Incorporating long-range interactions and polarization in machine learning potentials with explicit electrons

De beschrijving van langeafstandsinteracties en polarisatie in machinaal geleerde potentialen met expliciete elektronen

Maarten Cools-Ceuppens

Supervisors: prof. dr. ir. T. Verstraelen, prof. dr. ir. J. Dambre
Dissertation submitted in fulfillment of the requirements for the degree of Doctor (Ph.D.) of Science: Physics

Department of Physics and Astronomy
Department chair: prof. dr. Dimitri Van Neck
Ghent University
Academic year 2021–2022

GHENT UNIVERSITY

Members of the Examination Committee

Chair

prof. dr. Natalie Jachowicz (Ghent University)

Reading Committee

dr. Matti Hellström (Software for Chemistry & Materials, The Netherlands)
prof. dr. Paul W. Ayers (McMaster University, Canada)
prof. dr. ir. Louis Vanduyfhuys (Ghent University)
dr. Jelle Vekeman (Ghent University, *secretary*)

Supervisors

prof. dr. ir. Toon Verstraelen (Ghent University)
prof. dr. ir. Joni Dambre (Ghent University)

This research has been conducted at the **Center for Molecular Modeling**.

Voorwoord

If I have seen further it is by standing on the shoulders of Giants.
Sir Isaac Newton (1642–1727)

Ongeveer zes jaar geleden heb ik als student de keuze gemaakt om mijn masterthesis af te leggen in het Centrum voor Moleculaire Modellering. Hier kwam ik zowel in aanraking met de wonderlijke wereld van atomaire simulaties als met een geweldig team van begeleiders en promotoren. Toen ik de kans aangeboden kreeg om daar mijn doctoraat uit te voeren, heb ik onmiddellijk toegestemd. Nu, na een periode van vijf jaar, ben ik met het schrijven van deze thesis aan het einde van mijn doctoraat gekomen. Het was een periode in mijn leven die ik me altijd zal blijven herinneren vanwege het interessant onderzoek maar ook vanwege de aangename interacties met mijn vele collega's.

Dit doctoraat zou ik nooit tot een goed einde hebben gebracht zonder de hulp van mijn vele collega's, vrienden en familie. Dit is dan ook het ideale moment om jullie allemaal te bedanken. In eerste instantie mijn beide promotoren Toon en Joni. Vanaf het begin, bij het schrijven van mijn FWO aanvraag, hebben jullie mij bijgestaan en de mogelijkheid gegeven om zelfstandig onderzoek te verrichten. Toon, bij jou kon ik altijd terecht als ik vast zat bij het ontwikkelen van één of ander model en door je zeer uitgebreide computationele kennis en ervaring, kon je me telkens weer de meest geschikte oplossing aanreiken. Je kwam keer op keer af met vernieuwende ideeën die uiteindelijk hebben geleid tot onze eMLP paper. Bovendien zal de avond waarop je voor al je doctoraatsstudenten hotpot gemaakt had, altijd een aangename beleving blijven. Joni, ook al zagen we elkaar niet zo veel vanaf de coronaperiode, toch hielp je me op weg met alles wat met machine learning te maken had en las je ons werk kritisch na vanuit jouw wetenschappelijk perspectief, wat een zeer nuttig aanvulling bleek te zijn.

Verder kan ik natuurlijk mijn bureaugenoten niet vergeten. Alexander, Aran en Sander, we kenden elkaar al als student, maar ik durf te zeggen dat onze vriendschap alleen maar sterker is geworden door vijf jaar bijna dagelijks in hetzelfde kantoor te vertoeven. Als ik werkgerelateerde vragen had, was er altijd wel iemand van jullie die wist hoe ik verder kon. Maar ook bedankt voor de sfeer die jullie in ons kantoor brachten. Dit wordt alleen maar bevestigd door sommige andere collega's die te pas en te onpas onze bureau komen binnenwandelen om eens te babbelen. Buiten de werksfeer, heb ik met jullie een paar onvergetelijke Italiëreizen beleefd met talloze herinneringen die me ongetwijfeld altijd zullen bijblijven.

Bedankt Michael en Michiel. Samen introduceerden we SchNet op het CMM en konden we soms lange tijd discussiëren over allerhande machine learning zaken. Bedankt Sander. Mede dankzij jou zijn machine learning potentialen ingeburgerd geraakt op het CMM en ik hoop dat onze paper binnenkort (eindelijk) gepubliceerd zal raken. Ook bedankt aan Sven om op gepaste tijdstippen leven in de brouwerij te brengen met of zonder happy socks. Veronique, ik wil je ook expliciet bedanken voor de samenwerking en hulp vanuit jouw groep. Uiteraard wil ik ook nog alle andere collega's binnen het CMM bedanken. Tijdens de middag was het altijd amusant om met jullie samen te eten en te discussiëren. Ik wens jullie allemaal nog zeer veel geluk (en papers) in de komende jaren.

Bedankt Kurt en Jelle. Ik leerde jullie kennen als ervaren collega's binnen het CMM die altijd klaar stonden met goede raad. Samen met mijn bureaugenoten, waren jullie een aangenaam gezelschap tijdens de talloze spelletjesavonden. Tijdens de coronaperiode gingen deze echter noodgedwongen online door, maar dit maakte ze niet minder plezant.

Ten slotte wil ik hier ook mijn ouders bedanken voor de steun die ze mij hebben gegeven. Zonder jullie zou ik dit nooit gerealiseerd hebben. Bovendien weet ik nu dat het eenvoudig uitleggen van mijn onderzoek, soms lastiger is dan het lijkt.

Maarten Cools-Ceuppens
Gent, september 2022

Contents

Samenvatting

Moleculair modelleren is de theoretische en computationele studie van atomaire systemen onder verschillende fysische omstandigheden. Gedurende de laatste vijftig jaar heeft het een enorme impact op de wetenschap gehad omdat, behalve experimenten, nu ook numerieke simulaties kunnen helpen door het gedrag van materie op atomaire schaal te voorspellen. Experimentele waarnemingen kunnen nu theoretisch verklaard worden of volledig worden vervangen door numerieke simulaties in omstandigheden die niet toelaten om kwantitatieve of nauwkeurige experimenten uit te voeren. Bovendien laten nieuwe geavanceerde technieken zoals high-throughput screening ons toe om fysische eigenschappen te voorspellen voor allerlei atomaire systemen, waaronder hypothetische materialen. Moleculair modelleren is tegenwoordig een vast begrip in de fysica, chemie en biologie door toepassingen zoals het ontwikkelen van nieuwe materialen en medicijnen of de studie van biomoleculaire processen.

In moleculair modelleren worden atomaire systemen beschreven op het niveau van individuele atomen. Op deze lengteschalen zijn kwantum mechanische effecten niet meer te verwaarlozen. Daarvoor moet de elektronische Schrödingervergelijking opgelost worden, wat een moeilijke en tijdrovende opdracht is omdat de oplossing, namelijk de elektronische golffunctie, een ingewikkelde multidimensionale functie is. Als deel van de oplossing, kunnen de energie en krachten die op de atoomkernen inwerken, gebruikt worden om het dynamische gedrag van atomaire systemen te simuleren. Zodoende moet men de Schrödingervergelijking herhaaldelijk oplossen om deze systemen gedurende een lange tijd te bestuderen. Doorgaans maken de meeste algoritmen een afweging tussen nauwkeurigheid en rekenefficiëntie, wat benadrukt zal worden in **Hoofdstuk 1** van dit werk.

Er bestaan verschillende benaderingstechnieken om de Schrödingervergelijking op te lossen. In first-principles-methoden wordt de elektronische golffunctie of dichtheid rechtstreeks berekend zonder experimentele informatie

te gebruiken. Deze methoden zijn relatief nauwkeurig maar niet echt toepasbaar op grote systemen vanwege hun rekenkost en schaling. Krachtvelden daarentegen beschrijven de elektronische vrijheidsgraden niet meer expliciet en benaderen de energie als een reeks van fysisch geïnspireerde analytische uitdrukkingen. Bovendien schaalt hun rekenkost lineair met het aantal kernen, waardoor simulaties van grote systemen en lange tijdsduur mogelijk worden. Polarisatie en ladingstransfer kunnen echter niet goed beschreven worden met conventionele krachtvelden waardoor ze niet nauwkeurig genoeg zijn voor meer geavanceerde toepassingen. Daarom worden in polarizeerbare krachtvelden de elektronische vrijheidsgraden op een benaderende manier terug ingevoerd, dikwijls door aan elke kern een fluctuerende lading toe te kennen. In **Hoofdstuk 2** zullen we echter aantonen dat fluctuerende ladingen leiden tot een ambigue definitie van de polarizeerbaarheid in periodieke systemen. Expliciete elektronenkrachtvelden vermijden dit probleem door de elektronen te modelleren als semi-klassieke deeltjes met een vaste en gehele lading. Het karakteriseren van de verschillende energiebijdragen tussen de elektronendeeltjes en de kernen in deze soort krachtvelden blijft echter moeilijk omdat ze gedomineerd worden door verschillende kwantumeffecten zoals uitwisselingsinteracties. De expliciete elektronenkrachtvelden vormen de inspiratie voor de elektron machine-learning potentiaal (eMLP), waar we in dit werk naartoe zullen bouwen.

In **Hoofdstuk 3** introduceren we machine-learning potentialen (MLP's). De laatste jaren zijn ze een revolutie aan het teweegbrengen op het gebied van moleculair modelleren omdat ze de nauwkeurigheid van first-principles methoden benaderen en bijna even snel zijn als conventionele krachtvelden. MLP's verwezenlijken dit door de energie en krachten te voorspellen met een machine-learning algoritme (kernelregressie, neurale netwerken . . .) dat duizenden of zelfs miljoenen leerbare parameters heeft, die allemaal gefit worden aan first-principles data. In tegenstelling tot krachtvelden, kunnen ze de meest complexe interacties, waaronder chemische reacties, automatisch aanleren zonder dat hiervoor nog een manuele tussenkomst nodig is. Hoewel ze korteafstandsinteracties nauwkeurig kunnen modelleren, blijft het een open vraagstuk hoe langeafstandsinteracties en polarisatie op een correcte manier kunnen geïntegreerd worden in MLP's. Indien men ze toevoegt, worden typisch fluctuerende ladingen gebruikt die natuurlijk lijden aan het hierboven genoemde probleem. Daarom zullen we expliciete elektronenkrachtvelden en MLP's combineren om de state-of-the-art te verleggen met de eMLP, als het eerste doel van dit werk.

De volledige werking van de eMLP zal uiteengezet worden in **Hoofdstuk 4**. De eMLP is een expliciet elektronenkrachtveld waar de ingewikkelde korteafstandsinteracties worden aangeleerd met een MLP. Naast de atoom-

kernen, zullen de elektronenparen (die een spin-up en een spin-down elektron bevatten) extra deeltjes zijn die niet begrensd worden tot de atoomkernen maar de mogelijkheid krijgen om zich vrij te bewegen over het hele systeem. Deze elektronenpaar-deeltjes zullen zich in de centra van gelokaliseerde orbitalen bevinden, die afgeleid zijn van first-principles berekeningen. Een gevolg hiervan is dat de MLP automatisch het juiste first-principle dipoolmoment en andere elektronische eigenschappen zal aanleren. De nauwkeurigheid van de eMLP zal getest worden door kleine organische moleculen en het kristallijne systeem β-glycine te modelleren. Daartoe hebben we twee nieuwe datasets ontworpen. Er zal aangetoond worden dat er slechts kleine kracht- en energiefouten gemaakt worden en dat infraroodspectra van ongekende moleculen voorspeld kunnen worden, wat de transfereerbaarheid van de eMLP demonstreert. Verder kunnen de elasticiteit, diëlektrische en piëzo-elektrische tensor van β-glycine met een hoge nauwkeurigheid worden berekend. Gedurende de ontwikkeling van de eMLP, merkten we ook een zwak punt van ons model op. Dynamische simulaties zijn niet stabiel voor een minderheid van de ongekende moleculen en er treedt een drift op in de behouden grootheid van de simulatie. Dit zal opgelost worden door data-augmentatie, een machine-learning techniek die geometrieën van elektronenparen uit evenwicht genereert terwijl het netwerk wordt getraind door te leren dat ze onfysisch zijn.

Als tweede doel van dit werk, zullen we een nieuw dataprotocol ontwikkelen om MLP's af te leiden voor metaal-organische roosters (MOF's) aan het eind van **Hoofdstuk 3**. Deze materialen krijgen steeds meer aandacht in de wetenschappelijke gemeenschap vanwege hun aantrekkelijke eigenschappen, waaronder absorptie van moleculen, katalyse of faseovergangen. State-of-the-art MLP's voor MOF's vereisen steeds een dataset van meer dan tienduizend first-principles berekeningen, wat veel rekenkracht vraagt voor deze grote systemen. We zullen echter de benodigde hoeveelheid data reduceren tot enkele honderden configuraties en tegelijkertijd de fouten met meer dan een factor drie verminderen ten opzichte van andere beschikbare modellen in de literatuur. Dit alles wordt mogelijk gemaakt door ons voorgestelde dataprotocol en de nieuwe klasse van data-efficiënte equivariante MLP's.

Als besluit vermelden we dat we de state-of-the-art van MLP's in twee domeinen hebben uitgebreid. Ten eerste is ons dataprotocol voor MOF's meer data-efficiënt en nauwkeuriger dan andere beschikbare MLP's voor MOF's. Ten tweede hebben we de eMLP ontwikkeld waarin langeafstandinteracties en polarisatie ingebouwd zijn. Deze bevindingen zullen worden samengevat in **Hoofdstuk 5**. Bovendien zullen we daar enkele toekomstperspectieven geven over hoe de eMLP kan worden uitgebreid om meer algemeen toepasbaar te zijn.

Summary

Molecular modeling is the theoretical and computational study of atomic systems under varying physical conditions. During the last fifty years, it had an enormous impact on science because, besides experiment, numerical simulations could finally aid the prediction of the atomistic behavior of matter. Experimental observations can now be explained theoretically or completely replaced by numerical simulations in circumstances that do not allow for quantitative or accurate experiments. Moreover, advanced tools such as high-throughput screening became available in the last decade, enabling the prediction of physical properties for an abundance of atomic systems, including hypothetical materials. Nowadays, molecular modeling is a well-known concept in physics, chemistry and biology thanks to applications such as the design of new materials and drugs or the study of biomolecular processes.

In molecular modeling, atomic systems are described at the level of individual atoms. At these length scales, quantum mechanical effects are no longer negligible. Thus, the electronic Schrödinger equation should be solved which is a difficult and time-consuming task because its solution, i.e. the electronic wave function, is an intricate multidimensional function. As part of the solution, the energy and forces acting on the nuclei, can be used to simulate the dynamic behavior of atomic systems. Therefore, the Schrödinger equation should be solved repeatedly to track these systems over a long time. Typically, a trade-off exists between the accuracy and computational efficiency of most algorithms, which will be highlighted in **Chapter 1** of this work.

Different approximation techniques exist to solve the Schrödinger equation. In first-principles methods, the electronic wave function or density is computed directly without using any experimental input. These methods are relatively accurate but are not really applicable to large systems because of their computational cost and scaling. Force fields on the other hand, discard

all the electronic degrees of freedom and approximate the energy as a series of physically-inspired analytical expressions. Moreover, their computational cost scales linearly in the number of nuclei, allowing simulations of large systems and long time scales. However, conventional force fields cannot properly describe polarization or charge transfer and lack the accuracy to be generally applicable. Therefore, in polarizable force fields, the electronic degrees of freedom are reintroduced in an approximate manner, often by assigning a fluctuating charge to every nucleus. However, in **Chapter 2**, we will show that fluctuating charges lead to an ambiguous definition of the polarizability in periodic systems. Explicit-electron force fields avoid this issue by modeling the electrons as semi-classical particles with a fixed integer charge. Characterizing the different energy contributions between the electron particles and nuclei in these types of force fields remains difficult, however, because they are dominated by a variety of quantum effects such as exchange contributions. The explicit-electron force fields are the inspiration of the electron machine learning potential (eMLP), where we will build towards in this work.

In **Chapter 3**, we will introduce machine learning potentials (MLPs). They are revolutionizing the field of molecular modeling because they achieve the accuracy of first-principle methods with almost the same computational efficiency of conventional force fields. MLPs accomplish this by predicting the energy and forces with a machine learning algorithm (kernel regression, neural networks . . .) that can have thousands or even millions of trainable parameters, all fitted to first-principles data. Unlike force fields, they are less susceptible to human bias and can automatically model chemical reactions. Notwithstanding their demonstrated capability to learn short-ranged interactions, it remains an open question how to properly incorporate long-range interactions and polarization in MLPs. If included, fluctuating charges are typically employed, which suffer from the issue defined above. Therefore, we will combine explicit-electron force fields and MLPs to push forward the state-of-the-art with the eMLP as the first goal of this work.

The full methodology of the eMLP will be discussed in **Chapter 4**. The eMLP is an explicit-electron force field where the intricate short-ranged interactions are learned with a MLP. Besides the nuclei, electron pairs (containing a spin-up and a spin-down electron) will be additional particles in the eMLP. These electron pair particles are not restricted to any nucleus but may potentially move through the whole system and they will be located at the centers of localized orbitals, which are derived from first-principles calculations. As a consequence, the eMLP will automatically learn the correct dipole moment and other electronic response properties. The accuracy of the eMLP is assessed by modeling small organic molecules and the crystalline

system of β-glycine. Two new data sets have been constructed for that purpose. It will be shown that small force and energy errors are achieved and that infrared spectra of unseen molecules can be predicted, demonstrating the transferability of the eMLP. Furthermore, the elasticity, dielectric and piezoelectric tensor of β-glycine can be computed with high accuracy. While developing the eMLP, we also encountered a weakness of the model. Dynamic simulations are not stable for a minority of the unseen molecules and a drift of the conserved quantity occurs during the simulation. This will be resolved by data augmentation, a machine learning technique which generates out-of-equilibrium electron pair geometries while training the network to teach the eMLP that they are unphysical.

As a secondary goal of this work, we will develop a new data protocol to derive MLPs for metal-organic frameworks (MOFs), at the end of **Chapter 3**. These materials have gained increasing interest in the scientific community due their attractive properties, including guest adsorption, catalysis or phase transitions. State-of-the-art MLPs for MOFs still require a data set of more than ten thousand first-principles configuration which is computationally demanding for these large systems. We will reduce the necessary amount of data to only a few hundred configurations and at the same time, reduce the errors with more than a factor of three compared to other available models in the literature. This is made possible by our proposed data protocol and the new class of data-efficient equivariant MLPs.

In conclusion, we have extended the state-of-the-art of MLPs in two domains. First, our data protocol for MOFs is shown to be more data efficient and accurate than other available MLPs for MOFs. Second, we have developed the eMLP where long-range interactions and polarization are naturally included. All these findings will be summarized in **Chapter 5**. Furthermore, some future perspectives will be given on how the eMLP may be extended to be more generally applicable.

List of Symbols

In this thesis, italicized variables indicate scalars, italicized and bold-faced variables indicate vectors or tensors, and regular variables indicate matrices. The indices $a, b, \ldots$ will be used to denote variables related to nuclei or atomic cores and the indices $i, j, \ldots$ will be used to denote variables related to the electronic degrees of freedom (orbitals, electron pairs).

Alphanumerical symbols	
A	Unit cell matrix
$\boldsymbol{C}$	Elasticity or stiffness tensor
C_n^{ab}	n^{th} order dispersion coefficient
$\boldsymbol{d}$	Piezoelectric strain tensor
$\boldsymbol{D}$	Electric displacement field
$D^{N\times M}$	A dense neural network layer without activation function
$\widetilde{D}^{N\times M}$	A dense neural network layer with activation function
$\boldsymbol{e}$	Piezoelectric charge tensor
E	Energy
E_0	ground state energy
E_n	The energy of the n^{th} eigenstate
$\boldsymbol{F}_a$	Force vector on the a^{th} nucleus or atomic core
$\boldsymbol{f}_i$	Force vector on the i^{th} electron pair
H	The Hessian matrix
$\hbar$	The Planck constant divided by 2π
$\hat{\mathsf{H}}$	The Hamiltonian operator
k_{b}	Boltzmann's constant
M_a	Mass of the a^{th} nucleus
n	Electron number density
N_{e}	Number of electrons or electron pairs
N_{n}	Number of atoms or atomic cores
N_{S}	Number of species
q_a	Charge of the a^{th} nucleus or atomic core

$\boldsymbol{r}_i$	Position vector of the i^{th} electron (pair)
r_{ia}	The interatomic distance between the i^{th} electron (pair) and a^{th} nucleus or atomic core
r_{ij}	The interatomic distance between electron (pair) i and j
$\boldsymbol{R}_a$	Position vector of the a^{th} nucleus or atomic core
R_{ab}	The interatomic distance between nucleus or atomic core a and b
$\boldsymbol{S}$	The strain tensor
S_a	The species of the a^{th} nucleus or atomic core
t	Time
T	Temperature
$\hat{\mathsf{T}}_{\mathsf{n}}$	Kinetic energy operator of the nuclei
$\hat{\mathsf{T}}_{\mathsf{e}}$	Kinetic energy operator of the electrons
V	Volume
V_{ext}	External potential
$\hat{\mathsf{V}}_{\mathsf{e,e}}$	Coulomb repulsion operator between the electrons
$\hat{\mathsf{V}}_{\mathsf{e,n}}$	Coulomb attraction operator between the electrons and nuclei
$\hat{\mathsf{V}}_{\mathsf{n,n}}$	Coulomb repulsion operator between the nuclei
$\hat{\mathsf{V}}_{\mathsf{eff}}$	Effective potential
$\boldsymbol{x}$	Vector containing all the degrees of freedom of a general system
$\boldsymbol{x}_a$	Atomic descriptor or features of the a^{th} nucleus or atomic core
Z_a	The atomic number of a^{th} nucleus

Greek symbols	
$\boldsymbol{\alpha}$	Polarizability tensor
$\boldsymbol{\varepsilon}$	Dielectric tensor
$\boldsymbol{\mathcal{E}}$	Electric field vector
$\boldsymbol{\theta}$	Vector with all trainable parameters
θ_k	The k^{th} trainable parameter
$\boldsymbol{\mu}$	Dipole vector
$\boldsymbol{\mu}_a$	Dipole vector of the a^{th} nucleus
$\boldsymbol{\sigma}$	The stress tensor
$\lvert\phi_n\rangle$	The n^{th} general eigenstate or orbital
$\lvert\varphi_n\rangle$	The n^{th} (localized) orbital
$\lvert\Psi\rangle$	General wave function

List of Abbreviations

BO	Born-Oppenheimer
CCSD(T)	Coupled-cluster theory with a full treatment of singles and doubles but perturbative triples
CPMD	Car-Parrinello molecular dynamics
DFT	Density functional theory
eMLP	Electron machine learning potential
ES	Ewald-summation
GGA	Generalized-gradient approximation
FB	Foster-Boys
FF	Force field
GPR	Gaussian process regression
HF	Hartree-Fock
IR	Infrared
LDA	Local-density approximation
MD	Molecular dynamics
MAE	Mean absolute error
MLP	Machine learning potential
MLWF	Maximally localized Wannier functions
MOF	Metal-organic framework
MSE	Mean squared error
NMS	Normal mode sampling
MPNN	Message passing neural network
MSE	Mean squared error
NequIP	Neural Equivariant Interatomic Potential
NN	Neural network
OQML	Operator quantum machine learning
PES	Potential energy surface
RBF	Radial basis function
SCF	self-consistent field

Part I

Incorporating long-range interactions and polarization in machine learning potentials with explicit electrons

1

Introduction

The science of today is the technology of tomorrow.
Edward Teller (1908–2003)

The technological advancement of humankind is driven by the ever-continuing quest to understand, process and develop new materials. As early as 3.3 million years ago, our prehistoric ancestors, the *Australopithecus* or *Kenyanthropus*, started making stone tools.[1] These developments may have been the driving force behind the origin of genus *Homo* and in turn the whole human species. From stone to bronze, from bronze to iron, we gradually discovered new materials and learned how to use them. It drastically changed our societies and anyone who had access to better materials, had a considerable edge on neighboring civilizations. More recently, the industrial revolution caused a boom in new technologies and inventions, including the mass production of steel and in the last century, the advent of silicon semiconductors and transistors led to the foundation of our own digital society. Today, innovations in material science are still ongoing. We are now confronted with the urgent issue of climate change which requires sustainable solutions and the refusal of fossil fuels. For instance, research for materials suited in next-generation batteries is ongoing.[2] They can provide the necessary storage solutions for energy in electric cars. Also the long awaited fusion reactors are a promising alternative to reduce the current fossil fuel consumption. Still, their capability to produce a netto amount of energy should yet be demonstrated. The engineering of innovative fusion

materials, who can withstand the hazardous environment plasma inside the reactor, is a critical component in the development of fusion reactors.[3]

Not only materials, but also our understanding of biological processes and chemical reactions changed our lives. In the Middle Ages, diseases were thought to be a divine punishment for immoral and sinful behavior. Fortunately, it is known today that they are caused by microscopical bacteria or viruses, which can be treated with rigorously designed medicines. The medical industry specifically targets the interplay between different cells and proteins in our body to ameliorate our health. Chemical reactions on the other hand, have had a major impact on the transportation sector. Trains and later cars rely on combustion engines, fueled with gasoline or diesel or in the near future with hydrogen. Photosynthesis is another biological chemical process which is all around us and its understanding may lead to artificial energy sources.[4]

The link between material science, chemical reactions and biological processes, is molecular modeling. It allows us to simulate atomic behavior under a set of physical conditions. This is important for many reasons. First of all, it provides a link between theory and experiment. Computer simulations can explain why certain observations are made and in what circumstances they appear. Secondly, reactions or biological processes which are difficult or simply impossible to observe experimentally, become available to study. For instance, the folding of proteins can be examined dynamically. Furthermore, computational modeling enables the high-throughput screening of drug or material discovery. The physical and chemical properties of thousands or even millions of materials can be predicted and stored in large databases. Experimentally synthesizing and measuring all the properties of an abundance of materials is simply not feasible. Finally, it opens up the possibility to study hypothetical materials. Now, the properties of newly proposed materials can be predicted before one has to synthesize the real material.

All these promising features of computational modeling, gained a lot of attention in research and industry. For this reason, the US Governement launched the Materials Genome Initiative one decade ago.[5] It aims to accelerate the discovery and manufacturing of advanced materials with the help of computational modeling to cross the bridge between theory and experiment. Among other things, the initiative led to the development of extended databases. One of them is the Materials Project,[6] which can be used to design materials in a data-driven way. In recent years, other platforms were also put in place, such as the NOMAD laboratory,[7] to encourage the distribution of computational data among researchers.

Nowadays, machine learning algorithms are a popular tool where computers

```
7
Properties=species:S:1:pos:R:3 energy=-4182.668
C   0.997081    0.019000   -0.030982
C   2.489095    0.060727    0.031757
O   3.177968    0.844924   -0.562606
H   0.676270   -0.967103   -0.379509
H   0.611668    0.794697   -0.689836
H   0.588203    0.140568    0.976362
H   2.955705   -0.706303    0.687004
```

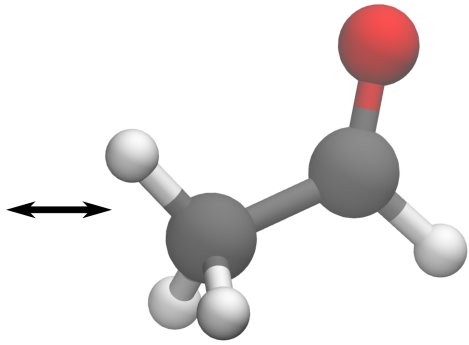


Figure 1.1: The ethanal molecule. Left: the geometry of ethanal is represented in an XYZ file. It numerically stores the positions (three coordinates for each atom) and species of the 7 atoms. Right: a visualization of the same ethanal molecule.

can automatically learn complex tasks from data. This technique has only recently been applied in molecular modeling and is revolutionizing the field because more and more researchers are using it to accelerate their simulations without compromising in accuracy. Therefore, in the following two sections, we will take a closer look at molecular modeling and how machine learning can improve this field, before stating the goal of this thesis at the end of the chapter.

1.1 A brief introduction to molecular modeling

Molecular modeling encompasses a variety of simulation techniques to study the behavior of molecular systems. Chapter 2 of this work will be fully devoted to this topic, but here we will briefly introduce this topic.

All the methods in this work will model the system at the atomistic level. This is visualized in Figure 1.1 for ethanal where it is shown that a molecule can simply be represented on a computer as a list of all the atomic species together with their positions and the total number of electrons. Ideally, for a given set of atoms, all relevant physical properties can be derived by solving the Schrödinger equation[8] for the electrons. The equation itself is completely known for almost a century and looks surprisingly simple on paper. The solutions on the other hand, are not simple at all but are intricate multidimensional functions, giving rise to emergent phenomena in chemical phase space. Solving the Schrödinger equation is simply not possible on current hardware for all but some trivial academic situations.

	Force field	Machine learning potential	First principles	
method	AMBER	eMLP	DFT/PBE0	CCSD(T)
Scaling	$\mathcal{O}(N_\mathrm{n})$	$\mathcal{O}(N_\mathrm{n})$	$\mathcal{O}(N_\mathrm{e}^3)$	$\mathcal{O}(N_\mathrm{e}^7)$
Time [s]	$< 1 \times 10^{-4}$	3.48×10^{-3}	18.6	6.32×10^{3}

Table 1.1: A comparison between force fields, machine learning potentials and first-principles methods. For each of the three categories, a typical method is chosen: AMBER,[9] eMLP,[10] DFT/PBE0[11] and CCSD(T).[12] The scaling depends on the number of nuclei N_n or the number of electrons N_e. The time required to perform one energy evaluation for ethanal is benchmarked on modern hardware (AMD Epyc 7H12 CPU at 16 cores, A100 GPU 80GB). For AMBER, the results are based on the TRPCage benchmark,[13] which is a larger system and hence the performance will be better than 1×10^{-4} s.

Hence, approximation techniques are introduced. Simply speaking, there are two major classes of methods for this purpose: first-principles methods and force fields. First-principles or ab initio methods are the most accurate models, i.e their predictions are closest to the true values when solving the Schrödinger equation. They approximately solve the Schrödinger equation without introducing any experimental input. However, it still turns out that first-principles methods require an unreasonable computational cost and furthermore, they scale poorly when applying the technique to large system sizes. Hence, extensive systems need more approximate solution strategies to speed up the calculations. This is where force fields come into play. They bypass the calculation of the electronic wavefunction or density and directly predict the energy as an analytical function of all the nuclear degrees of freedom. As an indication, the time required to solve the Schrödinger equation once for ethanal, is reported in Table 1.1 for a force field, machine learning potential (MLP) and two first-principles methods. It shows that force field methods are up to 6-7 orders of magnitudes faster than first-principles methods. Moreover, force fields scale linearly in the number of nuclei, unlike the first-principles methods, limiting them to smaller systems.

Together with High Performance Computing (HPC) infrastructure and even specialized hardware,[14] force fields allow us to model large-scale systems on long timescales. For instance, the details behind the phase-transition in MIL-53(Al), a metal-organic framework (MOF), can be unraveled.[15] This system with more than a million atoms requires the use of massive parallel

computing and GPUs. But also in the biomolecular world, force fields are opening up more possibilities. Micro- and millisecond simulations were performed for the viral envelope and other parts of the SARS-CoV-2 virus[16, 17] and simulations on a 64 million atom HIV-1 capsid were achieved.[18]

The analytical expression for the energy in a force field, is often a physically-inspired functional with a dozen of trainable parameters. They are fitted to highly accurate first-principles or experimental data. Of course, the goodness of the fit depends on the chosen energy functionals, which typically aim to describe a single system or single system class. For instance, there is no guarantee that a force field trained for liquid water, is a proper model to study ice. Hence, they are often not generally applicable and have a limited transferability. Furthermore, compared to first-principles methods, force fields lack accuracy. For instance, the predicted dynamic properties of peptide systems can differ orders of magnitude, depending on the specific force field chosen to perform the simulations.[19] Furthermore, in most force fields the energy expressions are constructed such that chemical bonds cannot be broken or formed. Thus, contrary to their incredible computational efficiency, conventional force fields lack accuracy, transferability and chemical reactivity.

The trade-off between computational efficiency and accuracy is schematically depicted in Figure 1.2 for both first-principles methods and force fields. Ideally, a methodology with the accuracy of first-principles methods and computational efficiency of force fields would solve many aforementioned issues. Additionally, if the method can naturally model reactivity, without requiring any human bias, it would be a major breakthrough. Already encountered in Table 1.1 and Figure 1.2, machine learning potentials[20–24] are the promising class of methods that can bridge that gap between force fields and first-principles methods. To understand how MLPs work, we will introduce machine learning in the next section.

1.2 Machine learning

Machine learning refers to computers and algorithms that improve automatically from experience and patterns based on data.[25] In contrast to conventional algorithms, no explicit knowledge about the problem is required. Programmers will not explicitly instruct a computer to solve a problem in an established sequence of successive operations but the computer will learn to solve the problem from the training data alone. That is why machine learning is often called the fourth paradigm of science[26] since besides experiments, scientific theory and computational modeling (the first, second and third

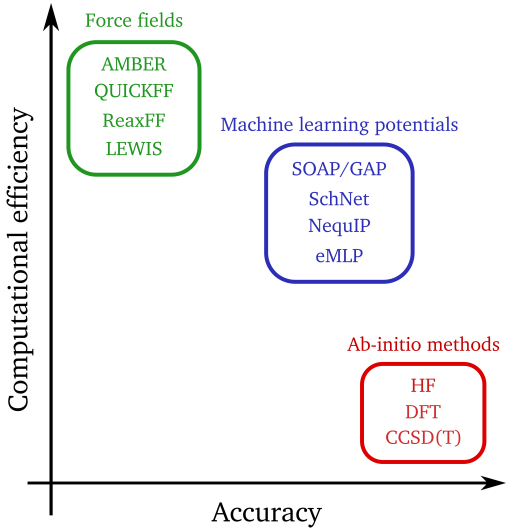


Figure 1.2: A schematic overview of the trade-off between computational efficiency and accuracy for first-principles methods, machine learning potentials and force fields. For each of the three classes, some specific methods are given, which will be encountered in this work.

paradigm respectively), big data can lead to new scientific discoveries and significant progress. Nowadays, machine learning is already extensively adopted in a variety of scientific and economic domains such as astronomy,[27] advertising,[28] physics,[29] biology and medicine[30] and of course molecular modeling.[31]

In general, most machine learning algorithms can be categorized in one of the following three subdivisions: supervised, unsupervised or reinforcement learning. In supervised learning, each data point has a certain input and output. The goal of the algorithm is to learn the relation between the inputs or features and the outputs, which are often called labels or targets. The labels are missing from the data in unsupervised learning. With only the input available, the algorithms try to recognize hidden patterns within the data. Common examples are clustering or dimensionality reduction. In reinforcement learning, the algorithm performs a complex task and based on the feedback it receives, it changes its strategy or method to solve the task. Thus, in this case there is no fixed data. For instance, OpenAI Five was able to beat the world champions in DOTA 2,[32] a popular esports game, which depends on more hidden and continuous variables than games like chess or Go.

In this work, we will only consider supervised learning. Depending on whether the labels are continuous or discrete, we are dealing with a regression or classification problem respectively. The classification of objects within images, is one of the most known use cases of machine learning. The construction of MLPs however, is a regression problem. A typical input is the set of positions of all the nuclei, see Figure 1.1, while the typical labels or targets are the energies and forces. Both are continuous variables. Fortunately, most techniques introduced and developed for computer vision,[a] can be used to train MLPs. These include artificial neural networks (NNs), linear and kernel regression, decision trees or support-vector machines.[33] The main focus in this work are (deep) neural networks.[34]

In a neural network, the input or features are transformed using a series of layers, each involving a number of neurons. It will be discussed in more detail in Chapter 3. The most important thing to know here, is that the performance of the model will drastically depend on the features going into the network. This makes feature engineering an important step in the development of neural networks. However, this a difficult process and it will inevitably require physical insight, which should ideally be avoided in MLPs. Deep learning or deep neural networks are developed to bypass feature engineering.[35] Here, trough multiple layers in the neural network (hence the preposition *deep*), the features are extracted automatically. This is a big advantage which led to its increase in popularity, first in computer vision, but also later in other scientific domains.

Strictly speaking, machine learning potentials are force fields. They still predict the energy as a predefined analytical function which depends on trainable parameters, fitted to first-principles data. However, contrary to the physically-inspired energy contributions, the analytical function is now a machine learning method. Instead of a few dozen parameters in conventional force fields, MLPs have thousands or millions of parameters. Furthermore, they require less human input and have no a priori constraints limiting bond breaking, making them an ideal tool to model more advanced and reactive systems. Thanks to these defining properties and the power of machine learning, MLPs can be trained with unprecedented accuracy. Hence, MLPs make predictions with nearly identical quality as the level of theory used to generate the training data. Their transferability still remains restricted to structures similar to the ones encountered in the training set, emphasizing the importance of data sampling techniques.

Although machine learning has been sporadically used to construct force

a. Computer vision is the scientific field where computers try to understand images or videos. This includes object recognition and tracking, image restoration, . . .

