Rendered representations of the p-type semiconductors considered in this study. Light gray: hydrogen, dark gray: carbon, blue: nitrogen. Visualization software: OVITO [93].

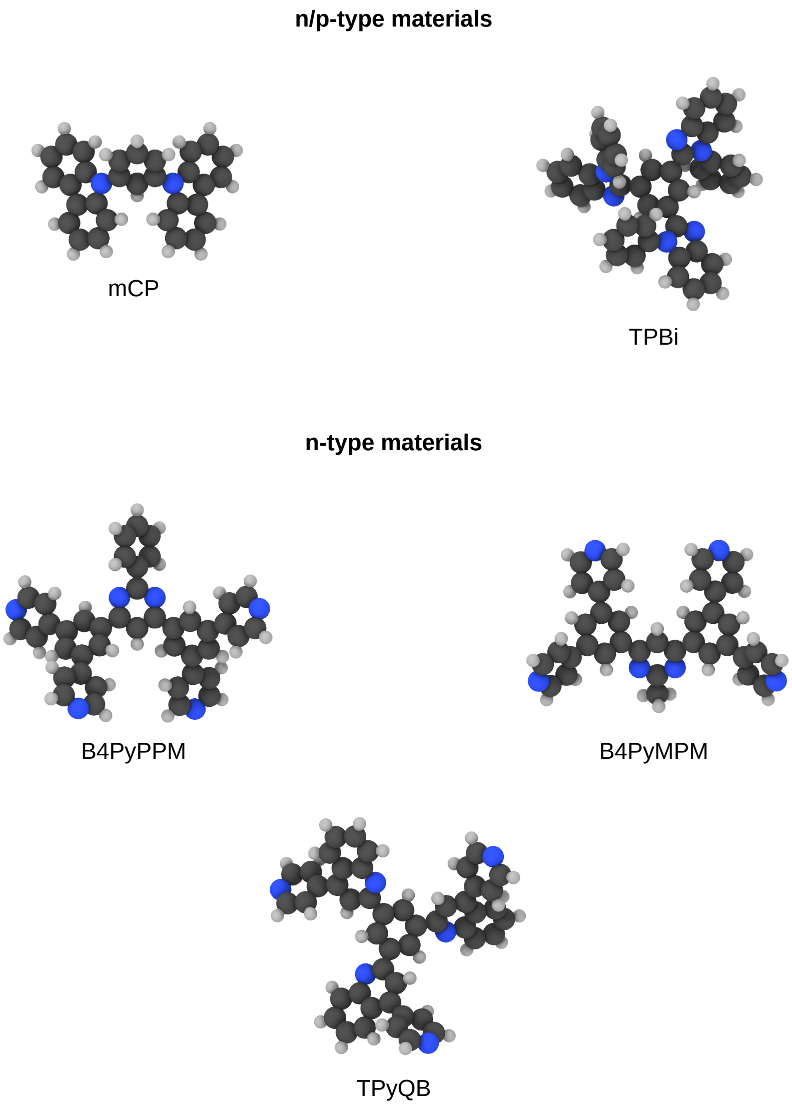


Fig. 3.3.: Rendered representations of the n-type and multi-purpose semiconductors considered in this study. Light gray: hydrogen, dark gray: carbon, blue: nitrogen. Visualization software: OVITO [93].

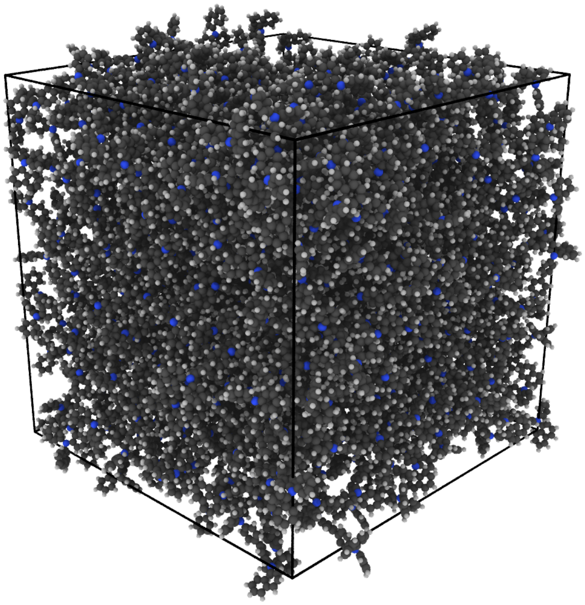


Fig. 3.4.: Rendered representation of the NPB bulk morphology resulting from the MD protocol presented in Section 3.2.1. Light gray: hydrogen, dark gray: carbon, blue: nitrogen. Visualization software: OVITO [93].

system towards its equilibrium mass density, several subsequent simulation intervals are performed. First, at a temperature of 300 K, the initial sparse structure is compressed to a dense amorphous morphology through a pressure ramp from 10 to 1 atm of the duration of 200 ps. After that, the long-range solver is switched on, and the simulation is continued for 200 ps at a pressure of 1 atm. To further improve the convergence of the density, a heating process lasting a total of 400 ps is simulated composed of a linear temperature ramp from 300 to 700 K followed by a cooling ramp of equal duration back to 300 K. The preparation of the morphology is completed by a final equilibration run of 500 ps at the target conditions of 300 K and 1 atm. From the resulting structure, a production simulation of 5 ns is performed during which the system coordinates are saved in an interval of 1 ps. For one of the materials, NPB, the rendered representation of the final morphology is shown in Fig. 3.4.

3.2.2 Results and discussion of disorder analysis

The calculations and analysis presented in the following were performed by Patrick Reiser [110].

A subset of the resulting trajectories of each molecule is used for the training of an ML model for the prediction of DFT energy levels. Details on the development of the ML approach are given in the original publication [110]. The trained ML model is used to evaluate the energy levels $U_i(t)$ of all monomers and snapshots. With that, it is possible to calculate the total conformational disorder of the materials and obtain separate values for the dynamic and static contribution, as illustrated in Fig. 3.5 and outlined in the following. The conformational disorder is one part of the total disorder next to electrostatic contributions, which are not treated in this study. The dynamic conformational disorder σ_{dyn} is defined by the ensemble average $\langle\cdot\rangle_i$ over the variances of fluctuating energies with time $\mathrm{Var}_t[U_i(t)]$ of the individual molecules:

$$\sigma_{\mathrm{dyn}} = \sqrt{\langle \mathrm{Var}_t[U_i(t)]\rangle_i} \tag{3.2}$$

The static contribution σ_{stat} is given by the variance of the per-molecule time-averages $\langle\cdot\rangle_t$ of energies:

$$\sigma_{\mathrm{stat}} = \sqrt{\mathrm{Var}_i[\langle U_i(t)\rangle_t]} \tag{3.3}$$

The total conformational disorder σ_{tot} is primarily computed by the variance of energies over all molecules and times:

$$\sigma_{\mathrm{tot}} = \sqrt{\mathrm{Var}_{i,t}[U_i(t)]} \tag{3.4}$$

If the dynamic and static distributions (Fig. 3.5 d) and e)) can be approximately described as independent normal distributions, an alternative expression for the total conformational disorder results:

$$\sigma_{\mathrm{tot}} = \sqrt{\sigma_{\mathrm{dyn}}^2 + \sigma_{\mathrm{stat}}^2} \tag{3.5}$$

Furthermore, as a measure for the deviation from the normal distribution, the population skewness γ_1 quantifies the asymmetry of the disorder distributions shown in Fig. 3.5 d-f):

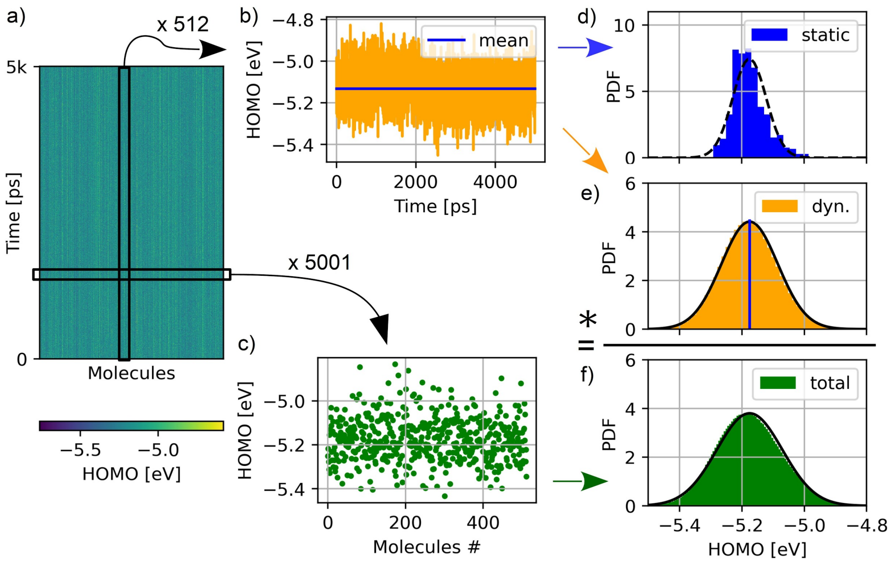


Fig. 3.5.: Breakdown of the conformational disorder into static and dynamic contributions shown for the example of the p-type material NPB. a) Heatmap of HOMO energies, each column represents one molecule of the morphology, each row a snapshot of the trajectory. b) Energy distribution during the trajectory of one specific molecule. c) Energy distribution of all molecules during one specific snapshot. d) Static disorder distribution of mean energy values of the molecular trajectories and the corresponding probability density function (PDF). e) Dynamic disorder distribution of energy fluctuations around the mean energy of the molecular trajectory, plotted for all trajectories. f) Total conformational disorder distribution of energies from all molecular trajectories and snapshots. This figure was created by Patrick Reiser. Reproduced with permission from [110]. Further permission requests related to the material excerpted should be directed to the American Chemical Society.

$$\gamma_1 = \frac{1}{N}\sum_{i}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3 \tag{3.6}$$

where N is the number of samples x_i, μ the mean and σ the standard deviation corresponding to the disorder. The results of the disorder components and skewness for the p-type and n-type materials are shown in Fig. 3.6, the corresponding numerical values are given in Tab. A.1 and A.2 of Appendix A.1. First, it can be seen that the total conformational disorder values are in good accordance with the quadratic summation of the static and dynamic disorder (Tab. A.1 and A.2). This agreement confirms the assumption of independent static and dynamic distributions, which allow a consistent separation. Therefore, it is possible to investigate the influence of molecular properties separately for both disorder contributions.

While the dynamic disorder makes up a significant part of the total disorder in all molecules, the ratio between the dynamic and static contribution shows considerable variations. Molecules with large rotatable side groups, such as TCTA, show a trend to larger static disorder values. Depending on the conformation, the HOMO can either be delocalized over the molecule or localized on the side groups (see Fig. 3.7). This trend is also observed for the BPD molecules. Here, the smallest static disorder is observed for the compact o-BPD with restricted rotational degrees of freedom. Another feature brought out by the separation of contributions is the slight asymmetry of the static distribution for some of the molecules. Depending on the sign of the skewness, this can be linked to an increased or decreased density of tail states in the energy gap. For example, the positively skewed HOMO distribution in the case of NPB corresponds to an increase in hole traps compared to a symmetric normal distribution.

In conclusion, the workflow provides a consistent breakdown of the conformational disorder into dynamic and static contributions and improves the insight into the charge transport conditions in organic semiconductors. Furthermore, the resulting time-resolved energy level fluctuations potentially allow a comparison with the time scales of hopping processes and could serve as a starting point for developing advanced hopping transport simulation methods with dynamic rates.

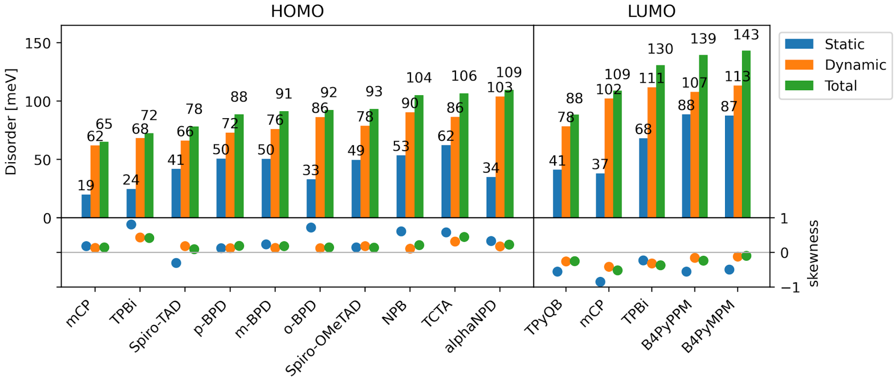


Fig. 3.6.: Results of the total conformational disorder and breakdown into static and dynamic contributions according to the method shown in Fig. 3.5. For the n-type materials, the disorder corresponds to the LUMO and for the p-type materials to the HOMO energies. The analysis involves 38.3 million energies obtained from ML model predictions on the MD geometries. The numerical results of all HOMO and LUMO disorder values are given in Tab. A.1 and A.2 of Appendix A.1. Additionally, the skewness is plotted for the static, dynamic, and total distributions. It reveals a tendency for asymmetric static distributions with a significant deviation from a normal distribution. This figure was created by Patrick Reiser. Reproduced with permission from [110]. Further permission requests related to the material excerpted should be directed to the American Chemical Society.

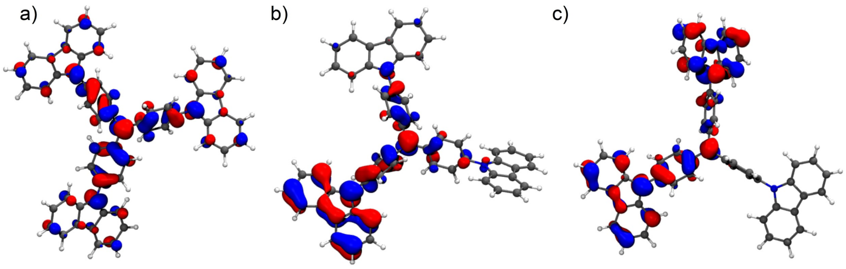


Fig. 3.7.: Representation of the HOMO orbital of TCTA computed in vacuum using DFT for different conformations: a) optimized geometry, b) and c) non-equilibrium conformers from the amorphous morphology. Compared to the delocalized orbital of the optimized structure in a), the twisted geometries in b) and c) show a localization on one or two sidegroups. This figure was created by Patrick Reiser. Reproduced with permission from the Supporting Information of [110]. Further permission requests related to the material excerpted should be directed to the American Chemical Society.

4 Multiscale modeling of separable non-covalent interactions

This chapter is about the development and application of the Component-separable Non-covalent Interaction Network (CONI-Net), an analytical model for the description of non-covalent interactions, which is trained on ab initio energies and applicable in molecular dynamics simulations.

In the first section, a detailed description of the method and all its components is given. In Section 4.2, the data efficiency of the model is examined by the computation of a learning curve. Subsequently, an application as a custom model for a set of organic molecules is shown in Section 4.3. Finally, in Section 4.4, the development of a transferable model for hydrocarbons is demonstrated.

This chapter is based on Konrad and Wenzel [146].

4.1 Method development

The method developed in this work can be described by a workflow as shown in Fig. 4.1, consisting of several steps discussed in the following subsections. In Section 4.1.1, several preprocessing steps for the involved molecules are described, such as geometry optimization of the monomers and the calculation of atomic descriptors and partial charges. The model training is performed on dimer samples. Therefore, in Section 4.1.2, a procedure for the automated generation of large dimer datasets is introduced. Section 4.1.3 gives an overview of the CONI-Net model and the training procedure. Finally, in Section 4.1.4, the protocol for the model application in molecular dynamics simulations for predicting thermodynamic properties is presented.

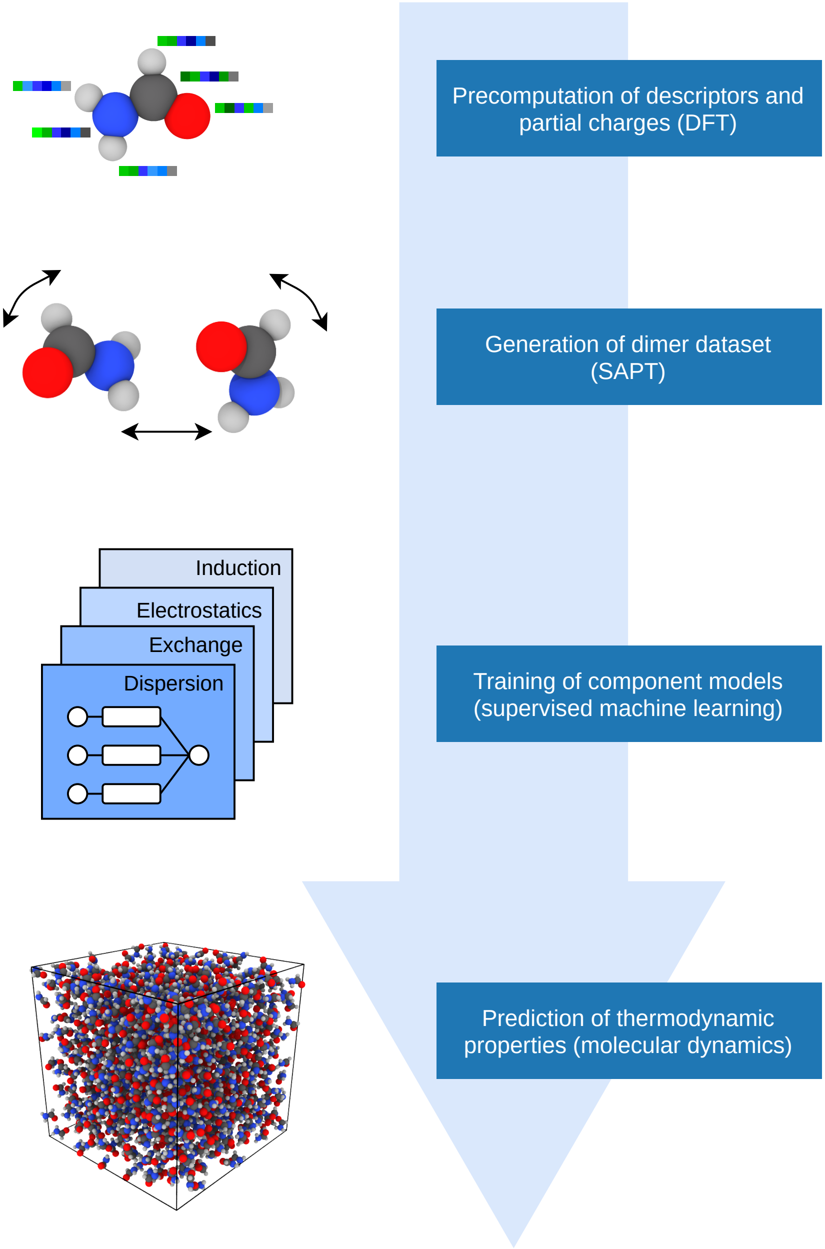

Precomputation of descriptors and partial charges (DFT)
Generation of dimer dataset (SAPT)
Induction
Electrostatics
Exchange
Dispersion
Training of component models (supervised machine learning)
Prediction of thermodynamic properties (molecular dynamics)

Fig. 4.1.: Overview of the workflow described in this section. Visualization software for the molecular renderings: OVITO [93].