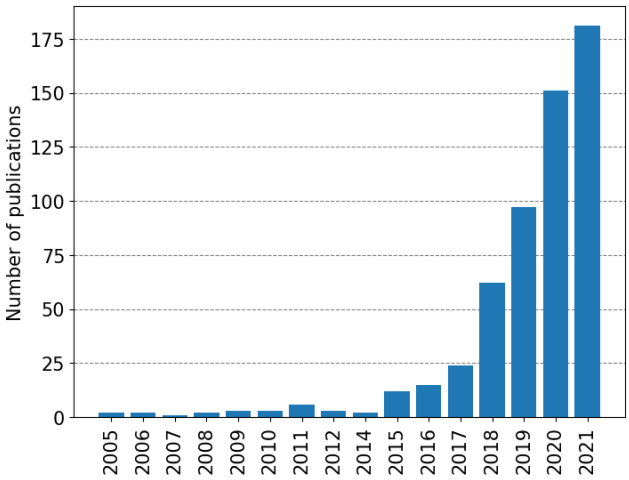


Figure 1.3: The yearly number of publications with *'machine learning potential'* in the title. This is an underestimation of the true number of publications on this topic because not all papers can be identified in this way. Data included herein are derived from Clarivate Web of Science. © Copyright Clarivate 2021. All rights reserved.

fields, it only recently gained traction. The development of the Behler-Parrinello or atom-centered symmetry functions (ACSF)[36, 37] was one of the initial sparks to ignite the research in MLPs. ACSFs are the typical engineered input features. Afterwards, major advancements of MLPs include Gaussian approximation potentials (GAP)[38], smooth overlap of atomic positions (SOAP)[39] features, the ANI-1 neural network[40] and gradient-domain machine learning (GDML).[41] Only around 2017, deep learning MLPs were introduced by using the concept of message passing neural networks (MPNNs).[42] The deep tensor neural network (DTNN)[43] was one of the first of many MLPs to explore the possibilities of deep learning. These breakthroughs, together with the advent of massive parallelism on GPUs in the last 10 years, led to an exponential increase in the usage of MLPs in the last few years. Figure 1.3, where the number of MLP-papers published in literature are counted, confirms this fact. Today, MLPs have entered the mainstream and have shown to model systems with first-principles accuracy for metallic systems,[44–46] MOFs[47, 48] or chemical reactions.[49–51]

1.3 Goal and overview of this work

Despite the enormous growth of MLP-applications in molecular modeling, most MLPs still lack a solid treatment of non-local charge transfer or long-range interactions. These complex phenomena are crucial to model advanced materials.[52, 53] Additionally, polarization or the influence of external electric fields remain non-trivial effects to include in MLPs. In this thesis, we will investigate whether we can improve MLPs by solving these issues. To accomplish this, we will work towards the following goal: developing an improved MLP, which naturally incorporates long-range interactions and polarization. It should not only be an improvement upon state-of-the-art MLPs but also upon existing polarizable force fields as well. If successful, it will spawn a new generation of MLPs to study advanced phenomena such as redox reactions and piezoelectricity.

A different but equally important issue is the absence of a widely applicable method to generate MLPs for metal-organic frameworks (MOFs). MOFs are porous materials, composed of inorganic clusters and organic ligands.[54, 55] They have gained a lot of interest lately due to their attractive properties, including catalysis,[56] CO_2 capture and conversion,[57] their use as a battery material,[58] chemical sensors[59] and much more. A variety of MOFs can undergo phase transitions induced by pressure, temperature or the presence of guest molecules. Because of these interesting characteristics, it is impossible to apply conventional sampling techniques to efficiently generate databases and eventually train an MLP. Some early steps have been taken to resolve that matter,[47] but a general protocol to generate MLPs for MOFs is still missing. The secondary goal of this work is to show that MOFs can be accurately described with MLPs in a data efficient way. This will finally allow us to study the large-scale behavior of MOFs without the crude approximations of conventional force fields.

In the remaining chapters of this work, we will further clarify these goals and establish the necessary context. In Chapter 2, we will introduce molecular modeling starting from the Schrödinger equation. Via the Born-Oppenheimer approximation, potential energy surfaces are defined. This will be a key concept in this work. Next, we will take a closer look at the first-principle methods and force fields. At the end of the chapter, the treatment of long-range interactions in force fields is discussed where we will demonstrate that fluctuating charges, the typical way to include long-range effects and polarization, are inherently flawed. Explicit-electron force fields, do not suffer from this issue and are the inspiration of the electron machine learning potential (eMLP), which will be discussed in Chapter 4.

Chapter 3 is fully devoted to MLPs. An overview of the whole process to derive MLPs is given, including data generation, choosing the right type of MLP and training it. Afterwards, we will return to the problem of modeling long-range interactions, now specifically in the context of MLPs. Finally, we focus on MLPs for MOFs, which is the second goal of this work. There, a data efficient sampling protocol will be proposed which is made possible by equivariant MLPs.

The eMLP is the sole focus of Chapter 4. After a quick introduction, the different particles in the eMLP will be discussed, together with all the energy contributions. Its accuracy will be validation on two newly created data sets, eQM7 and a data set for β-glycine. Finally, we will introduce data augmentation in the context of the eMLP, to improve the robustness and stability during MD simulations for unseen molecules.

In the final chapter, a brief summary of this work will be presented together with some perspectives how future work may build on top of this work.

2

Molecular modeling

It is a capital mistake to theorize before one has data.
Insensibly one begins to twist facts to suit theories,
instead of theories to suit facts.

Sir Arthur Conan Doyle (1859–1930)

Molecular modeling is a computational tool to understand the behavior of molecules or periodic systems at the nanoscale. Ideally, one wants to solve the Schrödinger equation since it contains all the knowledge required to model the interactions between atoms and molecules. Therefore, we start this chapter with a brief introduction of the theory of quantum mechanics and its key equations. Next, different simulation techniques will be studied. Afterwards, we will discuss the so-called first-principles methods, algorithms to solve the Schrödinger equation. They still allow for a detailed and generally applicable description of molecular systems but come at a large computational cost. Conventional force fields completely discard the electronic degrees of freedom and instead put forward an analytical expression for all the interatomic energy contributions. This enables the simulation of more extended systems. However, complex phenomena such as ionization, charge transfer or chemical reactions may require a more detailed description and reintroduce the electronic degrees of freedom in an approximate manner. This is made possible by polarizable or explicit-electron force fields, studied at the end of this chapter.

2.1 From the Schrödinger equation to potential energy surfaces

2.1.1 The Schrödinger equation

Ordinary matter is composed of electrons, protons and neutrons.[a] These particles have such a minuscule mass and size that the classical laws of motion are not valid anymore. Instead, they obey the Pauli exclusion principle and display wave-particle duality. One cannot measure both the position and velocity of the particles with infinite precision. This is intrinsically linked with the wavefunction description, which expresses the probability density to find a particle at a certain point in space. The equation that governs the dynamics of the wavefunction is the Schrödinger equation:[8]

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\left|\Psi(t)\right\rangle = \hat{\mathsf{H}}\left|\Psi(t)\right\rangle \tag{2.1}$$

with $|\Psi(t)\rangle$ the abstract state vector of the system and $\hat{\mathsf{H}}$ the Hamiltonian. For a general system with N_n nuclei and N_e electrons, the Hamiltonian in atomic units is given by:

$$\begin{aligned}\hat{\mathsf{H}} =& \hat{\mathsf{T}}_\mathsf{n} + \hat{\mathsf{T}}_\mathsf{e} + \hat{\mathsf{V}}_\mathsf{e,n} + \hat{\mathsf{V}}_\mathsf{n,n} + \hat{\mathsf{V}}_\mathsf{e,e} \\ =& -\frac{1}{2}\sum_{a=1}^{N_\mathsf{n}}\frac{\nabla_a^2}{M_a} - \frac{1}{2}\sum_{i=1}^{N_\mathsf{e}}\nabla_i^2 \\ & -\sum_{a=1}^{N_\mathsf{n}}\sum_{i=1}^{N_\mathsf{e}}\frac{Z_a}{r_{ia}} + \sum_{a\neq b}^{N_\mathsf{n}}\frac{Z_a Z_b}{R_{ab}} + \sum_{i\neq j}^{N_\mathsf{e}}\frac{1}{r_{ij}}\end{aligned} \tag{2.2}$$

with M_a and Z_a the mass and charge of the nucleus a respectively, while r_{ia} is the distance between electron i and nucleus a, R_{ab} the distance between nuclei a and b, and r_{ij} the distance between electrons i and j. If the Hamiltonian is not explicitly time-dependent, the solution $|\Psi(t)\rangle$ can always be written as:

$$\left|\Psi(t)\right\rangle = \sum_n c_n \exp(-\mathrm{i}E_n t)\left|\phi_n\right\rangle \tag{2.3}$$

with the eigenstates $|\phi_n\rangle$ of the time-independant Schrödinger equation:

$$\hat{\mathsf{H}}\left|\phi_n\right\rangle = E_n\left|\phi_n\right\rangle. \tag{2.4}$$

Hence, for a given set of initial conditions, only this eigenvalue equation has to be solved to know the full dynamic behaviour of the system. Although

a. In principle, protons and neutrons are not elementary particles but consist of three quarks. A whole zoo of elementary particles do exist but they are not of importance in this work.

the equation looks simple, solving it analytically is only possible for a small subset of physical situations (hydrogen atom, quantum mechanical oscillator ...).

2.1.2 The Born-Oppenheimer approximation

The Hamiltonian couples the nuclear and electronic degrees of freedom through the terms in r_{ia}. For that reason, the nuclear and electronic degrees of freedom in the wavefunction can not be separated. However, electrons are significantly lighter than the nuclei, which means that the nuclei move very slowly compared to the electrons. The Born–Oppenheimer approximation[60] takes advantage of this fact by demanding that the nuclear and electronic degrees of freedom can be separated:

$$\psi_{kn}(\{\boldsymbol{r}\},\{\boldsymbol{R}\}) = \chi_{kn}(\{\boldsymbol{R}\})\phi_n(\{\boldsymbol{r}\};\{\boldsymbol{R}\}). \tag{2.5}$$

Consequently, the electronic wavefunction $\phi_n(\{\boldsymbol{r}\};\{\boldsymbol{R}\})$ satisfies the following eigenvalue equation:

$$\left(\hat{\mathsf{T}}_\mathsf{e} + \hat{\mathsf{V}}_\mathsf{e,n} + \hat{\mathsf{V}}_\mathsf{n,n} + \hat{\mathsf{V}}_\mathsf{e,e}\right)\phi_n(\{\boldsymbol{r}\};\{\boldsymbol{R}\}) = E_n(\{\boldsymbol{R}\})\phi_n(\{\boldsymbol{r}\};\{\boldsymbol{R}\}), \tag{2.6}$$

which is the Schrödinger equation for electrons in the field of static nuclei where the resulting energy levels $E_n(\{\boldsymbol{R}\})$ depend on the nuclear geometry. If we assume that the systems remains instantaneously in a single electron eigenstate (this is the adiabatic approximation), then the nuclear wave function can found be solving the following equation:

$$\left(\hat{\mathsf{T}}_\mathsf{n} + E_n(\{\boldsymbol{R}\})\right)\chi_{kn}(\{\boldsymbol{R}\}) = E_{kn}\chi_{kn}(\{\boldsymbol{R}\}) \tag{2.7}$$

This assumption and the Born–Oppenheimer approximation is only valid when the set of energy levels $E_n(\{\boldsymbol{R}\})$ are sufficiently separated. Under normal conditions, the electronic energy levels are typically separated by at least a few electronvolts, larger than all the nuclear rotational and vibrational energy differences (having a magnitude of approximately 0.1 eV or below). Hence, in this work we will assume that the Born–Oppenheimer approximation is valid such that we should only focus on the ground state energy $E_0(\{\boldsymbol{R}\})$, also called the potential energy surface (PES).

In general, the PES is a complex function of multiple variables (the nuclear coordinates) and it potentially has numerous minima and maxima, visualized in Figure 2.1. A wealth of information is contained within the PES: the location of the minima correspond to the geometry of (meta)stable states, knowledge about reactions can be inferred from different pathways between

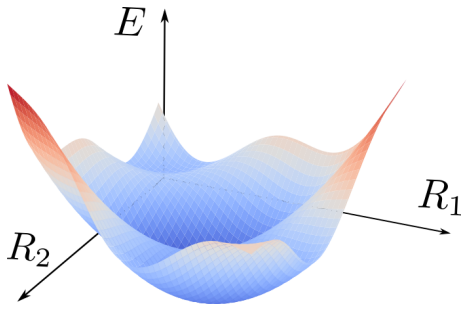


Figure 2.1: An illustration of a two dimensional PES $E(R_1, R_2)$. The minima and regions having low energies are colored blue, while areas with high energies are colored red.

minima while the height of the energy barriers correlate with the reaction rates. For all those reasons, the PES is the central function in molecular modeling. In principle, to understand the dynamics of the nuclei on the PES, Eq. (2.7) should be solved. However, one resorts to classical physics to model the nuclei by making once again use of fact that the nuclei are heavy enough such that nuclear quantum effects are negligible. In the classical regime, the PES is just an ordinary potential surface for all nuclei. Once known, the PES can be sampled and the system under consideration can be simulated in different thermodynamic ensembles.

2.2 Simulation techniques

Modeling the behavior of materials or chemical reactions requires more than only the potential energy surface. One should know how the nuclei move on the PES and how the system behaves under finite temperatures and pressures. In this work, three major simulation techniques will be utilized: geometry optimizations, the harmonic approximation and molecular dynamics (MD) simulations. The first two methods are exactly valid at absolute zero (0 K), while MD simulations allow us to explore the system at finite temperatures and pressures. An overview of the three methods is given in Figure 2.2.

2.2.1 Geometry optimization

In a geometry optimization, one is interested in the global minimum of the PES. This is the state of the system at 0 K because at that temperature no kinetic energy is left to visit higher energy configurations. Local minima

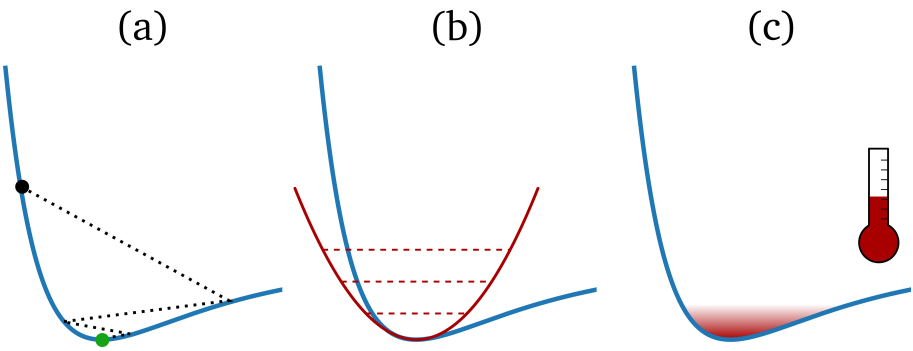


Figure 2.2: An overview of the three major simulation techniques used in this work: (a) geometry optimizations, (b) the harmonic approximation and (c) MD simulations. The first two methods are only exact at absolute zero (0 K or -273.15°C) while MD simulations can be done at finite temperatures.

of the PES correspond to metastable states instead. Finding the minima is important because it provides a useful approximation to the true geometry of the system at low temperatures. Furthermore, it serves as the starting point of the harmonic approximation or MD simulations. In practice, the optimized geometry is found after minimizing the energy with respect to all the degrees of freedom. They are the set of positions $\{\boldsymbol{R}_a\}$ and optionally, the lattice vectors of the unit cell (in periodic simulations). The lattice vectors are stored as the rows in the 3 by 3 matrix A.[b] Using those definitions, the following equations should be solved:

$$\frac{\partial E}{\partial \boldsymbol{R}_a} = 0 \qquad \frac{\partial E}{\partial \mathrm{A}} = 0 \tag{2.8}$$

where the PES $E(\{\boldsymbol{R}_a\}; \mathrm{A})$ is a function of the unit cell and nuclear positions. The energy minimization is typically done with an iterative algorithm such as the Quasi-Newton method.[61] This is visualized in Figure 2.2 (a): starting from an initial guess (the black point), the PES is minimized step by step until the optimum is found (green point).

2.2.2 Harmonic approximation

In the harmonic approximation, the true PES is approximated by a second order Taylor expansion, see Figure 2.2 (b), in the degrees of freedom $\boldsymbol{x}$

b. In this work, the lattice vectors will be stored in the rows. This is not a unique definition in the literature. Other work may use the rows to store the vectors.

(positions and/or unit cell):

$$E = E_0 + \frac{1}{2}\boldsymbol{\Delta x}^T \mathsf{H} \boldsymbol{\Delta x} \tag{2.9}$$

where $\boldsymbol{\Delta x} = \boldsymbol{x} - \boldsymbol{x}_0$ with $\boldsymbol{x}_0$ the degrees of freedom of the minimum. Hence, one should first perform a geometry optimization. The Hessian matrix H contains all the second order derivatives of the PES and should be positive definite. Otherwise, $\boldsymbol{x}_0$ does not correspond with the optimized geometry. Since in the harmonic approximation, the PES is a simple quadratic function, a variety of dynamical properties can be computed analytically, without the need for expensive dynamic simulations. For low temperatures, the computed properties will almost coincide with the true values and at 0 K they will match exactly. For instance, vibrational frequencies, infrared spectra, the elasticity tensor or piezoelectric tensor can all be derived from an (extended) Hessian.

2.2.3 Molecular dynamics

Molecular dynamics simulations are a tool to model the dynamic behavior of a system.[62] One assumes that the nuclei are classical particles and obey Newton's equations:

$$M_a \frac{\mathrm{d}^2 \boldsymbol{R}_a}{\mathrm{d}t^2} = \boldsymbol{F}_a = -\boldsymbol{\nabla}_a E \tag{2.10}$$

to propagate them forward in time. Consequently, the nuclei continuously explore a region of the PES, visualized in Figure 2.2. Finite temperatures, pressures and volume constraints can be imposed within certain statistical ensembles. For example, the microcanonical ensemble (NVE) studies the system with a fixed number of particles N, energy E and volume V while the canonical ensemble (NVT) enforces a constant temperature T instead of the energy. This is realized by modifying the equations of motion in the presence of a thermostat or barostat to control the pressure or temperature respectively.[62] By tracking the nuclei for a certain amount of time, one can study chemical processes or compute time-averages of certain properties of interest under real physical conditions. The ergodic hypothesis assumes that the averages over time correspond to the same properties averaged over all relevant microstates within a statistical ensemble.

Characterizing the nuclei as classical particles is a crude approximation for the lighter atoms like hydrogen. The wave-nature of the nuclei can be restored by taking nuclear quantum effects (NQE) into account by performing path integral molecular dynamics (PIMD) simulations.[63] For some systems

and properties, it has been shown that NQEs play a crucial role. For instance, when accurate structural properties of MOF-5 (a metal-organic framework) are required, NQEs are necessary.[64]

Newton's equations (2.10) are typically solved by a Verlet integrator.[62] In each iteration of the algorithm, the forces of the PES are evaluated and the velocities and positions of the nuclei are updated accordingly. The optimal time step between two force evaluations should be chosen such that the equations of motions are still solved accurately. This is determined by the fastest oscillating nuclei in the system, which are generally the hydrogen atoms. A time step of 0.5 fs is sufficient for most applications. Hence, if a system requires multiple nanoseconds of simulation time, more than two million energy and force evaluations are needed. Therefore, it is essential that computing the PES is a as cheap as possible without losing significant accuracy.

2.3 First principles methods

First principles or ab initio methods are computational methods and approximations to solve the electronic Born-Oppenheimer equation (2.6). *Ab initio* is the Latin expression for "from the beginning", denoting that first-principles methods all start from the Schrödinger equation and do not use any experimental input.[c] Highly accurate first-principles calculations are still prohibitive for all but only a subset of small molecules. Therefore, different first-principles methods have been developed which trade accuracy for computational efficiency and better scaling to larger systems. In this section, we will start reviewing the Hartree-Fock method, and gradually move up to more accurate methods.

2.3.1 Hartree-Fock method

The central approximation in the Hartree-Fock (HF) method,[66, 67] is that the electronic wavefunction is a single Slater determinant:

$$\psi(\{\boldsymbol{x}_i\}) = \frac{1}{\sqrt{N_e!}} \begin{vmatrix} \phi_1(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_1(\boldsymbol{x}_{N_e}) \\ \phi_2(\boldsymbol{x}_1) & \phi_2(\boldsymbol{x}_2) & \cdots & \phi_2(\boldsymbol{x}_{N_e}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{N_e}(\boldsymbol{x}_1) & \phi_{N_e}(\boldsymbol{x}_2) & \cdots & \phi_{N_e}(\boldsymbol{x}_{N_e}) \end{vmatrix} \tag{2.11}$$

c. The term *ab initio* was first coined in Ref. 65 with a slightly different meaning. There, different authors computed and verified their results independently from each other starting from zero or from ab initio.

where the generalized coordinate $\boldsymbol{x}_i = (\boldsymbol{r}_i, \sigma_i)$ of electron i also includes the spin σ_i. By construction, the wavefunction is antisymmetric under interchanging two electrons, which is a necessary condition for a fermionic wavefunction. The single particle orbitals are expanded into a set of N_b basisfunctions:

$$\phi_i(\boldsymbol{x}) = \sum_{k=1}^{N_b} c_k^{(i)} b_k(\boldsymbol{x}). \tag{2.12}$$

Typically, Gaussian-type basis sets are used.[68, 69] The ground state energy and wavefunction with corresponding expansion coefficients $c_k^{(i)}$ are found by applying the variational method:

$$E_0 = \min_{c_k^{(i)}} \langle\psi|\hat{\mathrm{H}}|\psi\rangle\,. \tag{2.13}$$

Hence, the Hartree-Fock energy overestimates the exact ground state energy even in the complete basis set limit, because only single slater determinant wavefunctions are considered which are a subset of all possible antisymmetric wavefunctions. The minimization yields a set of N_e single particle differential equations:

$$\left(-\frac{1}{2}\nabla_i^2 + \hat{V}_\mathrm{eff}(\boldsymbol{x})\right)\phi_i(\boldsymbol{x}) = \varepsilon_i\phi_i(\boldsymbol{x}) \tag{2.14}$$

where the effective potential,

$$\begin{aligned}\hat{V}_\mathrm{eff}(\boldsymbol{x})\phi_i(\boldsymbol{x}) = V_\mathrm{ext}(\boldsymbol{x})\phi_i(\boldsymbol{x}) &+ \sum_{j=1}^{N_\mathrm{e}} \int \frac{\phi_j^*(\boldsymbol{x}')\phi_j(\boldsymbol{x}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\phi_i(\boldsymbol{x})\mathrm{d}\boldsymbol{x}' \\ &- \sum_{j=1}^{N_\mathrm{e}} \int \frac{\phi_j^*(\boldsymbol{x}')\phi_i(\boldsymbol{x}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\phi_j(\boldsymbol{x})\mathrm{d}\boldsymbol{x}',\end{aligned} \tag{2.15}$$

is a combination of the external potential V_ext (the electrostatic interaction with the nuclei) and the Coulomb and exchange operators, which are the latter two terms respectively. As a consequence of the single Slater determinant wavefunction, the interactions with all other electrons $\hat{\mathrm{V}}_\mathrm{e,e}$ are averaged and combined into an effective operator. Thus, the Hartree-Fock method is a mean field theory and the missing energy is called the correlation energy. Forces and stresses are determined by applying the Hellmann-Feyman theorem.[70] For instance, the force on nucleus a is

$$\boldsymbol{F}_a = -\boldsymbol{\nabla}_a E_0 = -\langle\psi|\boldsymbol{\nabla}_a\hat{\mathrm{H}}|\psi\rangle\,. \tag{2.16}$$

Because the Coulomb and exchange operators depend on the other single particle orbitals (in an averaged manner), Eq. (2.14) is usually solved in a self-consisted manner. For that reason, the Hartree-Fock method is also called a self-consistent field (SCF) method. A general SCF algorithm starts with an initial guess for the expansion coefficients in Eq. (2.12). For instance, one can take the atomic orbitals of the isolated atoms. Next, the effective potential (2.15) is calculated using the current set of single particle orbitals. Afterwards, Eq. (2.14) is solved, yielding a new set of single particle orbitals, for which an updated version of the effective potential can be computed. This iterative procedure will continue until the single particle orbitals needed to calculate the effective potential and the single particle orbitals which solve Eq. (2.14) are consistent with one another. This is a necessary condition for Eq. (2.13).

2.3.2 Density functional theory

Density functional theory (DFT) is a very successful and popular method to model molecules and solid states because of its favorable computational efficiency and scaling. Unlike the Hartree-Fock method, which solves the Schrödinger equation for $\psi(\{\boldsymbol{r}_i\})$, the electron number density $n(\boldsymbol{r})$ is the central variable here:

$$n(\boldsymbol{r}) = N_\mathrm{e} \int \psi^*(\{\boldsymbol{r}_i\})\psi(\{\boldsymbol{r}_i\})\mathrm{d}\boldsymbol{r}_2 \dots \mathrm{d}\boldsymbol{r}_{N_\mathrm{e}}. \tag{2.17}$$

At first, it appears that information is lost when dealing with the number density alone. It looks like the correlation between different electrons in the wavefunction disappears after integration. Fortunately, the Hohenberg–Kohn theorems [71] show that the ground state wavefunction and ground state number density contain the same information. The first Hohenberg–Kohn theorem proves that the ground state electron density uniquely determines the external potential $V_\mathrm{ext}(\boldsymbol{r})$ and consequently the Hamiltonian $\hat{\mathrm{H}}$. This is a powerful property since the Hamiltonian in its turn determines the ground state wavefunction as the solution of the Schrödinger equation with the given Hamiltonian. Hence, the energy $E[n(\boldsymbol{r})]$ is a well-defined but unknown functional of the ground state electron density. The second Hohenberg–Kohn theorem proves that the ground state electron density minimizes the unknown energy functional:

$$E_0 = \min_{n(\boldsymbol{r})} E[n(\boldsymbol{r})] \tag{2.18}$$

with the additional constraint that there are N_e electrons:

$$N_\mathrm{e} = \int n(\boldsymbol{r})\mathrm{d}\boldsymbol{r}. \tag{2.19}$$

Finding the ground state energy requires the unknown energy functional. Although one can easily determine the contribution of the external potential in terms of the electron number density, the kinetic and electron-electron repulsion $\hat{V}_{e,e}$ part of the energy remains problematic. To partially solve this problem, Kohn and Sham introduced a fictitious non-interacting particle system,[72] modeled with a single Slater determinant wavefunction of Eq. (2.11). The resulting electron density

$$n(\boldsymbol{r}) = \sum_{i=1}^{N_e} |\phi_i(\boldsymbol{r})|^2 \tag{2.20}$$

still represents the density of the real system. The wavefunction on the other hand, has no real physical meaning but simplifies the kinetic energy functional. With this assumption, Eq. (2.18) leads to the Kohn-Sham equation for non-interacting orbitals:

$$\left(-\frac{1}{2}\nabla_i^2 + V_{\text{eff}}(\boldsymbol{r})\right)\phi_i(\boldsymbol{r}) = \varepsilon_i\phi_i(\boldsymbol{r}). \tag{2.21}$$

It has exactly the same form as Eq. (2.14), but here the effective potential is

$$V_{\text{eff}}(\boldsymbol{r}) = V_{\text{ext}}(\boldsymbol{r}) + \int \frac{n(\boldsymbol{r})}{|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{d}\boldsymbol{r}' + \frac{\delta E_{\text{XC}}[n(\boldsymbol{r})]}{\delta n(\boldsymbol{r})}. \tag{2.22}$$

The second term is the classical Coulomb energy, the known part of $\hat{V}_{e,e}$ expressed as a function of the electron density. The third term includes the exchange-correlation functional $E_{\text{XC}}[n(\boldsymbol{r})]$ which describes all the missing many-particles interactions, i.e. the part of the energy that is missing when only using the first two terms. Thus, up until this point, DFT is still an exact method. However, the true exchange-correlation functional is not known. Hence, approximations of the one true functional should be utilized. One can construct them using empirical parameters, fitted to first-principles energies of higher accuracy or experimental results.[d] A different approach is to design functionals purely with the help of physical principles.

Various exchange-correlation functionals have been developed, each with their own strengths and weaknesses.[73] Every functional can be classified on Jacob's ladder,[74] visualized in Figure 2.3. The most simple approximation can be found at the bottom of the ladder. They lack the accuracy of the more advanced functionals but they are computationally more efficient. Moving towards the top of the ladder, one approaches 'heaven', where perfect

d. In principle, when using experimentally fitted parameters, DFT is not an first-principles method anymore.

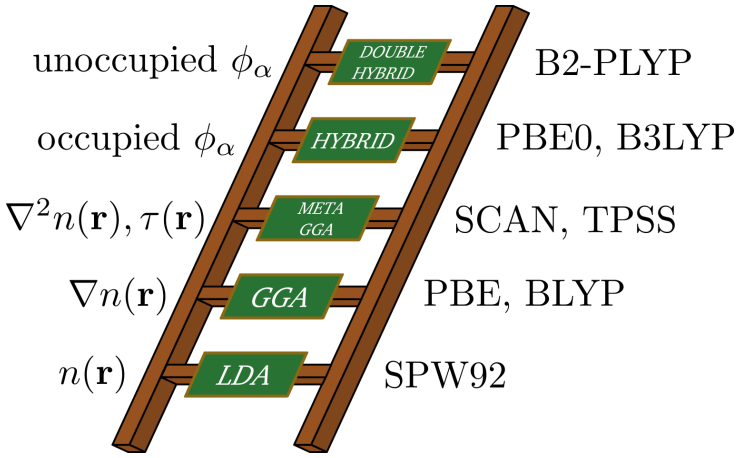


Figure 2.3: A visual classification system for the exchange-correlation functionals in DFT. Every rung is labeled with the name of the exchange-correlation class. On either side of the ladder, the input needed to compute the functional and some examples are indicated.

chemical accuracy can be found. At the left of every step, the information necessary to compute the functional, is indicated. At the base, one finds the local-density approximation (LDA), which only uses the electron density $n(\boldsymbol{r})$ itself. Moving one step up, the generalized-gradient approximation (GGA) methods also include the gradient of the density $\boldsymbol{\nabla} n(\boldsymbol{r})$. Meta GGA-functionals add the kinetic energy density $\tau(\boldsymbol{r})$ or the laplacian $\nabla^2 n(\boldsymbol{r})$ of the electron density. Moving one step further, the so-called hybrid functionals incorporate exact exchange energy which is a nonlocal functional of the occupied Kohn-Sham orbitals. Finally, double hybrid functionals include also unoccupied or virtual orbitals to model exact partial correlation. At the right side of the ladder, examples of exchange-correlation functionals belonging to every rung are indicated. In this work, mainly the PBE[75] and PBE0[11, 76] functionals are used.

Dispersion interactions are the results of dynamic correlation effects. Instantaneous dipoles in one molecular fragment induce electronic dipoles in other molecular fragments, resulting in a small attractive force. It can be shown that the leading term decays as R_{ab}^{-6} between two atoms a and b.[77] In principle, exchange-correlation functionals should be able to correctly capture these long-range electron correlations. However, most exchange-correlation functionals remain semi-local, making them inadequate to model dispersion.[78] For that reason, it is common practice to include an additional dispersion method on top of DFT. For instance, DFT-D3[79] is a popular tool,

where the pairwise two-body dispersion contributions are given by:[e]

$$E_{\text{DFT-D3}} = -\frac{1}{2}\sum_{a\in A}\sum_{b\in B}\sum_{n=6,8,10,\ldots} f_n(R_{ab})\frac{C_n^{ab}}{R_{ab}^n} \tag{2.23}$$

where $f(R_{ab})$ is a damping-function and C_n^{ab} are the n$^{\text{th}}$-order dispersion coefficients between atoms a and b in molecular fragment A and B respectively.

DFT computations are typically performed on medium sized systems because of their $\mathcal{O}(N_\text{e}^3)$ scaling. Still, it is one of the most accurate methods compared to its computational cost. A more approximate but efficient DFT-based method is linear scaling DFT.[80, 81] There, the exponential tail of the density matrix is truncated. This leads to a linear scaling, just like in force field methods. However, for systems with up to 10000 atoms, the computation time still takes several hours, trailing orders of magnitudes behind conventional force fields.

2.3.3 Post-Hartree–Fock methods

Post-Hartree-Fock methods are not explicitly used in this work but are of great importance when training accurate machine learning potentials. Therefore, we will briefly introduce them here. Essentially, these methods go beyond a single Slater determinant to represent the wavefunction. This enables the treatment of electron correlation besides the exact exchange, which already was included in the Hartree-Fock method.

There are several Post-Hartree-Fock methods available and implemented in quantum chemistry codes. One of them is Møller–Plesset (MP) perturbation theory[82] which, like the name suggests, uses perturbation theory up to a certain order to include correlation effects. For instance, the second order method is denoted as MP2, the third order method as MP3. The configuration interation (CI) method[83] defines the wavefunction as a linear combination of Slater determinants, which are determined variatonally. The popular CCSD(T) theory,[12] which stands for coupled-cluster (CC) with a full treatment of singles and doubles but perturbative triples, is often called the 'gold standard' of computational chemistry. It employs an exponential ansatz for the wavefunction:

$$|\psi\rangle = \mathrm{e}^{\hat{T}}\,|\psi_0\rangle \tag{2.24}$$

where $|\psi_0\rangle$ is typically the Hartree-Fock Slater determinant, serving as the reference wavefunction. The cluster operator $\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \ldots$ includes

e. Three-body terms are also included in DFT-D3, see Ref. 79.

the excitation operators $\hat{T}_n$. For instance, $\hat{T}_2$ generates all double excitations. In the CCSD(T) method, the series is truncated after the third term but the triples are only accounted for in perturbation theory.

In general, Post-Hartree-Fock methods will be more accurate than Hartree-Fock or DFT calculations, but it comes at a price. For instance, Hartree-Fock or DFT scales[f] as $\sim N_e^4$ while CCSD(T) scales as $\sim N_e^7$. This is the reason why only small molecular databases are constructed with post-Hartree-Fock methods like CCSD(T).

2.4 Force fields

Although first-principles methods yield accurate results for a variety of chemical environments, the computational cost still remains prohibitive for extended systems. For instance, on modern hardware,[g] a DFT calculation on a system with 540 atoms can last up to 40 minutes. Hence, large biomolecular simulations with more than a million atoms are simply not possible within this level of theory and more crude approximations to the Schrödinger equation should be made. Force fields (FF) bypass the computation of the electron wavefunction and directly calculate the PES[84], which tremendously reduces the computational cost. Essentially, the PES is now approximated by an analytical function of all the nuclear coordinates of the system. The functional form of the potential is fixed, physically inspired and only depends on a number of free parameters, representing bond distances, bond strengths and other chemical properties. With the help of force fields, longer timescales become available to simulate and larger chemical systems become computationally feasible. Among others things, they have become an established tool to model macromolecular biomolecules[85] or the mechanisms behind phase-transitions in MOFs.[15]

2.4.1 Conventional force fields

There is a diverse set of available force fields in literature, each having their own energy decomposition and domain of applicability. Here, we give a brief overview of the energy contributions of the typical force field. In general, the energy is the sum of the bonded and nonbonded interactions:

$$E_{\text{FF}} = E_{\text{bonded}} + E_{\text{nonbonded}}. \tag{2.25}$$

f. Evaluating the 4-center electron repulsion integrals causes a computational cost of $\sim N_e^4$. It can be reduced by screening the near-zero integrals or a sparse representation such that in practice a scaling of $\sim N_e^3$ is possible.

g. Two Xeon Gold 6132 CPU's at 2.6GHz, each having 14 cores.

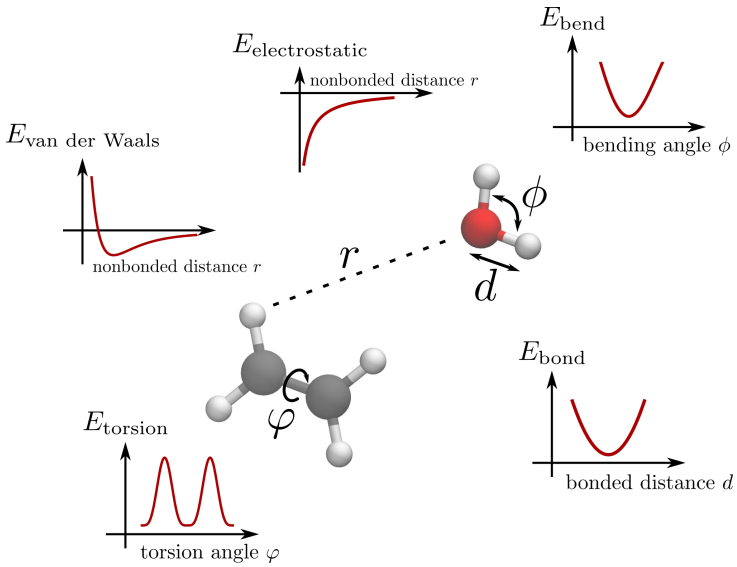


Figure 2.4: A visualization of the typical force field energy contributions for a ethene-water dimer. The three covalent interactions are E_{bond}, E_{bend} and E_{torsion}, which are a function of the bonded distance d, the bending angle ϕ and torsion angle φ respectively. The noncovalent interactions $E_{\text{electrostatic}}$ and $E_{\text{van der Waals}}$ act between the atoms of the two dimers and are a function of their intermolecular distance r.

The bonded or covalent interactions are short-range interactions, modeling the chemical bonds between atoms. Contrarily, the nonbonded or noncovalent interactions can act over long distances. The systems described with a nonreactive force field, have a fixed topology. Before the simulations starts, one defines which atoms are bonded and those who are not. Hence, the covalent interactions will only act on the atoms of the first group and no other chemical bonds can form or break up during the simulation. Reactive force fields on the other hand, do not have a fixed topology and can undergo chemical reactions. They will be discussed later in this section.

The covalent interaction is often modeled as the sum of a bond distance term, a bending angle term and a torsion angle term:

$$E_{\text{bonded}} = E_{\text{bond}} + E_{\text{bend}} + E_{\text{torsion}}. \tag{2.26}$$

The regular functional form of these terms is visualized in Figure 2.4. For instance, the bond distance term is typically a quadratic function in the

bonded distance d:

$$E_{\text{bond}} = \frac{1}{2} \sum_i k_i (d - d_{0,i})^2 \tag{2.27}$$

where the sum runs over every bonded atom pair. The two parameters k_i and $d_{0,i}$ are the force constant and rest length[h] of the bond i respectively. The bond distance term resembles the energy of a spring between the two atoms, keeping them together. Bond breaking is impossible since the energy becomes infinite for large bonded distances d. The functional form is a simplification of the true (first-principles) PES between two atoms, which is not symmetric, has a steeper wall when $d \to 0$ and goes asymptotically to zero when $d \to +\infty$. However, for many use cases which do not involve bond breaking and extreme temperatures, a second-order Taylor approximation around the rest length proves to be enough. The bend and torsion terms make similar simplifications and have mostly a functional form similar to the ones depicted in Figure 2.4. Some force fields also include other terms such as out-of-plane distances and cross terms.