4.1.1 Fingerprint descriptor and partial charges

For the model developed in this study, the representation of the system is divided into atomic pairwise descriptors, which characterize intermolecular pairs of atoms. As a first step, for each atom in the system, several quantities are precomputed in QM calculations. On the one hand, properties that describe the local chemical environment of the atom inside the molecule and, on the other hand, partial charges that approximate the long-range electrostatic potential.

Initially, a geometry optimization is conducted for each monomer with the DFT module of the ORCA package [147]. For this task, the B3LYP functional [86–88] is used in combination with the aug-cc-pVTZ basis set [148]. From the results, partial charges are extracted in a one-stage RESP fit [36] using the postprocessing tool Multiwfn [149]. On the relaxed geometry, another single-point calculation with the cc-pVTZ basis set is performed [150], followed by a calculation of Hirshfeld charges [94] and Mayer bond orders [95] using the postprocessing tools of ORCA.

With the calculated properties, an atomic fingerprint is defined by concatenating the Hirshfeld charge and the four highest bond order values. Thereby, each pair of atoms can be described by a ten-digit pair fingerprint constructed from the two atomic fingerprints (Fig. 4.2). The order of the two atomic fingerprints is defined by rules which ensure the invariance of the pair fingerprints to index exchanges in the Cartesian description. For pairs of different chemical elements, the order is determined by a fixed rule. Namely, the atomic fingerprint of the lighter chemical element is put in the first place. For pairs of identical elements, the order is deducted from comparing the entries of the two atomic fingerprints one by one. The first unequal entries determine the order such that the atomic fingerprint corresponding to the higher entry value is placed at the front. In addition to the pair fingerprint, for each pair, the distance and RESP charges are stored.

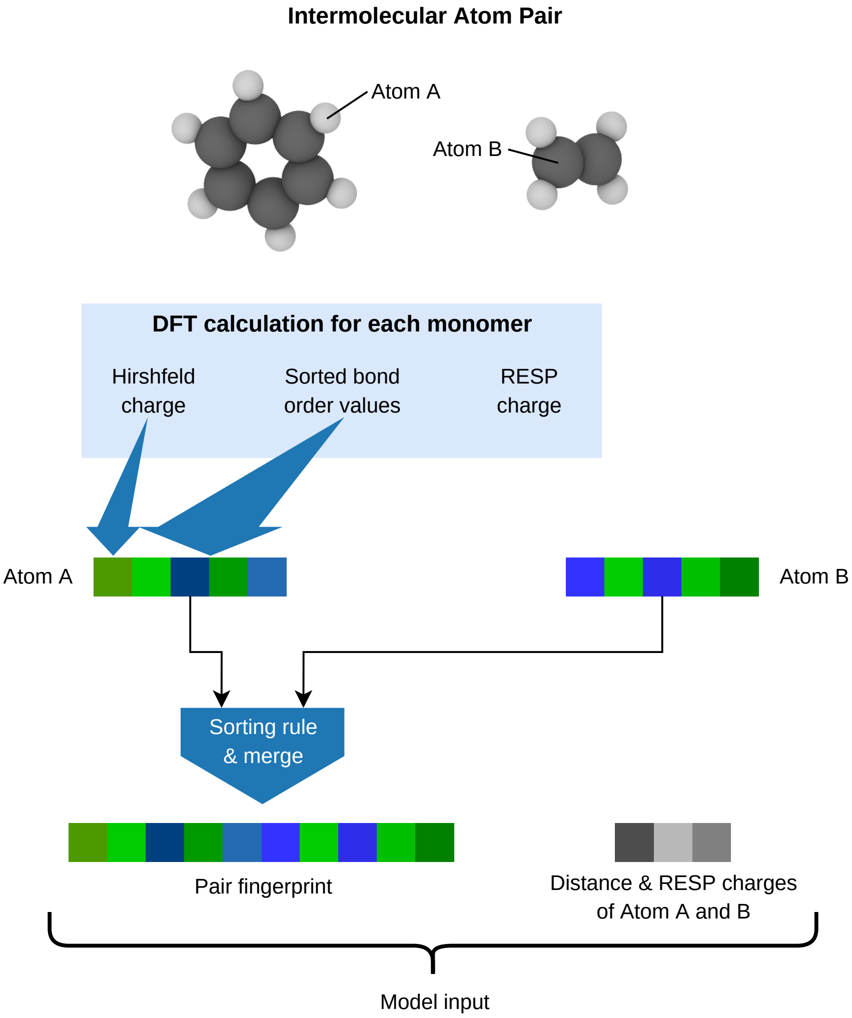


Fig. 4.2.: Definition of the pair fingerprint descriptor of an intermolecular atom pair. From DFT calculations of each monomer, the Hirshfeld charge and the four highest Mayer bond order values are obtained from postprocessing and combined to an atomic fingerprint. The pair fingerprint is constructed by the concatenation of both atomic fingerprints according to a sorting rule which ensures invariance towards index permutation. Furthermore, the RESP charges for the electrostatic baseline model and the pair distance are stored alongside the pair fingerprint. Visualization software for the molecular renderings: OVITO [93].

4.1.2 Dimer sampling method

The training of the CONI-Net model requires a dataset of dimer samples to learn the relationship between geometric arrangement and interaction energy. The reference energy for a dimer geometry is obtained from symmetry-adapted perturbation theory (SAPT), which provides a physically motivated grouping into the interaction components dispersion, exchange, electrostatics, and induction [7, 18]:

$$E_{\text{dimer}} = E_{\text{disp}} + E_{\text{exch}} + E_{\text{el}} + E_{\text{ind}} \tag{4.1}$$

For the dataset in the subsequent applications, the calculations are performed on SAPT2+3 level by the Psi4 quantum chemistry package using the aug-cc-pVDZ basis set and the density fitting approximation [10, 148, 151–153]. For the molecules in the subsequent applications, this combination yields an affordable computational expense that allows the generation of several thousand samples. For that purpose, an automatic procedure was developed to construct dimer arrangements from the monomer geometries obtained during the preparation step described in the previous subsection. In order to steer the sampling into a relevant distance region, an estimate for the dimer interaction energy E_{estimate} is defined by Lennard-Jones interactions based on GAFF parameters which are assigned automatically by the AmberTools software [28, 101] and Coulomb interactions of the precomputed RESP charges. This estimate enables the following sampling procedure:

1. Independent random rotations of both monomers

2. Identification of the dimer distance d_{rep} in the repulsive regime where $E_{\text{estimate}} = 1\ \text{kcal/mol}$

3. Random choice of a dimer distance in the range of d_{rep} to $d_{\text{rep}} + 5$ Å, with a selection probability proportional to $\exp(-\beta E_{\text{estimate}})$

The sampling parameter β can be related to a sampling temperature T_{s} by $\beta = \frac{1}{k_{\text{B}} T_{\text{s}}}$, where k_{B} is the Boltzmann constant. In the subsequent applications, T_{s} is set to room temperature which results in $\beta = \frac{1.69}{\text{kcal/mol}}$.

4.1.3 Network model

The CONI-Net model is comprised of four independent models, one model for each energy component provided by SAPT: dispersion, exchange, electrostatics, and induction. Each component model has the structure as shown in Fig. 4.3, which is described in the following. The overall network structure is based on the Behler-Parrinello approach, where the total network is divided into submodels that compute partial contributions of the total quantity [53]. In the present model, the submodels emerge from an atomic pairwise partitioning of the total interaction component [55, 57]. They are represented by pair networks which are composed of the following modules:

- Fully-connected artificial networks (ANNs): Calculation of exponents and prefactors from pair fingerprint
- Function layer: Calculation of pair contribution using several power laws constructed from ANN outputs
- Baseline model (only for the electrostatic component): Partial charge interaction according to Coulomb's law

The fully-connected ANNs for the prefactors and exponents contain two hidden layers, each with four neurons using ReLU activation functions, followed by an output layer with three neurons. The neurons, also referred to as nodes, contain the parameters of the model, the weights and biases. The execution of the total model involves one pair network instance per intermolecular pair. Instances for pairs of the same combination of chemical elements use the same parameter set, i.e., they share weights and biases. Furthermore, individual transformations are performed on the output values of the exponent and prefactor networks before they are passed on to the function layer. The direct output k_i^{out} of the exponent network is mapped to a finite value range around the bias value k_i^{bias} whereby the exponent k_i resulting from output node i is given by

$$k_i = k_i^{\text{bias}} + 2.0 \cdot (\text{sig}(k_i^{\text{out}}) - 0.5) \tag{4.2}$$

where $\text{sig}(x)$ is the simoid function which is defined as

$$\text{sig}(x) = \frac{1}{1 + e^{-x}} \tag{4.3}$$

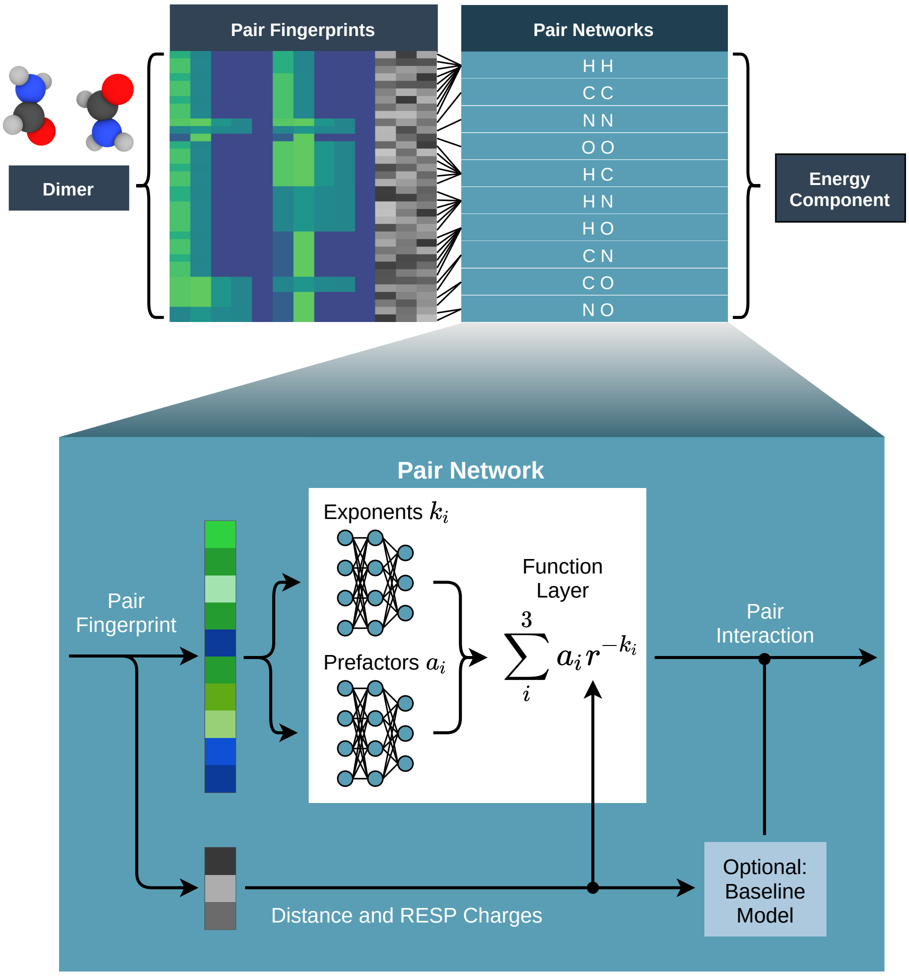


Fig. 4.3.: Structure of the CONI-Net model for any energy component. The pair fingerprint, distance, and RESP charges of each pair are the input of a pair network. Parameter sharing is applied for pair network instances that belong to the same combination of chemical elements. The pair network comprises two artificial networks for the computation of prefactors and exponents and a function layer that uses both values and the distance to evaluate the pair contribution to the energy component. Additionally, the electrostatic component includes a baseline model based on the Coulomb interaction between the RESP partial charges. Visualization software for the molecular renderings: OVITO [93]. Reproduced with permission from [146]. Copyright 2021 American Chemical Society.

The values for k_i^{bias} are hyperparameters of the model. Throughout this study, they are defined for the exchange component as

$$k_1^{\text{bias}} = 8,\ k_2^{\text{bias}} = 10 \text{ and } k_3^{\text{bias}} = 12, \tag{4.4}$$

and for the dispersion, electrostatics, and induction components as

$$k_1^{\text{bias}} = 6,\ k_2^{\text{bias}} = 8 \text{ and } k_3^{\text{bias}} = 10. \tag{4.5}$$

The output values of the prefactor network are also modified before passing on. On the one hand, the signs are replaced by a predefined value depending on the energy component. For the exchange component, the prefactors are constrained to positive values, and for the dispersion, electrostatics, and induction components, to negative values. On the other hand, the output values are scaled by a factor of $1000\ \text{kcal}/\text{mol}$, which defines the energy units and is comparable to a change of parameter initializations and learning rates locally inside the prefactor network. Subsequently, the resulting exponents k_i, prefactors a_i, and pair distance r are inserted into the function layer to calculate the pair contribution to the energy component:

$$E_{\text{p}}(r) = \sum_i^3 a_i r^{-k_i} \tag{4.6}$$

In order to prevent the pair contribution from diverging to negative infinite values, a taper function is applied to attractive pair contributions in the region below r_{min}, the shortest distance in the training data for the considered element combination. This ensures stable MD simulations, even if the pair distance falls below the known range in case of a rare close encounter. The taper factor f is implemented as proposed in Ref. [154]:

$$r_0 = r_{\text{min}} - 0.5 \tag{4.7}$$

$$x(r) = \frac{r - r_0}{r_{\text{min}} - r_0} \tag{4.8}$$

$$f(x) = (1-x)^3(1+3x+6x^2) \tag{4.9}$$

If the distance falls below r_{min}, the pair contribution E_{p} is scaled down smoothly. The resulting final pair interaction $\widetilde{E}_{\mathrm{p}}(r)$ is given by

$$\widetilde{E}_{\mathrm{p}}(r) = \begin{cases} 0, \ r \leq r_0 \\ (1-f)E_{\mathrm{p}}(r), \ r_0 < r < r_{\mathrm{min}} \\ E(r), \ r \geq r_{\mathrm{min}} \end{cases} \tag{4.10}$$

Finally, for the electrostatic component, a baseline energy is added to the pair contribution based on the partial charge interaction:

$$E_{\mathrm{baseline}}(r) = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r} \tag{4.11}$$

where ε_0 is the vacuum permittivity and q_1 and q_2 are the partial charges of the two pair atoms. The implementation of the model was realized with the Python library PyTorch [155]. The weights and biases in the hidden and output layers of the exponent and prefactor networks are initialized from the uniform distribution

$$\mathcal{U}\left(-\sqrt{n_{\mathrm{f}}^{-1}}, \sqrt{n_{\mathrm{f}}^{-1}}\right) \tag{4.12}$$

where n_{f} is the total number of input features of the corresponding layer. The model training is performed with the Adam optimizer using a learning rate α and other settings as proposed in the original study [107] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, $\alpha = 0.001$) and the mean squared error as loss function. In order to avoid overfitting, an early stopping protocol is applied. Therefore, the total dataset is randomly divided into a training and a validation set. The training set is used for model optimization, while at checkpoints every 100 epochs, the model loss on the validation set is checked. If the validation loss does not improve for 800 steps, the training is stopped. The split ratio of training to validation samples is 3:1 for all following applications. Before the training, a normalization is performed for the entries of the pair fingerprints. Furthermore, random noise from a normal distribution with a standard deviation of 0.1 is added to the entries during training. When the early stopping criterion is reached, the noise is turned off, and the training is continued for another 100 epochs from the checkpoint with the lowest validation loss.

4.1.4 Molecular dynamics protocol

In some of the following applications, the trained CONI-Net model is applied in molecular dynamics (MD) simulations of the liquid phase to predict the thermodynamic properties enthalpy of vaporization and mass density, which allows the comparison with experimental values and predictions from other force fields. Below, a description of the setup, procedure, and analysis of the simulations is given. The simulations are conducted with the LAMMPS molecular dynamics package [31] and initially require the preparation of the force field. Since the CONI-Net model is restricted to non-covalent interactions, the intramolecular energies and forces are modeled by the GROMOS force field [38]. The parameters for the bond, angle, and dihedral terms are generated via the web-accessible Automated Topology Builder (ATB) [144, 145]. In order to obtain the intermolecular part of the force field, first, the energy curves are extracted from the CONI-Net models of the four components. For each pair fingerprint occurring in the simulation, the energy components are evaluated at 500 discrete distances ranging from 0.5 to 15 Å and summed to a total energy curve. The force curve results from the numerical derivative of the total energy curve using the Python library NumPy [156] and the fundamental relation:

$$F_{\mathrm{p}}(r) = -\frac{\partial}{\partial r} E_{\mathrm{p}}(r) \tag{4.13}$$

Finally, the energy and force curves yield a tabulated force field directly usable with the LAMMPS package. Besides the tabulated interactions, which are truncated with a cutoff of 15 Å, the electrostatic baseline model based on the precomputed RESP charges is added through the corresponding LAMMPS module, which allows treatment of long-range interactions by the Ewald summation method.