Electrostatic and van der Waals interactions are included in the nonbonded part of the force field:

$$E_{\text{nonbonded}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}. \tag{2.28}$$

In principle, the computation of the electrostatic energy requires the electronic charge density. As discussed before, force fields neglect the electronic wavefunction or density. Therefore, static partial charges are introduced, centered at every nucleus. Theoretically, a partial charge is not a quantum mechanical observables, which means that a unique assignment of partial charges or partitioning scheme does not exist. Thus, a variety of partitioning schemes have been developed, each with their individual properties and advantages. The oldest method is the Mulliken population analysis.[86] It starts from the density matrix to compute the so-called Mulliken charges. Another class of methods try to reproduce to electrostatic potential by fitting the partial charges. For instance, the restrained electrostatic potential (RESP)[87] method belongs to this class. The Hirshfeld[88] method or the recently developed Minimal Basis Iterative Stockholder MBIS[89] method partition the electron density. An illustration and short description of this process is depicted in Figure 2.5 where MBIS charges are fitted for the water molecule. These methods are typically run only once on the optimized geometry, when creating the force field. From that instant, the charges are fixed and do not

h. In general, the rest length is not equal to the equilibrium bond distance of that specific bond in a larger molecular system.

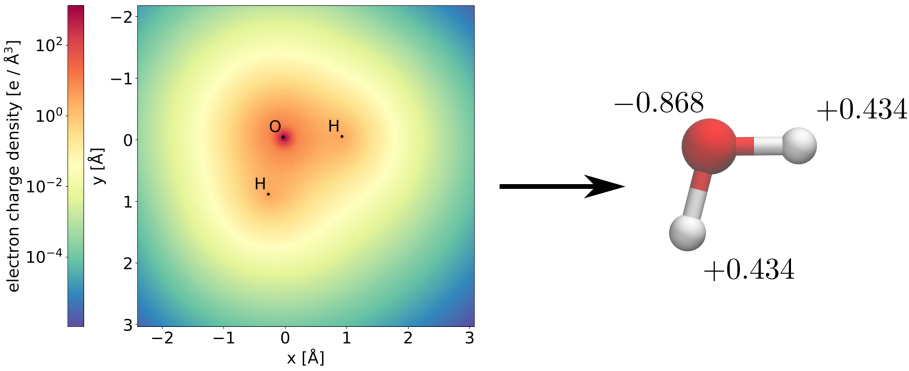


Figure 2.5: An illustration of the MBIS method to partition the electron density for the water molecule. First, the first-principles electron density is calculated. In this example, a cross section of the DFT density is shown, using the PBE0 functional. Next, a series of Slater functions $\sim \exp(-\alpha|\boldsymbol{r} - \boldsymbol{R}_a|)$ are fitted to the first-principles density. Their optimal populations (the linear coefficients belong to the Slater functions) are related the MBIS partial charges, after also taking the nuclear charge into account. In this case, the partial charges (in e) are depicted at the right for the water molecule.

change anymore, even when the geometry changes when using a force field. The electrostatic energy is simplified even further by modeling the charges as a point distribution or Gaussian distribution. This leads to the following expression for the electrostatic energy for Gaussian charges:

$$E_{\text{electrostatic}} = \frac{1}{2}\sum_{a\neq b}\frac{q_a q_b}{R_{ab}}\text{erf}\left(\frac{1}{\sqrt{2}}\frac{R_{ab}}{\sigma_{ab}}\right) \quad \text{with} \quad \sigma_{ab} = \sqrt{\sigma_a^2 + \sigma_b^2} \tag{2.29}$$

where q_a and σ_a is the partial charge and width of the Gaussian distribution of atom a and R_{ab} is the distance between atom a en b.

The van der Waals interaction has both an attractive and repulsive contribution. The dispersion interaction (see section 2.3.2) is the attractive contribution from which only the leading term proportional to R_{ab}^{-6} is retained. The repulsive contribution is inspired by the Pauli exclusion principle which forbids that two electrons occupy the same quantum state. Without the exclusion principle, ordinary matter would collapse and form a high-density phase.[90] Thus, when two atoms closely approach each other, they feel a repulsive force to push them apart. Several functional forms can be used to represent the repulsive contribution analytically in a force field. One can pick

an exponential function, inspired by the exponential scaling of the overlap of molecular orbitals, which is done in the Buckingham potential.[91] Alternatively, the Lennard-Jones potential combines a repulsive term proportional to R_{ab}^{-12} together with the leading order of the dispersion interaction:

$$E_{\text{van der Waals, LJ}} = \varepsilon \left[\left(\frac{R_0}{R_{ab}} \right)^{12} - 2 \left(\frac{R_0}{R_{ab}} \right)^{6} \right]. \tag{2.30}$$

In the discussion about force fields so far, a series of unknown parameters have been introduced such as the rest lengths $d_{0,i}$, force constants k_i, Lennard-Jones parameters ε and R_0. Before one can use the force field, the precise values of those parameters should be determined. One can choose to fit the parameters to first-principles data or experimental data. Huge amounts of first-principles data can be generated with relative ease for known and hypothetical systems. There is typically less experimental data available but here one can directly train to the measurable properties of interest. The fitting procedure itself can happen in multiple ways. One can try to optimize the force field parameters by demanding that the optimized geometries, Hessians or other properties are reproduced exactly. Hence, a unique recipe to derive a conventional force field does not exist. Recently, some tools have been developed to automatically generate force fields for molecules and condensed phases with QMDFF[92] or for MOFs with MOFFF[93] and QuickFF.[94, 95] For instance, QuickFF force fields are based on the first-principles equilibrium Hessian and geometry.

The transferability of a force field is also heavily dependent on the training data. Some force fields are widely applicable and can be used to simulate a variety of molecular systems. Examples are the universal force field (UFF)[96], the general AMBER force field (GAFF),[97] the Open Force Field[98] or CGenFF[99] for small molecules. Other force fields, only aim to accurately predict a certain class of systems. For instance, AMBER[9] and CHARMM[100, 101] are used to model biomolecules such as DNA or proteins. Even more specific force fields are only designed for one unique system. They are not transferable at all, but they are built with the highest accuracy in mind. Force fields for MOFs are usually constructed in this way, e.g. for MIL-53(Al).[102]

2.4.2 Reactive force fields

Reactive force fields are not considered conventional force fields since they allow bond breaking. Their energy contributions do not include the common analytic expressions of the previous section, like the bond energy of Eq. (2.27). A well-known reactive force field for describing metallic systems and

alloys is the embedded atom model (EAM).[103] Its functional form consists of pair-potential functions between every atom and an atomic embedding function, acting on a modeled electron density. A more recent and polarizable method is ReaxFF[104] which is also applicable to study reactions in molecules. It makes use of the bond order concept to compute the energy. The bond order is a function of the interatomic distance and can continuously change between single, double and triple bonds. When atoms move further apart, the bond order approaches zero and the chemical bond is broken. As a result of the more complex analytical expressions, reactive force fields are typically slower than the conventional non-reactive force fields. However, they are still magnitudes of orders faster than DFT and long nanosecond simulations with more than a million atoms are certainly possible.[105]

2.4.3 Polarizable force fields

All the conventional force fields in the previous section utilize static charges placed at the nuclei or no charges at all. This is a limitation if one wants to simulate a system under a different dielectric environment or when applying electric fields. Under these conditions, the charges or electron density in the system can change under the response of the environment. This is a complex effect since the already polarized charge density in the two molecules of a dimer will change in a non-additive way when a third molecule is placed in the vicinity. The force fields which account for this, are called polarizable force fields.[106–108] They have been developed as extensions to previously existing force fields, such as for AMBER[109] and CHARMM[110], or as completely new force fields, e.g. AMOEBA[111]. Typically, the polarization of the force field is realized by one of the following three methods, schematically illustrated in Figure 2.6: drude oscillators, induced dipoles or fluctuating charges. Depending on the chosen method, polarizable force fields can be a few times to more than a order of magnitude slower than conventional force fields. However, the electronic properties (dipoles, polarizabilities, …) will be predicted more accurately.

I. Drude oscillators

In the Drude oscillator model[112, 113] every atom with partial charge q_a is modeled as a core charge $q_a + \Delta q_a$, representing the nucleus and a negative shell $-\Delta q_a$, representing the electron cloud. A polarization energy

$$E_{\text{pol}} = \frac{1}{2} k d^2 \tag{2.31}$$

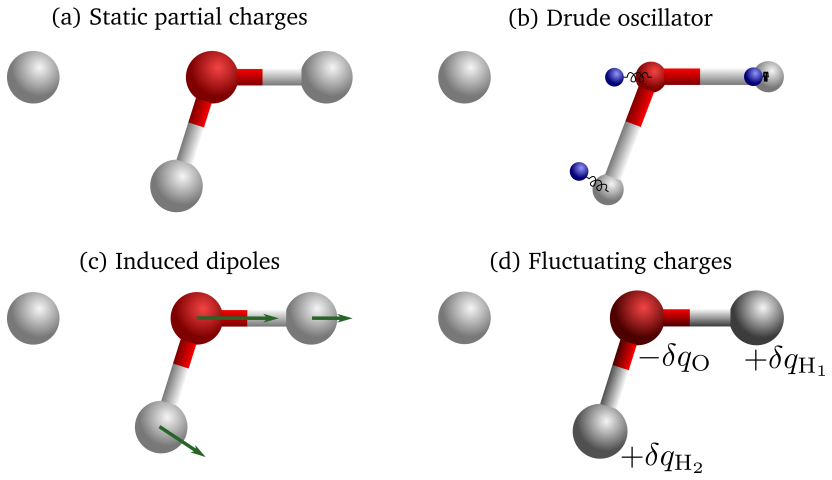


Figure 2.6: An overview of the three major polarizable force field methods. A proton (H^+) is placed within the neighborhood of a water molecule. In the non-polarizable situation of panel (a), the partial charges of the water atom are unchanged. In the Drude oscillator model, panel (b), negative charges are connected to the nuclei with a spring to model polarizability. In panel (c), the static field of the proton generates atomic dipoles in the induced dipole method. Finally, one can model polarizability with fluctuating charges δq, depicted with a color change in panel (d).

is needed to separate the two charges by a distance d. This means that a spring or harmonic oscillator connects the two charges as visualized in Figure 2.6 (b). Hence, the polarization remains a local effect since the negative electron cloud cannot move far away. Moreover, all the charges in the system interact with each other via the classical electrostatic interaction. However, the two charges, connected with the spring, are typically excluded.

Two different strategies can be followed to treat the drude oscillator model in dynamic simulations. The first approach assumes that electrons are significantly lighter than the core charges or nuclei. This leads to an adiabatic treatment of the charges, i.e. in every timestep the total energy is minimized with respect to the positions of the negative shells. This can be done with an iterative optimizer such as the BFGS algorithm.[61] In most situations, a dozen energy evaluations are necessary per timestep, which makes it an expensive approach. Alternatively, one can assign a certain mass to the negative electron cloud and perform MD simulations as if they were additional particles. Small masses (< 1 amu) are more appropriate, since they do not distort the dynamics of the system, but require smaller timesteps. Hence,

also this method will be slower than conventional force fields.

II. Induced dipoles

As the name suggests, the induced dipole method introduces a dipole moment $\boldsymbol{\mu}_a$ at the position of every nucleus a (visualized in Figure 2.6). They are induced by the local electric field $\boldsymbol{\mathcal{E}}_a$ at that point:[114, 115]

$$\boldsymbol{\mu}_a = \boldsymbol{\alpha}_a \cdot \boldsymbol{\mathcal{E}}_a = \boldsymbol{\alpha}_a \cdot \left[\boldsymbol{\mathcal{E}}_{0,a} - \sum_{b \neq a} \mathsf{T}_{ab} \boldsymbol{\mu}_b \right] \tag{2.32}$$

where $\boldsymbol{\alpha}_a$ is the polarizability tensor. The local electric field $\boldsymbol{\mathcal{E}}_a$ is the sum of the electric field $\boldsymbol{\mathcal{E}}_{0,a}$ from the permanent static charges and the field generated by the other induced dipoles. The dipole-dipole interaction tensor,

$$\mathsf{T}_{ab} = \frac{1}{R_{ab}^3}\mathsf{I} - \frac{3}{R_{ab}^5} \begin{pmatrix} x^2 & xy & xz \\ yx & y^2 & yz \\ zx & zy & z^2 \end{pmatrix} \tag{2.33}$$

with I the identity matrix, and $\boldsymbol{R}_{ab} = (x, y, z)$ the interatomic vector from a to b, occurs in the latter interaction. The total induced dipole energy is given by

$$E_{\text{ind}} = -\sum_a \boldsymbol{\mu}_a \cdot \boldsymbol{\mathcal{E}}_{0,a} + \frac{1}{2} \sum_{a \neq b} \boldsymbol{\mu}_a \mathsf{T}_{ab} \boldsymbol{\mu}_b + \frac{1}{2} \sum_{a \neq b} \boldsymbol{\mu}_a \boldsymbol{\alpha}_a^{-1} \boldsymbol{\mu}_b \tag{2.34}$$

which is minimized by the values of the dipole moments in Eq. (2.32). The induced dipoles can be exactly determined since they are the solution of the linear matrix equation (2.32): $\boldsymbol{\mu} = \mathsf{A}^{-1}\boldsymbol{\mathcal{E}}_0$ with A containing the inverse polarizabilities on its diagonal and the dipole-dipole interaction tensors on the off-diagonal. Hence, in dynamical simulations, the matrix A should be inverted to determine the induced dipoles which is a costly operation (it scales as $\sim N_n^3$). Alternatively, a cheaper strategy is to solve Eq. (2.32) self-consistently. Other more efficient schemes do also exist.[116, 117] These methods are still more expensive than conventional force fields however.

III. Fluctuating charges

Fluctuating charge methods encompasses a collection of different techniques, which all model changes in polarization by allowing the atomic partial charges to fluctuate (see Figure 2.6). They are now functions of the local or global environment of the atom. The electronegativity equalization

method (EEM)[110, 118] and charge equilibration methods (QEq)[119, 120] are based on the second order Taylor expansion of the energy to create the set of fluctuating charges $\{q_a\}$:

$$E(\{q_a\}) = \sum_a \left(E_a^0 + \chi_a^0 q_a + \frac{1}{2} J_{aa}^0 q_a^2 + \sum_{b \neq a} J_{ab} q_a q_b \right) \tag{2.35}$$

where χ_a^0 is the electronegativity of a single atom and J_{aa}^0 the hardness of the atom a. The coefficients J_{ab} depend on the interatomic distances R_{ab} and should approximate the regular Coulomb scaling $\sim R_{ab}^{-1}$ for large distances. To determine these atomic charges, the total energy should be minimized while constraining the total charge with a Langrange multiplier λ:

$$\frac{\partial}{\partial q_a} \mathcal{L}(\{q_a\}) = 0 \quad \text{with} \quad \mathcal{L}(\{q_a\}) = E(\{q_a\}) - \lambda \left(\sum_a q_a - q_{tot} \right) \tag{2.36}$$

This results in

$$\lambda = \frac{\partial E(\{q_a\})}{\partial q_a} = \chi_a^0 + J_{aa}^0 q_a + \sum_{b \neq a} J_{ab} q_b. \tag{2.37}$$

Since $\frac{\partial E(\{q_a\})}{\partial q_a}$ is equal to a constant, it should take the same value for every atom. Furthermore, the first order derivative of the energy with respect to an atomic charge is the electronegativity χ of that atom. Hence, the electronegativity is equalized for every atom, giving rise to the name of this method. Since Eq. (2.37) is a linear equation, the partial charges are found by inverting the matrix of all the linear coefficients. Just like the Drude oscillator and induced dipole method, this increases the computational cost and scaling of the algorithm.

Fluctuating charge methods are also a widespread approach to incorporate long-range interactions in MLPs.[121–128] In the so-called third generation of MLPs,[129] the atomic partial charges are a function of the local atomic environment, see Chapter 3. More advanced approaches, where the charges depend on the global structure, belong to the fourth generation of MLPs. For instance, after learning environment dependent electronegativities, the charges are computed with variations on the charge equilibration neural network technique (CENT)[125–127] or kernel Charge Equilibration (kQEq).[128] Both make use of the charge equilibration scheme described above. Besides MLPs, fluctuating charges also appear in a variety of conventional and reactive force fields such as ReaxFF.[104]

Fluctuating charge methods and specifically the ones who are using derivatives of the charge equilibration method, can suffer from nonphysical issues.

At large distances, there can be substantial charge transfer[112] which in turn can lead to serious over- or underestimation of dipole moments. A related problem is the dissociation limit of ionic compound such as NaCl. There is no constraint on the partial charges of the single Na and Cl atoms at infinity. Only the sum of the two charges should be zero. A correct physical model knows that electrons are inseparable and predicts an integer charge on both atoms, which is not necessarily the case here.

2.4.4 Ambiguous polarization with fluctuating charges

One of the key concepts in the modern theory of polarization,[130] is that the dipole moment itself has no physical meaning in periodic systems. The only thing which is well-defined, has a physical meaning and can be measured experimentally, is the change of the dipole moment. In the beginner's guide to the modern theory of polarization by Spaldin,[131] this was illustrated with a one-dimensional example for fixed partial charges. In the **eMLP Paper**, the same example is generalized to fluctuating charges. This led to the identification of a major issue: in periodic systems, the macroscopic polarization is not well-defined.

In this section, we will repeat the example of the **eMLP Paper** because it is a not well-known fact and it invalidates the use of most fluctuating charge methods in periodic systems, which also is the most popular method to include long-range interactions in MLPs. Most fluctuating charge methods simply model the changes of the charges or electron density directly. They do not explicitly define the charge transfer in the system, i.e. specifying how the charges flow between the nuclei or where the electrons move to. This will be the main culprit, causing the ambiguity. Note that not all fluctuating charge methods suffer from this issue. For instance, in the split-charge model[120] or electron-passing neural network (EPNN)[123] the movements of the charges are defined. Furthermore, in the following example, the fluctuating charges are not necessarily derived from a charge equilibratium method, such that there is no relation with the problem of superlinear polarizability scaling for these methods.[132]

In Figure 2.7, the same one-dimensional periodic system with lattice length a is described within two different configurations. In configuration A, the particles with charge $-q$ and $+q$ are located at $a/4$ and $3a/4$ respectively, at time t_0. In configuration B, the origin of the unit cell is shifted over a distance $a/2$ such that the negative particle is now located at $3a/4$ and the positive particle at $a/4$. Again, we emphasize that both configurations corresponds to exactly the same physical situation due to the periodic boundary conditions.

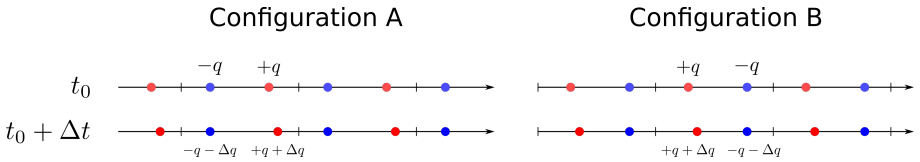


Figure 2.7: A one-dimensional example to illustrate that fluctuating charge methods lead to ambiguous polarization in periodic systems. In configuration A and B, exactly the same system is described, only the origin of the unit cell is shifted over half the lattice length. When going from t_0 to $t_0 + \Delta t$, the positive charge moves to the right and gains Δq in charge, while the negative charge remains at the same position but has lost a charge of Δq. Reprinted with permission from Ref. 10 Copyright 2022 American Chemical Society.

The dipole moments of both configurations at t_0 are:

$$\mu_A(t_0) = \frac{3a}{4}q - \frac{a}{4}q = \frac{a}{2}q \qquad \mu_B(t_0) = -\frac{3a}{4}q + \frac{a}{4}q = -\frac{a}{2}q \quad (2.38)$$

Two different results are obtained.[i] This is why the dipole moment itself is not well-defined. To calculate the change in polarization $\Delta\mu = \mu(t_0 + \Delta t) - \mu(t_0)$, the positions and charges of the particles at $t_0 + \Delta t$ are required. At that instant, the positively charged particle has become even more positive by gaining an extra charge Δq and it moved a distance Δr to the right. The negative particle did not move but it lost a charge of Δq. Hence, the change in dipole moment is:

$$\Delta\mu_A = q\Delta r + (\Delta r + \frac{a}{2})\Delta q \qquad \Delta\mu_B = q\Delta r + (\Delta r - \frac{a}{2})\Delta q \quad (2.39)$$

Again, two different values are acquired, showing that even the change in polarization is not well-defined for fluctuating charges. When dealing with permanent charges $\Delta q = 0$, the procedure would yield the same result $\Delta\mu_A = \Delta\mu_B = q\Delta r$ as required. This one-dimensional example can be generalized to three dimensions and multiple particles, but the same conclusions will be valid. Naively using fluctuating charges, leads to ambiguous polarization in periodic systems and should be avoided. Note that the problematic term is equal to the product of the charge Δq and the distance $\Delta r \pm \frac{a}{2}$ over which the non-permanent charge has moved. In configuration

i. Even for non-periodic systems, the dipole moment depends on the specific origin of choice for a system with a nonzero charge. This is a constant difference, which vanishes when computing the polarizability.

A, it moves to the right, while in configuration B, it moves to the left. When a force field uniquely defines the charge transfer, i.e. the direction of the displacement in this example, the problematic terms would be the same. Hence, one should only work with fluctuating charges if the charge transfers are uniquely defined.

2.4.5 Explicit-electron force fields

The computation of some useful properties remains challenging, even with polarizable force fields. Polarization is inherently local in the induced dipole or drude oscillator method, which makes it difficult to model long-range charge transfer. Furthermore, ionization energies or electron affinities are inaccessible. On the other hand, fluctuating charge methods suffer from ambiguous polarization or may possess nonphysical superlinear scaling.[132] Explicit-electron force fields provide an alternative approach to model polarizability and chemical reactions in molecular and periodic systems. In these methods, the electrons are additional and explicit semi-classical particles in the force field, which are allowed to move freely throughout the whole system.[133] This is visualized in Figure 2.8 for the water molecule for the eMLP[10] force field. Because of their nonlinear response and free movement, explicit electrons allow for a natural treatment of complex phenomena such as redox reactions, piezoelectricity or charge transfer.[134–137] Moreover, they carry integer and permanent charges, -e for single electrons and -2e for electron pairs, such that correct dissociation limits are possible and the ambiguity of fluctuating charges are avoided. They are also able to introduce spin at the force field level because of the distinction between spin-up and spin-down particles.

There is no unique recipe or procedure to construct explicit-electron force fields. First of all, one can choose to model all electrons separately as different particles or group them together in one way or another. Grouping them in so-called electron pairs, containing both one spin-up and spin-down electron, is a popular technique to reduce the computational cost. It has the obvious disadvantage that radicals cannot be described anymore. Optionally, one can only take the valence electrons into account, in analogy with the frozen core approximation of first-principles methods. In more drastic simplifications, only a single electron or hole is modeled, representing the excess negative or positive charge, or one can work with positive and negative shells with integer charges for each atom.[j] Another relevant choice is whether the electrons have a constant or variable finite extent. In practice, the electrons

j. Here, the negative charges should be able to move freely. This is in contrast with the drude oscillator model, where they are linked with an harmonic bond term.

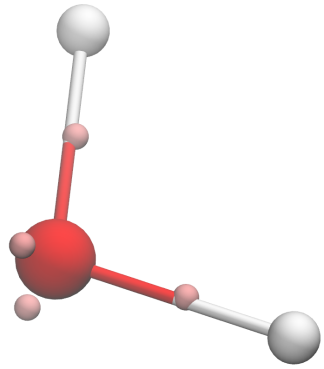


Figure 2.8: A visualization of the water molecule in an explicit-electron force field (eMLP)[10]. Here, only the valence electrons are included and spin-up and spin-down electrons are grouped into electron pairs. The positions of the electron pairs correspond to the well-known Lewis structure[138] of water: two single bonds and two lone pairs.

are represented using Gaussian charges and the width is varied for the latter choice. In theory, variable widths allow a better fit to the true (first-principles) electron density. Finally, the functional form of the energy and all its contributions determine the accuracy and transferability of the force field. It is important to stress that the electrons behave completely different than the nuclei in conventional force fields. Therefore, one cannot simply copy the energy expressions from conventional forces. Instead, there should be exchange contributions since the electrons obey the exclusion principle.[139] In general, the electron interactions are considerably more complex which makes deriving explicit-electron force fields a challenging task. As a final note, the force field parameters can be fitted to experimental or first-principles results, just like conventional force fields.

Some extra considerations are necessary when using explicit-electron force fields in dynamic simulations. There are two common approaches, which are similar to the ones encountered in the drude oscillator discussion. The first and most commonly used approach is to treat the electrons as active degrees of freedom in MD simulations, analogous to Car–Parrinello molecular dynamics (CPMD).[140] A small mass is chosen for the explicit-electron particles (< 1 amu), together with a small timestep, for instance 0.1 fs and the electrons are propagated further in time, just like the nuclei in conventional MD simulations. Special care should be taken when dealing with a thermostat or barostat. The electrons and nuclei should be decoupled to avoid nonphysical energy transfer. The alternative approach is to treat the electrons adiabatically. This is the zero mass limit of the previous approach,

i.e. in every timestep, one should minimize the total force field energy with respect to all the electronic degrees of freedom to determine the locations of the electron particles. This corresponds with the normal Born-Oppenheimer approximation. An iterative process of consecutive energy and force evaluations is required to minimize the energy, which is typically more expensive than the first approach.

Most explicit-electron force fields are at most only fifteen years old. The electron force field (eFF) [141, 142] was originally constructed to model warm dense mater, i.e. high temperatures and pressure, and can predict electron ejections. Modifications to this force field, lead to the eFF with effective core potentials (eFF-ECP),[143, 144] which is also applicable to model molecules. Around the same time, the LEWIS force field was developed to model liquid water and its acid-base behavior.[145, 146] Its successor, LEWIS•, extends the area of applicability to diatomic molecules and atoms from the second and third period.[134, 147, 148] It can predict the correct order of different spin states, ionization energies and electron affinities. A more recent explicit-electron force fields is eReaxFF.[135] Besides electrons, it also allows the inclusion of holes as additional particles and it shows promising results to model advanced systems, such as redox reactions in lithium-ion batteries.[136] Another ReaxFF modification uses the coarse-grained electron model (C-GeM),[137, 149] for which every atom is composed of a negative and positive shell and has also been applied to water. The quantum mechanical polarizable force field (QMPFF)[150] also makes use of explicit electron clouds which are constraint to their corresponding nucleus.

3

Machine learning potentials

I do not fear computers. I fear the lack of them.
Isaac Asimov (1920–1992)

In this chapter, we will explore why machine learning potentials (MLPs) are revolutionizing the field of molecular modeling. Nowadays, MLPs start to replace conventional force fields for more and more applications. This is all made possible by machine learning techniques originally developed for computer vision or other scientific purposes. We will start off with an introduction to MLPs and proceed with data generation techniques, which is an essential step to create accurate and transferable MLPs. Next, the invariances and typical atomic descriptors of MLPs will be discussed, together with the most common cost functions. Afterwards, an overview of the various MLP-methodologies will be given. The emphasis is on (deep) neural networks, although a short introduction to kernel regression techniques will not be omitted. Here, all the details about the necessary building blocks for neural networks (NNs) will be listed. Starting from descriptor-based NNs, we will move towards the more complex architectures of message passing NNs and the recently developed equivariant NNs. Afterwards, we will examine how MLPs incorporate long-range interactions. Finally, with all the available information, the development of MLPs for MOFs will be discussed.

3.1 The basics of machine learning potentials

Fundamentally, machine learning potentials[20–24] are force fields that are trained on first-principles data using machine learning methods. The underlying quantum mechanical laws are automatically extracted and learned without human bias. The accuracy of the MLP will in first instance depend on the level of theory chosen to generate the data set and how well the targeted phase space is sampled. Using the available data, the goal of an MLP is to learn the PES, i.e. the relation between the energy E and the atomic positions $\{\boldsymbol{R}_a\}$ and corresponding species $\{Z_a\}$:

$$E = f(\{\boldsymbol{R}_a\}; \{Z_a\}; \{\theta_k\}). \tag{3.1}$$

Not all possible functions f are valid. The energy should be invariant to global translations and rotations. Furthermore, all nuclei belonging to the same species are indistinguishable particles and hence, the MLP should also be invariant under exchange of such identical nuclei. These invariances and their consequences will be studied in section 3.1.2. Just like conventional force fields, MLPs depend on a number of parameters $\{\theta_k\}$, which are fitted to first-principles data. However, most MLPs have substantially more parameters than conventional force fields. It is not uncommon for deep learning potentials to have more than one million parameters. The huge number of parameters give MLPs the sought for flexibility to describe chemical reactions or any other challenging PES. If the PES is known, the force on atom a is simply the derivative of the energy with respect to corresponding nuclear coordinates:

$$\boldsymbol{F}_a = -\boldsymbol{\nabla}_a E, \tag{3.2}$$

while the stress tensor for a periodic system is given by:

$$\boldsymbol{\sigma} = \frac{1}{V}\frac{\mathrm{d}E}{\mathrm{d}\boldsymbol{S}} = \frac{1}{V}\mathrm{A}^T \cdot \frac{\mathrm{d}E}{\mathrm{dA}} \tag{3.3}$$

with $\boldsymbol{S}$ the strain tensor, $V = \det(\mathrm{A})$ the volume of the system and A the cell matrix.

The exact shape of the function f in Eq. (3.1) is MLP-specific. However, the majority of MLPs use an atomic energy decomposition:

$$E = \sum_a E_a(\boldsymbol{x}_a) \tag{3.4}$$

where $\boldsymbol{x}_a$ are the atomic descriptors of atom a. MLPs are only trained to the total energy and the resulting atomic energies are merely a consequence of the fit and do not have a physical meaning per se. Moreover, first-principles

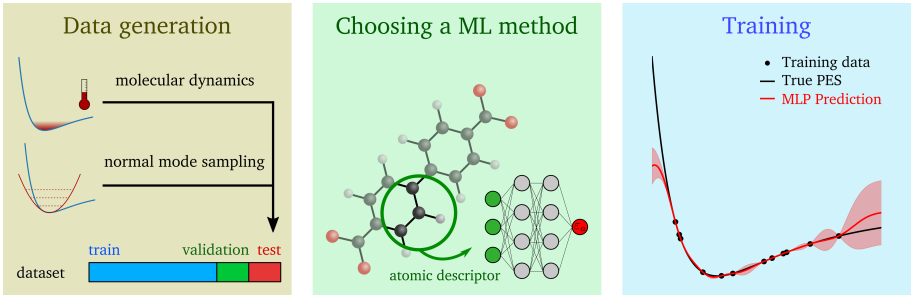


Figure 3.1: A schematic overview of the three tasks to be completed when constructing a new MLP. (i) Data generation: a single or combination of sampling strategies should be chosen to cover the targeted phase space as efficiently as possible. (ii) A specific MLP suited for the system at hand should be chosen. This includes picking the right atomic descriptor and machine learning method. (iii) The MLP should be trained on the available data.

atomic energies are also not unique. The decomposition is based on the concept of nearsightedness.[151, 152] It tells us that electronic properties are primarily dependent on the effective external potential[a] in neighboring points and are almost insensitive to changes far away. Therefore, the dependence of the atomic descriptors $\boldsymbol{x}_a$ is limited to a sphere with cutoff radius r_{cutoff} around a central atom a. Typical cutoff radii vary from 3 Å to 8 Å. MLPs with smaller cutoff radii are computationally more efficient but lack the ability to model more long-range effects. In subsection 3.3, additional ways to incorporate long-range effects in MLPs will be discussed.

The following three steps should always be completed when constructing a new MLP: data generation, choosing a machine learning method and training the MLP. These steps are schematically visualized in Figure 3.1. The first and potentially most important step is data generation. It requires knowledge of the system under consideration and will typically be the most time-consuming step. Moreover, the final accuracy and transferability of the MLP will only be as good as the data being provided. All the essential details to generate excellent data will be discussed in the next subsection. Picking the right MLP for the system at hand requires knowledge of the strengths and weaknesses of the available state-of-the-art MLPs. The size of the system, whether molecules or condensed phases are being described, the number of different chemical species and the size of the data set will all play a crucial role in that decision. A review of a number of specific machine learning

a. The effective external potential is the sum of the external potential and effects from the self-consistent electric fields generated by the other electrons.

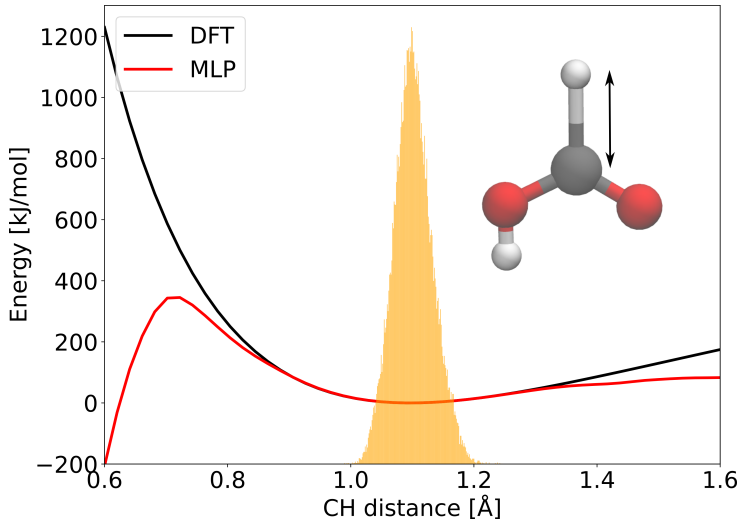


Figure 3.2: A potential energy scan of the CH-bond in formic acid with an MLP (SchNet[153]) and DFT (PBE0[11, 76]). In the training set of the MLP 51480 different CH-bonds are included, all lying approximately between 1.0 and 1.2 Å and indicated by the orange histogram. Outside this range or when extrapolating, the MLP displays erratic behavior.

methods and their advantages or disadvantages will be given in Section 3.2. Finally, when the data set is constructed and the MLP is chosen, one can start to train. Once these three steps are completed, one should always validate the MLP and if necessary, adjust one of the three steps in the process. Only then, the MLP will be ready to use.

3.1.1 Data generation

The general performance and accuracy of any MLP critically depends on the quality of the training data. Therefore, proper data generation techniques are of utmost importance. The level of theory of the data, the amount of data and the configurations that should be sampled all depend on the goal that the MLP should accomplish. The main idea is that the configurations which will be encountered in later simulations with the MLP, are similar to the ones stored in the data set. Otherwise, the MLP is said to be extrapolating. This should be avoided since no physical constraints are built into MLPs, leading to unpredictable and erratic behavior when extrapolating. This is made clear in Figure 3.2. Here, a potential energy scan of an MLP and the first-principles DFT reference is plotted along the CH-bond in formic acid. In the

training set only CH-distances between 1.0 and 1.2 Å are included, visualized by the orange histogram. In this area, the MLP and DFT predictions are in almost perfect agreement. However, at short distances the energy of the MLP becomes negative, contradicting the true repulsive wall in DFT. Unlike conventional force fields, MLPs do not have any physical constraints, leading to these unpredictable potential energy surfaces. Negative errors are exceptionally troublesome:[154] simulation will eventually visit these regions more than they normally would. Hence, the simulation will stay longer in these ill-fitted regions, potentially worsening all the results. This example highlights the necessity of a well sampled data set for the goal at hand.

Several strategies exist to thoroughly sample the phase space. The most straightforward but time-consuming method is to perform first-principles MD simulations. The sampled phase space is then exactly equal to the target phase space for many applications since the MLPs are typically used for MD simulations. Hence, all structures in the data set will be relevant. For certain systems, enhanced sampling techniques[155] may be necessary to obtain information about regions that are not visited by conventional MD simulations. Here, the system is biased along a certain collective variable (e.g. the volume) such that also less probable states are encountered. MD simulations have one major disadvantage however: consecutive MD steps are highly correlated. Hence, long simulations are necessary to sample all possible configurations. This is too expensive for many extended systems such that first-principles MD simulations are avoided. To overcome this issue, one can perform them at a lower level of theory and periodically recompute a configuration at the desired level of theory. This speeds up the data generation process but there is no guarantee that the MD simulation samples the correct distribution.

An alternative approach is normal mode sampling (NMS), which already has been successfully applied to small molecules.[156, 157] It makes use of the harmonic approximation (see section 2.2.2). For small perturbations or low temperatures, an harmonic PES will be an adequate approximation. In that case, the Boltzmann distribution is proportional to:

$$p(\boldsymbol{x}) \sim \exp\left(-\frac{E}{k_\mathrm{b}T}\right) = \exp\left(-\frac{\boldsymbol{x}^T \mathrm{H} \boldsymbol{x}}{2k_\mathrm{b}T}\right) \tag{3.5}$$

with $\boldsymbol{x}$ the degrees of freedom of the system (positions or cell vectors), H the Hessian and T the temperature. This is a multivariate normal distribution which can be easily sampled for $\boldsymbol{x}$, avoiding the need for expensive MD simulations. All the samples drawn from the distribution will be uncorrelated by construction, resulting in a data-efficient sampling strategy. Often, NMS is combined with dimer and torsion sampling for small molecules.[157] In dimer

sampling, a second molecule is placed in the vicinity of the first molecule with a random orientation and relative distance with the aim to learn intermolecular interactions. In torsion sampling on the other hand, a rotatable bond of the molecule is selected at random and one part of the molecule, separated by that bond, is randomly rotated along that axis. Unfortunately, NMS has some disadvantages. First of all, one can only thoroughly sample a single phase. Multiple phases require multiple Hessians. Moreover, one cannot sample non-quadratic regions of the PES such as chemical reactions or phase transitions. A final disadvantage of NMS is that, contrary to the efficiency to sample configurations, the computation of the Hessian itself is not cheap. Typically, it requires $3N_{\mathrm{n}}$ additional first-principles force calculations since many codes do not have an analytical implementation of the Hessian. If available, the latter approach should in principle be faster and generate less numerical noise.

The amount of data that is necessary to construct an MLP depends on a variety of elements. In the first place, every MLP will have a different data efficiency. This aspect of MLPs will be covered in section 3.2. Second, the size of the chemical phase space being sampled will matter. For instance, a universal MLP for multiple molecules will require more data than an MLP for one specific molecule if the same accuracy is desired in both cases. The available computational budget will limit the number of configurations that can be computed. Moreover, it can restrict the level of theory that can be used. This is especially critical in use-cases such as modeling reactions or radicals, which require post-Hartree-Fock methods. At the end, the computational cost, the amount of data, level of theory and final accuracy of the MLP are intricately related. Typical data set sizes start from a few hundred structures for MOFs to thousands and even millions for small molecules.

Once the data is computed and ready to use, a few preprocessing steps have to be completed. All the data should be divided into a training, validation and test set. Using the training set, the parameters of the MLP are fitted by minimizing the cost functions of section 3.1.3. After (or during) training, the performance of the MLP is assessed on the validation set. This gives an indication of the accuracy on unseen data. The optimal set of hyperparameters is determined by minimizing the validation accuracy obtained with different values of hyperparameters. Hence, the validation set is not used to optimize the parameters but to optimize the hyperparameters of the MLP. Possible hyperparameters of MLPs are the cutoff radius, number of layers in a neural network or the number of atomic descriptors in kernel regression. The final performance of the MLP is assessed on the test set after the MLP is trained with the best possible set of hyperparameters. The test set contains unseen data so far, providing a clear indicative measure of the generalization

accuracy. This number is often published and can be compared with other MLPs.

Regularly, the train, validation and test set contain 80%, 10% and 10% of all the data respectively. Other data partitions are certainly possible but the training set usually contains the majority of the data. It is crucial however that the three sets sample from the same data distribution in chemical space. However, the three sets should remain uncorrelated and independent. This means that two consecutive steps in an MD simulations should not be in two separate data sets. If this were the case, the performance on the test or validation set would be very similar to the training set, undermining their very purpose, since overfitting issues cannot be detected anymore.

One essential parameter we skipped so far in this discussion, is the temperature T to perform the MD simulations or NMS. Naively, one would think that one has to generate data at the same temperature as the target temperature. This is not the case. Ideally, one should generate data at higher temperatures. To confirm this, we constructed four data sets of MIL-53(Al) at four different temperatures from 300 K to 600 K. Next, four different SchNet[153] MLPs were trained on each set and their final performance was validated on the test set of each temperature. All the other hyperparameters (amount of training data, MLP, cost function ...) remain unchanged. In Figure 3.3 the mean absolute error (MAE) on the forces are tabulated for each different combination of train and test set temperature. Just by looking at the diagonal, top left to bottom right, it looks like the test error increases with temperature from 38.2 to 47.5 meV/Å. Based on these numerical results, one can erroneously conclude to keep the temperature low. However, when all possible temperature combinations are taken into account, the model trained at 600 K outperforms the model trained on 300 K on all temperatures. This is even the case at 300 K, the target temperature of the latter. This is explained by noticing that MLPs can make huge errors at the edge of the sampled data distribution (see also Figure 3.2), which are only occasionally visited in MD simulations. By increasing the temperature, the MD simulations of the training data will more frequently visit these regions, ultimately decreasing the error if the test set remains fixed.

To end this section, we will list some advanced data generation techniques. The first one is active learning.[158–161] In this technique, the training data set is automatically selected and generated. This is often done for neural network MLPs with a query by committee (QBC) strategy.[162] First, a small initial data set is constructed and several different MLPs are trained to this data set. This is the committee. In the next step, the relevant chemical space is being explored with the committee via MD simulations out of which new configurations are selected where the committee is most uncertain about.

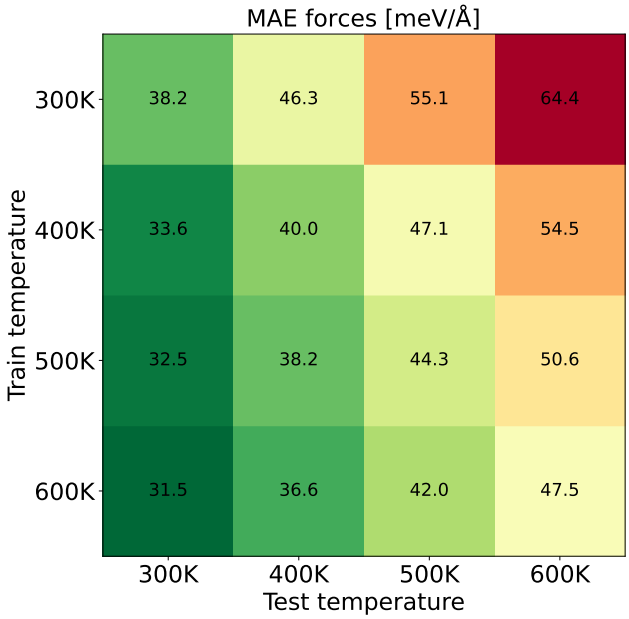


Figure 3.3: Four different models are trained on four training sets, generated at 300 K, 400 K, 500 K and 600 K (horizontal) and validated on test sets at the same temperatures (vertical). For each of the combinations the resulting mean absolute errors on the forces are depicted. SchNet[153] has been used here for MIL-53(Al).

Ideally, these structures correspond to configurations where the MLP is making the largest errors and hence, they are added to the data set to improve the MLP. Finally, the whole procedure is repeated. As such, there is an intricate feedback loop between the MLP and the data itself, resulting in a very data efficient protocol. Another advanced technique is the so-called on-the-fly learning.[163, 164] Here, the MLPs also provide an uncertainty measure while in production. If their uncertainty is too high, a first-principles calculation is done instead. In Δ-learning (delta learning),[165] there are two or more MLPs in play. Basically, the first MLP is trained in the conventional way with a large amount of data at the DFT level of theory. However, the second MLP will learn the difference between the predictions made by the first and a higher reference level such as CCSD(T). In this way, less computationally expensive data is necessary for the high level of theory calculations.

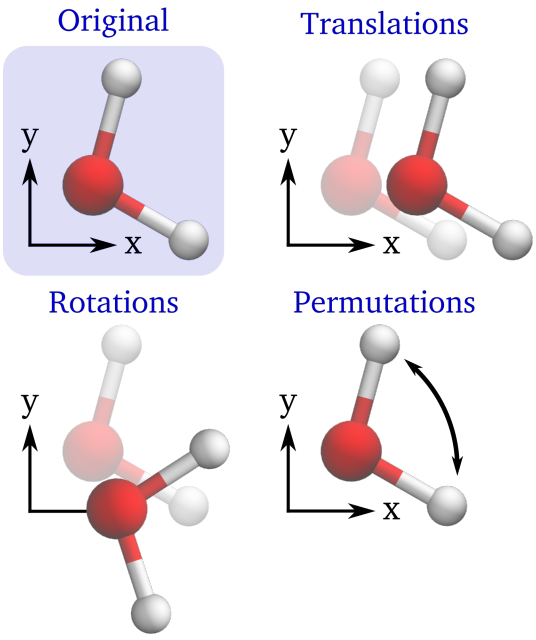


Figure 3.4: A visualization of the three invariances of MLPs. After a translation (top right), rotation (bottom left) or permutation of equivalent atoms (bottom right), the energy of the system should be equal to the original system (top left).

3.1.2 Invariances and atomic descriptors

The energy of a molecule or periodic system is invariant under global translation, rotations and permutations of equivalent atoms. These invariances are visualized in Figure 3.4. The first two are the cause and effect of the conservation of momentum and angular momentum. The invariance under perturbations is a consequence of the indistinguishability of particles belonging to the same species. Unlike conventional force fields, there are no atom types and every atom is treated equivalently. The functional form of the MLP should obey these three principles. Otherwise, there will be no conserved quantities during MD simulations. Furthermore, these invariances also increase the data efficiency of the MLP. Instead of all possible functions for the energy to learn, only the subset of invariant functions should be taken into account.

The energy of an atomic system is automatically invariant if the atomic descriptors $\boldsymbol{x}_a$ in Eq. (3.4) also obey the same invariances. The most straightforward approach is to make these descriptors only dependent on the

interatomic distances R_{ab}. These remain unchanged under translations and rotations. Moreover, constructing descriptors by summing over equivalent atoms also fulfills the invariance under permutations.

A variety of atomic descriptors have been designed with these ideas in mind. Their purpose is to numerically represent the chemical environment of every atom, which is the input of an MLP. They should locally encode as much chemical information as possible to learn the PES in an efficient way. The first descriptors were feature engineered. Their functional form was determined by human intuition and numerical experimentation. The first descriptor for general atomic systems was introduced by Behler and Parrinello in 2007: the atom-centered symmetry functions (ACSF).[36, 37] They encoded radial information by using the following radial descriptors for atom a:

$$G^{\text{rad}}_{a,\eta\mu} = \sum_b \mathrm{e}^{-\eta(R_{ab}-\mu)^2} f_{\text{cutoff}}(R_{ab}) \tag{3.6}$$

where the sum runs over every atom b within the cutoff radius r_{cutoff} of atom a, only providing local information. To avoid discontinuities in the energy and forces when an atom leaves the cutoff radius, the atomic descriptors and their first derivatives should be made continuous as well. This is realized by multiplying the descriptors with a smooth cutoff function. Often, the following functional form is used:

$$f_{\text{cutoff}}(r) = \begin{cases} \frac{1}{2}\left[\cos\left(\pi \frac{r}{r_{\text{cutoff}}} + 1\right)\right] & \text{if} \quad r \leq r_{\text{cutoff}} \\ 0 & \text{if} \quad r_{\text{cutoff}} < r \end{cases}, \tag{3.7}$$

introducing a smooth transient region from one at the center to zero at cutoff radius. The hyperparameters η and μ are real numbers and label the specific descriptor elements. For instance, descriptors where $\mu = 1\,\text{Å}$ will approximately count the number of neighbors lying at that distance and will be important to the model covalent bonds. Angular information is included by extending the set of descriptors with angular descriptors:

$$\begin{aligned} G^{\text{ang}}_{a,\eta\zeta\lambda} = 2^{1-\zeta} \sum_{bc} &(1 + \lambda \cos\theta_{abc})^{\zeta}\, \mathrm{e}^{-\eta(R^2_{ab}+R^2_{ac}+R^2_{bc})} \\ &\times f_{\text{cutoff}}(R_{ab}) f_{\text{cutoff}}(R_{ac}) f_{\text{cutoff}}(R_{bc}) \end{aligned} \tag{3.8}$$

Here, the sum runs over the neighbors b and c, while θ_{abc} is the angle between these neighbors and the central atom a. Again, the hyperparameters ζ, η and λ define the specific descriptor. The value of λ can be ± 1 and shifts the maximum value of the descriptor to $\theta = 0^\circ$ or 180°. The total atomic descriptor is the concatenation of the chosen set of radial and angular

descriptors $\boldsymbol{x}_a = (\boldsymbol{G}_a^{\text{rad}}, \boldsymbol{G}_a^{\text{ang}})$. Afterwards, it is the task of the MLP to learn the relations between these descriptors and the energy and forces of the system. Besides the ACSFs, many other atomic descriptors have been developed for kernel-based MLPs as well as neural networks.[39, 40, 166–168] All these approaches fulfill the three invariances.

The descriptors defined above implicitly assume an atomic system with only one species. The equations (3.6) and (3.8) cannot distinguish between different species as the sum runs indiscriminately over every atom. Therefore, the ACSFs are extended to incorporate that information. For instance, the angular descriptors become:

$$G_{a,\eta\zeta\lambda}^{\text{ang},BC} = 2^{1-\zeta} \sum_{b\in B, c\in C} (1 + \lambda \cos\theta_{abc})^{\zeta}\, \mathrm{e}^{-\eta(R_{ab}^2+R_{ac}^2+R_{bc}^2)} \times f_{\text{cutoff}}(R_{ab}) f_{\text{cutoff}}(R_{ac}) f_{\text{cutoff}}(R_{bc}) \tag{3.9}$$

where the sums runs over every atom b and c, belonging to the species B and C respectively. Hence, for a system with N_S species, there are $\frac{N_\text{S}(N_\text{S}+1)}{2}$ times more descriptors if all possible combinations are included. The radial descriptors can be extended analogously, resulting in N_S more descriptors. The quadratic increase of number of descriptors and therefore in computational cost (and memory[b]) is a major disadvantage since it often limits their applicability to systems with only a handful of species. Some developments have been made to avoid the unfavorable scaling but they are not widely used yet. For instance, the weighted atom-centered symmetry functions (wACSF) introduce weight factors in the definition of the descriptors to incorporate information about different species.[169]

The descriptors discussed above are all feature engineered. Human experience is required to select them, which undoubtedly introduces human bias in the system. This can potentially be beneficial but to avoid any bias and to simplify picking the right descriptors, automatically generating descriptors is preferred. This is one of the reasons why end-to-end or message passing neural networks (MPNNs) were developed.[42] Additionally, they do not suffer from the quadratic scaling in number of species, making them an ideal tool to model complex systems with multiple species. The only input required in MPNNs are the interatomic distances, often expanded in a radial basis. This will be discussed in greater detail in section 3.2.4.

b. This only becomes an issue if all the descriptors are stored in one array or tensor. This is typically done on GPUs to parallelize the calculations.

3.1.3 Cost function

The parameters $\boldsymbol{\theta}$ in an MLP are fitted by minimizing a cost function. Typically, a mean squared error (MSE) cost function is chosen. In that case, the optimized set of parameters of a linear MLP can be analytically determined, which is exploited in kernel-based MLPs. Moreover, a MSE cost function implicitly assumes that the errors being made are normally distributed. This is the most reasonable assumption if no additional information about the data set is known. One of the weaknesses of the MSE however, is that the cost function is extremely sensitive to outliers. If this should be avoided, one can switch to a less sensitive cost function such as the mean absolute error (MAE).

In general, the MSE cost function to train MLPs has the following form:

$$\begin{aligned}\mathcal{C}(\boldsymbol{\theta}) = \frac{\lambda_E}{D}\sum_{d=1}^{D}\left(\frac{E^{(d)}(\boldsymbol{\theta}) - \hat{E}^{(d)}}{N_b}\right)^2 + \frac{\lambda_F}{3N}\sum_{d=1}^{D}\sum_{a=1}^{N_\mathrm{n}^{(d)}}\|\boldsymbol{F}_a^{(d)}(\boldsymbol{\theta}) - \hat{\boldsymbol{F}}_a^{(d)}\|^2 \\ + \frac{\lambda_\sigma}{9D}\sum_{d=1}^{D}\|\boldsymbol{\sigma}^{(d)}(\boldsymbol{\theta}) - \hat{\boldsymbol{\sigma}}^{(d)}\|^2 \quad (3.10)\end{aligned}$$

where $E^{(d)}(\boldsymbol{\theta})$, $\boldsymbol{F}_a^{(d)}(\boldsymbol{\theta})$ and $\boldsymbol{\sigma}^{(d)}(\boldsymbol{\theta})$ are the MLP predictions of the energy, force on atom a and stress respectively. Their explicit dependence on the MLP parameters $\boldsymbol{\theta}$ is highlighted. The corresponding training targets are denoted with a hat on top of the respective symbol. The sum runs over every single configuration d, with $N_\mathrm{n}^{(d)}$ atoms, in the data set of size D with a total of $\sum_d N_\mathrm{n}^{(d)} = N$ atoms. The hyperparameters λ_E, λ_F and λ_σ determine the relative weight of the energies, forces and stresses in the cost function. Off course, if there are no forces or stresses available, the corresponding terms in the cost function are discarded.

It is important to stress the benefits of training to the forces. For a single molecule or periodic system, the final MLP errors decrease per training target, whether they are a single energy or single force component.[170] However, in a single configuration there are $3N_\mathrm{n}^{(d)}$ force components and only one energy label. Hence, by using forces, there is a lot more training data available which results in lower errors. Furthermore, larger systems (supercells) inherently contain more information, lessening the need for many different configurations in the data set.

3.2 Machine learning potential methods

In this section, an overview of different machine learning methods and MLPs is provided. We will explain how the atomic descriptors are utilized to predict the atomic energies in Eq. (3.4). We will discuss kernel-based methods, descriptor-based neural networks and message passing neural networks. In Table 3.1, a selection of different MLPs is listed and classified in one of these categories. Not all of them will be explicitly encountered in this section.

3.2.1 Kernel-based methods

In general, the atomic energies are intricate and nonlinear functions of the atomic descriptors $\boldsymbol{x}_a$. However, by making use of the *kernel-trick*, the dependence on the descriptors can be linearized in an abstract feature space, without ever explicitly mapping the descriptors to that feature space.[189, 190] Only the inner product of two abstract vectors in the features space is known and given by the kernel $K(\boldsymbol{x}_a, \boldsymbol{x}_b)$. One can think of the kernel as a measure of similarity between the two atomic descriptors. An example of a kernel is the Gaussian kernel with hyperparameter η:

$$K(\boldsymbol{x}_a, \boldsymbol{x}_b) = \mathrm{e}^{-\eta||\boldsymbol{x}_a - \boldsymbol{x}_b||^2}, \tag{3.11}$$

but many others exist.[190]

Different regression approaches are possible for kernel-based MLPs. Here, we will follow the operator quantum machine learning (OQML) approach.[191] There, the atomic energies are a linear function of the trainable parameters or regression weights α_k:

$$E_a = \sum_k \alpha_k K(\boldsymbol{x}_a, \boldsymbol{x}_k) \tag{3.12}$$

where $\boldsymbol{x}_k$ are the atomic descriptors of predefined representative points. These points are atoms in the training set and form a basis in which the energy is expanded. It can consist of all atoms available or a subset to reduce the computational cost. Automatic selection algorithms are readily available to extract the most representative points.[192] To calculate the forces, derivatives are taken with respect to the query atoms (the first argument of the kernel):

$$\boldsymbol{F}_b = -\sum_{ak} \alpha_k \nabla_b K(\boldsymbol{x}_a, \boldsymbol{x}_k) \tag{3.13}$$

The predicted energies and forces should match the training labels. This can be written down as one big matrix multiplication:

$$\mathsf{K}_{\mathrm{OQML}} \cdot \boldsymbol{\alpha} = \boldsymbol{y} \tag{3.14}$$

	Reference
Kernel methods	
FCHL	168
(s)GDML	41, 171
kQEq	128
SOAP	39
λ-SOAP/SA-GPR	172
Descriptor-based neural networks	
4G-HDNNP	127
Allegro[†]	173
ANI-1	40
Behler-Parrinello HDNNPs	36
DeepMD	174
GM-NN	175
Message passing neural networks	
AIMNET-NSE	124
BpopNN	122
DIMENET	176
DTNN	43
EPNN	123
FieldSchNet	177
GemNet	178
HIP-NN	179
NequIP	180
NewtonNet	181
PaiNN	182
PiNet	183
PHYSNET	121
SchNet	153
SpookyNet	184
UNiTE	185
Others	
ACE	186
Moment tensor potentials	187
SNAP	188

Table 3.1: A selection of different MLPs. Some entries have been named with respect to the atomic descriptor they are using.
[†] The recently published Allegro MLP works does not use feature engineered descriptors but a technique similar to message passing. The features itself are strictly local however such that we classify the MLP in this category.

where $\boldsymbol{\alpha}$ is a vector containing all the trainable parameters, $\boldsymbol{y}$ a vector concatenation of all the energy and force targets in the training set and $\mathrm{K_{OQML}}$ a rectangular matrix with coefficients gathered from Eq. (3.12) and (3.13). The optimal set of parameters α_k of the overdetermined system, can be analytically determined by substituting the previous equation in the cost function of Eq. (3.10) and minimizing with respect to the parameters α_k:

$$\boldsymbol{\alpha} = \left(\mathrm{K}_{\mathrm{OQML}}^T \mathrm{K}_{\mathrm{OQML}} + \lambda \mathrm{I}\right)^{-1} \mathrm{K}_{\mathrm{OQML}}^T \boldsymbol{y} \tag{3.15}$$

Here, λ is the strength of the L2 regularization, proportional to $\lambda \boldsymbol{\alpha}^T \boldsymbol{\alpha}$, which can be added to the cost function. Moreover, a small value of λ ensures that the matrix inverse is numerically stable.

Besides OQML, other regression approaches are Gaussian process regression (GPR)[193] or gradient-domain machine learning (GDML).[41] They have a specific meaning in the context of MLPs where also derivatives (forces or stresses) are fitted. A good overview of the differences of these approaches can be found in Ref. 168. From another point of view, they are all variants on GPR. They differ in the set of representative points. For instance, in comparison with OQML, GPR also uses derivatives of atomic descriptors and these derivatives have their own regression coefficients α_k. There is also symmetry-adapted GPR (SA-GPR), where the kernel itself is a tensor quantity which transforms equivariantly with global rotations of the system.[172]

Kernel-based methods are still a popular MLP methodology in a variety of applications ranging from metallic systems to chemical reactions due to their attractive properties.[23, 193] First of all, the cost function can be exactly minimized. Secondly, training is often fast since only the kernel matrix should be constructed and afterwards inverted. However, for large training sets, the matrix inverse can become computationally expensive. This is why neural network potentials are suggested for large data sets. Furthermore, the kernel-based methods inherit all advantages and disadvantages of the atomic descriptors being used. Hence, systems with a large number of different species remain prohibitive.

3.2.2 A short introduction to neural networks

Before immediately diving into neural network MLPs, we will first introduce the basics of neural networks in this section. Neural networks can learn highly dimensional functions by applying a series of linear combinations and nonlinear activation functions.[33] The most basic building block of a neural network is a dense layer, denoted in this work by $D^{N\times M}$. It acts upon an input vector $\boldsymbol{x} \in \mathbb{R}^M$ and yields the output vector $\boldsymbol{y} \in \mathbb{R}^N$:

$$\boldsymbol{y} = D^{N\times M}(\boldsymbol{x}) = \mathrm{W} \cdot \boldsymbol{x} + \boldsymbol{b} \tag{3.16}$$

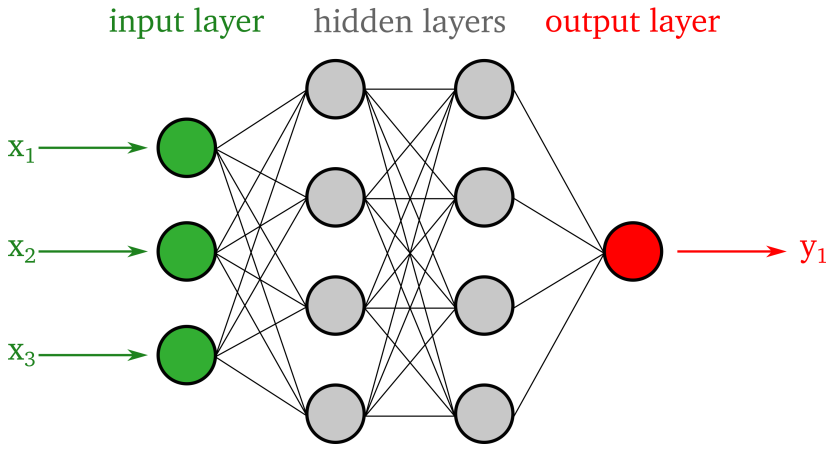


Figure 3.5: A schematic depiction of a dense neural network. By a series of hidden dense layers, the input vector $\boldsymbol{x}$ is transformed to the output $\boldsymbol{y}$ in a nonlinear fashion. By using the notation in the text, this network implements the following function: $\boldsymbol{y} = \left[D^{1\times4} \circ \widetilde{D}^{4\times4} \circ \widetilde{D}^{4\times3}\right](\boldsymbol{x})$. Typically, an activation function is not used in the last layer for regression.

where the weight matrix $W \in \mathbb{R}^{N\times M}$ and bias vector $\boldsymbol{b} \in \mathbb{R}^{N}$ are both trainable parameters. Hence, the outputs are just a linear combination of the inputs. Nonlinearities can be introduced by applying an activation function. We will write a tilde on top of the respective dense layer to indicate this:

$$\widetilde{D}^{N\times M}(\boldsymbol{x}) = f_{\text{act}}\left(\mathrm{W}\cdot\boldsymbol{x} + \boldsymbol{b}\right) \tag{3.17}$$

To ensure that the energy and forces predicted by an MLP are continuous, the activation function and its derivatives should be continuous as well. For instance, one can use the shifted softplus activation function:[153]

$$f_{\text{act}}(\boldsymbol{x}) = \log\left[1 + \exp\left(\boldsymbol{x}\right)\right] - \log(2). \tag{3.18}$$

By combining a series of dense layers, complicated and nonlinear functions of the inputs can be learned. This is visualized in Figure 3.5. There, every layer is a dense layer as it is fully connected to the previous layer. The output can be a single number (the energy) or a vector.

All the parameters in the neural network can be optimized by minimizing a certain cost function. Unlike kernel-based methods, the optimal set of parameters can no longer be calculated analytically. Therefore, most algorithms take iterative steps along the direction opposite of the first derivative to minimize the cost function. These first derivatives can be efficiently

computed with the backpropagation algorithm.[194] As the name suggests, the derivatives are obtained by propagating backwards through the neural network. Via the chain rule, the derivatives in layer l of the network can be written as a function of those in layer $l+1$. Hence, at first, there is a forward pass through the neural network (to compute the output $\boldsymbol{y}$) and later, the derivatives are computed from the last layer propagating back to the first layer. This is an efficient and fast method which is implemented in popular machine learning libraries such as TensorFlow[195] and PyTorch.[196] In these libraries, the user should only implement the forward pass while the derivatives can be requested in one line of code by automatic differentiation (via backpropagation).

Once the derivatives are known, the optimizer can take a step in the direction of steepest descent:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \lambda \frac{\mathrm{d}\mathcal{C}(\boldsymbol{\theta}^{t})}{\mathrm{d}\boldsymbol{\theta}} \tag{3.19}$$

where $\boldsymbol{\theta}^{t}$ are the trainable parameters at iteration t, $\mathcal{C}(\boldsymbol{\theta}^{t})$ is the cost function and λ is the learning rate. By using higher learning rates, the set of parameters will move faster to their optimum. However, there is a risk of overshooting the minimum, possibly culminating in numerical fluctuations and instabilities. Therefore, the learning rate should be carefully selected. To achieve faster computation times, the cost function and its derivatives are not evaluated for every sample d in the data set in each iteration. Only a random selection of data points are taken into account. These form a single *batch*, while the number of points included is called the *batch size*. In that case, one is employing the stochastic gradient descent algorithm.[33] If all data points are encountered exactly once (this can take multiple iterations or batches), one *epoch* has passed. Through the years, more advanced algorithms have been developed, the ADAM optimizer[197] being among the most popular. This algorithm keeps track of running averages of the first derivative and its square to obtain faster convergence.

Since the cost function cannot be exactly minimized, the algorithm itself should decide when to stop training without overfitting. Early stopping[198] is a common approach to fulfill this goal. It is best explained visually by looking at the learning curves of Figure 3.6. There, the training error is plotted for every epoch while training the network. Naively interpreting these results, it seems that the best possible model is found at the right, after approximately 25000 epochs. Additionally, after every epoch, the accuracy of the current MLP was assessed on the validation set. We observe that the validation error decreases at the start but starts to increase again after 2500 epochs. This is a clear indication of overfitting. Since neural networks can have millions of trainable parameters, they are extremely flexible, making them susceptible

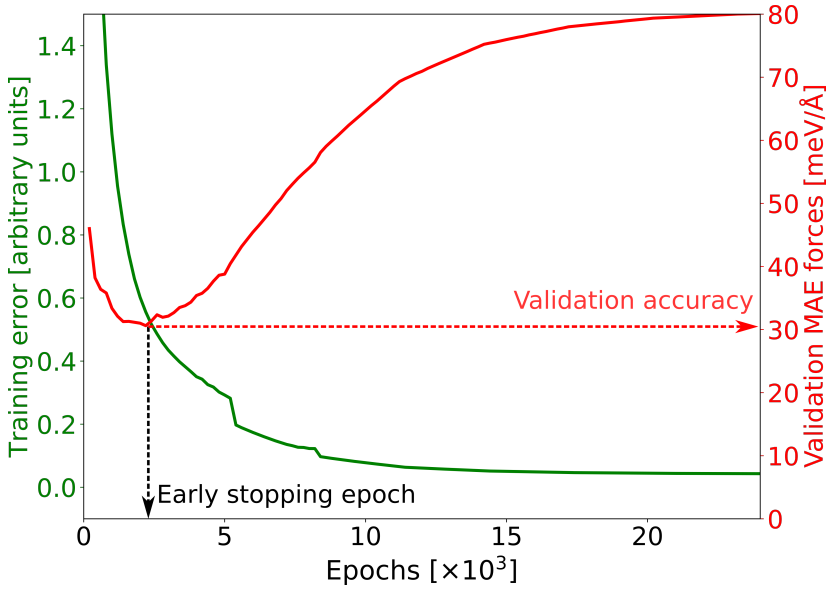


Figure 3.6: Training and validation learning curves of SchNet.[153] The training error keeps decreasing and has multiple small jumps where the learning rate decreases. The validation error increases after a while, which is an indication of overfitting. The best possible model and early stopping point of this training run is indicated on the figure.

to overfit on the training set. Hence, the training should be aborted when the validation error starts to increase. This is called early stopping.

Variants on early stopping are also occasionally employed. Instead of completely stopping the training, one can decrease the learning rate when the validation error does not decrease for a certain number of epochs, called the *patience*. A smaller learning rate can sometimes help to start finding better parameters sets again by increasing the numerical stability. For extremely large training sets, overfitting is not an issue. In these cases, early stopping cannot be used and the only factor limiting the training duration is the walltime one has available. There, an exponentially decreasing learning rate is a more appropriate choice.

3.2.3 Descriptor-based neural networks

The atomic energies in a descriptor-based neural network are computed with the help of a species-dependent neural network:

$$E_a = \mathrm{NN}_{S_a}(\boldsymbol{x}_a) \tag{3.20}$$

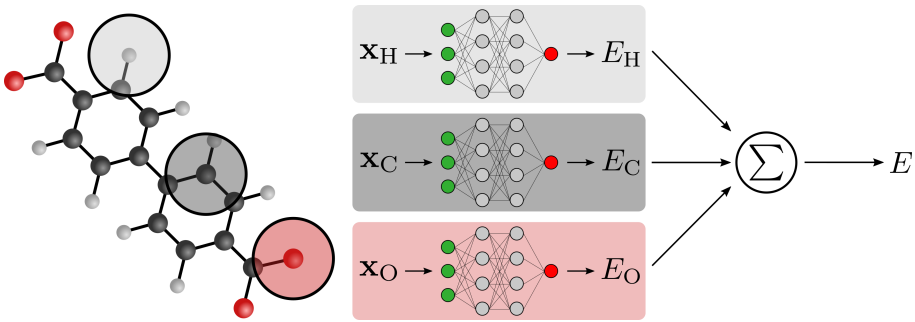


Figure 3.7: The architecture of a descriptor-based neural network illustrated with a biphenyl-4,4'-dicarboxylate (BPDC) linker molecule. For every atom, atomic descriptors are computed. They are the inputs of a dense neural network which is species dependent. The neural networks yield the atomic energies.

where S_a is the species of atom a. Hence, the same neural network is utilized for every atom of the same species. The inputs of the neural network are the atomic descriptors of choice. The neural network itself is composed of a series of fully connected dense layers and yield the atomic energies. The architecture of a descriptor-based neural networks is schematically depicted in Figure 3.7. Most descriptor-based neural networks only differ in the choice of atomic descriptor. In Table 3.1, a selection of those MLPs are listed. Compared to kernel-based MLPs, NN methods have no poor scaling with increased data set sizes. Hence, they are the method of choice if there is an abundance of data. However, they are still reliant on manually designed descriptors, inheriting their disadvantages.

3.2.4 Message passing neural networks

In the previous sections, the disadvantages of descriptor-based MLPs were listed. They require human input to carefully select the descriptors and they scale badly in the number of atomic species. Only half a decade ago, the first non-descriptor based MLPs were developed. They automatically learn their own descriptors or atomic representations by using the concept of message passing.[42] In message passing neural networks (MPNNs), each atom is characterized by an F-dimensional atomic feature vector $\boldsymbol{x}_a^t \in \mathbb{R}^F$. These vectors are not manually designed or feature engineered but are automatically refined using T interactions. In each interaction, the features are updated by an update function:

$$\boldsymbol{x}_a^{t+1} = U_t(\boldsymbol{x}_a^t, \boldsymbol{m}_a^{t+1}) \tag{3.21}$$

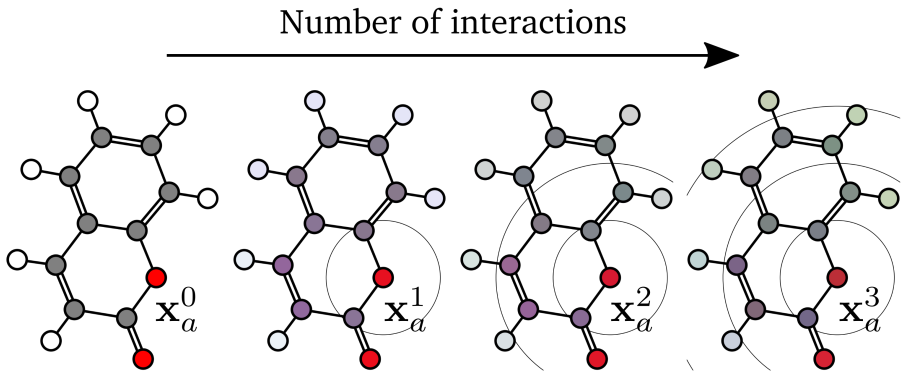


Figure 3.8: A schematic depiction of the iterative refinement of the atomic features in message passing neural networks, illustrated with a coumarin molecule. The RGB color of an atom, a three-dimensional feature, is updated based on the local environments lying within the cutoff radius (the inner black circles). After every iteration the effective cutoff radius increases since information can flow by message passing.

where the messages $\boldsymbol{m}_a^{t+1}$ are computed as sums or convolutions over the set of neighboring atoms $\mathcal{N}_a$, lying within the cutoff radius r_{cutoff}:

$$\boldsymbol{m}_a^{t+1} = \sum_{b \in \mathcal{N}_a} M_t(\boldsymbol{x}_a^t, \boldsymbol{x}_b^t, e_{ab}) \tag{3.22}$$

Here, e_{ab} are the *edge features*. In its most simple form, the edge features are just the interatomic distances R_{ab}. Finally, after T iterations, the energy is predicted by using a readout function R, typically as a sum over atomic contributions:

$$E = \sum_a E_a = \sum_a R(\boldsymbol{x}_a^T) \tag{3.23}$$

To visualize this iterative process, four message passing iterations have been schematically depicted in Figure 3.8. Here, the atomic features are represented with an RGB color (this is a three-dimensional feature vector). They are initialized purely based on their species. For instance, before the first interaction, every hydrogen atom still has the same color. Next, the first interaction takes place and the features of every atom are updated based on the local chemical environment, depicted with the black circle with radius r_{cutoff}. Now, there is already a difference between some of the carbon atoms. After the second interaction, the features are refined again, only by looking at the atoms within the same local chemical environment (the inner black circle

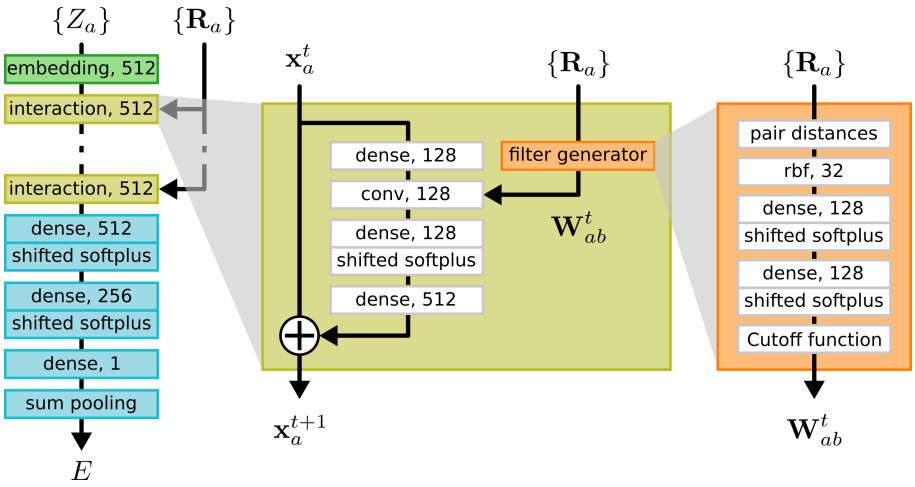


Figure 3.9: The SchNet architecture.[153] The hyperparameters are taken from the eMLP model.[10]

around $\boldsymbol{x}_a^2$). However, the atoms lying within that circle have already been updated in the first interaction and carry information about their own local environment. Hence, the *effective* cutoff radius, which influences the features in the second interactions, is twice the normal cutoff radius. This whole process is repeated for T interactions such that, in the end, the effective cutoff radius is T times the normal cutoff radius r_{cutoff}. Thus, information can flow over much longer distances as long as there is an uninterrupted connection between the receiving atoms.

While the eMLP model was in development, SchNet[153], a MPNN, achieved state-of-the-art results on different benchmarks and use-cases,[168, 199, 200] comparable to other non-MPNNs, especially for large data sets. For that reason, we chose SchNet to model the short-range interactions in the **eMLP paper**. Only recently, modern MLPs improved upon SchNet by making use of equivariant features.[180] Still, SchNet remains an accurate and fast MLP to predict atomistic properties. Therefore, we will now take a closer look at the SchNet architecture and how it implements the message and update functions of Eq. (3.21) and (3.22).

A schematic overview of the SchNet architecture is given in Figure 3.9. Everything starts at the atomic embedding, which initializes the F-dimensional feature vector:

$$\boldsymbol{x}_a^0 = \boldsymbol{a}_{S_a} \tag{3.24}$$

The embedding $\boldsymbol{a}_{S_a} \in \mathbb{R}^F$ is a trainable parameter and depends on the species S_a of the atom. This is the species-dependent part in SchNet. All

other layers and interaction blocks are the same for every species. The message function (3.22) is implemented as follows:

$$M_t(\boldsymbol{x}_a^t, \boldsymbol{x}_b^t, e_{ab}) = \frac{1}{J} D^{G\times F}(\boldsymbol{x}_b^t) \circ \boldsymbol{W}_{ab}^t(R_{ab}) \tag{3.25}$$

where $D^{G\times F}$ is a dense layer (3.16), projecting the feature vectors into a G-dimensional filter space. This is multiplied element wise with the filter-generating network $\boldsymbol{W}_{ab}^t$. The constant hyperparameter J normalizes the sum of the convolution (3.22). The average number of neighbors in the cutoff radius around every atom is selected as a typical value for J. In the eMLP model, we use $F = 512$ and $G = 128$, see Figure 3.9. The most computationally expensive part of SchNet is the convolution such that reducing G with respect to F yields a considerable speedup. We empirically observed that this does not significantly reduce the accuracy of the MLP. The filter-generation network,

$$\boldsymbol{W}_{ab}^t(R_{ab}) = \left[\widetilde{D}^{G\times G} \circ \widetilde{D}^{G\times N}\right](\boldsymbol{\phi}(R_{ab})) f_{\text{cutoff}}(R_{ab}), \tag{3.26}$$

requires N radial basis functions (rbfs) as input and f_{cutoff} is the cutoff function (3.7). In SchNet, a Gaussian basis is chosen for the rbfs:

$$\phi_n = \exp\left(\frac{(R_{ab} - \mu_n)^2}{2\sigma^2}\right) \tag{3.27}$$

with $\sigma = \frac{r_{\text{cutoff}}}{N}$ and the N centers $\mu_n = n\sigma$ are uniformly spaced within the cutoff radius.