For the preparation of the MD simulation, a total of 1000 molecules are first individually randomly rotated and then placed on a spacious cubic lattice to prevent initial overlaps. From this initial structure, the system is steered towards its equilibrium density by several successive runs with Nosé-Hoover style thermostatting and barostatting as implemented in LAMMPS. In Tab. 4.1 an overview of the MD parameters for the three preparation runs and the production run of the liquid phase system is given. Furthermore, for the following analysis, a simulation of a single molecule in vacuum is performed

Tab. 4.1.: Overview of MD simulation settings during the preparation steps 1-3 and the production step 4, with the simulated time, the integration timestep, the pressure p, the pressure coupling parameter τ_p, the temperature coupling parameter τ_T, and the relative target error in forces of the Ewald summation method. The target temperature T of the thermostat is constant for all steps and set to the value used in the experiment. Adapted with permission from [146]. Copyright 2021 American Chemical Society.

step	1	2	3	4
time (ps)	200	200	400	2000
timestep (fs)	1	1	0.2	0.2
p (atm)	$100 \to 1$	1	1	1
τ_p (ps)	1	1	5	5
τ_T (ps)	0.1	0.1	1	1
kspace rel. err.	1e-4	1e-4	1e-5	1e-5

with a Langevin thermostat using a timestep of 0.2 fs and a coupling parameter τ_T of 1 ps. After an equilibration of 100 ns, a production run of equal duration is performed. The simulation settings in the production runs of the liquid and vacuum simulation are chosen similarly to those used in the benchmark study of Caleman et al. [106], which will be used for comparison in one of the subsequent applications.

For the calculation of the thermodynamic properties, the average potential energies $E_{\text{pot}}^{\text{vac}}$ and $E_{\text{pot}}^{\text{liq}}$ are obtained from the production runs of the vacuum and liquid simulation. From the latter, the average volume $\langle V \rangle$ is also extracted. The predictions for the mass density ρ and enthalpy of vaporization ΔH_{vap} of the liquid can then be calculated using expressions which are equivalent to those applied in Ref. [106]:

$$\rho = \frac{M}{\langle V \rangle} \tag{4.14}$$

$$\Delta H_{\text{vap}} = E_{\text{pot}}^{\text{vac}} - E_{\text{pot}}^{\text{liq}} + k_{\text{B}}T \tag{4.15}$$

with the Boltzmann constant k_{B}, the total mass M and the temperature T.

4.2 Learning curve

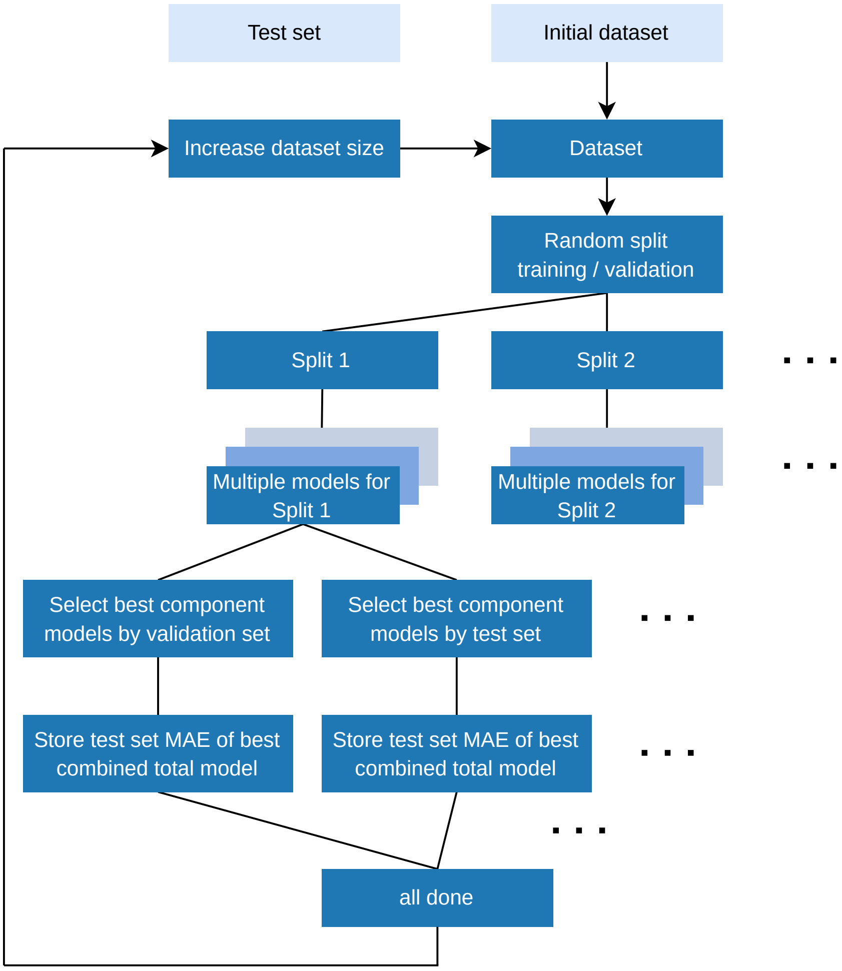
To evaluate the data efficiency of the model, in this section, the dependency of the dataset size on the model performance to interpolate between different dimer geometries of a single molecule is examined. The resulting relationship between performance and dataset size is also referred to as a learning curve. As a performance measure, the mean absolute error (MAE) on an independent test set is used. Furthermore, the model selected by the validation set from repeated training executions with random weight initializations is compared to the actual best model for the test set. The entire investigation is conducted for the organic molecule methanol.

4.2.1 Procedure

The following steps for calculating the learning curve are also outlined in Fig. 4.4. For different total dataset sizes, five unique random splits into training and validation set in the ratio 3:1 are generated. For each split, the model training is executed five times for each energy component. The best component models for each split are then selected by two alternative criteria. On the one hand, they are selected by the validation set MAE specific to the split. On the other hand, by the MAE on an independent test set containing 2000 dimer samples. For each split and selection criterion, the best component models are combined to a best total energy model, resulting in two sets of five total energy models for each dataset size.

4.2.2 Results and discussion

For both selection criteria, Fig. 4.5 shows the dependence of the mean and range of the test set MAEs for the different dataset splits on the training set size. From the best models for the test set (blue), it is clear to see that an increasing training set size results in a decreasing average MAE. On the one hand, this is due to the growing variety in the training set that comes with adding more dimer geometries. On the other hand, with increasing training set size, the number of validation samples increases proportionally and becomes more and



Test set
Initial dataset
Increase dataset size
Dataset
Random split training / validation
Split 1
Split 2
. . .
Multiple models for Split 1
Multiple models for Split 2
. . .
Select best component models by validation set
Select best component models by test set
. . .
Store test set MAE of best combined total model
Store test set MAE of best combined total model
. . .
. . .
all done

Fig. 4.4.: Procedure for the calculation of the learning curve. Five unique random splits are created for each dataset size, and for each split, five models per energy component are trained. For clarity, only a subset is shown in this figure.

more representative for many dimer arrangements. Since the validation set MAE controls the early stopping criterion of the model training, a growing number of validation samples can make the training more robust against over- and underfitting and less sensitive to changing the dataset split. This dependence is also reflected in the decreasing MAE ranges and convergence of the two mean MAE values of both selection criteria for large training and validation set sizes (Fig. 4.5).

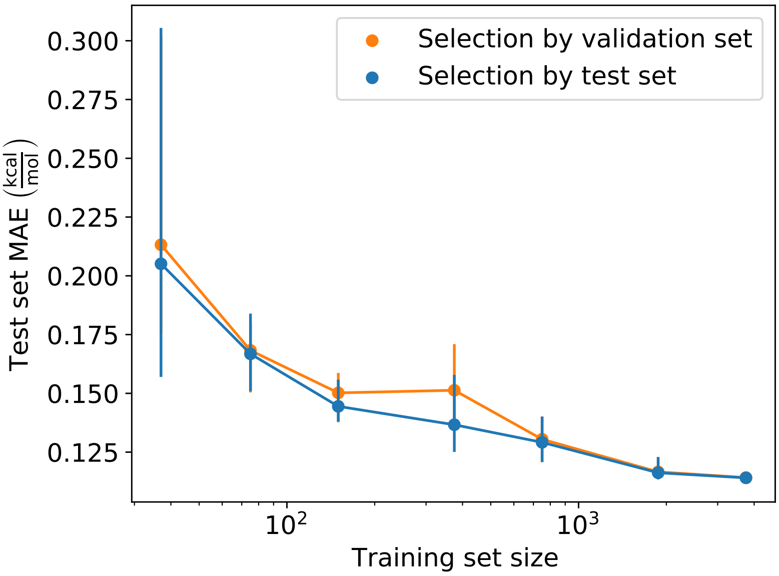


Fig. 4.5.: Learning curve (methanol dimers). The arithmetic mean and range of test set errors of the best combined total energy models are indicated by points and bars. The best models of each split and component are selected by the validation set (orange) and for comparison by the test set itself (blue). Reproduced with permission from [146]. Copyright 2021 American Chemical Society.

4.3 Custom force field

In this section, the developed workflow (Fig. 4.1) is applied for a set of organic molecules to test the interpolation capabilities of the CONI-Net model and fingerprint descriptor. Furthermore, the trained model is applied in molecular dynamics simulations, and the predicted values for thermodynamic observables are compared to values from experiments and conventional force fields.

4.3.1 Molecules and dataset

The molecules for this application consist of the chemical elements carbon, hydrogen, nitrogen, and oxygen. Due to the high computational cost of the SAPT method, they contain only between two and four non-hydrogen atoms each. Furthermore, for this application, all molecules involved in the model

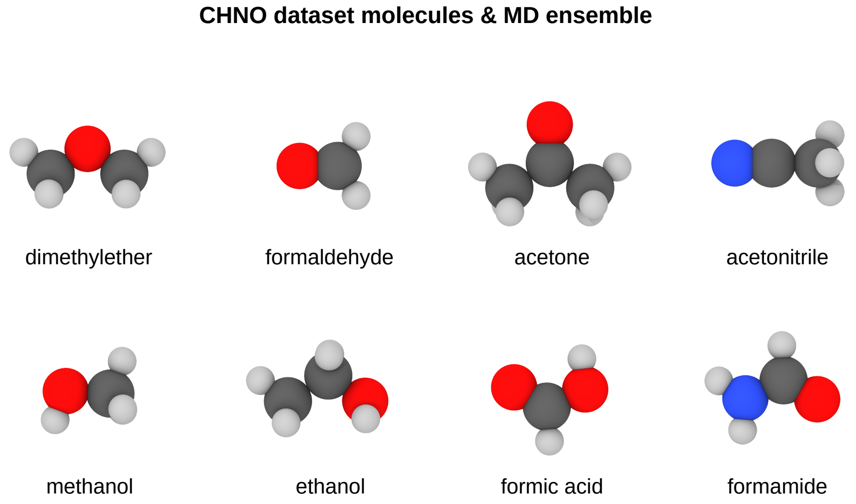


Fig. 4.6.: Renderings and names of the molecules in the CHNO dataset used for model training. They also represent the MD ensemble used for the property prediction simulations. Visualization software: OVITO [93].

training are also used in the MD simulations for the property predictions. Therefore, to enable a comparison, molecules were chosen for which literature values from experiments are available and which are included in the OPLS-AA and GAFF benchmark study of Caleman et al. [106]. The renderings and names of the molecules are shown in Fig. 4.6.

The dimer dataset for the model training, hereafter referred to as the CHNO dataset, consists of 2000 homodimers for each molecule and is divided into training and validation sets with the ratio 3:1, which totals in 12000 training and 4000 validation samples. Furthermore, an independent test set of 4000 samples with the same composition is prepared to evaluate the performance of the final model.

4.3.2 Results of model training and property predictions

For a fixed choice of the training and validation set, the model training procedure for each component (Sec. 4.1.3) is conducted in five independent executions with random weight initializations. Then, for each component, the model with the best MAE on the validation set is chosen for the final total energy model. The results for the components and the total energy of the final model on the test set are shown in Fig. 4.7, 4.8 and 4.9.

The final model is then applied in molecular dynamics simulations for all molecules in the dataset (Fig. 4.6). The predicted enthalpies of vaporization and mass densities are shown in Fig. 4.10 and Fig. 4.11. They are compared to experimental values and predictions of the OPLS-AA and GAFF force fields as published in Caleman et al. [106]. All numerical values related to Fig. 4.10 and Fig. 4.11 and references for the experimental values can be found in Tab. A.3 of Appendix A.1.

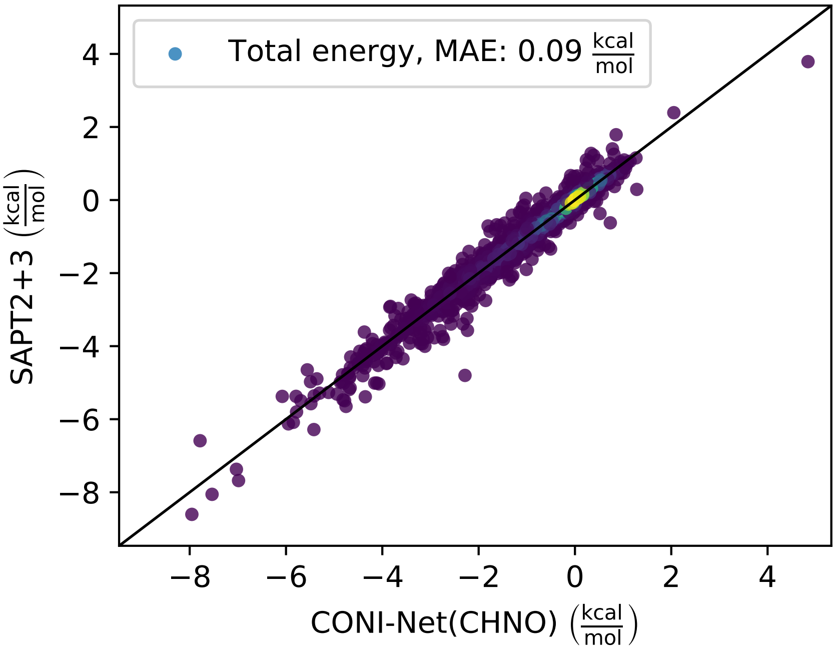


Fig. 4.7.: Total non-covalent interaction energies of the CHNO test set samples predicted by the CONI-Net model vs. SAPT2+3 reference. The color coding corresponds to the local point density calculated via a Gaussian kernel-density estimate as implemented in SciPy [157], violet represents low and yellow high values. Adapted with permission from [146]. Copyright 2021 American Chemical Society.

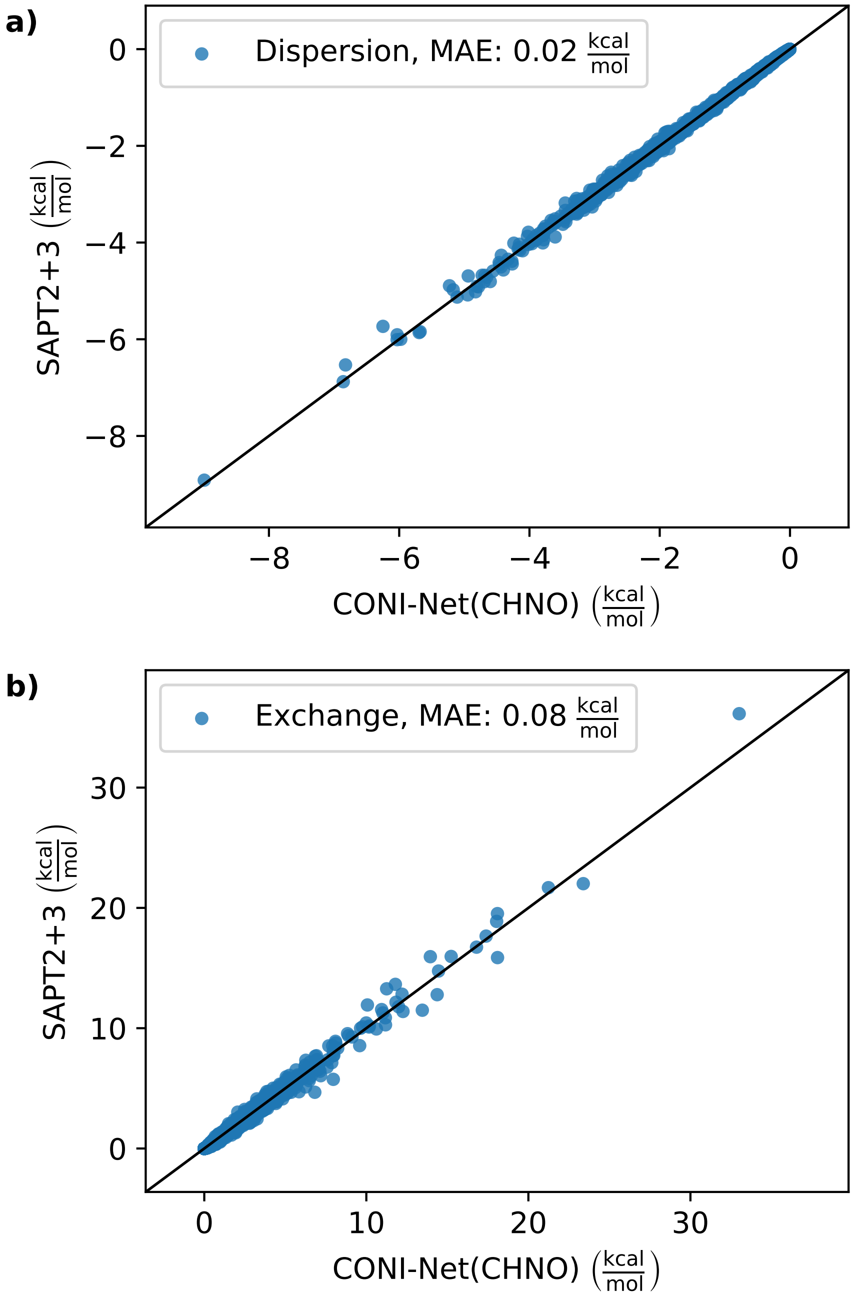


Fig. 4.8.: a) Dispersion and b) exchange energies of the CHNO test set samples predicted by the CONI-Net model vs. SAPT2+3 reference. Adapted with permission from [146]. Copyright 2021 American Chemical Society.