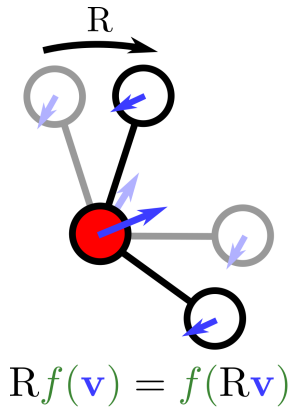
The update function (3.21) is rather straightforward in SchNet,

$$U_t(\boldsymbol{x}_a^t, \boldsymbol{m}_a^{t+1}) = \boldsymbol{x}_a^t + \left[D^{F\times G} \circ \widetilde{D}^{G\times G}\right](\boldsymbol{m}_a^{t+1}), \tag{3.28}$$

the messages are just transformed back to the feature space using a dense neural network. Finally, after T iterations, the readout function determines the energy as follows:

$$E_a = \left[D^{1\times\frac{F}{2}} \circ \widetilde{D}^{\frac{F}{2}\times F} \circ \widetilde{D}^{F\times F}\right](\boldsymbol{x}_a^T) \tag{3.29}$$

In principle, the MPNNs should be more accurate than the descriptor based NNs since their large effective cutoff can learn long-range interactions. Furthermore, the network itself can decide which description of the chemical environment is best suited to accurately model the PES. There are some subtleties however, SchNet and other MPNNs only make use of the interatomic distances R_{ab}. Angular information is not included. MPNNs can only indirectly learn the dependence on bond angles and higher many-body terms after a number of interaction blocks. This is harder to learn and might be the reason why they excel when more data becomes available.



Equivariant $\ell = 1$ operations	
Linear combinations	$\sum_k c_k \mathbf{v}_k$
Dot product	$\mathbf{v} \cdot \mathbf{w}$
Cross product	$\mathbf{v} \times \mathbf{w}$
Norm activation	$f_{\text{act}}(v)\hat{\mathbf{v}}$

Figure 3.10: Left: vector features $\boldsymbol{v} = v\hat{\boldsymbol{v}}$, depicted in blue, with rotation order $\ell = 1$ should transform equivariantly under global rotations with rotation matrix R. Right: a list of equivariant operations acting on $\ell = 1$ features.

3.2.5 Equivariant neural networks

All the atomic features $\boldsymbol{x}_a$ encountered so far, are scalar features. By construction, they are invariant under translations and rotations. On the other hand, vectors are not invariant under these transformations. However, they might be beneficial to include in MLPs since they can naturally incorporate angular information. Moreover, relevant physical properties such as atomic dipoles and forces are vectors. Atomic vector features that can learn these properties directly, can improve the accuracy of MLPs. Recently, this is accomplished with the new class of equivariant neural networks MLPs.[173, 180–182, 201] In these MLPs, the vector features are not invariant but rotate *equivariantly* with the system. For instance, if a molecule is rotated, the forces should also rotate with the same angle. This is visualized in Figure 3.10.

More generally, a learnable function $\boldsymbol{f}$, is equivariant under a group g if

$$\mathcal{T}_g^*(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{f}(\mathcal{T}_g(\boldsymbol{x})) \tag{3.30}$$

where $\mathcal{T}_g$ and $\mathcal{T}_g^*$ are the representations of the group in the input and output space respectively. If $\mathcal{T}_g^*$ is the identity operator, it is said that the function $\boldsymbol{f}$ is invariant under the transformation. For three-dimensional atomic systems, the transformation group is $E(3)$ and includes rotations, inversions and translations. In the Neural Equivariant Interatomic Potential (NequIP)[180] and Allegro,[173] the atomic features $\boldsymbol{x}_\ell^m$ are irreducible representations of the $O(3)$ symmetry group and are indexed by an arbitrary rotation order $\ell \in \mathbb{N}$ and

$-\ell \leq m \leq \ell$. Scalar features have rotation order $\ell = 0$ and are invariant. Features with $\ell = 1$ have three components and are equivalent with ordinary vectors such as atomic dipoles or forces.

Not every operator $\boldsymbol{f}$ in a neural network is equivariant and obeys Eq. (3.30). For instance, adding bias weights in a dense layer or element wise activation functions $f_{\text{act}}(\boldsymbol{x}_\ell^m)$, do not satisfy the constraints. On the other hand, the dot product or cross product, acting on two $\ell = 1$ features and resulting in a $\ell = 0$ and $\ell = 1$ feature respectively, are acceptable operators. For arbitrary rotation orders, equivariant MLPs make use of the following tensorproduct:[173]

$$(\boldsymbol{x} \otimes \boldsymbol{y})_{\ell_{\text{out}}}^{m_{\text{out}}} = \sum_{m_1, m_2} \begin{pmatrix} \ell_1 & \ell_2 & \ell_{\text{out}} \\ m_1 & m_2 & m_{\text{out}} \end{pmatrix} \boldsymbol{x}_{\ell_1}^{m_1} \boldsymbol{y}_{\ell_2}^{m_2} \tag{3.31}$$

which combines two features $\boldsymbol{x}_{\ell_1}^{m_1}$ and $\boldsymbol{y}_{\ell_2}^{m_2}$ and results in an output feature ℓ_{out}, where $|\ell_1 - \ell_2| \leq \ell_{\text{out}} \leq |\ell_1 + \ell_2|$. The coefficients appearing in the sum, are the Wigner $3j$ symbols.[202]

Instead of only using the norm R_{ab} of the interatomic distances $\boldsymbol{R}_{ab} = R_{ab}\widehat{\boldsymbol{R}}_{ab}$ as edge features, spherical harmonics $Y_\ell^m(\widehat{\boldsymbol{R}_{ab}})$, acting on the angular part, are utilized to construct equivariant features with arbitrary rotation order. They appear in the message function (3.22) as more general filter-generating networks $\boldsymbol{W}_{ab}^t(R_{ab})Y_\ell^m(\widehat{\boldsymbol{R}_{ab}})$, where they are combined with the existing features in layer t using Eq. (3.31). The exact architecture and implementations may differ between the equivariant MLPs. In this work, we only utilized NequIP as an invariant MLP. For more details on the specific architecture, consult Ref.180.

The biggest strength of equivariant MLPs such as NequIP is their data-efficiency.[180] Unlike conventional MPNNs, they can achieve highly accurate results with a limited amount of data, rivaling or coming out ahead of kernel-based MLPs. This is realized by the incorporation of angular information directly. One of the weaknesses of equivariant MLPs is that they can be a lot slower. For instance, SchNet typically outperforms NequIP by one order of magnitude. For this reason, equivariant MLPs are often constructed with less features and less layers, reducing the number of trainable parameters and complexity of the model. This side-effect is advantageous since it reduces the risk of overfitting, also potentially improving the data-efficiency and accuracy of equivariant MLPs. Recent developments have been made towards highly parallelizable MLPs with Allegro.[173] It employs equivariant features but their dependence remains strictly local. This new architecture has all the benefits of equivariant MLPs at the computational efficiency of descriptor-based NNs.

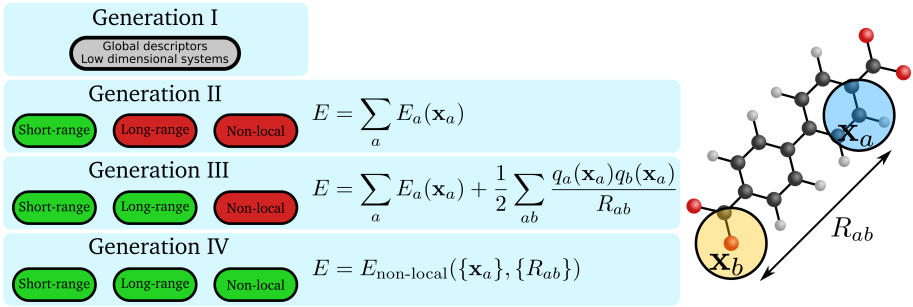


Figure 3.11: A schematic overview of the four different MLP generations. A description of all the different categories is given in the text of this section. The figure is adapted with permission from Ref. 129 Copyright 2021 American Chemical Society.

3.3 Long-range interactions

The PES of the majority of MLPs is given by Eq. (3.4). The energy is a function of local atomic environments which are encoded by local descriptors. Only interactions up to the cutoff radius are included.[c] Hence, there are no long-range interactions, except if the cutoff radius is extremely large, limiting the usefulness of the MLP by increasing the computational cost. Some well-known physical interactions are thus completely absent, including long-range electrostatics, induction and dispersion. These effects play a significant role to model advanced materials such as batteries.[52, 53] Therefore, multiple ways to incorporate long-range information in MLPs have been developed. In this section, we will review some of them. In general, MLPs can be categorized in four different generations, depending on whether and how they include long-range interactions.[129] They are schematically depicted in Figure 3.11.

The first generation of MLPs are not discussed in this work. They predict the energy using global descriptors, which are specifically designed for a single system. Furthermore, they have an unfavorable scaling with system size such that they are only applicable to low dimensional systems. The second generation of MLPs on the other hand are generally applicable and partition the energy in atomic contributions, see Eq. (3.4). In turn, these contributions depend on the local atomic descriptors, described in section 3.1.2, or are learned by an MPNN. Hence, they are inherently local or short-ranged because only neighboring atoms within the (effective) cutoff radius

c. In MPNNs, the interaction radius is given by the cutoff radius multiplied with the number of layers. For more information, see section 3.2.4.

can influence the descriptors. For instance, if two identical water molecules are placed 10 Å apart, and the cutoff radius is 5 Å, the total energy is just twice the energy of a single water molecule.

From the third generation and onward, long-range interactions are included. Typically, an electrostatic energy contribution is added to the usual short-range energy:

$$E_{\text{electrostatic}} = \frac{1}{2} \sum_{a \neq b} \frac{q_a(\boldsymbol{x}_a) q_b(\boldsymbol{x}_b)}{R_{ab}} \tag{3.32}$$

where the atomic charges depend on the local environments. Hence, the atomic descriptors are utilized to predict both the atomic energies E_a and the partial charges q_a. The charges themselves are still local functions and do not depend on the environments far away. Going back to the example of the two identical water molecules above, the charges of both molecules are still the same as they were in an isolated molecule but the total energy will deviate from the sum of the two water molecules due to the electrostatic interaction term.

Finally, in the fourth generation of MLPs, the charges themselves are not local anymore. In the example of the two water molecules, the charges in both molecules are different from the charges in the isolated molecules. Each molecule now polarizes the other molecule. Hence, only this generation of MLPs belongs to the class of fully polarizable force fields. There are different techniques to construct a fourth generation MLP. Just like in the third generation, one can use fluctuating charges. For instance, besides their dependence on local atomic descriptors, they are used in combination with the charge equilibration neural network technique (CENT)[125–127] or kernel Charge Equilibration (kQEq).[128] Both techniques are similar to the charge equilibration scheme of section 2.4.3. In the Becke Population Neural Network (BpopNN)[122] atomic populations are learned. There, all the different energy contributions depend parametrically on these atom populations and the total energy is minimized with respect to them. The eMLP[10] and the self-consistent field neural network (SCFNN)[203] also belong to the fourth generation. These are both explicit-electron MLPs and will be discussed in more detail in Chapter 4.

There is some ambiguity to place the MPNNs in the correct generation. In principle, with extremely large cutoffs and in the limit of infinite interaction blocks, the atomic energies depend on atoms far away and can be fully self-consistent. Moreover, all atoms in the system should be connected with one another (no gaps greater than the cutoff radius) to allow the transfer of messages throughout the entire system. Hence, only in this limit, they truly

belong to the fourth generation. In practice, MPNNs belong to the second (or third) generation because only a finite number of interactions are present.

As a final note, most but not all MLPs over the four different generations, make use of fluctuating charges. This means that polarization is not well defined, as described in section 2.4.4. Thus, only explicit-electrons MLPs or variations such as EPNN[123] are the most suitable and widely applicable methods to model long-range interactions and polarization.

3.4 Machine learning potentials for metal-organic frameworks

In the first chapter of this work, we introduced metal-organic frameworks as one of the most promising material classes in recent years. This is due to their atypical topology. As a part of their crystal structure, they contain regular nanometer-sized pores that play a central role in various interesting phenomena. These may include phase transitions, guest adsorption or negative thermal expansion. Most MOFs have unit cells with more than hundred atoms, making them very expensive to study with first-principle methods. Force fields on the other hand, may lack the reactivity and accuracy to describe phase transitions or guest adsorption. Therefore, there is a need for MLPs to advance MOF simulations towards greater accuracy and larger time and length scales.

Some first steps have already been taken to develop MLPs for MOFs. More specifically, Eckhoff *et al.*[47] constructed an MLP for MOF-5 and Yu *et al.*[48] for MOF-808. Both approaches make use of a descriptor-based neural network. More than 20000 configurations were sampled to train these models. Since MOF-5 and MOF-808 contain 424 and 1208 atoms in their unit cells respectively, sampling all of that data using periodic DFT calculations turns out to be extremely expensive. Therefore, atomic clusters were used to reduce the computational cost. These clusters should be larger than the cutoff radius such that the chemical environments of atoms lying in the center of these clusters are similar to the equivalent bulk atoms. Hence, only a fraction of all the atoms within the cluster can be used to train the MLP. Furthermore, the clusters themselves are terminated by adding hydrogen atoms. This approach works but has several shortcomings in other circumstances. First of all, the computational gains are lost when MPNNs are considered. They have effective cutoff radii which could span more than 30 Å (e.g. six interactions and $r_{\text{cutoff}} = 5\,\text{Å}$), requiring clusters of that size. Hence, these clusters are not small anymore, losing the benefits of the approach. Secondly, the termination of the clusters and saturation with hydrogen atoms, will introduce

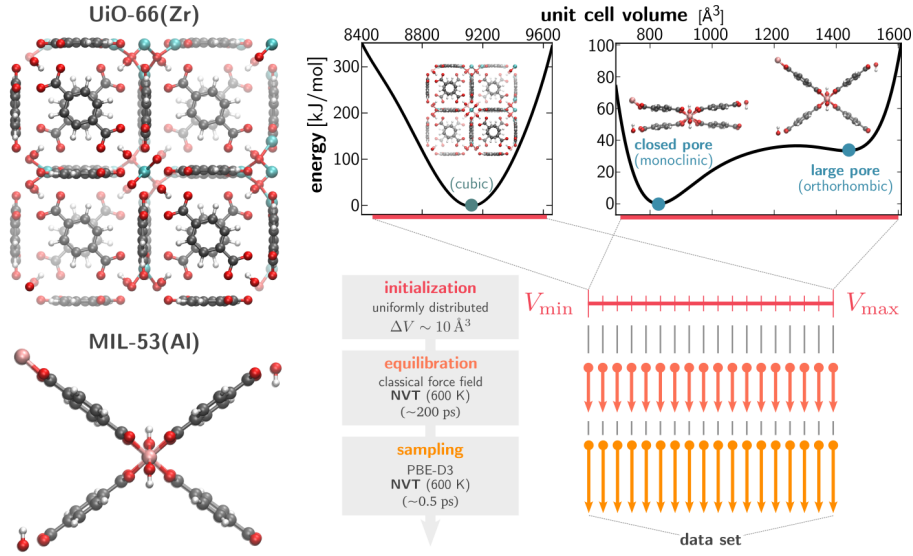


Figure 3.12: Left: a schematic depiction of the inorganic bricks and organic linkers in UiO-66(Zr)[204] and MIL-53(Al).[205] Right: energy-versus-volume profiles and the data generation protocol. Inspired by the profiles, the region of interest for the collective variable V is determined. Structures along the collective variable are generated and equilibration using force fields and finally sampled with first-principle methods. Figure courtesy of Sander Vandenhaute.

artifacts and noise on the training labels. It is unclear how large these effects are.

We have proposed a new strategy to derive MLPs for MOFs which relies on the data efficiency of NequIP[180] and a generally applicable data sampling protocol. The method was applied to two representative MOFs: UiO-66(Zr)[204] and MIL-53(Al),[205] both are visualized on the left side of Figure 3.12. With NequIP, an equivariant MPNN, only 240 and 451 training structures are necessary to obtain an highly accurate MLP for UiO-66(Zr) and MIL-53(Al) respectively, with force MAEs lower than 20 meV/Å. Without the data efficiency of the equivariant MLPs, this could not be achieved. For instance, in earlier unpublished work, we derived a SchNet MLP for MIL-53(Al) with force MAEs of 21 meV/Å and stress MAEs of 20 MPa but 30000 first-principle structures were required. Furthermore, the SchNet MLP did occasionally crash during long MD simulations for small volumes ($\leq$ 750 Å^3). These crashes or instabilities when extrapolating would sometimes only occur after a nanosecond of simulation time. This issue is also resolved by using NequIP

instead.

To pick the optimal 240 or 451 structures, we developed the data generation protocol depicted at the right side of Figure 3.12. Only a proper collective variable q and its range of interest are expected as input from the user, which ultimately requires some prior knowledge about the system. The protocol consists of three phases: initialization, equilibration and sampling. In the first two phases, proper initial structures for the sampling phase are obtained by using classical force fields. These are obtained via QuickFF or other tools.[94, 206, 207] Only in the sampling phase, first-principle MD calculations are performed. Since MOFs are systems with multiple metastable phases, the relevant regions in the whole phase space cannot be explored by ordinary MD simulations. It is impossible to cross high transition barriers separating two phases such that enhanced sampling methods should be utilized. In general, enhanced sampling takes place along a collective variable q, describing the transition. For MIL-53(Al), the unit cell volume V is ideally suited for that purpose and since UiO-66(Zr) has only one single phase, we choose $q = V$. In the initialization phase, structures along the region of interest of the collective variable V are generated. In practice, they were extracted from short NPT simulations. In the equilibration phase, all other degrees of freedom besides the collective variable are equilibrated. This is done by NVT simulations. Finally, the equilibrated systems are the starting point for first-principles MD simulation at the specific collective variable. Again, these are NVT simulations in our case. Moreover, they are subsampled to obtain uncorrelated geometries. Notice that the total length of the first-principles simulations remains limited (0.5 ps) since NequIP does not require many data points.

After training on the efficiently generated data sets, the MLPs were validated in multiple ways. First of all, static force and stress errors were computed on trajectories generated by the MLP itself. Thus, the values will represent the true errors when applying the MLP in practice. At 300 K, force MAEs hover around 20 meV/Å for all but the smallest volumes. Even at 600 K (the temperature of the training data) errors on the forces remain below 30 meV/Å. Although the MLPs were not fitted to the stress, stress MAEs are approximately 30 MPa, which is highly accurate. This is confirmed by predicting the elasticity tensor and comparing with the first-principles results. The largest deviations are only 6 GPa (on a first-principles value of 87.4 GPa). To further show the usefulness of the MLPs, the transition pressure for MIL-53(Al) was predicted. This can only be realized with a large supercell containing more than 1000 atoms. The predicted value of 18-20 MPa is in excellent agreement with experimental values.

These results show that our data protocol is a promising tool to derive MLPs

for MOFs. There is still some room for improvement however. In its current version, a conventional force field is required to equilibrate the system. When dealing with guest adsorption or chemical reactions in MOFs, finding a proper force field can be troublesome. Furthermore, in some instances, even short first-principles MD simulations are too costly. Active learning can aid us in these cases. Our data protocol remains valid but several iterations will be necessary. Instead of using a force field and first-principles MD, the MLP itself is being used to sample new data. Only the selected structures will be calculated with first-principles methods and added to the training data.

4

The electron machine learning potential

Protons give an atom its identity, electrons its personality.
Bill Bryson (⋆1951)

This chapter is fully devoted to the electron machine learning potential or eMLP. It has been developed as an improvement upon existing explicit-electron force fields. In Chapter 2, we saw how force fields completely discard all electronic degrees of freedom to speed up molecular simulations. Nevertheless, explicit-electron force fields reintroduce them in an approximate manner. This enables modeling of polarization, ionization and more complex phenomena. However, describing the electron interactions with physically inspired energy expressions remains difficult because they obey the exclusion principle and a variety of quantum effects. In the eMLP, we solve this issue by making use of MLPs. These automatically learn the complex energy landscape of the electrons without any human bias. This can lead to a much greater accuracy and flexibility to describe chemical reactions.

The eMLP does also improve upon MLPs in general. In section 3.3, each successive MLP generation did incorporate more and more long-range effects and polarization. The eMLP belongs to the fourth generation where the particles and charges in the system are fully self-consistent and globally interact with one another. Furthermore, the eMLP avoids fluctuating charges and hence also the problematic polarization ambiguity of section 2.4.4. All

the particles carry integer charges, just as the elementary particles, such that charge excess, ionization or dissociation into molecular fragments can be modeled naturally.

Another innovation of the eMLP is that the electron particles will be localized at the centers of localized orbitals. These positions can be extracted from Hartree-Fock or DFT calculations via Foster-Boys (FB) localization[208] in molecules or maximally localized Wannier functions (MLWF)[209, 210] in periodic systems. Furthermore, they impose the correct electronic dipole. Recently, modeling long-range interactions with centers of localized orbitals also gained traction in other state-of-the-art MLPs. These are the self-consistent field neural network (SCFNN)[203] and deep potential long-range (DPLR) model,[211] published barely one month after the eMLP. Together with the eMLP, these may potentially start a new class of 'localized' explicit-electron MLPs.

A python library of the eMLP is available online at https://github.com/mcoolsce/eMLP. It is built on top of TensorFlow[195] and does support GPU as well as multi-core CPU training and inference. MD simulations are available out-of-the-box via an interface with Yaff.[212]

A complete description of the eMLP can be found in the **eMLP paper**. In the following sections however, the most important aspects of the model will be revisited. We will start by giving an overview of all the particles and energy contributions in the model. Also the electron localization will be discussed which plays a central part in the eMLP. Afterwards, an overview of key results on the newly constructed eQM7 and β-glycine data set will be given. Finally, we will show that the robustness of the eMLP can be significantly improved by data augmentation, which is the topic of the last section in this chapter.

4.1 Methodology

The eMLP belongs to the class of explicit-electron force fields, discussed in section 2.4.5. Besides the nuclei, electrons will be described as semi-classical particles. They are not bound or restricted to a host nucleus but can move freely throughout the whole molecule. In the current version of the eMLP, only electron pairs are considered. They are the combination of a spin-up and spin-down electron grouped into a single particle with a charge of -2 e. This restricts the applicability of the eMLP to systems with an even number of electrons. Moreover, we will assume that all the core electron pairs are located approximately on top of the corresponding nuclei. In section 4.1.2, we will numerically confirm that this is a valid assumption. Therefore, only the valence electron pairs are explicitly described. Hence, in the eMLP, there

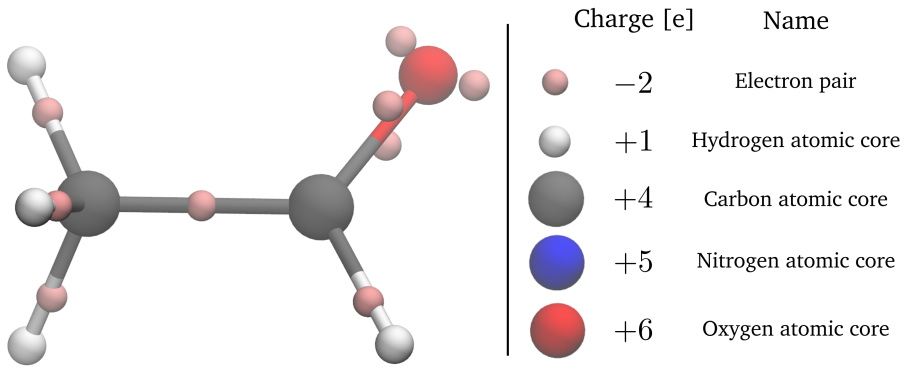


Figure 4.1: Left: the ethanal molecule in the eMLP model. The electron pairs, located at the centers of the Foster-Boys localized orbitals, correspond with the Lewis structure of the molecule. Right: all the different particles and their charges in the current eMLP model.

are two types of particles: *electron pairs*, where only valence electrons are included, and *atomic cores*, which are a combination of the nuclei and core electrons. This means that an atomic core of a species in the second row of the periodic table with atomic number Z_a does have a charge $q = Z_a - 2$ since they contain two core electrons. Hydrogen, an element of the first row, only has valence electrons such that the hydrogen atomic core has a charge $q = +1$. In the **eMLP paper**, applications were limited to systems only containing the elements H, C, N and O. In that case, an overview of all the particles is given in Figure 4.1. However, the eMLP is certainly not limited to these elements. In practice, it can be extended with elements from the third row and beyond by including the strongly bound electrons into the atomic cores and again only describing the valence electrons.

The eMLP aims to model the *extended* potential energy surface of all atomic cores and electron pairs:

$$E_{\text{eMLP}} = f(\{\boldsymbol{R}_a\}, \{Z_a\}; \{\boldsymbol{r}_i\}) \tag{4.1}$$

Therefore, besides the positions of the nuclei $\{\boldsymbol{R}_a\}$ and their species $\{Z_a\}$ in conventional MLPs, also the positions of the electron pairs $\{\boldsymbol{r}_i\}$ are required. The extended PES is fitted by using the total energy, stresses, forces on the atomic cores[a] and the forces on the electron pairs. The latter contribution is not present in conventional MLPs and hence, the cost function of Eq. (3.10)

a. The force on an atomic core is exactly the same as the force on the corresponding nucleus.

is extended with the following term:

$$\Delta\mathcal{C}(\boldsymbol{\theta}) = \frac{\lambda_f}{3N_\mathrm{e}} \sum_{d=1}^{D} \sum_{i=1}^{N_\mathrm{e}^{(d)}} \| \boldsymbol{f}_i^{(d)}(\boldsymbol{\theta}) - \hat{\boldsymbol{f}}_i^{(d)} \|^2 \tag{4.2}$$

where $\boldsymbol{f}_i^{(d)}(\boldsymbol{\theta})$ is the force on electron pair i in configuration d. There are $N_\mathrm{e}^{(d)}$ electron pairs in that configuration with a total of $\sum_d N_\mathrm{e}^{(d)} = N_\mathrm{e}$ electron pairs in the current batch to train the eMLP. The relative weight parameter λ_f controls the relative importance of this term in the cost function. Typically, the same weight is assigned as to the forces on the atomic cores, i.e. $\lambda_f = \lambda_F = 1$. In section 4.1.2, we will show how the training targets $\hat{\boldsymbol{f}}_i^{(d)}$ can be extracted from first-principles calculations.

The extended PES contains more information than a conventional PES due to the inclusion of the electronic degrees of freedom. For instance, one can easily compute the polarizability, dielectric or piezoelectric tensor of a system. These properties require knowledge of how the dipole vector changes under the influence of electric fields, strains or stresses, which is contained in the extended PES via the positions of the electron pairs. Furthermore, the eMLP can also predict static or dynamic infrared spectra. In section 4.2, we will show how some of these properties can be computed.

In conventional force fields, performing an MD simulation is straightforward since the dynamics are dictated by the forces on the nuclei in each time step. In the eMLP, there are also forces acting upon the electron pairs. There are two main approaches to deal with these. In the first approach, one assigns a mass to the electron pairs such that they have their own dynamics, similar to Car–Parrinello molecular dynamics (CPMD).[140] Thus, the electron pairs and atomic cores are treated on equal footing in CPMD. The electron pair mass is typically chosen to be equal or slightly lighter than the mass of an hydrogen atom.[142, 145] In principle, an electron pair is 918 times lighter than a proton but in that case, an unreasonably small time step is required for stable dynamics. Furthermore, thermostat or barostats in MD simulations should only interact with the atomic cores and not with the electron pairs. Otherwise, the wrong ensemble is sampled. Therefore, the electron pairs should be simulated at a different temperature with their own thermostat which is decoupled from the other particles. Because the CPMD approach does not work with the true electron mass, artifacts will inevitably be introduced in the simulation. For instance, this can affect moments of inertia or dynamic properties like correlation times, diffusion constants or reaction kinetics.

To eliminate the complexity and artifacts of CPMD, one can choose to follow the Born-Oppenheimer (BO) approach. In this approximation, the extended

PES is minimized with respect to the electron pair positions in every time step because the electron pairs are significantly lighter than the atomic cores and instantaneously occupy their equilibrium positions $\{\boldsymbol{r}_i^{\mathrm{BO}}\}$. The resulting BO PES, i.e.

$$\begin{aligned} E_{\mathrm{eMLP,BO}}(\{\boldsymbol{R}_a\},\{Z_a\}) &= \min_{\{\boldsymbol{r}_i\}} E_{\mathrm{eMLP}}(\{\boldsymbol{R}_a\},\{Z_a\};\{\boldsymbol{r}_i\}), \\ &= E_{\mathrm{eMLP}}(\{\boldsymbol{R}_a\},\{Z_a\};\{\boldsymbol{r}_i^{\mathrm{BO}}\}) \end{aligned} \tag{4.3}$$

is now perfectly suitable for ordinary MD simulations since it only depends on the coordinates of the atomic cores. The forces on the electron pair positions $\{\boldsymbol{r}_i^{\mathrm{BO}}\}$ which minimize the eMLP energy are zero:

$$\boldsymbol{f}_j = -\frac{\partial E_{\mathrm{eMLP}}(\{\boldsymbol{R}_a\},\{Z_a\},\{\boldsymbol{r}_i^{\mathrm{BO}}\})}{\partial \boldsymbol{r}_j} = 0 \tag{4.4}$$

such that

$$\begin{aligned} \boldsymbol{F}_b^{\mathrm{BO}} &= -\frac{\mathrm{d}E_{\mathrm{eMLP,BO}}(\{\boldsymbol{R}_a\},\{Z_a\})}{\mathrm{d}\boldsymbol{R}_b} \\ &= -\frac{\partial E_{\mathrm{eMLP}}(\{\boldsymbol{R}_a\},\{Z_a\},\{\boldsymbol{r}_i^{\mathrm{BO}}\})}{\partial \boldsymbol{R}_b} \\ &\qquad -\sum_{j=1}^{N_\mathrm{e}} \frac{\partial E_{\mathrm{eMLP}}(\{\boldsymbol{R}_a\},\{Z_a\},\{\boldsymbol{r}_i^{\mathrm{BO}}\})}{\partial \boldsymbol{r}_j^{\mathrm{BO}}} \frac{\partial \boldsymbol{r}_j^{\mathrm{BO}}}{\partial \boldsymbol{R}_b} \\ &= -\frac{\partial E_{\mathrm{eMLP}}(\{\boldsymbol{R}_a\},\{Z_a\},\{\boldsymbol{r}_i^{\mathrm{BO}}\})}{\partial \boldsymbol{R}_b} \end{aligned} \tag{4.5}$$

because the second term in the second line of the equation is identically zero. Hence, the forces on the atomic cores in the BO approach can be evaluated by taking the partial derivative of the extended PES when the electron pairs occupy their equilibrium positions. These forces are readily obtained via automatic differentiation and they are required to perform the BO MD simulations. To compute the equilibrium BO positions of the electron pairs, the BFGS algorithm[61] in SciPy[213] has been used in the **eMLP paper**. On average, a dozen of energy evaluations are necessary using this algorithm, making the BO approach at least an order of magnitude slower than CPMD. However, one can utilize the time step of an ordinary MD simulation, avoiding the small time steps of CPMD.

4.1.1 Energy contributions

The eMLP energy or extended PES is the sum of a short-ranged MLP, long-range classical electrostatics and a reference energy:

$$E_{\mathrm{eMLP}} = E_{\mathrm{MLP}} + E_{\mathrm{electrostatics}} + E_{\mathrm{ref}} \tag{4.6}$$

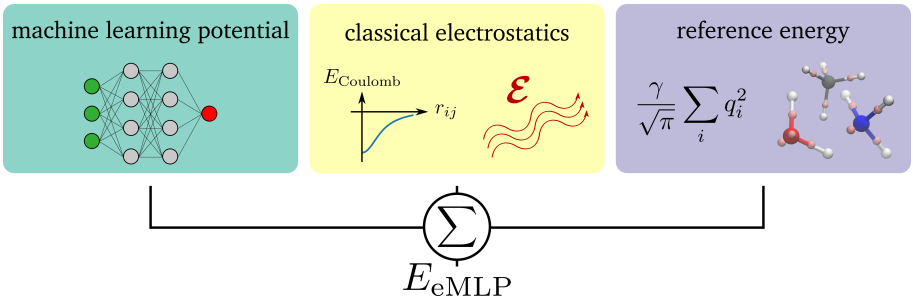


Figure 4.2: A schematic overview of the different energy contributions in the eMLP. The short-range interactions are modeled with an MLP. In the **eMLP paper**, SchNet has been used.[153] At long distances, the particles interact with each other and with external fields via classical electrostatics. Therefore, each particle has a Gaussian charge distribution with a fixed width. Finally, a reference energy is added and for small organic molecules, the eMLP energy is computed with reference to the four hydrides H_2, CH_4, NH_3 and H_2O (H_2 is not shown).

These three energy contributions are visualized in Figure 4.2. All contributions have their own specific purpose. The machine learning part models the short-range interactions, which is one of the main novelties of the eMLP in comparison with other explicit-electron force fields. The flexibility of MLPs make it possible to accurately learn the intricate short-ranged interactions, where quantum effects dominate. Besides the atomic cores, also the electron pair positions are used as the input of the eMLP. Practically, an unused atomic number is assigned to the electron pairs such that the MLP sees them as just another 'species'. In principle, the short-range interactions may be learned with any MLP. In the **eMLP paper**, we chose SchNet,[153] to accomplish this task for the reasons specified in section 3.2.4. In future versions of the eMLP, it might be beneficial to switch to equivariant MLPs to improve the accuracy and data-efficiency. A cutoff radius of 4 Å was selected for the MLPs.

The long-range interactions in the eMLP are modeled via classical electrostatics. All particles interact with each other via the Coulomb interaction as

if they have a Gaussian distribution,

$$\begin{aligned} E_{\text{electrostatics}} =& \frac{1}{2}\sum_{ab} q_a q_b \frac{\text{erf}(\gamma||\boldsymbol{R}_A - \boldsymbol{R}_b||)}{||\boldsymbol{R}_A - \boldsymbol{R}_b||} + \frac{1}{2}\sum_{ij} 4\frac{\text{erf}(\gamma||\boldsymbol{r}_i - \boldsymbol{r}_j||)}{||\boldsymbol{r}_i - \boldsymbol{r}_j||} \\ &+ \sum_{ai}(-2q_a)\frac{\text{erf}(\gamma||\boldsymbol{R}_a - \boldsymbol{r}_i||)}{||\boldsymbol{R}_a - \boldsymbol{r}_i||} + \sum_a q_a V_{\text{ext}}(\boldsymbol{R}_a) \\ &- 2\sum_i V_{\text{ext}}(\boldsymbol{r}_i), \end{aligned} \tag{4.7}$$

and also feel the presence of an arbitrary external field $V_{\text{ext}}(\boldsymbol{r})$. When dealing with a homogeneous electric field $\boldsymbol{\mathcal{E}}$, the last two terms become $-\sum_a q_a \boldsymbol{\mathcal{E}}\cdot\boldsymbol{R}_a + 2\sum_i \boldsymbol{\mathcal{E}}\cdot\boldsymbol{r}_i$. In the previous expression, $\text{erf}(x)$ is the errorfunction, q_a is the charge of the atomic core under consideration and $\gamma = \frac{1}{2\sigma}$ is inversely proportional to the width σ of the Gaussian charges, which is fixed to a value of 1.2728 Å. The same width σ is assigned to every particle, both electron pairs and atomic cores. Note that the eMLP does not try to model the true spatial extent of the first-principles charge distribution. For periodic systems, the electrostatic interaction between all the particles is computed with the Ewald-summation (ES).[214] The 'screening' Gaussian charges in the ES are chosen with exactly the same width as the Gaussian charges in the eMLP. In that case, the real space contribution of the ES vanishes and only the reciprocal part (and the constant self-interaction) has to be computed which increases the computational efficiency of the eMLP.

The long-range interactions do not try to mimic the Coulomb integrals, appearing in HF or DFT. They only provide a sensible long-range limit, which is classical electrostatics without exchange, i.e. outside the cutoff radius of 4 Å, Eq. (4.7) approximately reduces to the interaction between point charges. Inside the cutoff radius, the magnitude of the long-range interactions are damped via a smooth continuation of the classical limit. If there were no damping, atomic cores would feel forces up to 2500 eV/Å.[b] This would be problematic since the typical magnitude of a force is 1-2 eV/Å in MD simulations at 600 K and the MLP should unlearn these large electrostatic interactions.

Finally, the long-range contribution does not have any trainable parameters. This is beneficial since the number of representative chemical environments needed to train these interactions, dramatically increases at longer distances. Moreover, long-range energies and forces are generally smaller, making them increasingly harder to learn since the total force vector will be dominated by

b. This is the magnitude of the force between the point charge of an electron lone pair and the corresponding oxygen atomic core separated over a distance of 0.3 Å.

the short-range contribution. Hence, the partitioning in a short-range MLP and classical parameter-free long-range part increases the transferability and data-efficiency of the eMLP.

The last eMLP energy contribution is merely a reference energy, which does not depend on the exact geometric conformation of the system. It is introduced to normalize the eMLP energy predictions, which in turn accelerates and stabilizes the training of the short-ranged MLP part. Furthermore, it will provide a correct reference point for the energy, similar to atomization energies or other reference schemes in MLPs. It consists of two terms,

$$E_{\text{ref}} = \frac{\gamma}{\sqrt{\pi}} \left(\sum_a q_a^2 + 4N_{\text{e}} \right) + E_{\text{ref,sys}}, \tag{4.8}$$

with q_a the charges of the atomic cores, N_{e} the number of electron pairs and γ the same parameter appearing in Eq. (4.7). For the neutral systems considered in this work, one can prove for overlapping Gaussian charges that

$$E_{\text{electrostatics}} + \frac{\gamma}{\sqrt{\pi}} \left(\sum_a q_a^2 + 4N_{\text{e}} \right) \approx 0. \tag{4.9}$$

The **eMLP** paper can be consulted for a complete derivation. Thus, the first term in Eq. (4.8) (this is the self-energy in the **eMLP** paper) counteracts the absolute value of the long-range energy. Similarly to the damping of the electrostatics in the long-range contribution, the reference term additionally damps the energy, which is beneficial to train the short-ranged MLP part.

The system-specific reference $E_{\text{ref,sys}}$ provides the correct reference point for the energy. In the most simple case, for systems with a fixed chemical composition,[c] it is just a constant value. Often, the mean target energy of all systems in the training set is selected for that purpose. Again, it ensures that the short-range contribution of the eMLP predicts energy values close to zero. For data sets with different chemical compositions (e.g. the eQM7 data set, see section 4.2.1), a different approach is necessary. In conventional MLPs, a per-species reference energy is utilized, which is assigned to every species of atomic cores and the electron pairs. However, in the context of the eMLP, per-species reference energies are ill-defined. For instance, a single carbon atomic core without any (valence) electron pairs has by itself no suitable meaning. Therefore, the reference energy is computed with respect to the

c. All systems in the data set contain the same number of atomic cores of each species and the same number of electron pairs.

four hydrides H_2, CH_4, NH_3 and H_2O:

$$E_{\text{ref,sys}} = \frac{1}{2}(n_{\text{H}} - 2n_{\text{O}} - 3n_{\text{N}} - 4n_{\text{C}})E_{\text{H}_2} + n_{\text{C}}E_{\text{CH}_4} + n_{\text{N}}E_{\text{NH}_3} + n_{\text{O}}E_{\text{H}_2\text{O}} \tag{4.10}$$

where E_X is the energy of hydride X and there are n_Z atomic cores in the system of species Z. This system-specific reference energy is only valid for systems containing the elements H, C, N, and O. It can be easily extended to more elements by including additional hydrides. Unlike what is done in conventional MLPs, the energies of the hydrides E_X are not determined before training the eMLP. They do not have a constant value but depend on the actual trainable parameters $\{\boldsymbol{\theta}\}$ during the training process. This guarantees that the reference energies are fully consistent with the actual predictions of the hydrides. This is accomplished by including the four hydrides in every batch such that their eMLP energies E_X are computed in every training step.

When using the eMLP python package,[d] a single energy and force evaluation takes about 3.1 ms for 1,3-dimethylazetidine (17 atomic cores and 18 electron pairs) on a A100 GPU. In the BO approach, this increases by one order of magnitude since approximately ten energy evaluations are necessary to minimize the total energy. These small values prove the usefulness of the eMLP since a DFT first-principles calculations takes approximately 50 s on eight cores of an AMD Epyc 7H12 CPU. The enormous speed up of three to four orders of magnitude will increase even more when considering larger systems due to the cubic scaling of DFT. However, compared to conventional (polarizable) force fields, the eMLP will be at least one order of magnitude slower. This has two main causes: (i) evaluating the MLP energy is computationally more expensive than just evaluating more simple analytical expressions and (ii) the eMLP operates with double the amount of particles because of the electron pairs.

4.1.2 Electron localization

To be able to train the eMLP, the electron pair positions $\boldsymbol{r}_i$ and training forces $\hat{\boldsymbol{f}}_i$ should be defined. In conventional explicit-electron force fields, the precise location itself is often not directly related to the first-principles electron density. In the eMLP however, they are directly derived from the restricted Kohn-Sham orbitals in DFT or HF calculations. The orbitals $|\phi_i\rangle$ are the solutions of the self-consistent Kohn-Sham equations of Eq. (2.21) and correspond to a single energy level. Unfortunately, they are delocalized

d. https://github.com/mcoolsce/eMLP

over the whole system, making them inadequate to describe a single electron pair. Luckily, any measurable physical property including the total energy, is invariant under unitary transformations of all the occupied canonical orbitals:

$$|\varphi_i\rangle = \sum_{j \in \text{occ}} \mathsf{U}_{ij} \, |\phi_j\rangle \,, \tag{4.11}$$

with $\mathsf{U}\mathsf{U}^T$=$\mathsf{U}^T\mathsf{U}$=I. These resulting orbitals $|\varphi_i\rangle$ do not necessarily correspond to a single energy anymore but they can be localized by tuning the unitary matrix. Then, they are called localized molecular orbitals. For non-periodic systems, several schemes exist such as Foster-Boys (FB),[208] Edmiston-Ruedenberg (ER)[215] or Pipek-Mezey (PM)[216] localization. In these localization schemes, the unitary matrix is obtained by minimizing a cost function measuring the amount of localization of the orbitals. For instance, in the FB cost function, the spatial variance or spread of each orbital is minimized:

$$\mathcal{C}_{\text{FB}}(\mathsf{U}) = \sum_{i \in \text{occ}} \langle \varphi_i | \, || \hat{\boldsymbol{r}} - \langle \varphi_i | \hat{\boldsymbol{r}} | \varphi_i \rangle \, ||^2 \, | \varphi_i \rangle \tag{4.12}$$

After performing the minimization, the electron pair positions can be determined as the centers of the localized orbitals:

$$\boldsymbol{r}_i = \langle \varphi_i | \hat{\boldsymbol{r}} | \varphi_i \rangle \tag{4.13}$$

These are well-defined and can be used to train the eMLP. Furthermore, the total electronic dipole moment of a non-periodic system $\boldsymbol{\mu}$ is exactly reproduced by only using the electron pair positions $\boldsymbol{r}_i$ because it is invariant under a unitary transformation:

$$\boldsymbol{\mu} = -2 \sum_i \boldsymbol{r_i} = -2 \sum_i \langle \varphi_i | \hat{\boldsymbol{r}} | \varphi_i \rangle = -2 \sum_i \langle \phi_i | \hat{\boldsymbol{r}} | \phi_i \rangle \tag{4.14}$$

Hence, if the eMLP is trained well, it will also accurately reproduce the dipole moments. For higher order multipoles, there is no one-to-one relation between the first-principles value and its classical equivalent calculated by the electron pair positions $\boldsymbol{r}_i$.

In the eMLP, we make use of the FB localization since the most compact orbitals[e] with respect to the variance are obtained with the cost function of Eq. (4.12). Thus, correlation, exchange and other quantum effects of a certain electron pair will be localized or restricted around the position of its center. This is exploited in the eMLP where the cutoff radius of the short-ranged MLP part is limited to only 4 Å. Moreover, the FB electron pair locations are

e. Minimizing the fourth moment, i.e. $\langle r^4 \rangle$, might provide even more compact and chemically intuitive orbitals.[217]

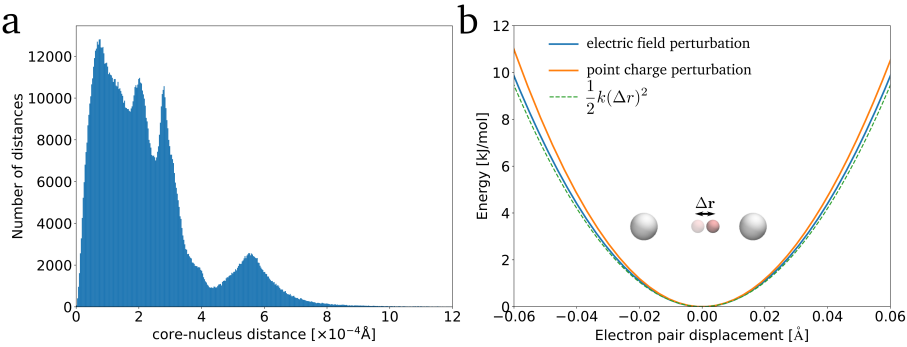


Figure 4.3: (a) A histogram of all the distances between a core electron pair and the corresponding nucleus in the eQM7 test set. Electric field perturbations are included herein. (b) An illustration of the non-uniqueness of the extended PES when displacing the single electron pair along the bond axis in H_2. It has been computed at the PBE0 level of theory by applying an electric field or point charge and subtracting the external field contribution (see text). The dashed line is a quadratic approximation of the extended PES with $k = 2.0\,\mathrm{Ha/Å^2}$.

almost indistinguishable from the expected Lewis structure of the molecule. One can identify single, double and triple bonds and lone pairs. In Figure 4.1, this is shown for the ethanal molecule. Finally, almost all FB positions of the core electron pairs, even when perturbed with electric fields, are located within approximately 5×10^{-4} Å from its corresponding nucleus. Figure 4.3 (a) demonstrates this numerically for all molecules within the eQM7 test set, which will be introduced in section 4.2.1. Thus, the assumption that the core electron pairs and nuclei can be described as one single particle, i.e. the atomic core, is now verified.

Maximally localized Wannier functions (MLWF)[209, 210] are utilized as reference positions for the electron pairs in periodic structures. The MLWFs are the periodic equivalent of the FB localized orbitals. They minimize the same cost function but the spread is now defined in terms of Wannier functions which are constructed from Bloch orbitals. Again, the electron pairs are placed at the centers of the MLWFs.

The localized orbitals $|\varphi_i\rangle$ are computed from the canonical orbitals from the ground state of the molecule. These are extracted from first-principles calculation within the Born-Oppenheimer approximation. Thus, the resulting electron pairs positions correspond to the BO equilibrium positions $\boldsymbol{r}_i^{\mathrm{BO}}$ of the eMLP. These minimize the extended PES of Eq. (4.3). Furthermore, in

DFT or HF, any perturbation to the ground state electron density will yield slightly displaced electron pairs, corresponding to energies higher than the ground state. Hence, the target forces, on the localized centers, extracted from a first-principles calculations in the ground state are zero:

$$\hat{\boldsymbol{f}}_i = 0 \tag{4.15}$$

This means that the eMLP is trained only using electron pair BO positions with a total force of zero acting upon them. These alone are certainly not enough to train the full extended PES of the eMLP, being a function of arbitrary electron pair positions $\boldsymbol{r}_i$. Fortunately, that kind of data can be generated by applying external perturbations. For instance, an homogeneous electric field perturbation,

$$\hat{V}_{\text{ext}}(\boldsymbol{r}) = 2\boldsymbol{\mathcal{E}} \cdot \boldsymbol{r}, \tag{4.16}$$

or point charge located at $\boldsymbol{r}_q$ with charge q,

$$\hat{V}_{\text{ext}}(\boldsymbol{r}) = -\frac{2q}{||\boldsymbol{r} - \boldsymbol{r}_q||} \tag{4.17}$$

can perturb the electron density and in turn the electron pair positions. In these cases, a DFT or HF calculation will still yield the ground state density or BO positions under the external field conditions. Again, the total force $\hat{\boldsymbol{f}}_i = \hat{\boldsymbol{f}}_i^{\text{int}} + \hat{\boldsymbol{f}}_i^{\text{ext}} = 0$ is zero but can be written as an external contribution, due to the external field, and an internal contribution, due to the polarization of the electron cloud. The external force acting on the electron pairs can be computed analytically as an expectation value, i.e.

$$\hat{\boldsymbol{f}}_i^{\text{ext}} = -\langle \varphi_i | \nabla \hat{V}_{\text{ext}}(\boldsymbol{r}) | \varphi_i \rangle \tag{4.18}$$

Therefore, if we consider the same perturbed electron density but without an external field, the total force would simply be $\hat{\boldsymbol{f}}_i = \hat{\boldsymbol{f}}_i^{\text{int}} = -\hat{\boldsymbol{f}}_i^{\text{ext}}$. Hence, applying an external field is a simple way to generate perturbed electron pair positions with nonzero forces.

There are still a number of difficulties and things to consider with the external field perturbations. The electron pair locations can only be perturbed in a collective manner. One cannot simply displace only one electron pair. Inevitably, the configurations will be correlated which is something to avoid when training MLPs. Moreover, the magnitude of the external fields is limited by the convergence of the SCF method. For instance, some of the larger molecules yield wrong results when going beyond field strengths of 0.01 au or 5.14×10^9 V/m. For this reason, all the electric field strengths were limited to that value. However, these strengths only manage to perturb the electron

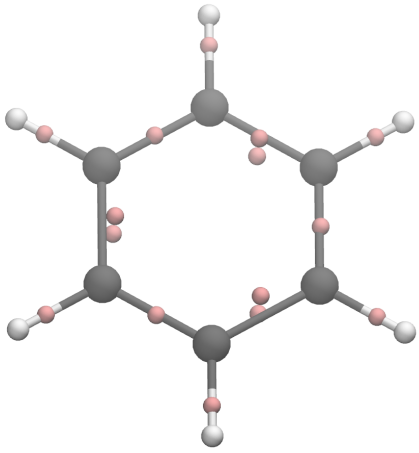


Figure 4.4: The FB reference structure for benzene. The electron pair positions are located at one of the two resonance structures, breaking the D6h point group symmetry of benzene.

pairs with a distance of approximately 0.02 Å. This is problematic because only a minor region around the BO equilibrium positions can be explored. Consequently, extrapolation will occur regularly when minimizing the BO PES with the eMLP. In section 4.3, we will introduce data augmentation to overcome this issue.

Multiple different electron densities or wavefunctions can also be associated with the same set of electron pair positions. They will likely have a different energy. Hence, the first-principles extended PES is not unique, if defined as a function of the electron pairs only. This is illustrated in Figure 4.3 (b) for H_2, where the single electron pair has been displaced along the bond by using an homogeneous electric field and point charge. Fortunately, the differences are almost negligible, especially for the sampled displacements up to 0.02 Å.[f] Furthermore, physical response properties such as polarizabilities which depend on the curvature of the PES are still well defined together with the BO PES. Hence, for all practical purposes of the eMLP, the non-uniqueness is not critical.

Finally, some specific molecules or materials cannot be properly modeled by using the centers of localized orbitals. One of the most simple examples is benzene, visualized in Fig. 4.4, where the electron pairs predict a structure with alternating single and double bonds. This is one of the two possible

f. Since H_2 is a small molecule, the SCF cycle will not break down that easily. Therefore, we can apply larger field strengths to drive the electron pair further than 0.02 Å in Figure 4.3 (b).

electron pair configurations that minimize the FB cost function. The other one is mirrored vertically, and they both correspond to the resonant structures of benzene. In the optimized structure, both are equivalent but when an external perturbation breaks the symmetry, one of the two structures might be preferred. Hence, the electron pairs might move discontinuously over a large distance. In practice, the eMLP cannot predict such displacements and the electron pairs remain in one of the electron pair configurations. In general, for molecules or materials in which the electron pair positions make discontinuous jumps, the eMLP will have trouble or is simply incapable to predict the correct structure. Also for highly delocalized orbitals, the centers might be very sensitive to small fluctuations of the external potential, resulting in difficulties when predicting their locations.

4.2 Results

To assess the accuracy and predictive power of the eMLP, we trained the eMLP for two specific use-cases. In the first use-case, we focused on modeling small organic molecules. For that purpose we have constructed a brand new data set, eQM7,[218] which will be described in the following subsection. The accuracy of dipole predictions, polarizabilities and infrared (IR) spectra will be explored for unseen molecules, verifying the transferability of the eMLP. Modeling periodic systems with the eMLP will be demonstrated in the second use-case. There, we aim to predict the response properties of β-glycine. This includes the elasticity, dielectric and most importantly the piezoelectric tensor, drawing a lot of attention in recent research due to large piezoelectric strain responses.[219, 220]

Most of the reported error values and results of the eMLP can be subdivided into two categories, which we have called *static errors* and *dynamic errors*. Static errors are evaluated by placing the electron pairs in the same positions as the reference electron pairs. Thus, the errors are computed with the extended PES of Eq. (4.1). These are the errors that are directly minimized in the cost function when training the eMLP. On the other hand, the BO PES of Eq. (4.3) is evaluated to compute dynamic errors. Here, the positions of the electron pairs are optimized and do not necessarily coincide anymore with the true first-principle positions. If the eMLP is trained well, the deviations will be small however.

In this section, only the most prominent and interesting results will be highlighted. For all other results, including training details and other numerical error values, we refer the interested reader to the **eMLP paper**.

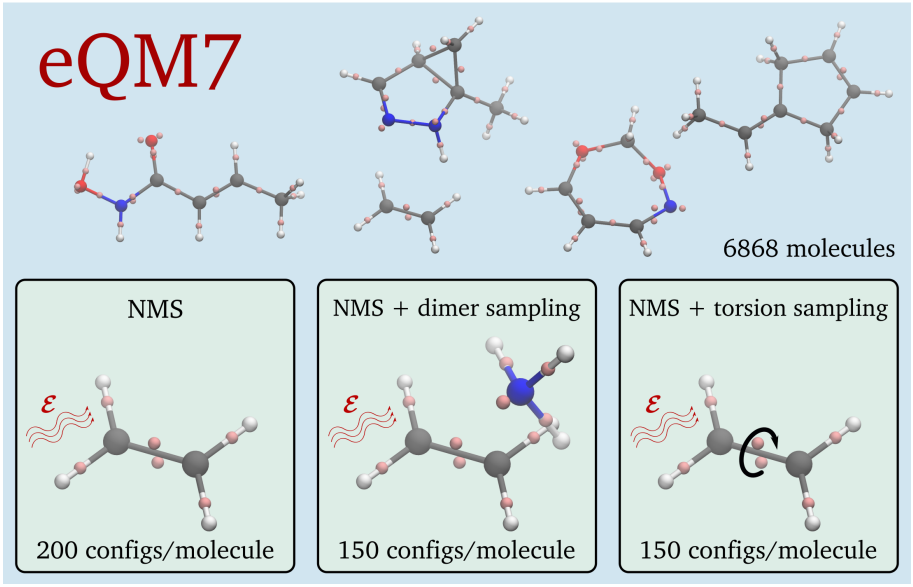


Figure 4.5: An illustration of the molecules and data generation techniques used in the eQM7[218] data set. Each of the 6868 molecules in the data set contains up to seven "heavy" atoms, i.e. carbon, nitrogen or oxygen. For each molecule, 500 different first-principles DFT calculations have been performed with the PBE0[11, 76] functional. The employed data generation techniques include normal mode sampling (NMS) at 600 and 800 K, dimer sampling and torsion sampling (see section 3.1.1). The electron pairs are perturbed with random homogeneous electric fields with a maximum strength of 0.01 au.

4.2.1 eQM7

The development of a transferable MLP for all kinds of small organic molecules, was one of the original goals of the eMLP. To accomplish this, the electron QM7 (eQM7) data set was constructed. The data set is published online (Ref. 218) at the Materials Cloud archive.[221] It contains 6868 molecules with up to seven "heavy" atoms. Some exemplary molecules are depicted in Figure 4.5. All the different molecules were extracted from the original QM7[166] data set. There, only equilibrium positions were considered, whereas in the eQM7 data set, also perturbed geometries are included. This was realized by normal mode sampling, dimer sampling and torsion sampling, which are described in section 3.1.1, and yielded 500 configurations per molecule. The number of samples for each perturbation technique is also

shown in Figure 4.5. Homogeneous electric field with a maximum field strength of 0.01 au and the previously mentioned dimer sampling were used to perturb the electron pairs. The locations of the electrons pairs themselves are computed via FB localization. Assessing the transferability towards small organic molecules is made possible by a train, validation and test split on a molecule by molecule basis. Hence, all 500 configuration of a certain molecule are stored in one of the three sets. All the results in this section are reported on the test set, which only contain unseen molecules, and directly verifies the potential transferability of the trained models.

After training the eMLP, the MAE of the forces (or atomic cores) are 48.8 meV/Å while the MAE on the energy is 4.45 meV/atom. These values are 20-30 times lower than the intrinsic fluctuations of these properties in the eQM7 data set. The intrinsic fluctuations are defined as the MAE of the best constant model, i.e. a model which yields a constant value for each configuration. This serves as a baseline value since "smart" models should definitely perform better. Hence, the eMLP does make accurate predictions given the context that the validation is done on unseen molecules. A quantitative comparison with other state-of-the-art MLPs is not straightforward since the eQM7 data set is only recently published and most other MLPs cannot deal with electron pair particles. However, to give some perspective to the quality of these results, the value of the errors can be compared with results on the ISO17[222] data set because, just like eQM7, the errors are also validated on unknown molecules. MAE on the forces ranging from 50 to 90 meV/Å are reported there for different modern MLPs.[121, 175, 191] This clearly indicates that the eMLP is capable of learning the extended PES with suitable accuracy.

Interpreting the dynamic errors require more care. Simply reporting the MAEs does not tell the whole story here. For instance, the MAE on the forces is 104 meV/Å, which only by itself would illustrate that the eMLP is not capable of learning the BO PES and performing MD simulations. However, the median error (50% of all the errors lie below this value) is only 30.0 meV/Å, comparable to the value of 27.9 meV/Å for the static median error. The full cumulative error distribution of both the static and dynamic forces, is plotted in Figure 4.6. Notice that the dynamic errors almost overlap with the static cumulative distribution for errors smaller than $\sim$ 30 meV/Å but they have a more heavy tail distribution. This skews the MAE to larger values which do not represent the full distribution anymore. Moreover, the dynamic distribution does not reach 100% anymore but plateaus at 97.6 %. For the remaining 2.4 % of all the configurations in the test set, the minimization of the extended PES simply fails. In that case, the BFGS algorithm[61] in SciPy[213] cannot find a proper solution anymore and the BO PES cannot be

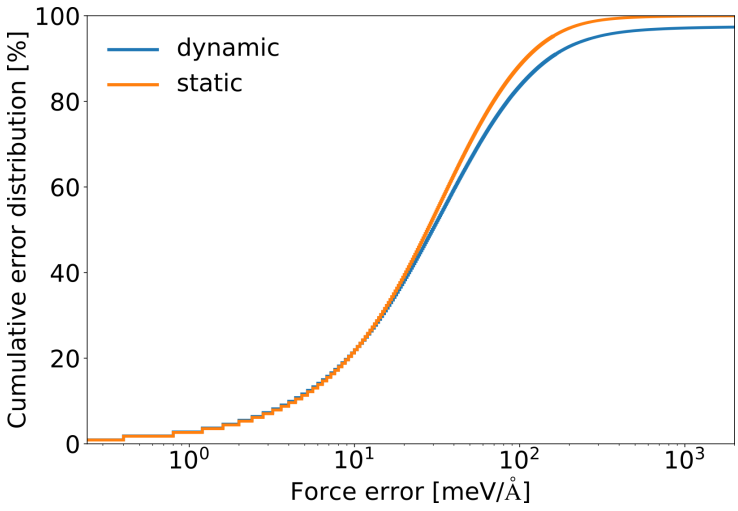


Figure 4.6: Cumulative error distributions of the static and dynamic force errors. The dynamic errors have a heavy tail distribution and plateau at 97.6 % because for 2.4 % of the configurations the BO minimization fails.

computed. This does not mean that the eMLP is not useful anymore. For the majority of configurations, energies and forces are predicted with high accuracy because the median error remains almost constant. Still, in section 4.3, we will introduce data augmentation, a technique which will reduce the error rate of 2.4 % to only 0.4 %, to increase the transferability of the eMLP even further.

Besides energies and forces, the eMLP can also predict dynamic errors of the dipole vector of a molecule.[g] The median error on the norm of the dipole is only 0.035 Debye, almost 40 times less than the average dipole norm of all the molecules in the test set. Another property related to the dipole moment is the polarizability tensor,

$$\alpha_{ij} = \frac{\partial \mu_i}{\partial \mathcal{E}_j} = -\frac{\partial^2 E_{\text{eMLP}}}{\partial \mathcal{E}_i \partial \mathcal{E}_j}, \tag{4.19}$$

which tells us how the dipole vector $\boldsymbol{\mu}$ changes under an electric field $\boldsymbol{\mathcal{E}}$. One of the strengths of the eMLP is that this property can be calculated analytically via automatic differentiation by computing the Hessian of the extended PES. The median error on the components of the polarizability tensor is 0.26 bohr^3 for the unseen molecules in the test set. This value is

g. Static errors of the dipole vector would be zero by definition because the electron pairs are exactly located at the true first-principles FB positions.

25 times lower than the intrinsic fluctuations on the polarizability tensor. Thus, the eMLP can describe both the dipole vector and polarizability tensor of a molecule with high precision.

Finally, we will predict infrared (IR) spectra with the eMLP. This is extremely useful since IR spectra are molecular fingerprints, providing a manner to identify molecules experimentally. This can be done with a static calculation in the 0 K limit or with dynamic simulations at finite temperatures. In static calculations, the spectra consist of a series of frequencies ω_q with intensities

$$I_q \sim \left|\left| \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}q} \right|\right|^2 \tag{4.20}$$

for each normal mode q of the molecule.[223] The intensities are related to the change of the dipole moment along a normal mode. Therefore, to obtain accurate predictions, the eMLP should reproduce both the mass weighted hessian (to compute the frequencies ω_q) and the dipole moments. Similarly to the polarizability tensor, the static IR spectrum can be calculated completely analytically in the eMLP. For dynamic spectra,

$$I(\omega) \sim \int \langle \dot{\boldsymbol{\mu}}(\tau) \cdot \dot{\boldsymbol{\mu}}(t+\tau) \rangle_\tau \, \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t \tag{4.21}$$

the autocorrelation of time derivatives of the dipole moment $\dot{\boldsymbol{\mu}}$ are required.[223] This can be computationally expensive at the first-principles level of theory since long MD simulations are required. Therefore, the eMLP will be a useful tool to accelerate these simulations. But first, the accuracy of the eMLP IR spectra are validated by comparing them to the first-principles spectra of all the unseen molecules in the test set. The mean absolute error between the sorted lists of frequencies is on average 10-15 cm^{-1}. Also by visually inspecting the spectrum, one can conclude that both the frequencies and intensities of each mode are in good agreement with the first-principles spectrum. This is illustrated in the left hand side of Figure 4.7 for one of the molecules in the test set. For all the other molecules in the test set, similar results are obtained, which validates the eMLP to predict spectra. Thus, we expect that spectra at finite temperatures will also be accurately predicted. The dynamic spectra of the same molecule is plotted on the right hand side of Figure 4.7, which can be used in practice to identify molecules at realistic temperatures.

4.2.2 Beta-glycine

To test the performance of the eMLP for a periodic structure, we focused on crystalline β-glycine. Its equilibrium structure and periodic unit cell are depicted in Figure 4.8. The ultimate goal is to predict its response properties and

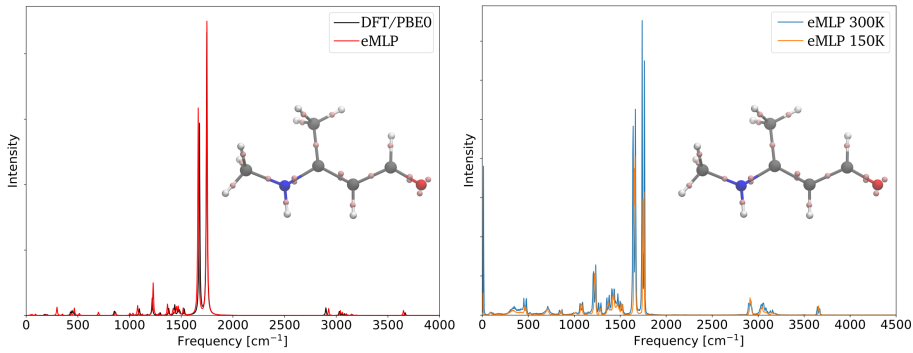


Figure 4.7: IR spectra of 3-(methylamino)but-2-enal. Left: static IR spectra, predicted by the eMLP and DFT with the PBE0 functional. Lorentzian line shape functions with a full width at half maximum of 10 cm^{-1} has been used to plot the spectra. Right: eMLP predictions of the dynamic IR spectra at 150 and 300 K. Adapted with permission from Ref. 10. Copyright 2022 American Chemical Society.

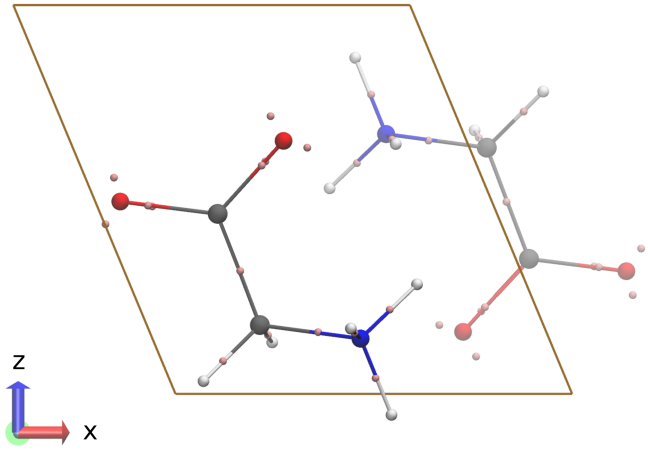


Figure 4.8: The equilibrium structure of β-glycine and its periodic unit cell. The electron pairs are located at the centers of maximally localized Wannier functions (MLWFs). Reprinted with permission from Ref. 10. Copyright 2022 American Chemical Society.

more specifically, the piezoelectric strain tensor. Therefore, a new data set was constructed and published online (Ref. 224) with 25676 first-principles DFT calculations with the PBE[75] functional. Normal mode sampling was used to generate uncorrelated geometries. Besides the nuclear positions, also the degrees of freedom of the cell vector were considered.[h] The electron pair positions are located at the centers of maximally localized Wannier functions and perturbed with homogeneous electric fields with a maximum strength of 0.01 au. In addition to the energies and forces, also a mean squared error on the stress tensor is included in the cost function to train the eMLP.

After the data set was constructed, we trained the eMLP and achieved static MAEs of 3.1 meV/atom, 35 meV/Å, and 0.063 GPa on the energies, forces and stresses respectively. Furthermore, after a geometry optimization with the eMLP, the structural properties (volume, lattice lengths, angles) were reproduced almost exactly with a maximum difference of only 0.3 %. Therefore, we concluded that we obtained a well-trained β-glycine model such that we are ready to validate the response properties of β-glycine (in the 0 K limit). More precisely, we are interested in the deformation of the system, measured by the strain tensor $\boldsymbol{S}$, due to stresses $\boldsymbol{\sigma}$ or electric fields $\boldsymbol{\mathcal{E}}$. Moreover, the electric displacement field $\boldsymbol{D}$ also varies under these perturbations. The linear relations between all these properties can be written down in the following coupled equations:[225]

$$\boldsymbol{S} = \boldsymbol{C}^{-1}_{\mathcal{E}=0} : \boldsymbol{\sigma} + \boldsymbol{d}^T \cdot \boldsymbol{\mathcal{E}} \tag{4.22}$$

$$\boldsymbol{D} = \boldsymbol{d} : \boldsymbol{\sigma} + \boldsymbol{\varepsilon}_{\boldsymbol{\sigma}=0} \cdot \boldsymbol{\mathcal{E}} \tag{4.23}$$

or alternatively

$$\boldsymbol{\sigma} = \boldsymbol{C}_{\mathcal{E}=0} : \boldsymbol{S} - \boldsymbol{e}^T \cdot \boldsymbol{\mathcal{E}} \tag{4.24}$$

$$\boldsymbol{D} = \boldsymbol{e} : \boldsymbol{S} + \boldsymbol{\varepsilon}_{\boldsymbol{S}=0} \cdot \boldsymbol{\mathcal{E}} \tag{4.25}$$

which is the strain-charge form. The response properties appearing in these equations are the elasticity (stiffness) tensor $\boldsymbol{C}$, dielectric tensor $\boldsymbol{\varepsilon}$, piezoelectric charge tensor $\boldsymbol{e}$ and piezoelectric strain tensor $\boldsymbol{d}$:

$$\varepsilon_{ij} = \left(\frac{\partial D_i}{\partial \mathcal{E}_j}\right)^{\boldsymbol{S}} \qquad C_{ijkl} = \left(\frac{\partial \sigma_{ij}}{\partial S_{kl}}\right)^{\boldsymbol{\mathcal{E}}} \tag{4.26}$$

$$e_{ijk} = \left(\frac{\partial D_i}{\partial S_{jk}}\right)^{\boldsymbol{\mathcal{E}}} = -\left(\frac{\partial \sigma_{jk}}{\partial \mathcal{E}_i}\right)^{\boldsymbol{S}} \qquad d_{ijk} = \left(\frac{\partial D_i}{\partial \sigma_{jk}}\right)^{\boldsymbol{\mathcal{E}}} = \left(\frac{\partial S_{jk}}{\partial \mathcal{E}_i}\right)^{\boldsymbol{\sigma}} \tag{4.27}$$

They are all related to second order derivatives of the total energy. Thus, within the framework of the eMLP, these can be calculated analytically via

h. Therefore, the Hessian matrix $\mathrm{H} \in \mathbb{R}^{(3N+9)\times(3N+9)}$ in Eq. (3.5) is an extended Hessian which also includes the nine elements of the unit cell.

automatic differentiation. This is explained in more detail in appendix A of the **eMLP paper**.

A complete overview the predicted and target values of all these response properties is given in the **eMLP paper**. Here, we will emphasize the most important results. First of all, the dielectric tensor is almost exactly reproduced, with deviations less than one percent. This means that the curvature of the extended PES with respect to the electron pair degrees of freedom is learned well. For the elasticity tensor, the results are more nuanced: most of the components have a minor absolute and relative error except for the C_{22} and C_{66} components (in Voigt notation). The reference C_{66} coefficient is 3.8 GPa, while the predicted value is 5.7 GPa. For such a small reference value,[i] a deviation of only 1.9 GPa is small in absolute value but is a considerable relative error. This is not surprising, considering the fact that the eMLP is trained with a mean squared error cost function where minimizing large errors is the priority. The error of the C_{22} coefficient (31.4 versus 22.2 GPa) has a different cause. The coefficient corresponds to the y-direction in Figure 4.8 where the zwitterions are stacked and interact with each other via $\pi - \pi$ bonds. These interactions are often dominated by dispersion which is not explicitly included in the eMLP.

The piezoelectric charge tensor $\boldsymbol{e}$ measures how the displacement field changes or how the electron pairs relax under varying strains. This behavior is captured within the eMLP since all components of the tensor are predicted accurately except for e_{22}, again probably due to missing dispersion. Finally, the piezoelectric strain tensor $\boldsymbol{d}$ measures the deformation of a material under the influence of an external field and is related to the piezoelectric charge tensor:

$$\boldsymbol{d} = \boldsymbol{e} : \boldsymbol{C}^{-1}_{\mathcal{E}=0} \tag{4.28}$$

Hence, the accuracy of the piezoelectric strain tensor is intrinsically linked with the accuracy of the elasticity tensor. This is also apparent in the results where some components have a significant relative error. Fortunately, we still observe a large d_{16} component of 46.5 pm/V which is one of the key features of β-glycine and critical in many technological applications.[219, 220] Hence, the predictions of all the response properties indicate that the eMLP can capture the correct behavior of β-glycine, although there is some room for improvement.

i. For instance, the C_{11} reference component is 62.4 GPa.

4.3 Data augmentation

Computing the BO PES of Eq. (4.3) requires minimizing the extended PES with respect to the electron pair positions. Therefore, the eQM7 and β-glycine data set were generated by applying random electric fields to sample the electronic degrees of freedom of the extended PES. For convergence reasons, the field strength was limited to only 0.01 au, however. These fields can only perturb the electron pair positions over a distance of at most 0.02 Å. Outside this region, the MLP will start to extrapolate and in the worst case scenario, it can display unpredictable behavior, similar to the PES in Figure 3.2. This is troublesome since the eMLP energy should be well-behaved around the BO equilibrium pair positions and certainly have a local minimum to compute the BO PES. In the worst-case scenario, the trained extended PES might have spurious minima close to the true BO equilibrium or no minimum at all. These effects have already been encountered in this chapter: computing the dynamic errors of the eQM7 data set failed for 2.4 % of the configurations. Additionally, we observed that MD simulations of a minority of the unseen test set molecules in the eQM7 data are not stable and simply crash after a certain number of steps. This is problematic and directly related to insufficiently sampling of the electronic degrees of freedom. A sampled region of 0.02 Å is too limited to learn the extended PES. Ideally, a first-principles method is preferred where the electron pairs can be perturbed at will over large distances and without any SCF issues. Unfortunately, we are not aware of such a method as of today.

To solve this issue, we will train the eMLP by applying data augmentation. In image classification, it is a well-known and common preprocessing step to increase the accuracy of a model by regularization.[226] There, the machine learning models are trained by presenting it with images which are transformed. Among others, these transformations include rotations, flipping the image or changing the contrast of the colors. The labels of the images remain the same however (e.g. a rotated cat is still a cat). Hence, the resulting models will be more robust to minor changes. These underlying ideas will be applied to make the eMLP more robust as well. In our case, the transformations are displacements of the electron pairs outside the 0.02 Å region. Unfortunately, the labels do not remain the same but the energies and forces do change in our context. Since these labels are not known a priori, a surrogate model is necessary to compute the target energies and forces. The only thing which is known with certainty, is that the total energy will increase when displacing an electron pair from its BO equilibrium positions. Thus, we assign a reference energy $\hat{E}_{\text{aug}} = \hat{E} + \Delta E_{\text{aug}}$ to the augmented structure, where $\hat{E}$ is its original first-principles reference energy and $\Delta E_{\text{aug}} > 0$. The

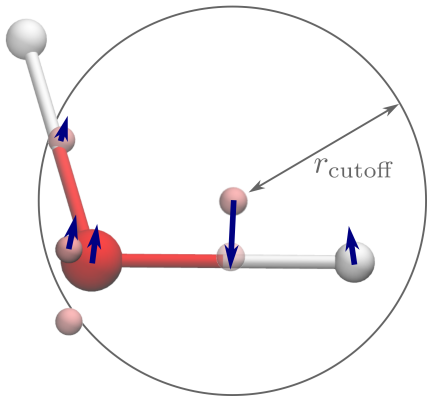


Figure 4.9: An example of data augmentation for the water molecule and the augmented force assignment. The central electron pair is displaced and feels the restoring force, depicted with a blue arrow. All the other particles within the cutoff radius r_{cutoff} of that electron pair do also feel the influence of the displacement and do receive an additional force.

precise value of ΔE_{aug}, computed by the surrogate model, is unimportant since these augmented configurations will never be visited in conventional simulations. We only want to teach the eMLP to predict higher energies in the neighborhood of the BO equilibrium positions such that spurious minima are avoided.