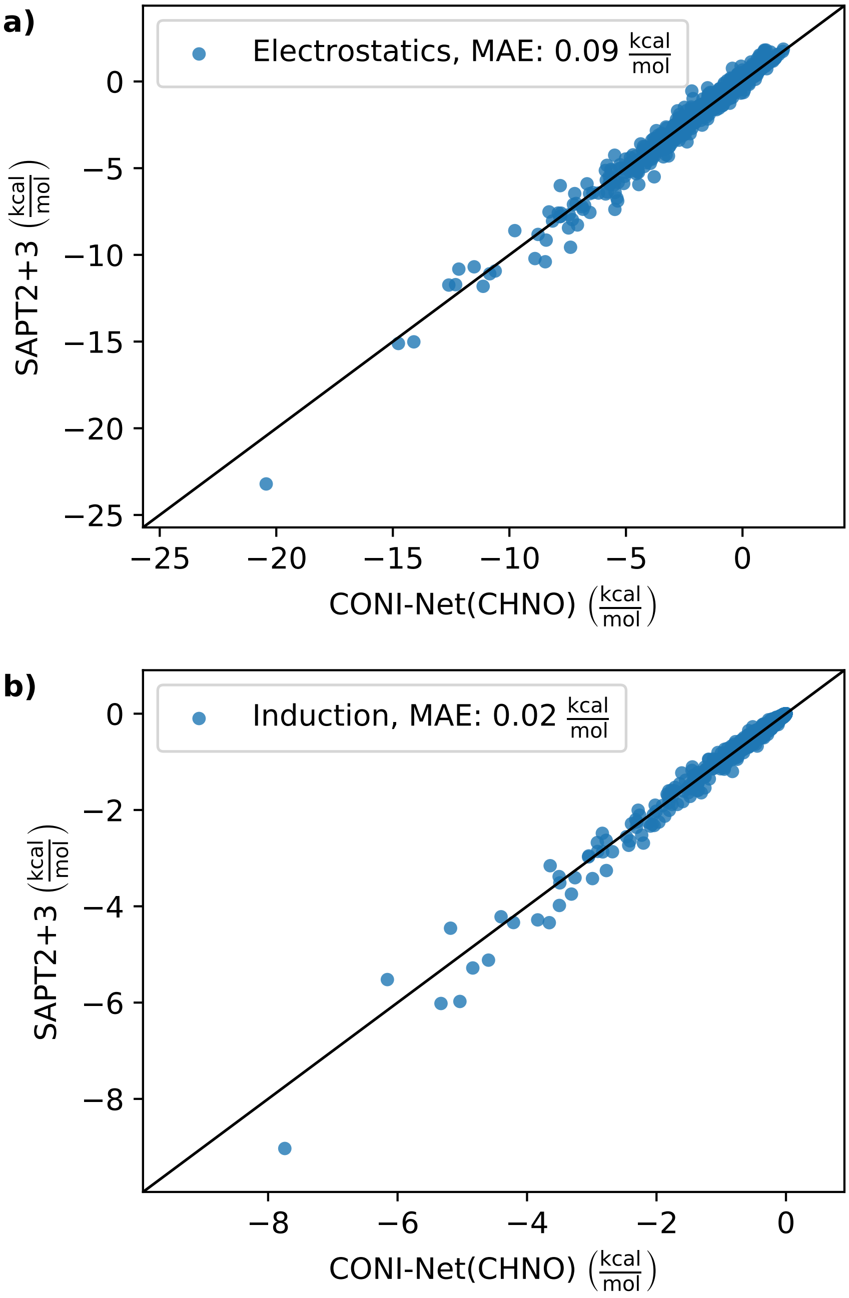


Fig. 4.9.: a) Electrostatic and b) induction energies of the CHNO test set samples predicted by the CONI-Net model vs. SAPT2+3 reference. Adapted with permission from [146]. Copyright 2021 American Chemical Society.

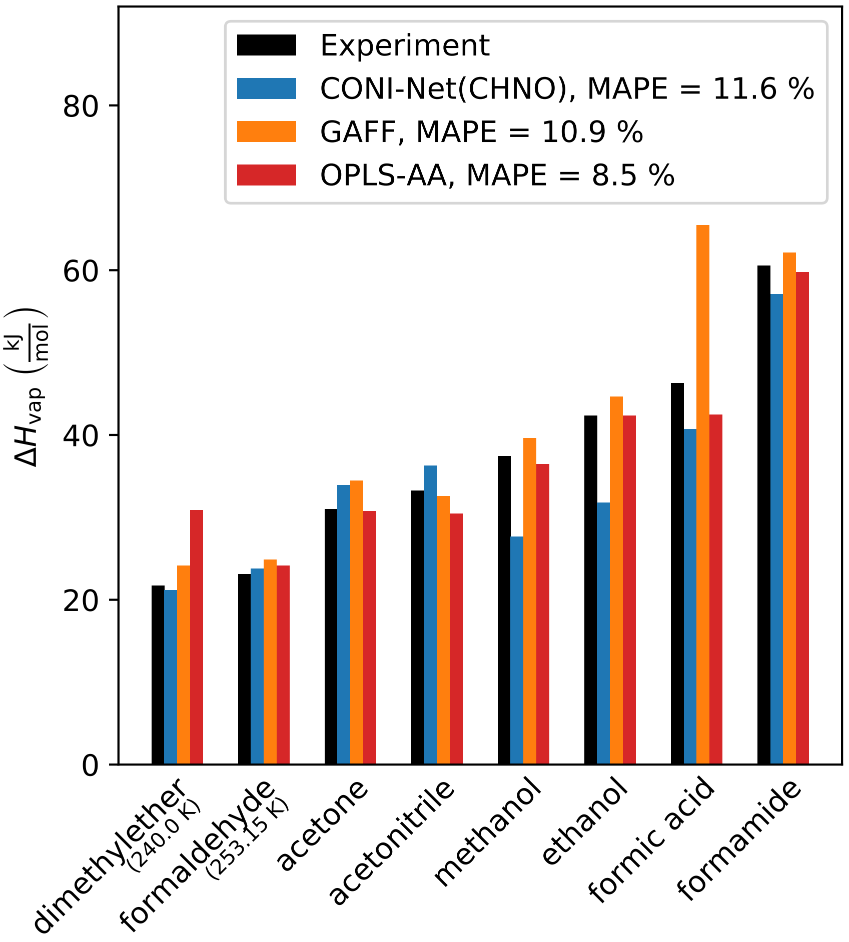


Fig. 4.10.: Predictions of the trained CONI-Net model for the enthalpy of vaporization of the molecules in the MD ensemble (Fig. 4.6) and comparison to experiments and conventional force fields. Furthermore, the mean absolute percentage error (MAPE) with respect to the experimental values is given for all models. The predictions of GAFF and OPLS-AA are taken from Ref. [106]. The corresponding numerical values and the literature references for the experimental values are given in Tab. A.3 of Appendix A.1. Adapted with permission from [146]. Copyright 2021 American Chemical Society.

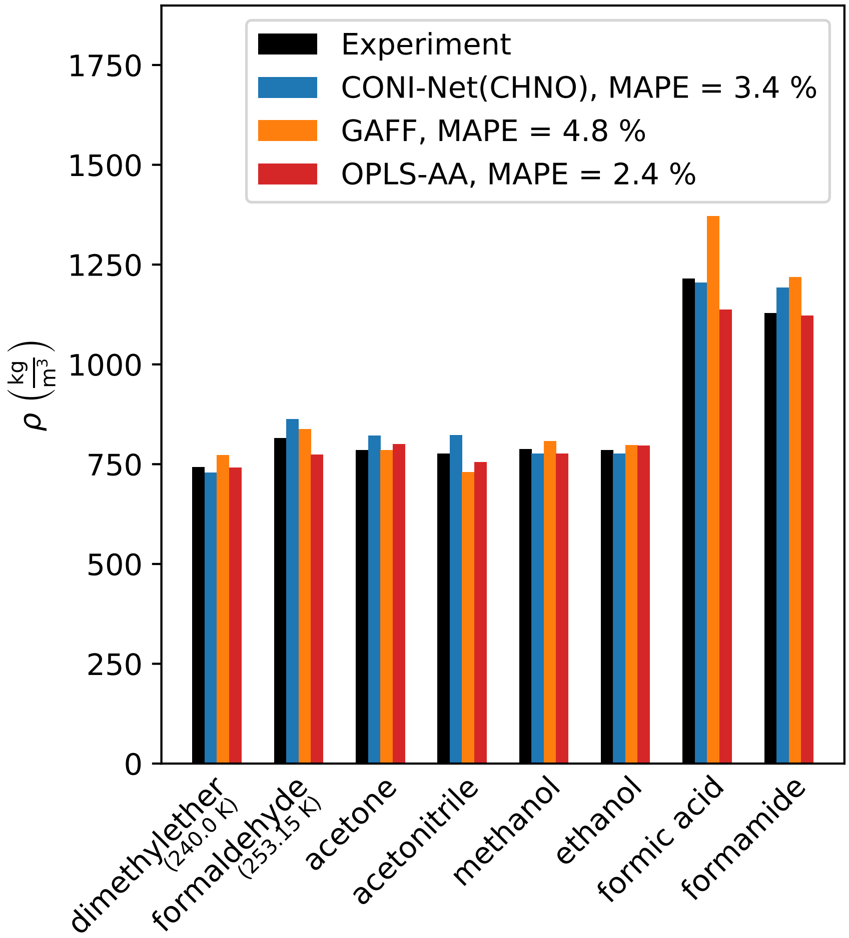


Fig. 4.11.: Predictions of the trained CONI-Net model for the mass density of the molecules in the MD ensemble (Fig. 4.6) and comparison to experiments and conventional force fields. Furthermore, the mean absolute percentage error (MAPE) with respect to the experimental values is given for all models. The predictions of GAFF and OPLS-AA are taken from Ref. [106]. The corresponding numerical values and the literature references for the experimental values are given in Tab. A.3 of Appendix A.1. Adapted with permission from [146]. Copyright 2021 American Chemical Society.

4.3.3 Discussion

The direct comparison of the test set MAEs of the different components shows that for the CHNO model, the prediction of exchange and electrostatic interactions is more challenging compared to the dispersion and induction components. On the one hand, the exchange and electrostatics models have to cover larger energy ranges, therefore relative deviations can have a stronger effect on the absolute error. On the other hand, due to the different nature of the interactions, the approximations of the model can have an unequal impact on the individual components. One of the strongest approximations is the isotropy of the pair potentials between the atoms, which mainly affects the modeling of short-range effects that arise from orbital overlap, particularly the charge penetration and Pauli repulsion of the electrostatic and exchange components. For these effects, the anisotropy of the interaction occurs only spherically averaged in the resulting model. This compromise can therefore introduce an uncertainty in the prediction of close configurations and increase the MAE. Whereas, for the dispersion, the isotropic model performs exceptionally well. Due to the flatter distance dependence of this component, the relevant distance range is larger compared to the effects mentioned above. The applicability of isotropic models for the description of dispersion interactions has also been demonstrated by dispersion correction schemes for DFT [97].

The combined model for the total energy shows an overall consistent performance to interpolate between the different molecules and dimer geometries, even though the components were trained independently, and therefore no error cancellation between the different interactions was enforced. For neural network potentials of the atomization energy, it is common to evaluate the model by its capability to achieve the so-called chemical accuracy of at least $1\ \mathrm{kcal/mol}$ for the MAE [54]. The MAE of the CHNO model is way below this threshold. However, since, in general, the absolute energies of NCIs are lower compared to atomization energies in most cases, the chemical accuracy may not be a reliable indicator. The obvious solution would be to define a comparable quality threshold for NCIs, but this is problematic since NCIs for different configurations can have a high relative variation. Since the absolute model error is expected to vary for the different regimes, the MAE for a set of samples is heavily dependent on its composition. As an example, dimer configurations at large distances may cope better with the approximation of isotropic pair interactions than close configurations, as discussed above. Therefore, while the MAE is a useful metric to compare models for a given dataset, its significance to assess the absolute model performance is limited.

Furthermore, an important reason why the test MAE has only partial relevance is that the application area of the model is molecular dynamics simulations of bulk systems. When transferring to extensive dynamic simulations, several approximations can influence the simulation behavior that are not directly visible in the model performance for dimer samples. One problem that arises from the molecular mechanics model is the forced symmetry of the intramolecular force field of specific functional groups. This issue particularly affects methyl groups, which generally have a dihedral potential that allows rotation into three equivalent minimum positions. Due to this symmetry, the partial charges and pair potentials for all hydrogen atoms also have to be equal. For the model in this work, this symmetry is introduced during the fingerprint and partial charge calculation stage. However, this forced symmetry introduces an artificial constraint that limits the modeling flexibility, especially if the rest of the molecule does not share the symmetry of the functional group.

On the level of model training, only dimers constructed from equilibrium monomer geometries are used in this study. However, when applying the model in MD, distortions and dihedral rotations are introduced that are unknown to the model, which can potentially lead to discrepancies in the intermolecular potential. In order to avoid this, the dataset generation procedure would have to include non-equilibrium monomer geometries, and the fit of partial charges would have to be a compromise for the different conformers of the molecule.

Another approximation occurs during the application of the model in MD. The model in this study is trained exclusively on dimer samples. Therefore, many-body effects between three or more molecules are neglected, introducing a source of error for describing the cohesive energy density. The significance of this neglect depends on the nature of the considered molecules, e.g., in polar or ionic liquids, many-body interactions can represent a significant share of the total interaction energy [158, 159].

The objective of the model is to predict ab initio energies from SAPT. However, to keep the computational demands for the generation of the datasets feasible, the accuracy of the ab initio reference method is limited. The resulting deviation between the reference energies for training and the true values is forwarded to the model and can affect the prediction of thermodynamic properties. Furthermore, since the reference SAPT calculations do not supply forces, a direct training of the force is not possible. Any potential artifacts related to the force would also first become apparent in the MD simulations.

When comparing the property predictions for the different molecules (Fig. 4.10 and 4.11), it shows that for all models, the results for the enthalpy of vaporization have higher deviations from the experimental values than the mass density predictions. On average, the OPLS-AA force field has the most accurate predictions for both tasks, which can be explained by the focus on the optimization of liquid properties during the development of the force field [27]. The GAFF force field shows similar performance for most molecules. However, mainly due to the poor prediction results for formic acid, the mean absolute percentage errors (MAPE) for both properties are slightly increased compared to the OPLS-AA force. For the enthalpy of vaporization, the average performance of the CONI-Net model lines up right behind the conventional force fields. The most considerable deviations arise for the two alcohols, methanol and ethanol. A possible source of error could be the enforced symmetry on the methyl groups, as discussed above. For the mass density, the CONI-Net model shows an overall good agreement with experimental data and positions between the OPLS-AA and GAFF performance.

To conclude, the results for the test set in Fig. 4.7 show, that the network structure combined with the fingerprint descriptor is capable of interpolating the dimer potential energy surface for a set of small organic molecules. Furthermore, the transferability to large arrangements is demonstrated in MD simulations of the liquid phase (Fig. 4.10 and 4.11), where despite the approximations, the performance to predict thermodynamic observables is comparable to force fields with empirically derived non-covalent interactions.

4.4 Transferable force field

In the following, the application of the CONI-Net model for the development of a transferable force field is investigated. Therefore, the MD ensemble to test the property prediction consists of molecules that were not used in model training.

4.4.1 Molecules and dataset

For computational feasibility, the chemical space for this study is limited to hydrocarbons consisting only of the chemical elements carbon and hydrogen. The training ensemble contains compounds with up to three carbon atoms, except for benzene. The MD ensemble contains larger hydrocarbon molecules since no expensive dimer dataset calculation is required for them. The renderings and names of the molecules for both ensembles are shown in Fig. 4.12.

The CH dataset contains 2000 homodimers for each molecule in the training ensemble and 400 heterodimers for every possible combination. The total dataset is randomly split into a training and validation set at the ratio of 3:1, resulting in a total of 20400 training and 6800 validation samples. Additionally, an independent test set with the same proportions is generated, which contains 6800 samples.

4.4.2 Results of model training and property predictions

As for the CHNO model, with a fixed training and validation set split, the model training is repeated five times for each component with random weight initializations. The component models with the best validation set MAE are then combined to the final total energy model. The test set results for the total energy and individual components are shown in Fig. 4.13, 4.14 and 4.15. Subsequently, the final model is used to conduct MD simulations for all molecules in the MD ensemble (Fig. 4.12). The resulting predictions for the thermodynamic properties and the corresponding experimental values are shown in Fig. 4.16 and 4.17. The related numerical values and references are given in Tab. A.4 of Appendix A.1.

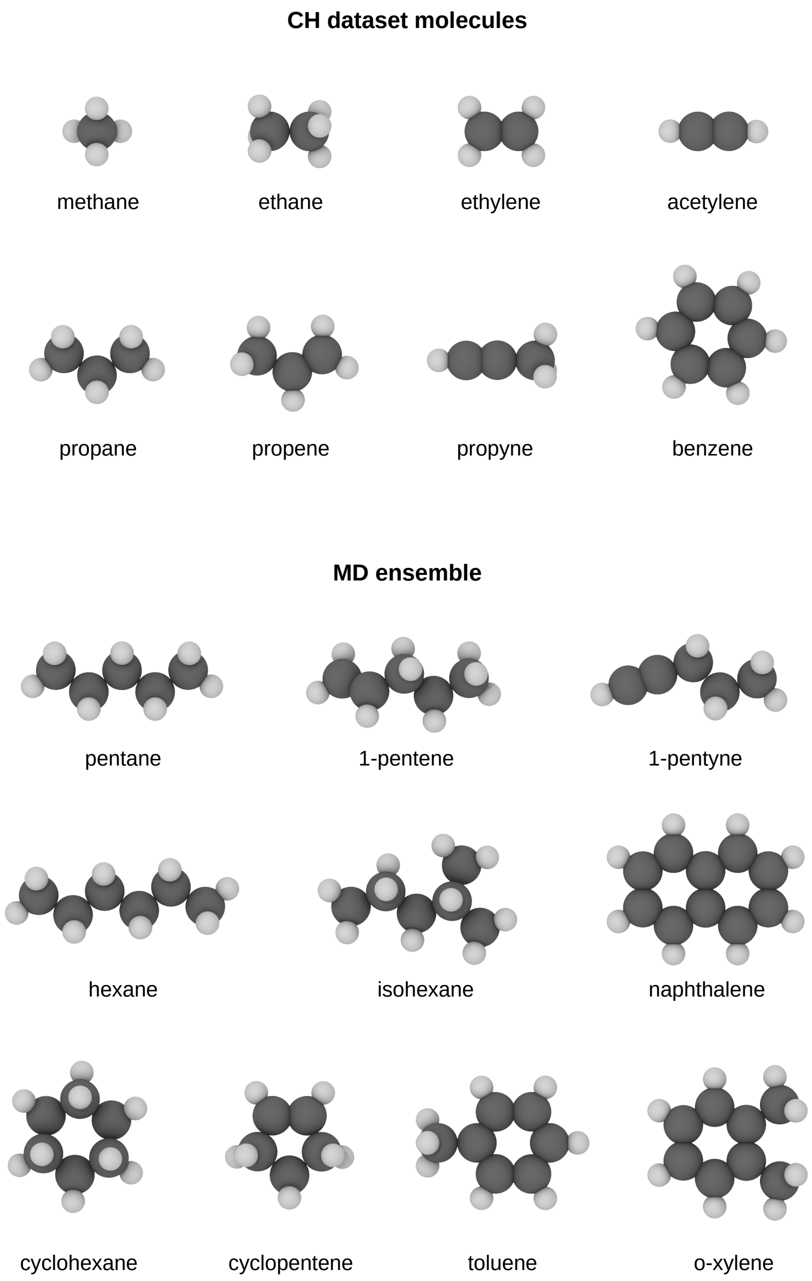


Fig. 4.12.: Renderings and names of the molecules in the CH dataset used for model training and the MD ensemble used in the property prediction simulations. Visualization software: OVITO [93].

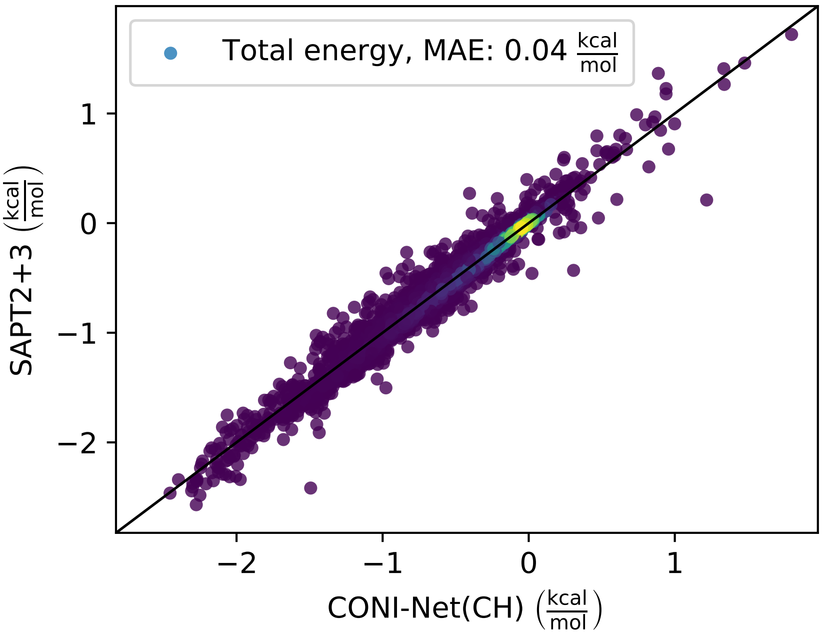


Fig. 4.13.: Total non-covalent interaction energies of the CH test set samples predicted by the CONI-Net model vs. SAPT2+3 reference. The color coding corresponds to the local point density calculated via a Gaussian kernel-density estimate as implemented in SciPy [157], violet represents low and yellow high values. Adapted with permission from [146]. Copyright 2021 American Chemical Society.

4.4.3 Discussion

In comparison to the model for the CHNO dataset, the resulting test set MAEs for the total energy and components are slightly lower. On the one hand, the total energy range is smaller compared to the CHNO model. On the other hand, the attractive component with the largest range is now the dispersion component compared to the electrostatic component for the CHNO model, which indicates that the nature of interactions is shifted for the CH dataset. Since the dispersion component copes very well with the approximations of the model, this shift could also lead to an improvement in the prediction of the total energy. Furthermore, an error cancellation effect cannot be ruled out for the combined total energy, even though the components were trained independently.

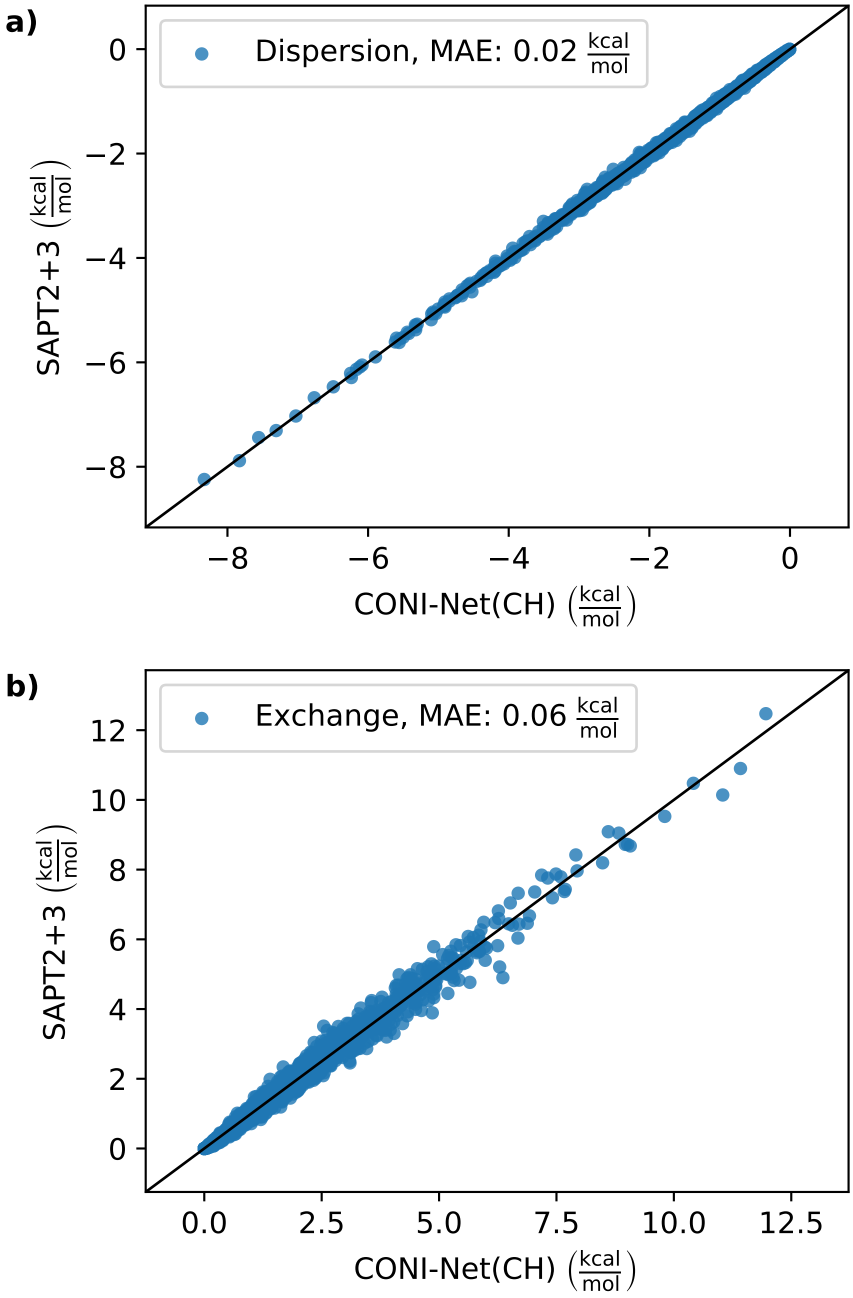


Fig. 4.14.: a) Dispersion and b) exchange energies of the CH test set samples predicted by the CONI-Net model vs. SAPT2+3 reference. Adapted with permission from [146]. Copyright 2021 American Chemical Society.

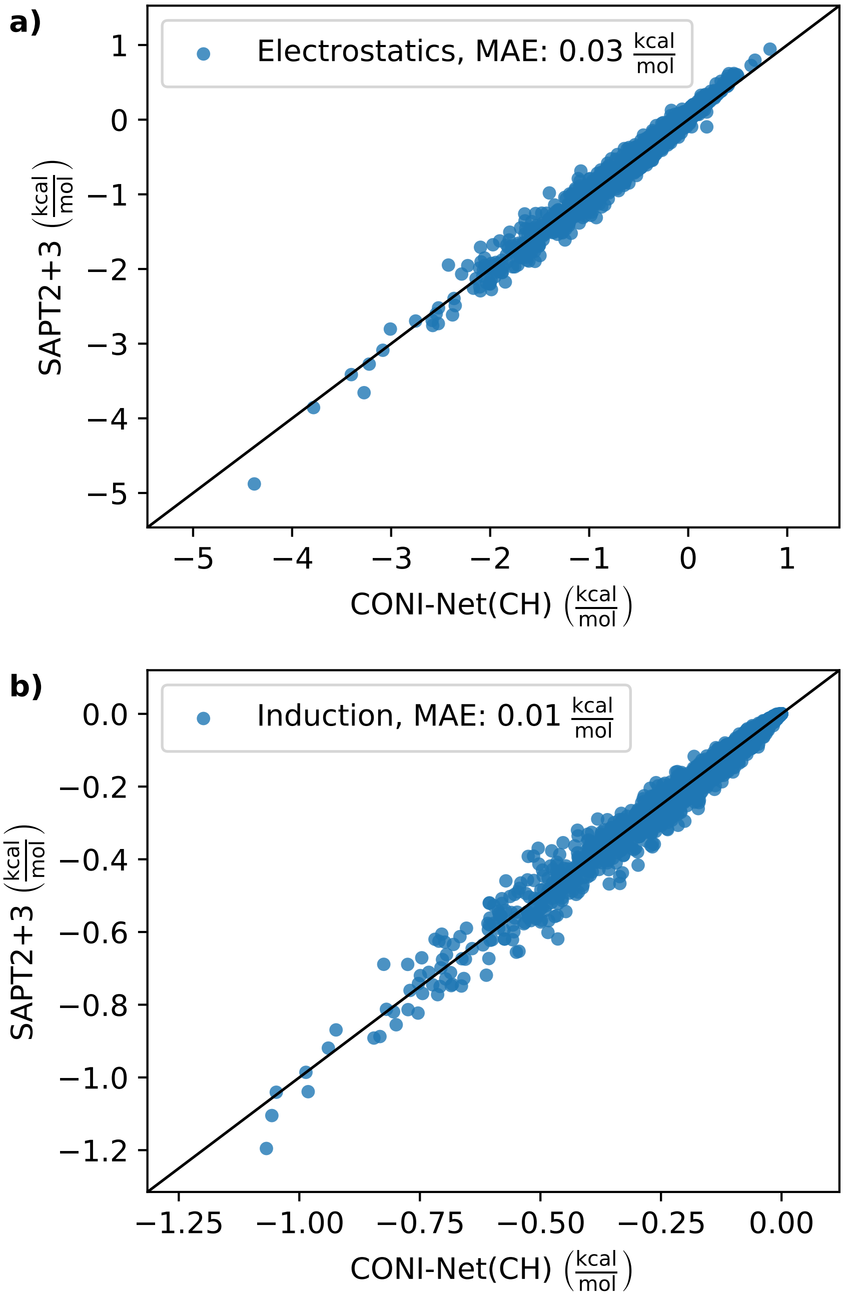


Fig. 4.15.: a) Electrostatic and b) induction energies of the CH test set samples predicted by the CONI-Net model vs. SAPT2+3 reference. Adapted with permission from [146]. Copyright 2021 American Chemical Society.

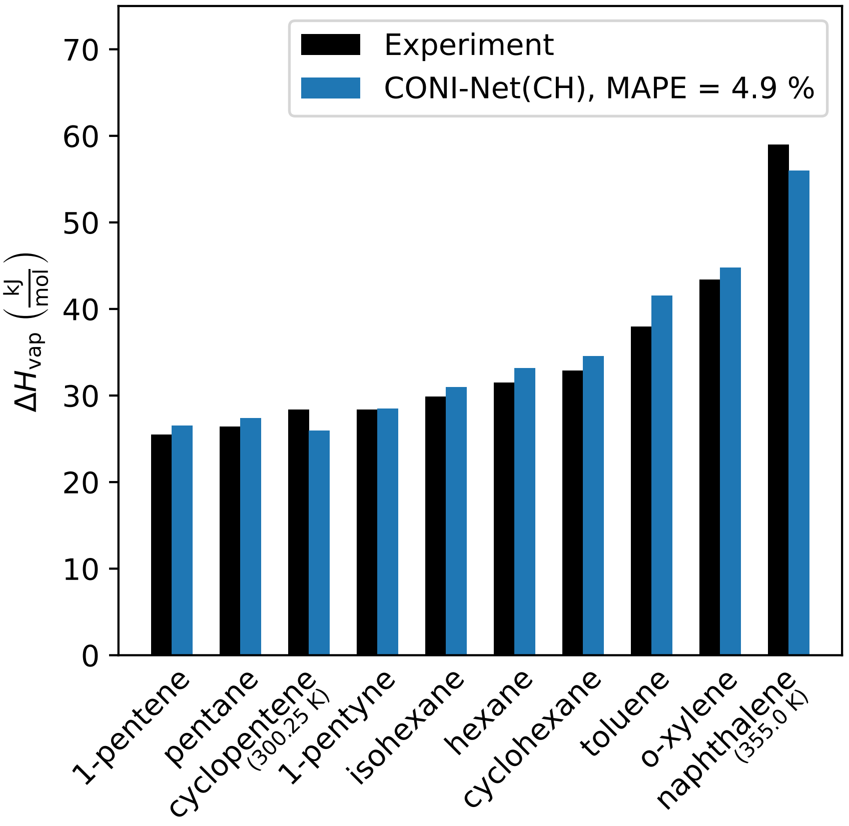


Fig. 4.16.: Predictions of the trained CONI-Net model for the enthalpy of vaporization of the molecules in the MD ensemble (Fig. 4.12) and comparison to experiments. Furthermore, the mean absolute percentage error (MAPE) with respect to the experimental values is given. The corresponding numerical values and the literature references for the experimental values are given in Tab. A.4 of Appendix A.1. Adapted with permission from [146]. Copyright 2021 American Chemical Society.

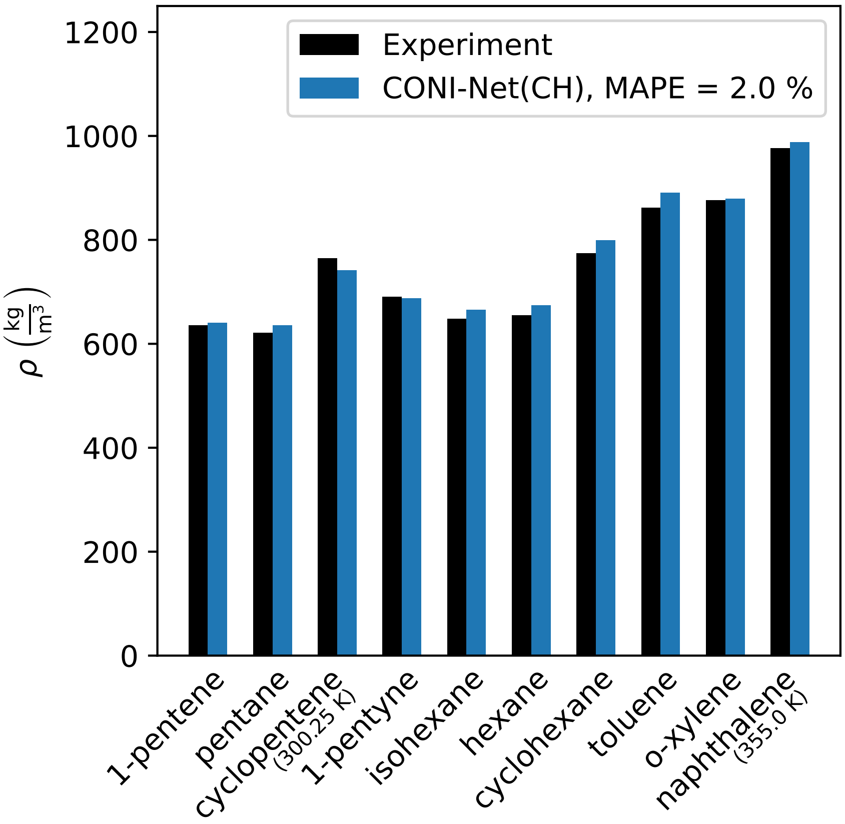


Fig. 4.17.: Predictions of the trained CONI-Net model for the mass density of the molecules in the MD ensemble (Fig. 4.12) and comparison to experiments. Furthermore, the mean absolute percentage error (MAPE) with respect to the experimental values is given. The corresponding numerical values and the literature references for the experimental values are given in Tab. A.4 of Appendix A.1. Adapted with permission from [146]. Copyright 2021 American Chemical Society.

The systematic errors at the different stages of the model development as described for the CHNO model in Sec. 4.3.3 also apply for this application. The main difference here is the use of separate sets of molecules for the model training and MD property prediction simulations. Therefore, the MD simulations can also include larger molecules that allow significant conformation changes due to soft torsional degrees of freedom, and hence the approximation to use only equilibrium monomer geometries for the partial charge calculations can have a more substantial effect. Particular affected could be linear chain molecules with a highly symmetric equilibrium geometry. However, as discussed in Sec. 4.3.3, this issue could be reduced by an extended fitting procedure that takes into account non-equilibrium conformers.