Practically, only 10 % of all the structures in a single batch are being augmented. For these systems, a single electron pair is chosen at random and displaced with a distance vector $\Delta \boldsymbol{r}_i$ between 0.06 and 0.12 Å. This range lies just outside the displacement range of 0.02 Å sampled with the electric field perturbations such that the eMLP will not encounter configurations with multiple labels. Afterwards, inspired by Figure 4.3 (b), we will assume a quadratic increase in energy:

$$\Delta E_{\text{aug}} = \frac{1}{2} k (\Delta r_i)^2 \tag{4.29}$$

which is the surrogate model. Furthermore, not only the energy but also the forces do change. The augmented force on the displaced electron pair is a restoring force $\Delta \boldsymbol{f}_i = -k \Delta \boldsymbol{r}_i$. This is shown in Figure 4.9 for a water molecule. All the other neighboring particles will also feel the influence of the displacement and receive an additional force contribution. However, there are two constraints: (i) the net force $\boldsymbol{F}_{\text{tot}}$ on the system should always be zero (for a neutral uncharged system) and (ii) the total torque $\boldsymbol{\tau}_{\text{tot}}$ should be

	no augmentation	augmentation	
$k\,[E_\mathrm{h}/\text{Å}^2]$	/	2.0	4.0
median error energy [meV/atom]	**3.15**	3.45	3.70
median error forces [meV/Å]	**30.0**	32.9	38.8
median error dipole norm [Debye]	**0.035**	0.078	0.083
median error polarizability [bohr^3]	**0.26**	0.31	0.40
error rate [%]	2.4	0.4	**0.07**

Table 4.1: A comparison of the dynamic errors with and without data augmentation on the eQM7 test set. Best values are put in bold.

equal to $\boldsymbol{\mu} \times \boldsymbol{\mathcal{E}}$ with $\boldsymbol{\mu}$ the dipole vector and $\boldsymbol{\mathcal{E}}$ the electric field:

$$\boldsymbol{F}_\mathrm{tot} = \sum_a \boldsymbol{F}_a + \sum_i \boldsymbol{f}_i = \boldsymbol{0} \tag{4.30}$$

$$\boldsymbol{\tau}_\mathrm{tot} = \sum_a \boldsymbol{R}_a \times \boldsymbol{F}_a + \sum_i \boldsymbol{r}_i \times \boldsymbol{f}_i = \boldsymbol{\mu} \times \boldsymbol{\mathcal{E}} \tag{4.31}$$

A unique weighted-least squares force assignment for all the neighboring particles has been derived in appendix B of the **eMLP paper**, which fulfills these two conditions.

The effect of data augmentation can immediately be seen by looking at the static and dynamic errors. In Table 4.1, the dynamic errors have been tabulated for a non-augmented model and two augmented models with different k-parameters on the eQM7 set. In general, the median errors do increase slightly when more data augmentation is being applied. However, the error rate decreases dramatically from 2.4 % to 0.4 % or even 0.07 %. This is exactly the reason why data augmentation was introduced, namely to avoid spurious minima in the extended PES and to stabilize the minimization of the electron pair positions. The force constant k is a hyperparameter of the eMLP, for which we have chosen the value $k = 2.0\,\mathrm{Ha/eV^2}$ in the **eMLP paper**, because its increase in robustness only comes at a minor cost in terms of accuracy.

Besides its effect on the dynamic errors, we also investigated whether data augmentation can further stabilize MD simulations. Therefore, we performed 500 fs long NVT simulations at 300 K for every unseen molecule in the test set. Instead of using the ordinary BFGS algorithm, the electron pairs positions were optimized with the L-BFGS-B algorithm,[227] where box constraints are imposed to limit the possible electron pair displacements from their initial guess. The size of the box was chosen to be three times the maximum displacement of an atomic core with respect to the previous MD step. In

almost all cases, the BO equilibrium positions are found within the box and not on the boundary such that the box constraints do not invalidate the correctness of the results. The algorithm was only introduced to speed up the minimization because some clues were added about the length scale of the problem. After the MD simulations were performed, we measure the stability of the simulations by keeping track of the conserved quantity per atom E_{cons}.[j] In the perfect scenario, where the BO equilibrium positions always vary smoothly with the nuclear degrees of freedom, the conserved quantity will remain conserved. In practice, the optimizer does not exactly find the minimum or it finds another physical but different local minimum, which is only separated by a small barrier. Then, the precise value of the conserved quantity will jump between two MD steps. If there are less spurious minima, the maximum jump $\max|\Delta E_{\text{cons}}|$ in an MD simulation will be small and thus, we expect data augmentation to limit the maximum jump and on average, the conserved quantity will increase or decrease less during the simulation.

Three types of MD simulations can be distinguished: stable MD simulations, depicted in Figure 4.10 (b), stable MD simulations with small electron pair jumps, depicted in Figure 4.10 (c), and unstable MD simulations. Unstable MD simulations are characterized with a maximum jump in conserved quantity per atom higher than 0.01 eV. These do correspond to situations where the BO optimization simply fails or where the solution is found on the boundaries of the box constraints. For these molecules, the eMLP is not able to perform MD simulations. Visually, this can be noticed by overlapping electron pairs or simply by a chaotic distribution of the electron pairs over the whole molecule. This typically happens in configurations which have an oxygen atomic core with two lone electron pairs. The latter two fall on top of each other after a certain amount of time. The unstable MD simulations happen in 9.8 % of the cases for non-augmented models while with augmentation, unstable MD trajectories were not observed. This is a major accomplishment and proves the effectiveness of data augmentation.

Stable MD simulations on the other hand are characterized with conserved quantity jumps smaller than 5×10^{-4} eV while the intermediate category has jumps with 5×10^{-4} eV < $\max|\Delta E_{\text{cons}}|$ < 0.01 eV. In the latter category, the conserved quantity increases or decreases during the simulation but the resulting MD simulation will still be stable because the thermostat in an NVT simulation can counteract the increase or decrease by adding thermal energy. Finally, in Figure 4.10 (a) every single MD simulation (one per molecule) is represented as a point where the first coordinate is the averaged loss or gain of the conserved quantity and the second coordinate is the maximum

j. In an NVE simulation, the conserved quantity is the total energy. In an NVT simulation, the thermostat itself also contributes to the conserved quantity.

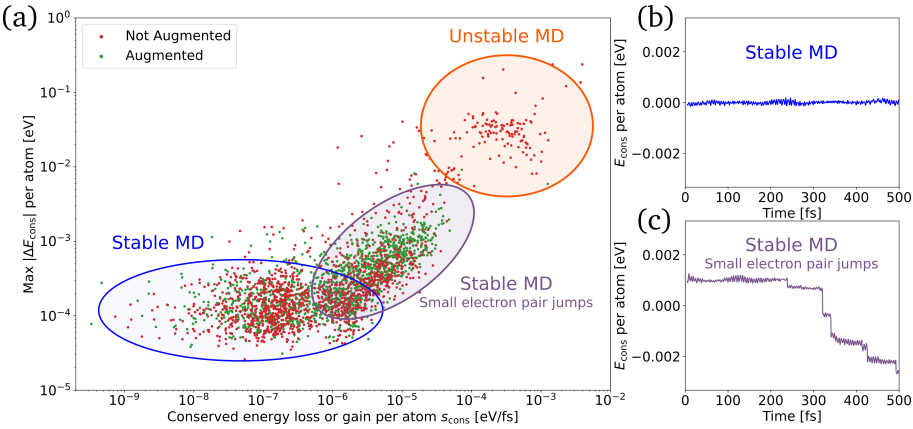


Figure 4.10: Panel (a): a scatter plot indicating the stability of MD simulations with and without data augmentation. Every point represents an MD simulation of one unseen molecule in the test set. The first coordinate of the point is the average loss or gain in conserved quantity per atom while the second coordinate is the maximum jump in conserved quantity per atom. The three different stability regions are encircled: stable MD, stable MD with small electron pair jumps and unstable MD simulations. Panel (b): an example of the conserved quantity during an stable MD simulation. Panel (c): an example of the conserved quantity during an MD simulations with small electron pair jumps. Reprinted with permission from Ref. 10. Copyright 2022 American Chemical Society.

jump. The three MD categories are also indicated. Again, the effect of data augmentation is clearly seen: it improves the stability and robustness of the eMLP.

5

Conclusions and Perspectives

There is no real ending.
It's just the place where you stop the story.
Frank Herbert (1920–1986)

5.1 Conclusions

Nowadays, molecular modeling has become an essential tool to understand and develop atomic systems. Its applications range from predicting properties of newly proposed but yet unsynthesized materials to the study of protein folding in biological processes. At the nanoscale, the forces acting on the atoms are determined by the quantum-mechanical electronic wave function. To simulate these systems numerically, the Schrödinger equation should be solved. First-principles methods accomplish this by solving for the electronic wave function or density without using any experimental input. Although these methods are relatively accurate, they come with a large computational cost and become unpractical to use when dealing with a large number of electrons due to poor system-size scaling. Force fields on the other hand are computationally cheap and applicable to massive atomic systems. They completely discard all electronic degrees of freedom and approximate the sought-for potential energy surface with an adequate analytical expression that is fitted to first-principles or experimental data.

However, these conventional force fields lack the accuracy, transferability and reactivity for a wide variety of applications.

Instead of utilizing physically inspired energy contributions, machine learning potentials make use of a machine learning algorithm to approximate the PES. They can have ten thousand to potentially ten million of trainable parameters, allowing them to fit the most intricate and multidimensional functions with unprecedented accuracy. Furthermore, MLPs are automatically able to model chemical reactions. These defining properties have caused their rise in popularity in recent years and currently, they are starting to replace conventional force fields.

Despite the fact that state-of-the-art MLPs can fully capture the short-range interactions in a data efficient manner, a solid treatment of long-range interactions or polarization is still missing. In this thesis, we have proposed a new methodology, the eMLP, to incorporate these important interactions. Therefore, we investigated how conventional force fields include these long-range energy contributions in section 2.4.3. The most popular method hereof employs fluctuating charges. However, in section 2.4.4, we have shown that the polarization in periodic systems is ambiguous or not well-defined when using fluctuating charges. This is a far-reaching observation which invalidates most conventional force fields and MLPs to properly predict physical properties such as dielectric or piezoelectric constants. Explicit-electron force fields, discussed in section 2.4.5, do not suffer from this issue since they reintroduce the electronic degrees of freedom by means of semi-classical electron particles. Therefore, they lie at the basis of the eMLP. Typically, the intricate short-range interactions between all the particles in an explicit-electron force field are hard to model because they are dominated by quantum effects. For this reason, we make use of machine learning in the eMLP to increase its accuracy. The eMLP is not only an improvement of explicit-electron force fields but advances the long-range interactions in MLPs as well. They were the topic of section 3.3 where the four MLP generations are described. The eMLP belongs to the fourth generation of globally polarizable MLPs.

Ultimately, the eMLP itself is the topic of Chapter 4. In section 4.1.2, we highlighted the fact that the electron pair particles are located at the centers of localized orbitals which are derived from Hartree-Fock or DFT calculations. These centers are appropriate descriptors of the electron density since they exactly reproduce certain expectation values such as the total dipole moment of the molecule. Furthermore, the locality of these orbitals allows for a short-ranged machine learning part. The long-range interactions on the other hand, are parameter-free and modeled via classical electrostatics. In the eMLP, the extended PES is a function of both the electron pairs and atomic cores and it should be minimized with respect to the electron pairs during dy-

namic simulations to obtain the Born-Oppenheimer PES. Unfortunately, the sampling of the electron pairs was shown to be inadequate to compute the BO PES in all circumstances. Therefore, we introduced data augmentation in the context of MLPs in section 4.3. By randomly displacing the electron pairs out of their BO equilibrium positions and stating that the total energy, predicted by a surrogate model, should be higher for such a configuration, the eMLP will become more robust and stable when performing molecular dynamics simulations. This is a major accomplishment which contributed to the success of the eMLP.

Finally, the eMLP is benchmarked on two newly constructed data sets. The first data set, eQM7, is introduced in section 4.2.1 and consists of a variety of nuclear and electronic perturbations for 6868 small organic molecules. The accuracy of the eMLP is assessed by predicting energies, forces, dipole moments and polarizabilities for unseen molecules. For all properties, the errors being made are more than twenty times smaller than the corresponding internal fluctuations and similar in accuracy with other state-of-the-art MLPs, proving the transferability of the eMLP. Additionally, we demonstrated the capability of the eMLP to predict correct static and dynamic infrared spectra. The second data set contains crystalline β-glycine and is described in section 4.2.2. There, several response properties were predicted, including the elasticity, dielectric and piezoelectric tensors. Most of the components were modeled with great accuracy, except for some directions in which dispersion interactions dominate. The latter interactions were not the focus and are not included in the currect version of the eMLP.

The secondary goal of this work was to propose a widely applicable data protocol to derive MLPs for MOFs, a promising class of materials due to their atypical topology. Data generation is challenging for these frameworks because of their behavior (phase transitions, adsorption . . .) and size. Therefore, we introduced a data generation protocol, consisting of three phases in section 3.4: initialization, equilibration and sampling. The first two phases explore and provide initial configurations along a certain collective variable using force field MD while in the third phase, short first-principles MD simulations are performed to train the MLP. All of this is made possible by the excellent data efficiency of equivariant MLPs, the subject of section 3.2.5. To validate our protocol, an MLP was derived for UiO-66(Zr) and MIL-53(Al) using only a training set with a few hundred structures. MAE errors around 20 meV/Å were obtained on trajectories generated by the MLPs themselves, which demonstrates their accuracy. Hence, in comparison to other state-of-the-art MLPs for MOFs, we have lowered the typical error values by more than a factor of three while significantly reducing the amount of training data. Furthermore, the transition pressure of MIL-53(Al), a property unattain-

able at the first-principles level of theory due the the computational cost, was predicted with the MLP and its value of 18-20 MPa is in perfect agreement with experiment.

In summary, the goal of this thesis was to improve two shortcomings of the current generation of MLPs: (i) the general description of long-range interactions and (ii) an efficient data generation protocol for MOFs. For both objectives, we have improved upon the state-of-the-art. First, the eMLP, an explicit-electron MLP, is an accurate and transferable model which naturally incorporates long-range interactions and is able to predict infrared spectra, dipole moments, piezoelectric constants and more. Second, our data generation protocol tested for UiO-66(Zr) and MIL-53(Al) and driven by equivariant MLPs, only requires a few hundred configurations to make highly accurate predictions. These accomplishments can provide a strong basis for future work, which may build on top of these frameworks.

5.2 Perspectives

In this work, we saw that the eMLP is already applicable to many small organic molecules and periodic systems. However, the current version of the eMLP remains limited to the elements H, C, N and O and systems with an even number of electrons. In future work, one could certainly extend the eMLP towards more general systems. Heavier elements, from the third row and beyond, can be easily included by lumping the lower bound electron pairs into the atomic cores. Again, only the valence electron pairs would be explicitly described. For systems with a small band gap, the electrons may be inherently delocalized since small external fields may displace electrons over large distances. Therefore, the BO equilibrium positions of the electron pairs could potentially be extremely sensitive to the exact configuration of the nuclei or external fields. Then, data augmentation might become more important to avoid the crossing of multiple shallow minima of the extended PES.

To properly describe atomic systems with an odd number of electrons, a variant with single electron particles should be developed. Consequently, each electron would be a spin up or spin down particle. The eMLP methodology should be modified at several places to account for this change. First, post-Hartree-Fock methods may be necessary since unrestricted DFT calculations may not reach an appropriate level of accuracy. The resulting increase in computational cost limits the amount of training data that one can generate but fortunately, the recent class of equivariant MLPs have proven to be highly data efficient. Moreover, one cannot derive the electron pair locations

from the centers of localized orbitals because they are only defined in Kohn-Sham DFT or Hartree-Fock. For that reason, one could use the centers of the geminals in a geminal-product wavefunction, where the most important electron correlation effects are already modeled. For other approaches, other localization techniques should be developed that only access the total electron density but they can be tuned and optimized with the specific purpose of the eMLP in mind, to increase its predictive power. Finally, the total eMLP energy should be made invariant with respect to relabeling the spin up particles as spin down particles and vice versa.

Other aspects of the eMLP may be improved as well. It is straightforward to replace the short-range machine learning part by more data efficient and accurate MLPs. Equivariant MLPs such as NequIP are the ideal candidate for this part. The long-range interactions on the other hand, may be modified as well by adding an explicit dispersion contribution or by accounting for charges with a variable spatial extent. Moreover, data augmentation might be more effective with more suitable surrogate models to compute the augmented energy and forces. Constrained DFT (CDFT) could also be a useful tool to generate local perturbations in the electron density to produce more relevant training data.

Finally, one should show that the eMLP is able to model more complex phenomena. For instance, one could use the eMLP in a continuum solvation model (e.g. COSMO) to study the interaction between solvents and solutes. Moreover, it still needs to be shown that the eMLP improves upon conventional polarizable force fields in application where explicit electrons have an added value. This includes ionization, (redox) reactions, long-range charge transport and more. Generating accurate training data in those circumstances will be the most essential and potentially most difficult step towards accomplishing these goals.

The modeling of MOFs does also have its own future prospects. MLPs are revolutionizing the field and allow us to study physical phenomena like defects or chemical reactions within MOFs at the accuracy of first-principles methods. This will be enabled by active learning, ideally incorporated within our proposed data generation protocol. Active learning will significantly reduce the number of required first-principles calculations since sampling with first-principles MD trajectories can be avoided. Furthermore, it will provide a pathway towards building a universal MLP for MOFs.

Part II

Published Paper(s)

A

Publications in International Peer–Reviewed Journals

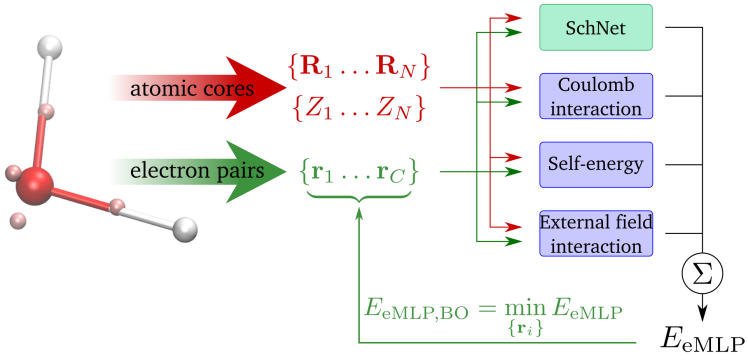
Paper I **Modeling Electronic Response Properties with an Explicit-Electron Machine Learning Potential**

M. Cools-Ceuppens, J. Dambre, and T. Verstraelen

Journal of Chemical Theory and Computation, **2022**, *18* (3), 1672–1691

Paper I

Modeling Electronic Response Properties with an Explicit-Electron Machine Learning Potential



M. Cools-Ceuppens, J. Dambre, and T. Verstraelen

Journal of Chemical Theory and Computation, **2022**, *18* (3), 1672–1691

M. Cools-Ceuppens performed the research and wrote the manuscript.

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

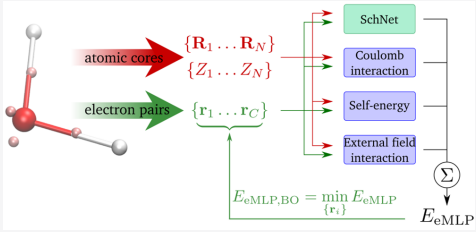
Modeling Electronic Response Properties with an Explicit-Electron Machine Learning Potential

Maarten Cools-Ceuppens, Joni Dambre, and Toon Verstraelen*

Cite This: J. Chem. Theory Comput. 2022, 18, 1672–1691 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Explicit-electron force fields introduce electrons or electron pairs as semiclassical particles in force fields or empirical potentials, which are suitable for molecular dynamics simulations. Even though semiclassical electrons are a drastic simplification compared to a quantum-mechanical electronic wave function, they still retain a relatively detailed electronic model compared to conventional polarizable and reactive force fields. The ability of explicit-electron models to describe chemical reactions and electronic response properties has already been demonstrated, yet the description of short-range interactions for a broad range of chemical systems remains challenging. In this work, we present the electron machine learning potential (eMLP), a new explicit electron force field in which the short-range interactions are modeled with machine learning. The electron pair particles will be located at well-defined positions, derived from localized molecular orbitals or Wannier centers, naturally imposing the correct dielectric and piezoelectric behavior of the system. The eMLP is benchmarked on two newly constructed data sets: eQM7, an extension of the QM7 data set for small molecules, and a data set for the crystalline β-glycine. It is shown that the eMLP can predict dipole moments, polarizabilities, and IR-spectra of unseen molecules with high precision. Furthermore, a variety of response properties, for example, stiffness or piezoelectric constants, can be accurately reproduced.



1. INTRODUCTION

A central problem in computational chemistry is finding approximate, yet sufficiently accurate, solutions for the quantum-mechanical electronic structure problem, given a configuration of nuclei in a molecule or a condensed system. Many microscopic properties of such a system can be derived once the wave function is solved, such as the potential energy surface, a molecular dipole moment, etc. Usually, the calculation of the electronic wave function is merely a required intermediate step toward those properties of interest. For this reason, many force field models have been developed, which compute properties of interest directly, bypassing the electronic wave function.[1,2] Their main advantage is a drastic reduction in computational cost, bringing much larger atomistic systems and their dynamics at longer time scales in reach of computer simulations. Still, (approximate) electronic structure calculations are widely used despite their higher computational cost. In general, they predict properties more accurately for a broad range of systems and they rely less on empirically adjusted model parameters. In addition, the electronic wave function and its response to external stimuli gives access to many properties of interest. In force fields, such electronic properties are not trivially available, due to the absence of a detailed model of electronic structure.

Despite the lack of an electronic wave function, force-field models can incorporate electronic features to some degree, such that a subset of the electronic properties can be derived. The most common feature is a fixed partial charge for each atom, which is mainly used to describe electrostatic properties and the corresponding long-range interactions.[3,4] Atomic partial charges are essentially a coarse-grained description of a frozen electron density. Polarizable force fields go beyond static charge distributions by also modeling the change in electron density due to an applied external field.[5,6] This is typically accomplished by introducing variables in the atomic multipole expansions, for example, induced dipoles, which are solved by an energy minimization. Most polarizable force fields rely on a linear-response approximation, optionally with nonlinear corrections,[7] limiting their applicability to small electronic rearrangements. A less appreciated limitation of most polarizable force fields is that they only attempt to

Received: September 28, 2021
Published: February 16, 2022

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672–1691

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

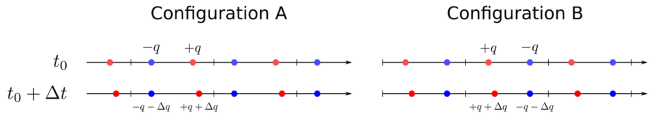


Figure 1. A one-dimensional system with lattice length a, containing two charged particles. At t_0, the particles have a charge of $-q$ and $+q$ and are separated by $a/2$. At $t_0 + \Delta t$, the positive particle has moved a distance Δr and gained a charge Δq while the negative particle stays fixed but loses Δq in charge. Both configurations describe the same system but the origin of unit cell is translated over a distance $a/2$.

approximate changes in electron density. This is problematic because a change in electron density alone cannot describe the change in macroscopic polarization of a periodic system.[8] Simply put, under periodic boundary conditions, a change in electron density is insufficient to derive from where to where electrons have moved, hampering a sound definition of macroscopic polarization. Unless additional assumptions are made, polarizable force fields inherit this ambiguity. Fortunately, most polarizable models avoid this difficulty by clearly specifying from where to where charge is displaced. For example, in a Drude oscillator[9] or induced dipole[10,11] model, polarization is local within each atom. However, for models with fluctuating atomic charges, e.g., EEM,[12] QEq,[7,13] or machine learned models for partial charges,[14,15] the polarization of periodic models remains ambiguous, which will be shown in the next paragraph. A completely different approach to incorporate electronic degrees of freedom in force fields, are the so-called semiclassical or explicit electron force fields.[16] These models introduce local electrons or electron pairs as negatively charged particles, in addition to the nuclei or positively charged atomic cores. Because all particles have a fixed charge, just as in quantum-mechanical electronic structure methods, they have no issues defining the polarization of a periodic model. Furthermore, by allowing electron (pair) particles to migrate away from their original bound configuration, processes such as ionization, redox reactions, and charge transport, can in principle be described.[17−21] Because these processes depart manifestly from a linear response regime, they remain challenging for conventional polarizable force fields. This work is motivated by the appealing prospects of explicit electron models and explores new approaches to develop such models.

Models with geometry-dependent atomic charges have been developed for several decades and are widely employed for molecular simulations.[7,12−15,22−27] To the best of our knowledge, it was never reported previously that such models are inconsistent with the modern theory of polarization and are therefore problematic for modeling dielectric systems. We will clarify this issue with a simple one-dimensional example. Note that the issue shown here is more general than the metallic polarizability scaling of charge equilibration models,[23,24] because we do not assume partial charges are found by charge equilibration. It is well-known in the modern theory of polarization that the dipole vector itself is not well-defined for a periodic system and that only the change in polarization has a physical meaning and can be measured experimentally. This was illustrated by Spaldin[28] for a one-dimensional lattice with fixed charges. Here, we take the same example and allow charges to fluctuate when they are displaced, to demonstrate that this results in an ill-defined change in polarization. Consider a one-dimensional lattice with lattice length a, as depicted in Figure 1. Two particles are present with opposite charges $+q$ and $-q$, separated by a distance $a/2$ at the initial time t_0. At a later time $t_0 + \Delta t$, the positively charged particles moves a distance Δr to the right while gaining an additional charge Δq. In configuration A, the unit cell is chosen such that the negative particle starts at $a/4$ and the positive particle at $3a/4$. Exactly the opposite is true in configuration B but they both describe the same system when periodic boundary conditions are taken into account. In configuration A, the change in dipole per lattice length is

$$\begin{aligned}\Delta p &= \frac{1}{a}\sum_i q_i(t_0+\Delta t)r_i(t_0+\Delta t) - \frac{1}{a}\sum_i q_i(t_0)r_i(t_0)\\ &= (q+\Delta q)\left(\frac{3}{4}+\frac{\Delta r}{a}\right) + (-q-\Delta q)\frac{1}{4} - q\frac{3}{4} + q\frac{1}{4}\\ &= q\frac{\Delta r}{a} + \left(\frac{\Delta r}{a}+\frac{1}{2}\right)\Delta q\end{aligned} \tag{1}$$

while for configuration B, the same equation yields

$$\begin{aligned}\Delta p &= (q+\Delta q)\left(\frac{1}{4}+\frac{\Delta r}{a}\right) + (-q-\Delta q)\frac{3}{4} - q\frac{1}{4} + q\frac{3}{4}\\ &= q\frac{\Delta r}{a} + \left(\frac{\Delta r}{a}-\frac{1}{2}\right)\Delta q\end{aligned} \tag{2}$$

Hence, the change in polarization depends on the chosen definition of the unit cell, which is a major concern since this quantity can be measured experimentally. Note that the problematic term $\left(\frac{\Delta r}{a} \pm \frac{1}{2}\right)\Delta q$ is related to the charge transfer. Within the unit cell of configuration A, a charge of Δq is transferred to the right over half the lattice length, while in configuration B, the charge moves to the left. The change in polarization only becomes well-defined if there are no fluctuating charges $\Delta q = 0$ or when one unambiguously defines from where to where charge is transferred, such as in the split-charge equilibration model.[22]

A variety of explicit electron force fields have been developed. The electron force field[29,30] (eFF) and its successor eFF-ECP[31,32] (eFF with effective core potentials) are mainly established to model materials under extreme conditions, whether they are high pressures or high temperatures. The LEWIS[17,18] force field is capable of simulating liquid water and its dielectric response. It has been extended to LEWIS•[19−21] to incorporate $2p$ and $3p$ elements and diatomic molecules and shows promising results in predicting electron affinities and ionization potentials. More recently, explicit electron extensions to the reactive force field ReaxFF have been developed. The inclusion of electrons or holes as additional particles is realized in eReaxFF,[33] whereas in ReaxFF/C-GeM[34,35] (ReaxFF and the coarse-grained electron model), each atom is characterized by a positive core and negative shell (not necessarily at the same location) and modeled as interacting Gaussian charges. All these methods share the same viewpoint of the electron as a pseudoclassical particle but differ in a

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

variety of aspects. One can model all electrons (both core and valence), or only model the valence electrons. In the latter scenario, the core electrons are simply frozen at the positions of the nuclei and not treated explicitly. Next, one can choose to separate the spin up and spin down electrons as different particles or group a spin-up and spin-down electron as an electron pair, represented by a single particle with charge $-2e$, which reduces the amount of particles, lowers the model complexity, and improves computational efficiency. Also the parameters in explicit electron force fields have been estimated in several different ways. Unless the models are trained directly to terms involving the positions of the electrons, like the dipole vector, the electrons will be located at positions which are a consequence of the fit and do not have a direct relation to the electronic wave function. One of the innovations in this work is that the electron pair particles will be positioned at centers of localized molecular orbitals.[36] These centers have a reasonable similarity to Lewis structures and provide ample microscopic data for training. Furthermore, by placing $-1e$ charges at centers of localized orbitals, one reproduces the molecular dipole moment exactly, guaranteeing proper long-range electrostatics.

A major difficulty in the process of developing explicit electron force fields is the characterization of the short-range interactions. At long distances, the classical Coulomb electrostatics are an adequate approximation for the interaction between charged particles but at short distances quantum effects may dominate. Electrons are fermions for which the Pauli exclusion principle is valid, resulting in exchange interactions, where its effect on explicit electron force fields have been already investigated in detail.[37] In this work, however, we do not attempt to derive the short-range theoretically or via heuristic approximations. Instead, they will be modeled with machine learning to avoid any assumptions on their functional form. In general, machine learning force fields[38−40] try to learn the relation between the geometry and chemical species in the system and the total energy and its derivatives. No physical insight is required, only a vast amount of first-principles data is necessary to fit the potential energy surface (PES). Most of the machine learning models can be subdivided in a few classes: neural networks[41−43] or message-passing[44] (deep) neural networks (MPNN),[45−50] and kernel-based methods.[51−55] In essence, a local representation of every atom in a multidimensional vector space is defined as a feature (for the kernel-based methods) or learned (for the message passing neural networks). This representation encodes the chemical environment around every atom in a certain cutoff radius and serves as the input to predict the atomic energies. We will utilize the SchNet[45,56] deep neural network to model the short-range interactions in an explicit electron force field. It is a well-proven and benchmarked[57,58] architecture, which performs equally well on relevant benchmarks compared to other state-of-the-art machine learning force fields, given that enough data are available. Furthermore, the prohibitive scaling of the number of features per atom, with respect to total amount of chemical elements in the system, is avoided because MPNNs learn their own representation.

Our new model is not only a refinement upon existing explicit electron force fields but is also an innovation in the treatment of electrostatic and polarization interactions in machine learning force fields. Typically, long-range interactions are modeled by learning partial charges for every atom based on the local representation, after which the charges then interact with classical electrostatics. These are nonpolarizable machine learning force fields because the charges only depend on the local environment and are insensitive to electric fields from more distant charge distributions. This is addressed in so-called fourth generation neural network potentials.[40] Another concern is that charges directly predicted by neural networks have an incorrect total charge and must be shifted *ad hoc*.[48] In the context of neural networks, only recently some advancements have been made to address these charge issues. Models based on the charge equilibration neural network technique (CENT)[15,25,26] do predict atomic electronegativities, which are employed to optimize the charges by minimizing the electrostatic energy. In the BpopNN model,[27] electronic populations are introduced, which serve as extra input variables in the neural network, where the optimal values are again calculated by minimizing the energy. In AIMNet-NSE,[59] an arbitrary molecular charge or spin can be imposed, while the electron passing neural network (EPNN)[14] iteratively updates partial charges in its message passing network, while constraining the total charge. Total charge and spin are constrained and serve as extra input in SpookyNet,[60] which also includes nonlocal interactions by using self-attention and analytic long-range corrections. The influence of electric fields are taken into account in FieldSchNet[61] by learning vectorial representations per atom and coupling it with the external field by taking scalar products. For kernel-based methods, progress has been made by incorporating the long-distance equivariant (LODE)[62] representation to describe long-range effects. The majority of these efforts to incorporate electronic polarization in machine-learned potentials rely on environment-dependent fractional charges. As shown above, this leads to nontrivial difficulties when describing the polarization of periodic systems. This motivated us to explore explicit electrons as an alternative approximate representation of the electronic structure.

In this work, we will present a new explicit electron force field, which we will call the electron machine learning potential (eMLP), where the short-range interactions are learned via machine learning. For now, only electron pairs will be considered by grouping up the spin-up and spin-down electrons in a single effective particle. Furthermore, only valence electron pairs are considered as a starting point. This simplifies the methodology and training of the neural network. In addition, this also improves the computational efficiency when making predictions. The electron pairs will be located at well-defined positions, derived from localized molecular orbitals. This will naturally impose the correct dipole moments, polarization, or piezoelectric behavior of the system. We will study the ability of the model to predict polarizabilities and IR spectra of unseen small molecules. For periodic systems, we will focus on β-glycine as a case study, since piezoelectricity in biomolecules has gained a lot of attention in research[63] in recent years due to the possibly large piezoelectric strain responses. It is shown that the eMLP will make it possible to accurately reproduce stiffness and dielectric and piezoelectric constants for β-glycine. Finally, to enable stable molecular dynamics (MD) simulations, data-augmentation will be introduced. During the training phase, additional out-of-equilibrium electron pair positions will be generated since they are poorly sampled by conventional techniques.

Section 2 discusses the mathematical structure of short- and long-range interactions in eMLP and the computational details

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

of the localization procedure. Next, the databases will be introduced. Finally, we will explain how the model is trained, with or without data augmentation. In section 3, the results will be discussed for eQM7, a data set of small molecules, and crystalline β-glycine, after which the main conclusions and outlook will be presented in section 4.

2. METHODOLOGY

In this section, the methodology followed in this work will be explained in detail. We will start by describing the particles in the eMLP, its overall structure, and the energy decomposition. In the next two consecutive subsections, the long-range and short-range contributions will be defined. Subsequently, it is described how the electronic positions are extracted from DFT calculations by making use of electron localization, both in periodic and nonperiodic systems. Next, we introduce two new data sets, created for this work: eQM7, a data set for the purpose of modeling a variety of small molecules and a data set for β-glycine. In the section thereafter, data-augmentation will be discussed to overcome sampling deficiencies with respect to the electronic degrees of freedom. Finally, the cost function to train the eMLP will be presented.

2.1. Description of the Model. *2.1.1. Overall Structure.* The eMLP treats the electrons as semiclassical particles in addition to the nuclei. In this work, only systems with an even number of electrons will be considered and all pairs of spin-up and -down electrons are grouped into a single pair particle with charge −2 e, similarly to restricted closed-shell Hartree−Fock (RHF). The eMLP in this work is developed for systems comprising elements H, C, N, and O. We assume that the 1s core electron pair in C, N, or O is located exactly on top of their corresponding nucleus, such that only the valence electron pairs need to be explicitly described. This will be validated in section 2.2. From this point onward, the term *electron pair* will be reserved for valence electron pairs and the term *atomic core* will be used to describe the ensemble of the nucleus and where applicable the core electrons. The terminology is summarized in Table 1.

Table 1. An Overview of the Particles Appearing in the eMLP and Their Symbols and Charges

	atomic core	electron pair
symbol of position	$\boldsymbol{R}_i$	$\boldsymbol{r}_i$
charge [e]	+1 for $Z_i = 1$ $(Z_i - 2)$ for $Z_i \in \{6, 7, 8\}$	−2

As an example, consider a single water molecule (a 10-electron system) where the particles and their charges are visualized in Figure 2. Four electron pairs are explicitly described by the eMLP: two electron pairs participate in the bond between the oxygen and hydrogen atomic cores and another two give rise to the lone pairs of the oxygen atom. The final two electrons are core electrons which are not treated explicitly but are incorporated in the atomic core of oxygen, giving it a charge of $+6e$.

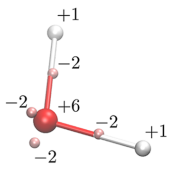


Figure 2. A single water molecule with 10 electrons, giving rise to four explicit electron pairs. The charge (in e) of each particle is indicated.

The eMLP models an extended potential energy surface (PES) including electronic degrees of freedom. Besides the positions of the atomic cores $\boldsymbol{R}_i$ and their atomic numbers Z_i, electron pair positions $\boldsymbol{r}_i$ serve as inputs for the potential energy of the system: $E_{\mathrm{eMLP}} = E_{\mathrm{eMLP}}(\{\boldsymbol{R}_i\},\{Z_i\};\{\boldsymbol{r}_i\})$.

The extended PES is modeled by combining machine learning and classical long-range electrostatics.

$$E_{\mathrm{eMLP}} = E_{\text{long-range}} + E_{\text{short-range}} \tag{3}$$

The complex short-range interactions are modeled with SchNet,[45] a deep neural network with over one million trainable parameters. The long-range interactions consist of the Coulomb repulsion or attraction between Gaussian charge densities centered at each particle (atomic cores and electron pairs), together with a self-energy term and the interaction with an external field. A schematic overview of the eMLP and its energy contributions is given in Figure 3. The energy partitioning has several advantages. First of all, machine learning force fields require no physical insight. This is helpful since the short-range interactions between the electron pairs and the atomic cores are nontrivial. Furthermore, making the long-range interactions parameter-free will help to overcome overfitting issues. The forces of the long-range interactions are after all harder to learn since they are generally smaller than those of the short-range contributions. Additionally, this removes the need for representative data points for the long-range interactions in the data set as the total amount of possible chemical environments drastically increases at longer length scales. The partitioning ultimately increases the transferability of the eMLP. In the next two subsections, both interaction types are described in more detail.