The property predictions in Fig. 4.16 and 4.17 show an overall good agreement with the experimental values for both examined observables, the enthalpy of vaporization and mass density. Compared to the results of the CHNO model, it is noticeable that the deviations have less variance. In particular, the reliable predictions suggest a solid performance of the modeling approach and ab initio method for aliphatic and aromatic functional groups.

In conclusion, the applicability of the CONI-Net model and fingerprint descriptor for the development of a transferable force field could successfully be demonstrated. The application of the model in MD simulations yields promising results for the prediction of thermodynamic properties of molecules that are not involved in the training set. Despite the minimalist dataset of small hydrocarbons for the training, the model is transferable to larger molecules and structures unknown to the model, such as cyclic or branched aliphatic compounds and new combinations of functional groups.

Summary and outlook 5

Summary

Analytical force fields are a useful concept to extend the length and time scales of molecular simulations of soft matter phenomena. General force field definitions provide parameter sets specialized for subsets of the chemical space, such as organic molecules. The non-covalent interactions in these models are represented by simple function expressions which are computationally efficient but require empirical target data for the top-down parameter fitting procedure.

One part of this work is the application of a general force field for molecular dynamics simulations of several organic semiconductors. The efficiency of the method allowed the simulation of long trajectories of the molecular vibrations inside amorphous bulk morphologies. These trajectories enabled a collaborative study on the development of a workflow to calculate the dynamic and static contributions to the conformational energy level disorder in the materials. However, the applicability of this approach is dependent on available force field parameters for the considered molecules. For instance, organic semiconductors with metal centers, which are common emitters in organic light-emitting diodes, could therefore not be included in this study. An extension of the conventional force field is not straightforward, and experimental properties of these exotic materials are in general not available. This challenge motivates the development of a bottom-up approach for the modeling of non-covalent interactions to eliminate the dependence on empirical target values and replace it with data from ab initio calculations.

A bottom-up model for predicting non-covalent interactions in large-scale molecular dynamics simulations has to fulfill several criteria. First of all, it has to provide smooth and consistent force curves over the whole relevant distance range to ensure stable integration of the equation of motions while providing adequate fitting flexibility to interpolate the ab initio interactions for different species and geometries. Furthermore, to achieve long simulation times and large system sizes, the energy and force calculations are required to be computationally efficient. Finally, an automated and universal parameter fitting procedure is essential for a straightforward extension into new areas of the chemical space.

In this work, I presented the development of a neural network potential for non-covalent interactions, which fulfills all of these requirements. The method is based on a modified Behler-Parrinello approach with pairwise energy contributions. As an alternative to symmetry functions, I developed a fingerprint descriptor that characterizes the chemical environment of atoms in a molecule by properties of the equilibrium geometry. Additionally, I introduced a new submodel structure that processes the fingerprint descriptor and the distance dependence of the pair contribution in separate steps. The resulting model is robust against overfitting and flexible enough to interpolate complex potential energy surfaces. Another advantage is the possibility to precompute the neural network inference for the application in molecular dynamics simulations, enabling length and time scales comparable to conventional force fields. As ab initio reference for the model training serves the interaction energy from symmetry-adapted perturbation theory, which provides a decomposition into physically motivated energy components. In particular, for each component, an independent model is trained, thereby preserving full separability. Furthermore, all required steps to build a model for a set of molecules from scratch, including the preparatory calculations of the descriptor and dataset, are implemented in a well-defined automated procedure.

I demonstrated the performance of the model in two separate applications. First, a custom model was trained for a series of small organic molecules. Subsequently, predictions for the enthalpy of vaporization and mass density of the molecules were obtained from molecular dynamics simulations of the liquid phase. The results agree well with the values from experiments and conventional force fields, despite the absence of empirical target values in the training procedure and moderate quality of the ab initio method. In the second application, a model was trained for a set of hydrocarbons. However, the prediction of thermodynamic observables was performed for molecules not included in the training set. Nevertheless, the model is in excellent agreement with experiments and represents a successful proof-of-concept for deploying the developed method as a transferable force field.

Outlook

The developed approach has possibilities for improvements at various points. One aspect that influences the predictive performance of the model is the quality of the ab initio reference calculations. If the computational resources allow, there are larger basis sets and higher-order truncations of the symmetry-adapted perturbation theory available [17]. Active learning schemes can compensate for higher computational demands by reducing the number of required training samples. For instance, the query-by-committee selection method was shown to facilitate the training of a neural network potential for predicting atomization energies [160, 161].

Another aspect related to the dataset is the dimer sampling algorithm, which in this study is limited to equilibrium monomer geometries. In order to improve transferability, the sampling procedure could also generate dimers from non-equilibrium monomers. Similarly, the partial charge fit for the electrostatic baseline model could also take into account non-equilibrium conformations of the monomers. Furthermore, the fingerprint descriptor could easily be extended by additional properties that characterize the atom pairs in more detail. The model structure also has the potential for extensions. On the one hand, a direct polarization model could be integrated, which introduces many-body interactions that do not require self-consistent treatment [162]. On the other hand, the method could be combined with a bottom-up parametrization for the intramolecular interactions to achieve an entirely consistent and complete model [33–35].

However, also in its current state, the method establishes new possibilities for molecular simulations. The separability of the interactions enables a contribution-resolved analysis of large arrangements such as solvent-solute systems, and the bottom-up approach enables force field development for new molecular materials where empirical data is not available.

Appendices

A

A.1 Numerical values

Tab. A.1.: Results for the dynamic, static, and total disorder values from the predicted HOMO energies. The mean absolute error (MAE) of the ML model on the validation set should be smaller than the dynamic disorder, so that the noise from the energy prediction is not dominant. To demonstrate the consistency of the disorder separation method, the squared sum of the static and dynamic disorder is given, which should be in agreement with the directly computed total disorder. The data of this table was provided by Patrick Reiser and is presented in the Supporting Information of [110].

Molecule (HOMO)	MAE (meV)	R^2	static disorder (meV)	dynamic disorder (meV)	total disorder (meV)	squared sum (meV)
a-NPD	17	0.96	35.0	104.0	109.7	109.7
B4PyMPM	29	0.85	14.4	91.7	92.8	92.8
B4PyPPM	36	0.69	16.2	80.4	82.0	82.0
m-BPD	17	0.95	50.5	76.0	91.3	91.3
mCP	9	0.96	20.0	62.1	65.2	65.2
NPB	15	0.97	53.5	90.4	105.0	105.0
o-BPD	18	0.94	33.0	86.4	92.4	92.4
p-BPD	17	0.94	50.6	73.0	88.8	88.8
Spiro-OMeTAD	34	0.81	49.7	79.0	93.2	93.2
Spiro-TAD	25	0.85	41.8	66.2	78.2	78.3
TCTA	20	0.94	62.2	86.6	106.6	106.6
TPBi	21	0.87	24.6	68.3	72.6	72.6
TPyQB	25	0.85	33.9	70.1	77.8	77.9

Tab. A.2.: Results for the dynamic, static, and total disorder values from the predicted LUMO energies. The mean absolute error (MAE) of the ML model on the validation set should be smaller than the dynamic disorder, so that the noise from the energy prediction is not dominant. To demonstrate the consistency of the disorder separation method, the squared sum of the static and dynamic disorder is given, which should be in agreement with the directly computed total disorder. The data of this table was provided by Patrick Reiser and is presented in the Supporting Information of [110].

Molecule	MAE	R^2	static	dynamic	total	squared
(LUMO)			disorder	disorder	disorder	sum
	(meV)		(meV)	(meV)	(meV)	(meV)
a-NPD	14	0.97	22.1	95.1	97.6	97.6
B4PyMPM	20	0.97	87.7	113.3	143.2	143.3
B4PyPPM	28	0.93	88.8	107.9	139.7	139.7
m-BPD	36	0.91	91.3	116.0	147.6	147.7
mCP	15	0.97	38.0	102.2	109.0	109.1
NPB	14	0.96	27.3	93.9	97.7	97.8
o-BPD	31	0.93	80.4	122.0	146.0	146.1
p-BPD	34	0.92	93.2	118.0	150.3	150.4
Spiro-OMeTAD	35	0.88	81.2	99.9	128.7	128.7
Spiro-TAD	33	0.89	73.5	101.4	125.2	125.3
TCTA	32	0.94	106.0	134.4	171.1	171.2
TPBi	32	0.91	68.2	111.8	130.9	130.9
TPyQB	20	0.92	41.2	78.4	88.6	88.6

Tab. A.3.: Numerical values of the thermodynamic properties predicted by the CONI-Net(CHNO) model, GAFF, and OPLS-AA and comparison with experimental data. The enthalpy of vaporization ΔH_{vap} and mass density ρ are given at the temperature $T = 298.15$ K (unless stated otherwise) and in the units of kJ/mol and kg/m^3. The mean absolute percentage errors (MAPE) are specified with respect to experimental values. The data of this table is presented in the Supporting Information of [146].

Name	Obs.	Exp.	CONI-Net(CHNO)	GAFF	OPLS-AA
dimethylether	ΔH_{vap}	21.72 [106][a]	21.19 ± 0.02[c]	24.12 ± 0.02 [106]	30.87 ± 0.03 [106]
(240.0 K)	ρ	742.1 [106][a]	728.2 ± 0.6[c]	773.0 ± 0.2 [106]	741.4 ± 0.1 [106]
formaldehyde	ΔH_{vap}	23.10 [106][a]	23.76 ± 0.01[c]	24.85 ± 0.02 [106]	24.11 ± 0.01 [106]
(253.15 K)	ρ	815.0 [163]	863.0 ± 0.4[c]	838.0 ± 0.2 [106]	773.7 ± 0.1 [106]
acetone	ΔH_{vap}	30.99 [164]	33.88 ± 0.03[c]	34.47 ± 0.02 [106]	30.76 ± 0.02 [106]
	ρ	784.9 [164]	821.1 ± 0.4[c]	785.6 ± 0.1 [106]	800.3 ± 0.2 [106]
acetonitrile	ΔH_{vap}	33.23 [164]	36.28 ± 0.02[c]	32.60 ± 0.01 [106]	30.42 ± 0.02 [106]
	ρ	776.0 [164]	823.0 ± 0.4[c]	729.6 ± 0.1 [106]	755.1 ± 0.2 [106]
methanol	ΔH_{vap}	37.43 [164]	27.67 ± 0.03[c]	39.62 ± 0.02 [106]	36.44 ± 0.01 [106]
	ρ	787.2 [164]	776.8 ± 0.6[c]	807.5 ± 0.2 [106]	776.8 ± 0.1 [106]
ethanol	ΔH_{vap}	42.32 [164]	31.76 ± 0.03[c]	44.62 ± 0.02 [106]	42.32 ± 0.02 [106]
	ρ	784.8 [164]	776.0 ± 0.5[c]	797.3 ± 0.1 [106]	796.3 ± 0.0 [106]
formic acid	ΔH_{vap}	46.30 [165][b]	40.69 ± 0.02[c]	65.46 ± 0.02 [106]	42.48 ± 0.01 [106]
	ρ	1214.5 [164]	1204.7 ± 0.5[c]	1371.0 ± 0.2 [106]	1136.8 ± 0.1 [106]
formamide	ΔH_{vap}	60.57 [164]	57.10 ± 0.03[c]	62.15 ± 0.01 [106]	59.76 ± 0.01 [106]
	ρ	1128.8 [164]	1192.0 ± 0.2[c]	1218.5 ± 0.1 [106]	1122.1 ± 0.1 [106]
	ΔH_{vap}	MAPE	11.6 %	10.9 %	8.5 %
	ρ	MAPE	3.4 %	4.8 %	2.4 %

[a] To enable a comparison at the same temperature as for the GAFF and OPLS-AA benchmark of Ref. [106], the experimental values given by this reference were also used here. The authors obtained them from the web database knovel.com.

[b] In the gas phase of formic acid, the molecules are partially present as dimers, and therefore the value for the enthalpy of vaporization strongly deviates from ideal gas behavior. In order to enable a comparison with the expressions used to calculate the enthalpy of vaporization from the molecular dynamics results, a literature value is chosen, which is corrected by the dissociation enthalpy of the dimers. The corrected value is related to the transition of a liquid to a monomeric gas.

[c] The averages and uncertainties are calculated through block averaging with a block size of 200 ps.

Tab. A.4.: Numerical values of the thermodynamic properties predicted by the CONI-Net(CH) model and comparison with experimental data. The enthalpy of vaporization ΔH_{vap} and mass density ρ are given at the temperature $T = 298.15$ K (unless stated otherwise) and in the units of kJ/mol and kg/m^3. The mean absolute percentage error (MAPE) is specified with respect to experimental values. The data of this table is presented in the Supporting Information of [146].

Name	Obs.	Exp.	CONI-Net(CH)
1-pentene	ΔH_{vap}	25.47 [163]	26.55 ± 0.05[b]
	ρ	635.3 [166]	640.7 ± 0.1[b]
pentane	ΔH_{vap}	26.41 [164]	27.41 ± 0.06[b]
	ρ	621.4 [164]	635.8 ± 0.2[b]
cyclopentene	ΔH_{vap}	28.37 [167]	25.97 ± 0.04[b]
(300.25 K)	ρ	764.4 [166][a]	741.3 ± 0.6[b]
1-pentyne	ΔH_{vap}	28.40 [168]	28.50 ± 0.05[b]
	ρ	690.7 [169]	687.8 ± 0.3[b]
isohexane	ΔH_{vap}	29.89 [163]	30.99 ± 0.06[b]
	ρ	648.5 [170]	666.0 ± 0.2[b]
hexane	ΔH_{vap}	31.48 [164]	33.16 ± 0.06[b]
	ρ	654.9 [164]	674.6 ± 0.2[b]
cyclohexane	ΔH_{vap}	32.89 [164]	34.55 ± 0.04[b]
	ρ	774.2 [164]	799.5 ± 0.3[b]
toluene	ΔH_{vap}	37.99 [164]	41.57 ± 0.04[b]
	ρ	861.9 [164]	890.8 ± 0.2[b]
o-xylene	ΔH_{vap}	43.43 [164]	44.77 ± 0.04[b]
	ρ	876.0 [164]	879.0 ± 0.2[b]
naphthalene	ΔH_{vap}	59.00 [171]	56.01 ± 0.05[b]
(355.0 K)	ρ	976.7 [172]	988.1 ± 0.1[b]
	ΔH_{vap}	MAPE	4.9 %
	ρ	MAPE	2.0 %

[a] The density at the temperature 300.25 K is obtained from an extrapolation using the value at 298.15 K and the temperature coefficient of density.

[b] The averages and uncertainties are calculated through block averaging with a block size of 200 ps.

A.2 Abbreviations

Adam	Adaptive moment estimation optimizer
ANN	Artificial neural network
CG	Coarse graining
CH	Label for dataset (carbon and hydrogen)
CHNO	Label for dataset (carbon, hydrogen, nitrogen, and oxygen)
CONI-Net	Component-separable Non-covalent Interaction Network
DFT	Density functional theory
GAFF	General Amber force field
HOMO	Highest occupied molecular orbital
LUMO	Lowest unoccupied molecular orbital
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MD	Molecular dynamics
ML	Machine learning
MM	Molecular mechanics
NCI	Non-covalent interaction
NNP	Neural network potential
NVE	Micro-canonical ensemble
NVT	Canonical ensemble
NPT	Isotherm-isobaric ensemble
OFET	Organic field-effect transistor
OLED	Organic light-emitting diode
OPLS-AA	Optimized Potentials for Liquid Simulations (all-atom)
OPV	Organic photovoltaics
QM	Quantum mechanics
ReLU	Rectified linear unit
RESP	Restrained electrostatic potential
SAPT	Symmetry-adapted perturbation theory

Bibliography

[1] A. Verma, A. Schug, K. H. Lee, W. Wenzel, “Basin hopping simulations for all-atom protein folding”, *The Journal of Chemical Physics* **2006**, *124*, 044515.

[2] J. Flick, F. Tristram, W. Wenzel, “Modeling loop backbone flexibility in receptor-ligand docking simulations”, *Journal of Computational Chemistry* **2012**, *33*, 2504–2515.

[3] J. Xu, R. L. B. Selinger, J. V. Selinger, R. Shashidhar, “Monte Carlo simulation of liquid-crystal alignment and chiral symmetry-breaking”, *The Journal of Chemical Physics* **2001**, *115*, 4333–4338.

[4] D. Danilov, C. Barner-Kowollik, W. Wenzel, “Modelling of reversible single chain polymer self-assembly: from the polymer towards the protein limit”, *Chemical Communications* **2015**, *51*, 6002–6005.

[5] P. Friederich, V. Meded, F. Symalla, M. Elstner, W. Wenzel, “QM/QM Approach to Model Energy Disorder in Amorphous Organic Semiconductors”, *Journal of Chemical Theory and Computation* **2015**, *11*, 560–567.

[6] J. Černý, P. Hobza, “Non-covalent interactions in biomacromolecules”, *Physical Chemistry Chemical Physics* **2007**, *9*, 5291–5303.

[7] B. Jeziorski, R. Moszynski, K. Szalewicz, “Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes”, *Chemical Reviews* **1994**, *94*, 1887–1930.

[8] C. Møller, M. S. Plesset, “Note on an Approximation Treatment for Many-Electron Systems”, *Physical Review* **1934**, *46*, 618–622.

[9] J. Čížek, “On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods”, *The Journal of Chemical Physics* **1966**, *45*, 4256–4266.

[10] E. G. Hohenstein, C. D. Sherrill, “Wavefunction methods for noncovalent interactions: Noncovalent interactions”, *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 304–326.

[11] R. J. Bartlett, G. D. Purvis, “Many-body perturbation theory, coupled-pair many-electron theory, and the importance of quadruple excitations for the correlation problem”, *International Journal of Quantum Chemistry* **1978**, *14*, 561–581.

[12] G. D. Purvis, R. J. Bartlett, “A full coupled-cluster singles and doubles model: The inclusion of disconnected triples”, *The Journal of Chemical Physics* **1982**, *76*, 1910–1918.

[13] J. A. Pople, M. Head-Gordon, K. Raghavachari, “Quadratic configuration interaction. A general technique for determining electron correlation energies”, *The Journal of Chemical Physics* **1987**, *87*, 5968–5975.

[14] K. Raghavachari, G. W. Trucks, J. A. Pople, M. Head-Gordon, “A fifth-order perturbation comparison of electron correlation theories”, *Chemical Physics Letters* **1989**, *157*, 479–483.

[15] P. Jurečka, J. Šponer, J. Černý, P. Hobza, “Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs”, *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

[16] J. Řezáč, K. E. Riley, P. Hobza, “S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures”, *Journal of Chemical Theory and Computation* **2011**, *7*, 2427–2438.

[17] T. M. Parker, L. A. Burns, R. M. Parrish, A. G. Ryno, C. D. Sherrill, “Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies”, *The Journal of Chemical Physics* **2014**, *140*, 094106.

[18] K. Patkowski, “Recent developments in symmetry-adapted perturbation theory”, *WIREs Computational Molecular Science* **2020**, *10*.

[19] A. D. MacKerell, D. Bashford, M. Bellott, et al., “All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins”, *The Journal of Physical Chemistry B* **1998**, *102*, 3586–3616.

[20] S. Takamori, M. Holt, K. Stenius, et al., “Molecular Anatomy of a Trafficking Organelle”, *Cell* **2006**, *127*, 831–846.

[21] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, K. Schulten, “Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus”, *Structure* **2006**, *14*, 437–449.

[22] M. Zink, H. Grubmüller, “Mechanical Properties of the Icosahedral Shell of Southern Bean Mosaic Virus: A Molecular Dynamics Study”, *Biophysical Journal* **2009**, *96*, 1350–1363.

[23] A. Minoia, L. Chen, D. Beljonne, R. Lazzaroni, “Molecular modeling study of the structure and stability of polymer/carbon nanotube interfaces”, *Polymer* **2012**, *53*, 5480–5490.

[24] N. R. Tummala, C. Risko, C. Bruner, R. H. Dauskardt, J.-L. Brédas, “Entanglements in P3HT and their influence on thin-film mechanical properties: Insights from molecular dynamics simulations”, *Journal of Polymer Science Part B: Polymer Physics* **2015**, *53*, 934–942.

[25] O. M. Roscioni, G. D’Avino, L. Muccioli, C. Zannoni, “Pentacene Crystal Growth on Silica and Layer-Dependent Step-Edge Barrier from Atomistic Simulations”, *The Journal of Physical Chemistry Letters* **2018**, *9*, 6900–6906.

[26] W. D. Cornell, P. Cieplak, C. I. Bayly, et al., “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules”, *Journal of the American Chemical Society* **1995**, *117*, 5179–5197.

[27] W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, “Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids”, *Journal of the American Chemical Society* **1996**, *118*, 11225–11236.

[28] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, “Development and testing of a general amber force field”, *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

[29] T. Strunk, M. Wolf, M. Brieg, et al., “SIMONA 1.0: An efficient and versatile framework for stochastic simulations of molecular and nanoscale systems”, *Journal of Computational Chemistry* **2012**, *33*, 2602–2613.

[30] H. J. C. Berendsen, D. van der Spoel, R. van Drunen, “GROMACS: A message-passing parallel molecular dynamics implementation”, *Computer Physics Communications* **1995**, *91*, 43–56.

[31] S. Plimpton, “Fast Parallel Algorithms for Short-Range Molecular Dynamics”, *Journal of Computational Physics* **1995**, *117*, 1–19.

[32] D. A. Pearlman, D. A. Case, J. W. Caldwell, et al., “AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules”, *Computer Physics Communications* **1995**, *91*, 1–41.

[33] J. M. Seminario, “Calculation of intramolecular force fields from second-derivative tensors”, *International Journal of Quantum Chemistry* **1996**, *60*, 1271–1277.

[34] A. E. A. Allen, M. C. Payne, D. J. Cole, “Harmonic Force Constants for Molecular Mechanics Force Fields via Hessian Matrix Projection”, *Journal of Chemical Theory and Computation* **2018**, *14*, 274–281.

[35] J. Wildman, P. Repiščák, M. J. Paterson, I. Galbraith, “General Force-Field Parametrization Scheme for Molecular Dynamics Simulations of Conjugated Materials in Solution”, *Journal of Chemical Theory and Computation* **2016**, *12*, 3813–3824.

[36] C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, “A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model”, *The Journal of Physical Chemistry* **1993**, *97*, 10269–10280.

[37] J. Huang, S. Rauscher, G. Nawrocki, et al., “CHARMM36m: an improved force field for folded and intrinsically disordered proteins”, *Nature Methods* **2017**, *14*, 71–73.

[38] N. Schmid, A. P. Eichenberger, A. Choutko, et al., “Definition and testing of the GROMOS force-field versions 54A7 and 54B7”, *European Biophysics Journal* **2011**, *40*, 843–856.

[39] X. Wang, S. Ramírez-Hinestrosa, J. Dobnikar, D. Frenkel, “The Lennard-Jones potential: when (not) to use it”, *Physical Chemistry Chemical Physics* **2020**, *22*, 10624–10633.

[40] J. G. McDaniel, J. Schmidt, “Physically-Motivated Force Fields from Symmetry-Adapted Perturbation Theory”, *The Journal of Physical Chemistry A* **2013**, *117*, 2053–2066.

[41] J. G. McDaniel, J. R. Schmidt, “First-Principles Many-Body Force Fields from the Gas Phase to Liquid: A “Universal” Approach”, *The Journal of Physical Chemistry B* **2014**, *118*, 8042–8053.

[42] J. G. McDaniel, J. Schmidt, “Next-Generation Force Fields from Symmetry-Adapted Perturbation Theory”, *Annual Review of Physical Chemistry* **2016**, *67*, 467–488.

[43] A. J. Stone, A. J. Misquitta, “Atom–atom potentials from *ab initio* calculations”, *International Reviews in Physical Chemistry* **2007**, *26*, 193–222.

[44] M. Tafipolsky, K. Ansorg, “Toward a Physically Motivated Force Field: Hydrogen Bond Directionality from a Symmetry-Adapted Perturbation Theory Perspective”, *Journal of Chemical Theory and Computation* **2016**, *12*, 1267–1279.

[45] P. Xu, E. B. Guidez, C. Bertoni, M. S. Gordon, “Perspective: *Ab initio* force field methods derived from quantum mechanics”, *The Journal of Chemical Physics* **2018**, *148*, 090901.

[46] Y.-P. Liu, K. Kim, B. J. Berne, R. A. Friesner, S. W. Rick, “Constructing *ab initio* force fields for molecular dynamics simulations”, *The Journal of Chemical Physics* **1998**, *108*, 4739–4755.

[47] R. Bukowski, K. Szalewicz, C. F. Chabalowski, “Ab Initio Interaction Potentials for Simulations of Dimethylnitramine Solutions in Supercritical Carbon Dioxide with Cosolvents”, *The Journal of Physical Chemistry A* **1999**, *103*, 7322–7340.

[48] M. Hloucha, A. K. Sum, S. I. Sandler, “Computer simulation of acetonitrile and methanol with ab initio-based pair potentials”, *The Journal of Chemical Physics* **2000**, *113*, 5401.

[49] R. Podeszwa, R. Bukowski, K. Szalewicz, “Potential Energy Surface for the Benzene Dimer and Perturbational Analysis of π-π Interactions”, *The Journal of Physical Chemistry A* **2006**, *110*, 10345–10354.

[50] K. Yu, J. G. McDaniel, J. R. Schmidt, “Physically Motivated, Robust, ab Initio Force Fields for CO $_2$ and N $_2$”, *The Journal of Physical Chemistry B* **2011**, *115*, 10054–10063.

[51] J. R. Schmidt, K. Yu, J. G. McDaniel, “Transferable Next-Generation Force Fields from Simple Liquids to Complex Materials”, *Accounts of Chemical Research* **2015**, *48*, 548–556.

[52] C. Liu, J.-P. Piquemal, P. Ren, “AMOEBA+ Classical Potential for Modeling Molecular Interactions”, *Journal of Chemical Theory and Computation* **2019**, *15*, 4122–4139.

[53] J. Behler, M. Parrinello, “Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces”, *Physical Review Letters* **2007**, *98*, 146401.

[54] J. S. Smith, O. Isayev, A. E. Roitberg, “ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost”, *Chemical Science* **2017**, *8*, 3192–3203.

[55] K. V. J. Jose, N. Artrith, J. Behler, “Construction of high-dimensional neural network potentials using environment-dependent atom pairs”, *The Journal of Chemical Physics* **2012**, *136*, 194111.

[56] D. P. Metcalf, A. Koutsoukas, S. A. Spronk, et al., “Approaches for machine learning intermolecular interaction energies and application to energy components from symmetry adapted perturbation theory”, *The Journal of Chemical Physics* **2020**, *152*, 074103.

[57] Z. L. Glick, D. P. Metcalf, A. Koutsoukas, et al., “AP-Net: An atomic-pairwise neural network for smooth and transferable interaction potentials”, *The Journal of Chemical Physics* **2020**, *153*, 044112.

[58] J. R. Boes, M. C. Groenenboom, J. A. Keith, J. R. Kitchin, “Neural network and ReaxFF comparison for Au properties”, *International Journal of Quantum Chemistry* **2016**, *116*, 979–987.

[59] J. M. Seminario, “Energetics using DFT: comparions to precise ab initio and experiment”, *Chemical Physics Letters* **1993**, *206*, 547–554.

[60] J. Heurich, J. C. Cuevas, W. Wenzel, G. Schön, “Electrical Transport through Single-Molecule Junctions: From Molecular Orbitals to Conduction Channels”, *Physical Review Letters* **2002**, *88*, 256803.

[61] H. Long, B. Pivovar, “Hydroxide Degradation Pathways for Imidazolium Cations: A DFT Study”, *The Journal of Physical Chemistry C* **2014**, *118*, 9880–9888.

[62] S. Yang, P. Olishevski, M. Kertesz, “Bandgap calculations for conjugated polymers”, *Synthetic Metals*, Michael J. Rice Memorial Festschrift **2004**, *141*, 171–177.

[63] C.-G. Zhan, J. A. Nichols, D. A. Dixon, “Ionization Potential, Electron Affinity, Electronegativity, Hardness, and Electron Excitation Energy: Molecular Properties from Density Functional Theory Orbital Energies”, *The Journal of Physical Chemistry A* **2003**, *107*, 4184–4195.

[64] D. Jacquemin, E. A. Perpète, G. E. Scuseria, I. Ciofini, C. Adamo, “TD-DFT Performance for the Visible Absorption Spectra of Organic Dyes: Conventional versus Long-Range Hybrids”, *Journal of Chemical Theory and Computation* **2008**, *4*, 123–135.

[65] R. J. Bartlett, M. Musiał, “Coupled-cluster theory in quantum chemistry”, *Reviews of Modern Physics* **2007**, *79*, 291–352.

[66] B. Aradi, B. Hourahine, T. Frauenheim, “DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method”, *The Journal of Physical Chemistry A* **2007**, *111*, 5678–5684.

[67] M. Karplus, J. A. McCammon, “Molecular dynamics simulations of biomolecules”, *Nature Structural Biology* **2002**, *9*, 646–652.

[68] Y.-T. Fu, C. Risko, J.-L. Brédas, “Intermixing at the Pentacene-Fullerene Bilayer Interface: A Molecular Dynamics Study”, *Advanced Materials* **2013**, *25*, 878–882.

[69] K. Chenoweth, A. C. T. van Duin, W. A. Goddard, “ReaxFF Reactive Force Field for Molecular Dynamics Simulations of Hydrocarbon Oxidation”, *The Journal of Physical Chemistry A* **2008**, *112*, 1040–1053.

[70] L. Yang, C.-h. Tan, M.-J. Hsieh, et al., “New-Generation Amber United-Atom Force Field”, *The Journal of Physical Chemistry B* **2006**, *110*, 13166–13176.

[71] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, A. H. de Vries, “The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations”, *The Journal of Physical Chemistry B* **2007**, *111*, 7812–7824.

[72] T. Sun, A. Mirzoev, V. Minhas, et al., “A multiscale analysis of DNA phase separation: from atomistic to mesoscale level”, *Nucleic Acids Research* **2019**, *47*, 5550–5562.

[73] C.-K. Lee, C.-W. Pao, “Nanomorphology Evolution of P3HT/PCBM Blends during Solution-Processing from Coarse-Grained Molecular Simulations”, *The Journal of Physical Chemistry C* **2014**, *118*, 11224–11233.

[74] N. Berton, F. Lemasson, A. Poschlad, et al., “Selective Dispersion of Large-Diameter Semiconducting Single-Walled Carbon Nanotubes with Pyridine-Containing Copolymers”, *Small* **2014**, *10*, 360–367.

[75] P. Español, P. B. Warren, “Perspective: Dissipative particle dynamics”, *The Journal of Chemical Physics* **2017**, *146*, 150901.

[76] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, et al., “The power of coarse graining in biomolecular simulations”, *WIREs Computational Molecular Science* **2014**, *4*, 225–248.

[77] F. Müller-Plathe, “Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back”, *ChemPhysChem* **2002**, *3*, 754–769.

[78] M. E. Tuckerman, “Ab initiomolecular dynamics: basic concepts, current trends and novel applications”, *Journal of Physics: Condensed Matter* **2002**, *14*, R1297–R1355.

[79] H. M. Senn, W. Thiel, “QM/MM Methods for Biomolecular Systems”, *Angewandte Chemie International Edition* **2009**, *48*, 1198–1229.

[80] H. Fujitani, A. Matsuura, S. Sakai, H. Sato, Y. Tanida, “High-Level ab Initio Calculations To Improve Protein Backbone Dihedral Parameters”, *Journal of Chemical Theory and Computation* **2009**, *5*, 1155–1165.

[81] F. Symalla, P. Friederich, A. Massé, et al., “Charge Transport by Superexchange in Molecular Host-Guest Systems”, *Physical Review Letters* **2016**, *117*, 276803.

[82] Z. Li, J. R. Kermode, A. De Vita, “Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces”, *Physical Review Letters* **2015**, *114*, 096405.

[83] P. Hohenberg, W. Kohn, “Inhomogeneous Electron Gas”, *Physical Review* **1964**, *136*, B864–B871.

[84] W. Kohn, L. J. Sham, “Self-Consistent Equations Including Exchange and Correlation Effects”, *Physical Review* **1965**, *140*, A1133–A1138.

[85] J. P. Perdew, W. Yue, “Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation”, *Physical Review B* **1986**, *33*, 8800–8802.

[86] A. D. Becke, “Density-functional thermochemistry. III. The role of exact exchange”, *The Journal of Chemical Physics* **1993**, *98*, 5648–5652.

[87] C. Lee, W. Yang, R. G. Parr, “Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density”, *Physical Review B* **1988**, *37*, 785–789.

[88] S. H. Vosko, L. Wilk, M. Nusair, “Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis”, *Canadian Journal of Physics* **1980**, *58*, 1200–1211.

[89] U. C. Singh, P. A. Kollman, “An approach to computing electrostatic charges for molecules”, *Journal of Computational Chemistry* **1984**, *5*, 129–145.

[90] C. M. Breneman, K. B. Wiberg, “Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis”, *Journal of Computational Chemistry* **1990**, *11*, 361–373.

[91] H. Hu, Z. Lu, W. Yang, “Fitting Molecular Electrostatic Potentials from Quantum Mechanical Calculations”, *Journal of Chemical Theory and Computation* **2007**, *3*, 1004–1013.

[92] C. Campañá, B. Mussard, T. K. Woo, “Electrostatic Potential Derived Atomic Charges for Periodic Systems Using a Modified Error Functional”, *Journal of Chemical Theory and Computation* **2009**, *5*, 2866–2878.

[93] A. Stukowski, “Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool”, *Modelling and Simulation in Materials Science and Engineering* **2009**, *18*, 015012.

[94] F. L. Hirshfeld, “Bonded-atom fragments for describing molecular charge densities”, *Theoretica Chimica Acta* **1977**, *44*, 129–138.

[95] I. Mayer, “Charge, bond order and valence in the AB initio SCF theory”, *Chemical Physics Letters* **1983**, *97*, 270–274.

[96] S. Grimme, “Accurate description of van der Waals complexes by density functional theory including empirical corrections”, *Journal of Computational Chemistry* **2004**, *25*, 1463–1473.

[97] S. Grimme, “Semiempirical GGA-type density functional constructed with a long-range dispersion correction”, *Journal of Computational Chemistry* **2006**, *27*, 1787–1799.

[98] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, “A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu”, *The Journal of Chemical Physics* **2010**, *132*, 154104.

[99] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, B. I. Lundqvist, “Van der Waals Density Functional for General Geometries”, *Physical Review Letters* **2004**, *92*, 246401.

[100] D. Frenkel, B. Smit, *Understanding molecular simulation: from algorithms to applications*, 2nd ed, Academic Press, San Diego, **2002**, 638 pp.

[101] D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, et al., *AMBER 2019, University of California, San Francisco*, **2019**.

[102] S. Nosé, “A unified formulation of the constant temperature molecular dynamics methods”, *The Journal of Chemical Physics* **1984**, *81*, 511–519.

[103] W. G. Hoover, “Canonical dynamics: Equilibrium phase-space distributions”, *Physical Review A* **1985**, *31*, 1695–1697.

[104] M. Parrinello, A. Rahman, “Polymorphic transitions in single crystals: A new molecular dynamics method”, *Journal of Applied Physics* **1981**, *52*, 7182–7190.

[105] W. Shinoda, M. Shiga, M. Mikami, “Rapid estimation of elastic constants by molecular dynamics simulation under constant stress”, *Physical Review B* **2004**, *69*, 134103.

[106] C. Caleman, P. J. van Maaren, M. Hong, et al., “Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant”, *Journal of Chemical Theory and Computation* **2012**, *8*, 61–74.

[107] D. P. Kingma, J. Ba, “Adam: A Method for Stochastic Optimization”, *arXiv: 1412.6980 [cs]* **2017**.

[108] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning”, *Physical Review Letters* **2012**, *108*, 058301.

[109] S. De, A. P. Bartók, G. Csányi, M. Ceriotti, “Comparing molecules and solids across structural and alchemical space”, *Physical Chemistry Chemical Physics* **2016**, *18*, 13754–13769.

[110] P. Reiser, M. Konrad, A. Fediai, et al., “Analyzing Dynamical Disorder for Charge Transport in Organic Semiconductors via Machine Learning”, *Journal of Chemical Theory and Computation* **2021**, *https://pubs.acs.org/doi/10.1021/acs.jctc.1c00191*.

[111] C. W. Tang, S. A. VanSlyke, “Organic electroluminescent diodes”, *Applied Physics Letters* **1987**, *51*, 913–915.

[112] B. Geffroy, P. l. Roy, C. Prat, “Organic light-emitting diode (OLED) technology: materials, devices and display technologies”, *Polymer International* **2006**, *55*, 572–582.

[113] G. Zhang, J. Zhao, P. C. Y. Chow, et al., “Nonfullerene Acceptor Molecules for Bulk Heterojunction Organic Solar Cells”, *Chemical Reviews* **2018**, *118*, 3447–3507.

[114] H. E. Katz, Z. Bao, S. L. Gilat, “Synthetic Chemistry for Ultrapure, Processable, and High-Mobility Organic Transistor Semiconductors”, *Accounts of Chemical Research* **2001**, *34*, 359–369.

[115] J. Czolk, D. Landerer, M. Koppitz, D. Nass, A. Colsmann, “Highly Efficient, Mechanically Flexible, Semi-Transparent Organic Solar Cells Doctor Bladed from Non-Halogenated Solvents”, *Advanced Materials Technologies* **2016**, *1*, 1600184.

[116] C. Groves, “Simulating charge transport in organic semiconductors and devices: a review”, *Reports on Progress in Physics* **2016**, *80*, 026502.

[117] A. Miller, E. Abrahams, “Impurity Conduction at Low Concentrations”, *Physical Review* **1960**, *120*, 745–755.

[118] R. A. Marcus, “Electron transfer reactions in chemistry. Theory and experiment”, *Reviews of Modern Physics* **1993**, *65*, 599–610.

[119] H. Bässler, “Charge Transport in Disordered Organic Photoconductors a Monte Carlo Simulation Study”, *physica status solidi (b)* **1993**, *175*, 15–56.

[120] H. Bässler, A. Köhler in *Unimolecular and Supramolecular Electronics I: Chemistry and Physics Meet at Metal-Molecule Interfaces*, (Ed.: R. M. Metzger), Topics in Current Chemistry, Springer, Berlin, Heidelberg, **2012**, pp. 1–65.

[121] H. Bässler, A. Köhler in *Organic Light-Emitting Diodes (OLEDs)*, (Ed.: A. Buckley), Woodhead Publishing Series in Electronic and Optical Materials, Woodhead Publishing, **2013**, pp. 192–234.

[122] P. Friederich, V. Meded, A. Poschlad, et al., “Molecular Origin of the Charge Carrier Mobility in Small Molecule Organic Semiconductors”, *Advanced Functional Materials* **2016**, *26*, 5757–5763.

[123] P. Friederich, V. Gómez, C. Sprau, et al., “Rational In Silico Design of an Organic Semiconductor with Improved Electron Mobility”, *Advanced Materials* **2017**, *29*, 1703505.

[124] D. Andrienko in *Handbook of Materials Modeling : Methods: Theory and Modeling*, (Eds.: W. Andreoni, S. Yip), Springer International Publishing, Cham, **2018**, pp. 1–12.

[125] A. Fediai, F. Symalla, P. Friederich, W. Wenzel, “Disorder compensation controls doping efficiency in organic semiconductors”, *Nature Communications* **2019**, *10*, 4547.

[126] P. Friederich, A. Fediai, S. Kaiser, et al., “Toward Design of Novel Materials for Organic Electronics”, *Advanced Materials* **2019**, *31*, 1808256.

[127] F. Symalla, A. Fediai, J. Armleder, et al., “43-3: Ab-initio Simulation of Doped Injection Layers.”, *SID Symposium Digest of Technical Papers* **2020**, *51*, 630–633.

[128] A. Fediai, A. Emering, F. Symalla, W. Wenzel, “Disorder-driven doping activation in organic semiconductors”, *Physical Chemistry Chemical Physics* **2020**, *22*, 10256–10264.

[129] T. Neumann, D. Danilov, C. Lennartz, W. Wenzel, “Modeling disordered morphologies in organic semiconductors”, *Journal of Computational Chemistry* **2013**, *34*, 2716–2725.

[130] P. Friederich, F. Symalla, V. Meded, T. Neumann, W. Wenzel, “Ab Initio Treatment of Disorder Effects in Amorphous Organic Materials: Toward Parameter Free Materials Simulation”, *Journal of Chemical Theory and Computation* **2014**, *10*, 3720–3725.

[131] J. Kirkpatrick, V. Marcon, J. Nelson, K. Kremer, D. Andrienko, “Charge Mobility of Discotic Mesophases: A Multiscale Quantum and Classical Study”, *Physical Review Letters* **2007**, *98*, 227402.

[132] W. F. Pasveer, J. Cottaar, C. Tanase, et al., “Unified Description of Charge-Carrier Mobilities in Disordered Semiconducting Polymers”, *Physical Review Letters* **2005**, *94*, 206601.

[133] M. Bouhassoune, S. L. M. v. Mensfoort, P. A. Bobbert, R. Coehoorn, “Carrier-density and field-dependent charge-carrier mobility in organic semiconductors with correlated Gaussian disorder”, *Organic Electronics* **2009**, *10*, 437–445.

[134] A. Massé, P. Friederich, F. Symalla, et al., “Ab initio charge-carrier mobility model for amorphous molecular semiconductors”, *Physical Review B* **2016**, *93*, 195209.

[135] V. Rodin, F. Symalla, V. Meded, et al., “Generalized effective-medium model for the carrier mobility in amorphous organic semiconductors”, *Physical Review B* **2015**, *91*, 155203.

[136] T. Vehoff, Y. S. Chung, K. Johnston, et al., “Charge Transport in Self-Assembled Semiconducting Organic Layers: Role of Dynamic and Static Disorder”, *The Journal of Physical Chemistry C* **2010**, *114*, 10592–10597.

[137] N. R. Tummala, Z. Zheng, S. G. Aziz, V. Coropceanu, J.-L. Brédas, “Static and Dynamic Energetic Disorders in the C60, PC61BM, C70, and PC71BM Fullerenes”, *The Journal of Physical Chemistry Letters* **2015**, *6*, 3657–3662.

[138] Z. Zheng, N. R. Tummala, T. Wang, V. Coropceanu, J.-L. Brédas, “Charge-Transfer States at Organic–Organic Interfaces: Impact of Static and Dynamic Disorders”, *Advanced Energy Materials* **2019**, *9*, 1803926.

[139] G. Kupgan, X.-K. Chen, J.-L. Brédas, “Low Energetic Disorder in Small-Molecule Non-Fullerene Electron Acceptors”, *ACS Materials Letters* **2019**, *1*, 350–353.

[140] P. Friederich, S. León, J. D. Perea, L. M. Roch, A. Aspuru-Guzik, “The influence of sorbitol doping on aggregation and electronic properties of PEDOT:PSS: a theoretical study”, *Machine Learning: Science and Technology* **2020**, *2*, 01LT01.

[141] C. Bannwarth, E. Caldeweyher, S. Ehlert, et al., “Extended tight-binding quantum chemistry methods”, *WIREs Computational Molecular Science* **2021**, *11*, e1493.

[142] P. Pracht, F. Bohle, S. Grimme, “Automated exploration of the low-energy chemical space with fast quantum chemical methods”, *Physical Chemistry Chemical Physics* **2020**, *22*, 7169–7192.

[143] S. Grimme, “Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations”, *Journal of Chemical Theory and Computation* **2019**, *15*, 2847–2862.

[144] A. K. Malde, L. Zuo, M. Breeze, et al., “An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0”, *Journal of Chemical Theory and Computation* **2011**, *7*, 4026–4037.

[145] M. Stroet, B. Caron, K. M. Visscher, et al., “Automated Topology Builder Version 3.0: Prediction of Solvation Free Enthalpies in Water and Hexane”, *Journal of Chemical Theory and Computation* **2018**, *14*, 5834–5845.

[146] M. Konrad, W. Wenzel, “CONI-Net: Machine Learning of Separable Intermolecular Force Fields”, *Journal of Chemical Theory and Computation* **2021**, *https://pubs.acs.org/doi/10.1021/acs.jctc.1c00328*.

[147] F. Neese, “The ORCA program system”, *WIREs Computational Molecular Science* **2012**, *2*, 73–78.

[148] R. A. Kendall, T. H. Dunning, R. J. Harrison, “Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions”, *The Journal of Chemical Physics* **1992**, *96*, 6796–6806.

[149] T. Lu, F. Chen, “Multiwfn: A multifunctional wavefunction analyzer”, *Journal of Computational Chemistry* **2012**, *33*, 580–592.

[150] T. H. Dunning, “Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen”, *The Journal of Chemical Physics* **1989**, *90*, 1007–1023.

[151] E. G. Hohenstein, C. D. Sherrill, “Density fitting of intramonomer correlation effects in symmetry-adapted perturbation theory”, *The Journal of Chemical Physics* **2010**, *133*, 014101.

[152] E. G. Hohenstein, C. D. Sherrill, “Efficient evaluation of triple excitations in symmetry-adapted perturbation theory via second-order Møller–Plesset perturbation theory natural orbitals”, *The Journal of Chemical Physics* **2010**, *133*, 104107.

[153] R. M. Parrish, L. A. Burns, D. G. A. Smith, et al., “Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability”, *Journal of Chemical Theory and Computation* **2017**, *13*, 3185–3197.

[154] J. Mei, J. W. Davenport, G. W. Fernando, “Analytic embedded-atom potentials for fcc metals: Application to liquid and solid copper”, *Physical Review B* **1991**, *43*, 4653–4658.

[155] A. Paszke, S. Gross, F. Massa, et al. in Advances in Neural Information Processing Systems, *Vol. 32*, (Eds.: H. Wallach, H. Larochelle, A. Beygelzimer, et al.), Curran Associates, Inc., **2019**, pp. 8026–8037.

[156] C. R. Harris, K. J. Millman, S. J. van der Walt, et al., “Array programming with NumPy”, *Nature* **2020**, *585*, 357–362.

[157] SciPy 1.0 Contributors, P. Virtanen, R. Gommers, et al., “SciPy 1.0: fundamental algorithms for scientific computing in Python”, *Nature Methods* **2020**, *17*, 261–272.

[158] A. Heßelmann, “Correlation effects and many-body interactions in water clusters”, *Beilstein Journal of Organic Chemistry* **2018**, *14*, 979–991.

[159] C. K. Egan, F. Paesani, “Assessing Many-Body Effects of Water Self-Ions. II: H3O+(H2O)n Clusters”, *Journal of Chemical Theory and Computation* **2019**, *15*, 4816–4833.

[160] H. S. Seung, M. Opper, H. Sompolinsky in Proceedings of the fifth annual workshop on Computational learning theory - COLT '92, the fifth annual workshop, ACM Press, Pittsburgh, Pennsylvania, United States, **1992**, pp. 287–294.

[161] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, "Less is more: Sampling chemical space with active learning", *The Journal of Chemical Physics* **2018**, *148*, 241733.

[162] L.-P. Wang, T. Head-Gordon, J. W. Ponder, et al., "Systematic Improvement of a Classical Molecular Model of Water", *The Journal of Physical Chemistry B* **2013**, *117*, 9956–9972.

[163] *CRC handbook of chemistry and physics: a ready-reference book of chemical and physical data*, 85. ed, (Eds.: C. R. Company, D. R. Lide), OCLC: 249875978, CRC Press, Boca Raton, **2004**.

[164] Y. Marcus, *The properties of solvents*, Wiley, Chichester ; New York, **1998**, 239 pp.

[165] J. Konicek, I. Wadsö, J. Munch-Petersen, R. Ohlson, A. Shimizu, "Enthalpies of Vaporization of Organic Compounds. VII. Some Carboxylic Acids.", *Acta Chemica Scandinavica* **1970**, *24*, 2612–2616.

[166] A. Forziati, D. Camin, F. Rossini, "Density, refractive index, boiling point, and vapor pressure of eight monoolefin (1-alkene), six pentadiene, and two cyclomonoolefin hydrocarbons", *Journal of Research of the National Bureau of Standards* **1950**, *45*, 406.

[167] M. W. Lister, "Heats of Organic Reactions. X. Heats of Bromination of Cyclic Olefins", *Journal of the American Chemical Society* **1941**, *63*, 143–149.

[168] R. C. Reid, "Handbook on vapor pressure and heats of vaporization of hydrocarbons and related compounds, R. C. Wilhort and B. J. Zwolinski, Texas A&M Research Foundation. College Station, Texas(1971). 329 pages.$10.00", *AIChE Journal* **1972**, *18*, 1278–1278.

[169] F. J. Krieger, H. H. Wenzke, "The Dielectric Properties of Acetylenic Compounds. X. Equipment for Measuring Dielectric Constants of Gases. The Polarity of Gaseous Monoalkyl Acetylenes", *Journal of the American Chemical Society* **1938**, *60*, 2115–2119.

[170] J. Ott, R. Grigg, J. Goates, "Excess enthalpies and excess volumes for n-hexane + 2-methylpentane, + 3-methylpentane and + 2,3-dimethylbutane at 283.15, 298.15 and 313.15 K", *Australian Journal of Chemistry* **1980**, *33*, 1921.

[171] G. Beech, R. M. Lintonbon, "The measurement of sublimation enthalpies by differential scanning calorimetry", *Thermochimica Acta* **1971**, *2*, 86–88.

[172] J. Hales, R. Townsend, “Liquid densities from 293 to 490 K of nine aromatic hydrocarbons”, *The Journal of Chemical Thermodynamics* **1972**, *4*, 763–772.

List of Publications

L

Articles

- P. Friederich, M. Konrad, T. Strunk and W. Wenzel
  Machine Learning of Correlated Dihedral Potentials for Atomistic Molecular Force Fields
  Scientific Reports, 2018, **8**, 2559

- C. N. Shyam Kumar, M. Konrad, V. S. Kiran Chakravadhanula, S. Dehm, D. Wang, W. Wenzel, R. Krupke and C. Kübel
  Nanocrystalline graphene at high temperatures: insight into nanoscale processes
  Nanoscale Advances, 2019, **1** (7), 2485–2494

- P. Friederich, A. Fediai, S. Kaiser, M. Konrad, N. Jung and W. Wenzel
  Toward Design of Novel Materials for Organic Electronics
  Advanced Materials, 2019, **31** (26), 1808256

- M. Konrad and W. Wenzel
  CONI-Net: Machine Learning of Separable Intermolecular Force Fields
  Journal of Chemical Theory and Computation, 2021
  https://pubs.acs.org/doi/10.1021/acs.jctc.1c00328

- P. Reiser, M. Konrad, A. Fediai, S. León, W. Wenzel and P. Friederich
  Analyzing Dynamical Disorder for Charge Transport in Organic Semiconductors via Machine Learning
  Journal of Chemical Theory and Computation, 2021
  https://pubs.acs.org/doi/10.1021/acs.jctc.1c00191

Submitted

- C. Degitz, M. Konrad, S. Kaiser and W. Wenzel
  Simulating the growth of amorphous organic thin films
  Submitted

- S. Bag, M. Konrad, T. Schlöder, P. Friederich and W. Wenzel
  Fast Generation of Machine Learning Based Force Fields for Adsorption Energies
  Submitted

Other Contributions

- M. Konrad, C. N. Shyam Kumar, W. Wenzel and C. Kübel
  Semantic Segmentation for High-Throughput Image Analysis
  STN Day 2018 (poster)

- M. Konrad, S. Kaiser, F. Symalla and W. Wenzel
  Influence of mesoscopic structure on device properties of organic solar cells
  PISACMS 2018: Paris International School on Advanced Computational Materials Science (poster)

- C. N. Shyam Kumar, M. Konrad, W. Wenzel, R. Krupke and C. Kübel
  Ostwald-like Ripening in Highly Defective Graphene
  Imaging & Microscopy 3/19 (non-peer-reviewed article)

- M. Konrad and W. Wenzel
  Silizium-Anoden für Lithium-Batterien
  Virtueller Forschungstag 2020, Baden-Württemberg Stiftung (digital poster)

- M. Konrad and W. Wenzel
  Separable intermolecular force fields from first principles
  Virtual DPG Spring Meeting 2021 of the divisions Biological Physics, Chemical and Polymer Physics, Dynamics and Statistical Physics, and Physics of Socio-economic Systems (digital poster)

- Project "Multi-Skalen-Modellierung von Materialien und Bauelementen für die Energieumwandlung" (MSMEE) of the Baden-Württemberg Stiftung: Collaboration meetings with the project partners of the University of Ulm and development of a parallelization algorithm to accelerate grand-canonical Monte Carlo simulations of battery electrodes.

Acknowledgment

In the midst of the Corona pandemic and after months in home office, I am especially thankful for all the support I could experience over the last years working at the KIT.

First, I want to thank Prof. Wolfgang Wenzel for supervising and enabling this work. Wolfgang's mentoring and support over the last years allowed me to follow my ideas, and he always found time for lively scientific discussions.

Special thanks also go to Prof. Lars Pastewka for being the second referee of this work.

I want to thank all my colleagues responsible for my wonderful experience during my time in the group. Particularly, my office mates Carl Degitz and Simon Kaiser, who always provided a good mood and were open for scientific and non-scientific discussions. I also want to thank all my other colleages who supported me during my time at the INT, especially Manuel Rommel, Pascal Friederich, Timo Strunk, Tobias Neumann, Franz Symalla, Velimir Meded, Modan Liu, Jorge Olivares Pena, Clemens Possel, Shahriar Heidrich, Tobias Schlöder, Montserrat Penaloza-Amion, Anna Mauri, Jörg Schaarschmidt, Yohanes Pramudya, Mariana Kozlowska, Heike Fliegl, Saientan Bag, Patrick Reiser, Artem Fediai, Deniz Özdemir and Praneeth Chilakalapudi. Also, I want to thank Shyam Kumar for our enjoyable collaboration. Furthermore, I want to thank the students and organizers of the RTG 2450 for the very nice workshops and events during my membership.

I want to thank all my friends for making my time in Karlsruhe an awesome and unforgettable experience. Furthermore, I am very grateful for my family's support during all stages of my life and career. Finally, I want to thank my girlfriend Esther for her unconditional support and encouragement during my studies and Ph.D. years.