This extended PES already enables us to optimize geometries (both atomic cores and electron pairs) and to predict the dipole moments, polarizabilities, infrared (IR) spectra and more. However, molecular dynamics (MD) simulations require some extra considerations for the dynamics of the electron pairs. We follow a similar reasoning as in the Born−Oppenheimer approach: the electron pairs are significantly lighter than the atomic cores, such that in each step of the MD simulation the electronic positions are being relaxed,

$$E_{\mathrm{eMLP,BO}}(\{\boldsymbol{R}_i\},\{Z_i\}) = \min_{\{\boldsymbol{r}_i\}} E_{\mathrm{eMLP}}(\{\boldsymbol{R}_i\},\{Z_i\},\{\boldsymbol{r}_i\}) \tag{4}$$

resulting in a conventional PES, suitable for MD simulations. This closely resembles an SCF optimization for DFT calculations. The positions of the electron pairs for which the energy is minimized $\boldsymbol{r}_i^{\mathrm{eq}}$ can be considered as a side-product in each MD step. In a system with C electron pairs, the forces on the atomic cores are simply given by

$$\begin{aligned}\boldsymbol{F}_{\mathrm{BO},j} &= -\frac{\partial E_{\mathrm{eMLP,BO}}(\{\boldsymbol{R}_i\},\{Z_i\}}{\partial \boldsymbol{R}_j} \\ &= -\frac{\partial E_{\mathrm{eMLP}}(\{\boldsymbol{R}_i\},\{Z_i\},\{\boldsymbol{r}_i^{\mathrm{eq}}\})}{\partial \boldsymbol{R}_j} \\ &\quad - \sum_{k=1}^{C} \frac{\partial E_{\mathrm{eMLP}}(\{\boldsymbol{R}_i\},\{Z_i\},\{\boldsymbol{r}_i^{\mathrm{eq}}\})}{\partial \boldsymbol{r}_k^{\mathrm{eq}}} \frac{\partial \boldsymbol{r}_k^{\mathrm{eq}}}{\partial \boldsymbol{R}_j}\end{aligned} \tag{5}$$

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

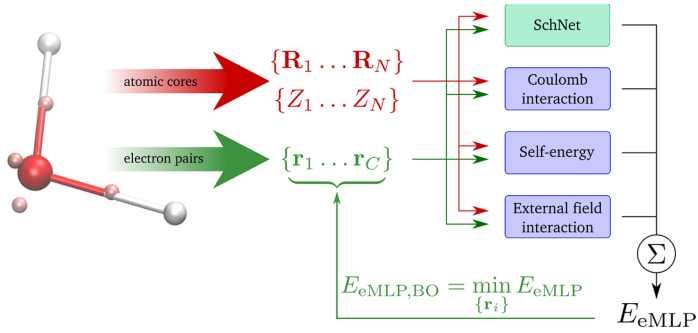


Figure 3. A schematic overview of the eMLP. Besides positions and species of the atomic cores, positions of the electron pairs are needed as inputs for the eMLP. The four building blocks of the eMLP are the short-range machine learning contribution (SchNet) and the three long-range contributions (the Coulomb interaction, self-energy, and external field interaction). In MD simulations, the Born–Oppenheimer eMLP energy is needed, for which the energy is minimized with respect to the electron pair positions.

of which the second term is zero since the forces on the electron pairs,

$$\boldsymbol{f}_j = -\frac{\partial E_{\mathrm{eMLP}}(\{\boldsymbol{R}_i\}, \{Z_i\}, \{\boldsymbol{r}_i^{\mathrm{eq}}\})}{\partial \boldsymbol{r}_j} \tag{6}$$

are zero for the equilibrium positions.

2.1.2. Long-Range Interactions. The long-range energy in eMLP consists of three contributions:

$$E_{\text{long-range}} = E_{\mathrm{Coulomb}} + E_{\mathrm{self}} + E_{\mathrm{ext}} \tag{7}$$

which will be explained in more detail below.

All the particles are modeled as Gaussian charge densities and they interact with each other through the Coulomb interaction. This includes electron–electron, core–electron, and core–core interaction:

$$\begin{aligned} E_{\mathrm{Coulomb}} = {} & \frac{1}{2} \sum_{i \in \mathcal{E}} \sum_{j \in \mathcal{E} \setminus \{i\}} q_i q_j \frac{\mathrm{erf}(\gamma \|\mathbf{r_i} - \mathbf{r_j}\|)}{\|\mathbf{r_i} - \mathbf{r_j}\|} \\ & + \sum_{i \in \mathcal{E}} \sum_{j \in C} q_i q_j \frac{\mathrm{erf}(\gamma \|\mathbf{R_i} - \mathbf{r_j}\|)}{\|\mathbf{R_i} - \mathbf{r_j}\|} \\ & + \frac{1}{2} \sum_{i \in C} \sum_{j \in C \setminus \{i\}} q_i q_j \frac{\mathrm{erf}(\gamma \|\mathbf{R_i} - \mathbf{R_j}\|)}{\|\mathbf{R_i} - \mathbf{R_j}\|} \end{aligned} \tag{8}$$

where $\mathcal{E}$ and C are the sets of electron pair and atomic core particles, respectively. The charges q_i of each particle are listed in Table 1. The parameter γ appearing in the error function is inversely proportional to the width of the Gaussian charges and controls the distance for which the electrostatic potential is damped compared to point-charge interactions. Note that γ is not intended to mimic the spatial extent of localized orbitals. The Coulomb energy is solely designed to model long-range classical electrostatics, with a smooth continuation of the long-range interactions within the cutoff sphere. In this work, the eMLP will use $\gamma = 0.3928/\text{Å}$, which is chosen such that the energy between particles greater than the cutoff radius $r_{\mathrm{cutoff}} = 4$ Å is approximately that of two point charges, while huge forces inside the cutoff radius are damped, as will be shown in more detail below. Because the Coulomb term is only intended to be smooth, as opposed to accurate, within the cutoff sphere, the short-range machine learning contributions will also include some Coulomb corrections as well as other non-classical electronic interactions.

The long-range energy also includes a self-energy term for each particle:

$$E_{\mathrm{self}} = \frac{\gamma}{\sqrt{\pi}} \sum_i q_i^2 \tag{9}$$

the main purpose of which is to normalize the magnitude of the total electrostatic energy:

$$E_{\mathrm{Coulomb}} + E_{\mathrm{self}} \approx 0 \tag{10}$$

When charged particles are sufficiently far apart, E_{Coulomb} would go to zero already. However, in all data sets in this work, charges are relatively close and E_{Coulomb} is large in magnitude. In the limit of overlapping charge distributions, making use of $\lim_{x \to 0} \mathrm{erf}(x)/x = 2/\sqrt{\pi}$, the sum of the Coulomb and self-energy becomes zero for neutral systems:

$$E_{\mathrm{Coulomb}} + E_{\mathrm{self}} \approx \frac{\gamma}{\sqrt{\pi}} \sum_{i \neq j} q_i q_j + \frac{\gamma}{\sqrt{\pi}} \sum_i q_i^2 = \frac{\gamma}{\sqrt{\pi}} \left(\sum_i q_i \right)^2 \tag{11}$$

For all systems considered in this work, the sum of these two terms is much closer to zero than the Coulomb energy alone.

By including the self-energy, not only are the long-range forces small in magnitude, but also the long-range energies. If they were not, the short-range interaction would need to cancel out these large forces and energies. For instance, a lone electron pair sitting around an oxygen atom at a distance of 0.3 Å will feel attractive forces up to 2000 eV/Å if their interaction was modeled via point charges. On the other hand, the average magnitude of the forces of particles in an regular MD simulation at 600 to 800 K are about 2 eV/Å. In this situation, our machine learning part should try to unlearn those large contributions which is an almost insurmountable challenge. Hence, the introduction of Gaussian charges and a self-energy term bypasses this issue by normalizing the force and energy targets and greatly simplifies the training process. The machine

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672–1691

learning short-range interaction is now solely responsible for learning all the subtle nonclassical details in the PES without having to compensate large systematic errors from the long-range energy.

Finally, a third term is optionally added for the interaction with a homogeneous external electric field $\boldsymbol{\mathcal{E}}_{\text{ext}}$:

$$E_{\text{ext}} = -\sum_i Z_i \boldsymbol{\mathcal{E}}_{\text{ext}} \cdot \boldsymbol{R}_i + 2\sum_i \boldsymbol{\mathcal{E}}_{\text{ext}} \cdot \boldsymbol{r}_i \tag{12}$$

or in the case of a more general external potential V_{ext}:

$$E_{\text{ext}} = \sum_i Z_i V_{\text{ext}}(\boldsymbol{R}_i) - 2\sum_i V_{\text{ext}}(\boldsymbol{r}_i) \tag{13}$$

Only classical electrostatics are explicitly present in the long-range interactions in the eMLP. In principle, dispersion interactions may still be described in such a framework, through the response of the electronic degrees of freedom to the presence of other molecules.[64] However, dispersion interactions are not yet the focus of this work and we will only employ the long-range term for modeling classical electrostatic interactions effectively.

2.1.3. Short-Range Interaction. The short-range interactions are modeled using a message passing neural network (MPNN).[44] More specifically, we make use of the SchNet architecture.[45] In this section, we briefly elucidate its mathematical structure. For a more detailed introduction or schematic overview of the architecture, we refer to the original SchNet paper.[45]

To describe the full architecture, some reoccurring operations are introduced. A single dense neural network layer is given by

$$D^{N\times M}(\boldsymbol{h}) = \mathbf{W}\cdot\boldsymbol{h} + \boldsymbol{b} \tag{14}$$

in which $\boldsymbol{W} \in \mathbb{R}^{N\times M}$ and $\boldsymbol{b} \in \mathbb{R}^N$ are respectively the weight matrix and bias vector of the dense layer, acting on a general vector $\boldsymbol{h} \in \mathbb{R}^M$. Additionally, an activation function can be added, which we will denote by a tilde: $\tilde{D}^{N\times M}(\boldsymbol{h}) = f(D^{N\times M}(\boldsymbol{h}))$ in which f is the softplus activation function

$$f(\boldsymbol{h}) = \log(1 + e^{\boldsymbol{h}}) - \log(2) \tag{15}$$

Next, the cutoff function is continuous and has a continuous derivative:

$$f_{\text{cutoff}}(r) = \begin{cases} 1, & r \le r_{\text{cutoff}} - r_\Delta \\ \frac{1}{2}\left[1 + \cos\left(\pi \frac{r - (r_{\text{cutoff}} - r_\Delta)}{r_\Delta}\right)\right], & r_{\text{cutoff}} - r_\Delta < r \le r_{\text{cutoff}} \\ 0, & r_{\text{cutoff}} < r \end{cases} \tag{16}$$

where the parameter r_Δ introduces a smooth transient zone.

MPNNs encode information about every particle i as a feature vector $\boldsymbol{h}_i^t \in \mathbb{R}^F$ and transfer it to all its neighbors within the cutoff radius r_{cutoff} in an iterative way. A look-up table or embedding $\mathbf{a}_{S_i}$, which is a trainable vector, serves as the starting point for the feature vector:

$$\boldsymbol{h}_i^0 = \boldsymbol{a}_{S_i} \tag{17}$$

Depending on the species S of the particle i (electron pair or atomic number of the atomic core), another initial vector is employed. This is the only place in the short-range contribution where a distinction between species is made.

Next, messages m_i^t, based on the geometry of neighboring particles j inside the cutoff radius, are generated in each iteration t

$$m_i^t = \frac{1}{J}\sum_{\substack{i\neq j \\ r_{ij}<r_{\text{cutoff}}}} M(\boldsymbol{h}_j^t, r_{ij}) \tag{18}$$

The message function M is implemented in SchNet[45] by taking the element-wise product of the feature vector in the filter space $\mathbb{R}^G$, the so-called filter-generating network $W_{\text{filter}}(r_{ij})$, and the cutoff function:

$$M(\boldsymbol{h}_j^t, r_{ij}) = D^{G\times F}(\boldsymbol{h}_j^t) W_{\text{filter}}(r_{ij}) f_{\text{cutoff}}(r_{ij}) \tag{19}$$

The constant J is introduced to normalize the sum over all neighbors, speeding up the training of the neural network. Ideally, the constant should take the value of the average amount of neighbors of a particle. In this work, we pick $J = 70$ which corresponds to the amount of neighbors in condensed systems. The filter-generating network is a simple two-layered dense neural network,

$$W_{\text{filter}}(r_{ij}) = [\tilde{D}^{G\times G} \circ \tilde{D}^{G\times N}](\boldsymbol{\varphi}(r_{ij}))$$

mapping N radial basis functions,

$$\varphi_n = \exp\left(-\frac{(r_{ij} - \mu_n)^2}{2\sigma^2}\right) \tag{21}$$

to the filter-space. The centers μ_n are uniformly spaced over the interval $[0, r_{\text{cutoff}}]$ and $\sigma = r_{\text{cutoff}}/N$. When all messages are calculated, the feature vectors are updated,

$$\boldsymbol{h}_i^{t+1} = \boldsymbol{h}_i^t + [D^{F\times G} \circ \tilde{D}^{G\times G}](\boldsymbol{m}_i^t) \tag{22}$$

and the whole process is repeated T times. Finally, after T iterations, each feature vector is sent through the output network, yielding the particle energies:

$$\epsilon_i = [D^{1\times\lfloor\frac{F}{2}\rfloor} \circ \tilde{D}^{\lfloor\frac{F}{2}\rfloor\times F} \circ \tilde{D}^{F\times F}](\boldsymbol{h}_i^T) \tag{23}$$

while the total short-range contribution of the energy is the sum of all the particle energies

$$E_{\text{short-range}} = \sum_i \epsilon_i \tag{24}$$

Note that the energy is translationally and rotationally invariant by construction as only the interparticle distances enter the equations.

2.2. Electron Localization. The positioning of the electron pairs is extremely important to reproduce quantities such as the dipole moment, polarizabilities, and so on. In this work, centers of localized restricted Kohn–Sham orbitals will be used as reference data for the electron pair positions, because they are well-defined, they exactly reproduce molecular dipole moments, they offer an intuitive picture of chemical features, and they provide extensive training data. The canonical Kohn–Sham orbitals, found by solving the self-consistent equations (SCF) of Kohn–Sham density functional theory (KS-DFT) equations,[36] are generally delocalized and have well-defined energy levels. However, any observable is invariant under unitary transformations of the occupied orbitals:

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

$$|\psi_i\rangle = \sum_j U_{ij}^{\dagger} |\phi_j\rangle \tag{25}$$

where U is a unitary matrix, $|\phi_j\rangle$ is the occupied canonical orbitals, and $|\psi_i\rangle$ is the new set of occupied orbitals. Any unitary transformation U produces an equally valid set of orbitals. This freedom can be exploited to construct maximally local occupied orbitals. One of the most popular electron localization schemes was developed by Foster and Boys (FB).[65] In this method, the unitary matrix U is chosen such that the new orbitals have a minimal spatial extent, which is equivalent to minimizing the spread of all the occupied orbitals:

$$C(U) = \sum_i \langle \psi_i | [\boldsymbol{r} - \langle \psi_i | \boldsymbol{r} | \psi_i \rangle]^2 | \psi_i \rangle \tag{26}$$

Because the orbitals are localized, corresponding centers are well-defined and easily computed as the expectation value of the position operator:

$$\boldsymbol{r}_i = \langle \psi_i | \boldsymbol{r} | \psi_i \rangle \tag{27}$$

These positions will be used as training data for the electron pair particles in eMLP.

The centers of 1s core orbitals in second-row elements fall almost exactly on top of the corresponding nuclei, that is, typically closer than 5×10^{-4} Å. This validates our choice to combine the core electrons and the nuclei into atomic cores as a single particle.

The FB method is ideally suited for use within the eMLP because it results in the most compact localized orbitals. Hence, all complex quantum effects (exchange, correlation, ...) are as short-ranged as possible, allowing a small cutoff radius for the machine learning part. Furthermore, they correspond in most situations with chemical intuition, as they agree with the Lewis structure of bonds and lone pairs in molecular systems. For instance, electron pairs can be classified as lone pairs or electrons participating in single, double, or triple bonds between atoms. The electron pairs of Figure 2 are generated using this localization procedure. Note that the dipole moment of a molecule calculated with point charges located at the centers, exactly reproduces the DFT dipole moment. There might exist discrepancies between higher order multipoles however.

The FB localized orbitals correspond to the ground state of the molecule. The electron pairs are situated in their equilibrium position, meaning that forces on the electron pairs are zero $\boldsymbol{f}_i = 0$. Training the eMLP to reproduce these vanishing forces will eventually lead to a model for which the equilibrium positions $\mathbf{r}_i^{\mathrm{eq}}$ correspond with the FB positions.

For periodic systems, maximally localized Wannier functions[66] (MLWFs) replace the FB centers. MLWFs can be constructed from Bloch orbitals, for which observables are also invariant under unitary transformations.[67] Again, the unitary transformation which minimizes the resulting spread of the orbitals is chosen, similar to the FB cost functions of eq 26. The resulting Wannier centers are used as reference locations of the electron pairs in periodic structures.

2.3. Data sets. *2.3.1. eQM7.* A new data set, which we will call electron QM7 (eQM7), is created with the purpose of training and validating polarizable force fields on non-equilibrium configurations of small molecules together with external field perturbations. The QM7 data set,[68] a subset of the more comprehensive GDB-13 database,[69] serves as the source of the molecules, from which 6868 out of the 7165 molecules are utilized. The remaining 297 molecules contain sulfur, an element of the third row, being beyond the scope of this work. Hence, the only elements appearing in the data set are hydrogen, carbon, nitrogen, and oxygen. For each molecule, 500 perturbations are constructed, described in detail below, resulting in 3 434 000 different configurations in total. Properties of these configurations are computed with Kohn–Sham density functional theory (DFT), using the PBE0 functional,[70,71] in line with the original QM7 data set, and aug-cc-pVTZ basis set[72,73] in the quantum chemistry program Psi4.[74] The FB localization is also performed using Psi4. After each DFT calculation, the following properties and arrays are stored: the total energy of the system, the positions $\{\boldsymbol{R}_i\}$ and atomic numbers Z_i of the nuclei, the FB centers $\{\boldsymbol{r}_i\}$ (core electrons included), the forces on the nuclei, the forces on the FB centers (which are zero by construction), and the electric field vector.

A total of 500 nonequilibrium configurations are generated by combining three different sampling techniques: normal mode sampling (NMS), torsion sampling, and dimer sampling, which have been already successfully applied in the literature.[75] Unlike MD sampling, these techniques yield independent and uncorrelated structures.

In normal mode sampling, the atoms are displaced along the normal modes of the molecule following the procedure by Smith et al.[76] In summary, the Hessian of each molecule is calculated, yielding the normal modes, and afterward the atoms are displaced along a few randomly selected modes at a temperature T according to the Boltzmann distribution. The sampling is performed at high temperatures of 600 and 800 K, at least double the target temperature of 300 K. The elevated temperatures broaden the distribution of the training data, such that it includes structures with a high potential energy and a low probability at 300 K. A machine learning force field trained without high-temperature data could erroneously underestimate the energy of structures with a low probability at 300 K, simply due to the lack of examples. Especially for MD applications, this would be problematic:[77] MD explores all low-energy regions, except those that are separated by a sufficiently high barrier from the starting point. The high-temperature data ensures such high barriers are consistently present between the regions of realistic and unrealistic structures. Because NMS cannot sample rotational barriers, torsional sampling was also employed. This is implemented by selecting a rotatable bond at random (if present) and rotating the fragment on one side of the rotatable bond over a random angle.

Since it is not simply possible to displace individual centers of localized orbitals at will in a KS-DFT calculation, the electronic degrees of freedom are sampled by applying a homogeneous electric field across the molecule in a random direction, on top of the geometric distortions. In this way, a force acts on all the electron pairs in the direction opposite to the electric field, displacing them out of their zero-field equilibrium positions. The perturbation by a homogeneous field has some limitations. First of all, the same force acts on all electrons together, inducing more or less a collective motion such that not all normal modes of the electron pair PES are sampled independently. Second, a too strong electric field breaks the SCF convergence, especially for the larger molecules, such that only small displacements of the centers can be achieved. Even when convergence is reached, the response to strong fields may go beyond the capabilities of the

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

local basis set used, resulting in poor wave functions with erroneous local orbital centers. An upper limit of 0.01 au ($=5.14 \times 10^9$ V m^{-1}) for the magnitude of the electric field was put in place such that every DFT calculation runs without any issue.

To address the first limitation of the homogeneous fields, dimer configurations are constructed, in which a small probe molecule, CH_4, NH_3, or H_2O, exerts a realistic and more local perturbation on a molecule in eQM7. The probe molecule is placed at a random distance and orientation next to the main molecule, such that the distance separating the two molecules is between 0.9 and 6 Å. In the same way as described above, both molecules are also subject to NMS sampling prior to creating the dimer, and an additional homogeneous field is included to maximize the diversity of the training data. As a side-effect, the dimer samples also contain some information on intermolecular interactions. However, our sampling is primarily designed to efficiently perturb centers of local orbitals, not for the calculation of precise interaction energies. The latter would require, for example, coupled-cluster calculations and energy differences of the form $E_{AB} - E_A - E_B$, to subtract out the relatively large intramolecular energy changes, such as those due to normal mode sampling.

An overview of the types and numbers of perturbations applied to each molecule is given in Table 2.

Table 2. Overview of the Perturbations and Number of Samples Per Molecule in the eQM7 Dataset

perturbation	no. of samples
NMS@600 K + elec. field	100
NMS@800 K + elec. field	100
NMS@800 K + torsion + elec. field	150
NMS@800 K + dimer + elec. field	150

2.3.2. Beta Glycine Data Set. The data set for β-glycine is constructed with Quantum ESPRESSO.[78,79] The self-consistent KS-DFT equations are solved using a plane-wave basis set and the PBE[80] functional with ultrasoft pseudopotentials.[81] Grimme's dispersion corrections[82] are included with Becke-Johnson damping (DFTD3-BJ).[83] A 3 × 3 × 3 Monkhorst–Pack grid[84] for the k-points is used together with a kinetic energy cutoff of 85 Ry. To make sure that the stresses correspond with the energies in a fixed k-point grid, a smooth penalty for the high energy Fourier components is introduced.[85] This is done in Quantum ESPRESSO with the following keywords: ECFIXED=80, QCUTZ=80, and Q2SIGMA=5. The convergence threshold for the SCF equations is 10^{-8} Rydberg. The positions of the electron pairs are calculated using Wannier90,[86] and uniform electric fields are applied within the modern theory of polarization[28] with the keyword LELFIELD.

Two different sets of single-point DFT calculations are performed. For the first set, a random electric field with a maximum strength of 0.01 au is applied in a random direction but the stress tensor is not computed since the calculation of the stress tensor in conjunction with a nonzero external field is not implemented in Quantum ESPRESSO. In the other set, stresses are computed but, as a consequence of the previous point, no external field is applied. In this way, a total amount of 25 676 first-principles calculations have been performed, 15 871 for the first set and 9805 for the second set.

In addition to the positions, the cell matrix should also be sampled extensively if the force field should be able to predict stresses. Therefore, we cannot simply use conventional normal mode sampling but follow a more general sampling strategy. First, we start by doing a first-principles optimization of β-glycine. Next, an *extended* Hessian is computed in the resulting energy minimum. This Hessian H_{ext} is a square $3N + 9$ by $3N + 9$ matrix with $N = 20$ the number of atoms in β-glycine. The extra degrees of freedom are the nine elements of the cell matrix. The extended Hessian has six zero frequencies, three translational and three rotational modes, which are discarded by projecting onto the other $3N + 3$ internal degrees of freedom. This newly created *internal* Hessian H_{int} defines the harmonic approximation of the PES:

$$E = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\mathrm{H}_{\mathrm{int}}\boldsymbol{x} \tag{28}$$

where $\boldsymbol{x}$ is a $3N + 3$ vector of the internal degrees of freedom. These degrees of freedom are sampled by making the connection with the Boltzmann-distribution:

$$p(\boldsymbol{x}) \sim \exp\left(-\frac{E}{k_{\mathrm{b}}T}\right) = \exp\left(-\frac{\boldsymbol{x}^{\mathrm{T}}\mathrm{H}_{\mathrm{int}}\boldsymbol{x}}{2k_{\mathrm{b}}T}\right) \tag{29}$$

Hence, the probability distribution of a sample $\boldsymbol{x}$ is a multivariate normal distribution with covariance matrix $\mathrm{H}_{\mathrm{int}}^{-1}k_{\mathrm{b}}T$. After a random sample $\boldsymbol{x}$ of the internal degrees of freedom is taken, it is transformed back to the original space of $3N$ elements for the positions and 9 for the cell matrix elements. Every configuration in our data set is sampled using this general framework at $T = 600$ K.

2.4. Data Augmentation. Finding the location of the electron pairs or minimizing eq 4 assumes that the electronic energy landscape around the equilibrium positions is well-known. Therefore, the structures in our data set were perturbed with homogeneous electric fields with randomly sampled direction and magnitude. The magnitude was limited to 0.01 au to avoid convergence problems in the SCF cycle. This upper limit corresponds to displacements of the electron pairs by about 0.02 Å compared to their zero-field positions. For a variety of applications, especially MD simulations, it was observed that those displacements are not large enough to sample the essential region of the electronic energy landscape. If the eMLP is trained to just these data, it will become ill-behaved when it tries to extrapolate outside the region of 0.02 Å electron pair displacements. Spurious minima and a rather chaotic PES would appear, leading to unreasonable results when minimizing the energy as a function of the electron positions with eq 4.

To overcome this problem, data augmentation is introduced as a preprocessing step when training the neural network. It is a popular technique in image classification to regularize the model and improve its performance.[87] In such applications, images are transformed in many different ways (flipping and rotating images, color transformations, etc.) to artificially increase the data set size. Every image maintains its true label after such a transformation. We utilize the same ideas to combat extrapolation of the electronic energy landscape outside the region of 0.02 Å. Essentially, by randomly displacing electron pairs further away from their equilibrium, a larger region of the electronic energy landscape is sampled and by training against these augmented data, the PES will become a well-behaving function outside the region of 0.02 Å

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

displacements. However, there is one main difference with data augmentation for image classification: our true labels will change after transforming the input data. The true labels, that is, energies and forces, of an augmented molecular system in which the electron pairs are displaced at random, are not known. Currently, no methodology is available to calculate the KS-DFT energy of a system where the centers of localized orbitals are chosen at will. The only thing that is known, due to the variational principle, is that the augmented system is higher in energy. Therefore, a heuristic estimate of the increase in energy, ΔE, is made and the neural network is trained with the target energy $E = E_{gs} + \Delta E$ as its new label. Hence, outside the region of 0.02 Å not the true PES is learned, but an approximate one. This approximation is only intended to inform the neural network that no spurious minima with low energies should be predicted for large displacements of the electron pairs. The exact value of the energy increase is not critical, because the high-energy region will practically never be visited in molecular simulations, thanks to the optimization of the electron centers in eq 4.

The data augmentation procedure starts by randomly selecting an electron pair and displacing it over a distance $\Delta \boldsymbol{r}$ which is uniformly sampled between 0.06 and 0.12 Å. The minimum displacement is large enough to not overlap with the sampled distribution in our data set (i.e., the region of 0.02 Å) and the maximum displacement is small enough to still be of use when minimizing the energy around the minimum. Next, an additional energy of

$$\Delta E = \frac{1}{2} k \|\Delta \boldsymbol{r}\|^2 \tag{30}$$

is added to the target energy of the system. The numerical value of k is set to be 2.0 Ha/Å^2 and is estimated from the first-principles results of displacing the single electron pair of H_2. Furthermore, an extra force $\Delta \boldsymbol{f}_i = -k\Delta \boldsymbol{r}$ pointing in the opposite direction of the displacement is assigned to the selected electron pair i. All the neighboring particles will also feel the influence of the displacement of the selected electron pair. Hence, the forces on these particles should also be modified such that two conditions are met: the total force is zero (for neutral systems under the influence of an constant electric field) and for nonperiodic systems, the total torque should also be consistent with the external field and the electron pair displacement. In general, many solutions satisfy these constraints, and in this work a unique weighted least-norm solution is always used. A full derivation of the expression of the augmented forces is given in Appendix B.

2.5. The Cost Function. The parameters $\boldsymbol{a}$ of the neural network are trained by minimizing the following cost function:

$$\begin{aligned} C(\boldsymbol{a}) = {} & \frac{\lambda_E}{B} \sum_{b=1}^{B} \left(\frac{E_{\mathrm{MLP}}^{(b)}(\boldsymbol{a}) - E_{\mathrm{tr}}^{(b)}}{N_b} \right)^2 \\ & + \frac{\lambda_f}{3N} \sum_{b=1}^{B} \sum_{i=1}^{N_b} \|\boldsymbol{F}_i^{(b)}(\boldsymbol{a}) - \boldsymbol{F}_{\mathrm{tr},i}^{(b)}\|^2 \\ & + \frac{\lambda_c}{3C} \sum_{b=1}^{B} \sum_{j=1}^{C_b} \|\boldsymbol{f}_j^{(b)}(\boldsymbol{a}) - \boldsymbol{f}_{\mathrm{tr},j}^{(b)}\|^2 \\ & + \frac{\lambda_\sigma}{9B} \sum_{b=1}^{B} \|\boldsymbol{\sigma}^{(b)}(\boldsymbol{a}) - \boldsymbol{\sigma}_{\mathrm{tr}}^{(b)}\|^2 \end{aligned} \tag{31}$$

where $E_{\mathrm{MLP}}^{(b)}(\boldsymbol{a})$ is the predicted energy, $\boldsymbol{F}_i^{(b)}(\boldsymbol{a})$ is the force on the ith atomic core, $\boldsymbol{f}_j^{(b)}(\boldsymbol{a})$ is the force on the jth electron pair, and $\boldsymbol{\sigma}^{(b)}$ is the stress tensor of the current system b, while $E_{\mathrm{tr}}^{(b)}$, $\mathbf{F}_{\mathrm{tr},i}^{(b)}$, $\mathbf{f}_{\mathrm{tr},j}^{(b)}$, and $\boldsymbol{\sigma}_{\mathrm{tr}}^{(b)}$ are their respective training targets. Each system b contains N_b atomic cores and C_b electron pairs such that $\sum_{b=1}^{B} N_b = N$ and $\sum_{b=1}^{B} C_b = C$ where B is the total amount of systems in the current mini-batch. The adjustable weights λ_E, λ_f, λ_c, and λ_σ determine respectively the relative importance between the energies, forces on the atomic cores, forces on the electron pairs, and the stress tensor.

The target energy $E_{\mathrm{tr}}^{(b)} = E_{\mathrm{tr,abs}}^{(b)} - E_{\mathrm{tr,ref}}^{(b)}$, appearing in the cost function, is not the absolute first-principles energy $E_{\mathrm{tr,abs}}^{(b)}$ but the difference between that value and a strategically chosen reference energy $E_{\mathrm{tr,ref}}^{(b)}$. In this way, all numerical training targets are normalized, improving the stability and speed of convergence while training the neural network. If only systems with the same chemical configuration are considered (the amount of each element and electron pairs stays the same), a single reference energy $E_{\mathrm{tr,ref}}$ for all systems in the whole data set will suffice. For instance, to train on the β-glycine data set, the mean value of all the first-principles energies $E_{\mathrm{tr,abs}}^{(b)}$ is used as the reference energy. The same approach would not work for the eQM7 data set because it comprises molecules with different chemical formulas. In this case, a reference energy must be defined for each chemical composition. The sum of isolated-atom energies is not a suitable reference: single atoms may have an uneven number of electrons, which is not supported by the current version of the eMLP. Instead, four reference hydrides are introduced: H_2, CH_4, NH_3, and H_2O, with corresponding energies E_{H_2}, E_{CH_4}, E_{NH_3}, and E_{H_2O}, which are meaningful in the eMLP framework. The energy of every neutral closed-shell molecule with n_H hydrogen, n_C carbon, n_N nitrogen, and n_O oxygen atoms can then be expressed relative to the energies of the reference hydrides:

$$\begin{aligned} E_{\mathrm{ref}} = {} & n_C E_{CH_4} + n_N E_{NH_3} + n_O E_{H_2O} \\ & + \frac{1}{2}(n_H - 2n_O - 3n_N - 4n_C) E_{H_2} \end{aligned} \tag{32}$$

This linear combination of reference hydrides contains the same amount of atomic cores and electron pairs as the molecule. This formula is used in two places. First of all, the target reference energy $E_{\mathrm{tr,ref}}^{(b)}$ is calculated in this way with the KS-DFT energies of the four reference molecules in eq 32. Second, the predictions being made by the eMLP are also adjusted: $E^{(b)}(\mathbf{a}) = E_{\mathrm{abs}}^{(b)}(\mathbf{a}) - E_{\mathrm{ref}}^{(b)}(\mathbf{a})$, where $E_{\mathrm{abs}}^{(b)}(\mathbf{a})$ is the sum of the long-range and short-range contributions explained in sections 2.1.2 and 2.1.3, but here, the reference energy $E_{\mathrm{ref}}^{(b)}(\mathbf{a})$ is dependent on the parameters of the neural network and is calculated with the actual predicted energies of the four reference hydrides in eq 32. Therefore, the four reference hydrides are included into every mini-batch and their energies are calculated in each training step. This procedure results in a consistent calculation of energy differences over the whole training set.

3. RESULTS AND DISCUSSION

3.1. Small Molecules. *3.1.1. General.* In total eight eMLP parametrizations for small molecules are trained and validated on the eQM7 data set, four with and four without data augmentation. Besides the augmentation, the training is carried out in exactly the same way, except for the random initialization of the weights in SchNet. The training, validation,

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

and test set consist of respectively 90%, 5%, and 5% of the molecules. As described in section 2.3.1, for each molecule 500 different DFT calculations were stored. After the split in train, validation, and test set, all 500 calculations belonging to a single molecule stay grouped together in a single set. Hence, the validation and test set contain unseen molecules, and a good performance on these sets indicates that the eMLP is not overfitting or extrapolating but instead generalizing well to other molecules.

The cost function of eq 31 (without stresses) is minimized with the ADAM optimizer[88] while the weights λ_E, λ_f, and λ_c are all equal to one (in units of electronvolt and angstrom). The initial learning rate is 3×10^{-4} and decays exponentially with a factor of 2 every 30 epochs. We never observe an increase in error on the validation set while training the network, mainly because the training set is large enough such that the risk of overfitting is negligible. Hence, the early stopping criterion is never triggered and each model is trained for 288 h on V100-GPUs after which the decrease in error on the validation set becomes inconsiderable. All mini-batches contain 64 systems (the reference hydrides not included) and if data augmentation is applied, 10% of all the systems in every mini-batch are augmented. The short-range SchNet network has $F = 512$ features, $G = 128$ filters, $N = 32$ radial basis functions and $T = 4$ interaction blocks or iterations. The cutoff radius r_{cutoff} is 4 Å, and the parameter γ of the long-range electrostatic interaction is 0.3928 /Å. An overview of all the hyperparameters can be found in Table 3. Their values have been adjusted empirically.

Table 3. Overview Table of the Hyperparameters of the eMLP

hyperparameter	value
initial learning rate	3×10^{-4}
batch size B	64
cost function weights λ_E, λ_f, λ_c	1/eV^2, 1 Å^2/eV^2, 1 Å^2/eV^2
cutoff radius r_{cutoff}	4 Å
cutoff transition width r_Δ	0.5 Å
Gaussian charge inverse width γ	0.3928/Å
features F	512
filters G	128
radial basis functions N	32
interaction blocks T	4
convolution normalization factor J	70
augmentation percentage (if applicable)	10%
augmentation strength k	2 Ha/Å^2
minimum augmentation displacement	0.06 Å
maximum augmentation displacement	0.12 Å

In the next two sections, we will first focus on the nonaugmented models by looking at two different categories of errors: *static* and *dynamic* errors. Static errors are reported with the atomic cores and electron pairs at the same location of the Foster-Boys centers. Training the model by performing gradient descent (or alternatives) on the cost function of eq 4 directly minimizes this type of error. Dynamic errors on the other hand, are reported after the electron pairs are relaxed. Thus, eq 4 is minimized and the errors on physical properties are reported with the electrons in their eMLP equilibrium positions, which can be slightly displaced from their true Foster-Boys positions. In section 3.1.4, we will show that data augmentation is necessary when performing MD simulations. The advantages and minor disadvantages of augmented models compared to nonaugmented models will be explored.

3.1.2. Static Errors. In Table 4 the static errors are reported on the test set as an average (±standard deviation) over four models, each optimized starting from a different set of random initial model variables. Mean absolute errors (MAE) and median errors (50% of the errors are lower than this value) are tabulated and can be compared to the intrinsic variability of the data set. The intrinsic variability is a measure of the variance of the training data. It can be understood as the error being made by the best possible *constant* model, a model that predicts the same value (i.e., the mean) irrespective of the input. The mean absolute error of that constant model is the intrinsic MAE. An accurate and well-performing model should have errors which are significantly lower than the intrinsic MAE. This is the case: the errors on the energies and forces are a factor 20−30 times smaller than the intrinsic MAE while the electron pair forces are about 1 order of magnitude smaller, showing that the eMLP is capable of making accurate predictions.

Table 4. Static Errors: Electron Pairs Are Located at the Foster-Boys Positions[a]

	energy [meV/atom]	forces [meV/Å]	electron pair forces [meV/Å]
	Intrinsic Variability		
MAE	170.55	1106.8	385.9
	Nonaugmented Models		
MAE	4.45 (±0.11)	48.8 (±0.3)	43.8 (±0.2)
median error	3.02 (±0.13)	27.9 (±0.2)	30.1 (±0.2)
	Augmented Models		
MAE	4.95 (±0.25)	52.1 (±0.3)	50.3 (±0.4)
median error	3.45 (±0.39)	29.8 (±0.2)	34.6 (±0.3)

[a]Mean absolute errors and median errors of augmented and nonaugmented models are reported on the test set. Results are averaged over four different models, and the value between parentheses is the standard deviation between those models. Note that the errors have been calculated after subtracting the external energies of eq 12 such that the intrinsic MAE of the electron pair forces is not zero, even though all the *total* electron pair forces in the test set are zero.

A direct comparison with other machine learning force fields (MLFFs) is not possible since we are dealing with a new database and the model itself is different due to the inclusion of electron pair particles. A similar data set however is the ISO17 data set,[89] on which several machine learning force fields were trained.[48,55,58] It is similar because training is performed on energies and forces while the test set contains unseen molecular isomers (not contained in the training set). In those works, the nuclear forces in the validation set were reproduced with an MAE on the range of 1 to 2 kcal/mol/Å. The eMLP has force errors just above 1 kcal/mol/Å, putting it alongside state-of-the-art machine learning models for reliable force estimations.

3.1.3. Dynamic Errors. To calculate dynamic errors, the minimization of eq 4 must be performed. Hence, here we make use of the Born−Oppenheimer eMLP energy $E_{\text{eMLP,BO}}$. The BFGS[90] algorithm in SciPy[91] is utilized to accomplish this task. The resulting dynamic errors are tabulated in Table 5. The first three rows correspond to the average over four models on a

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

Table 5. Dynamic Errors: Electron Pairs Are Optimized[a]

	energy [meV/atom]	forces [meV/Å]	dipole norm [Debye]	polarizability [bohr^3]
		Intrinsic Variability		
MAE	174.93	1177	1.229	6.70
		Nonaugmented Models		
MAE	40.8 (±21.1)	104.0 (±23.2)	0.169 (±0.042)	0.72 (±0.44)
median error	3.15 (±0.13)	30.0 (±0.6)	0.035 (±0.001)	0.26 (±0.01)
error rate	2.4% (±0.8%)			0.2% (±0.4%)
		Augmented Models		
MAE	49.3 (±33.2)	137.8 (±63.8)	0.286 (±0.014)	0.54 (±0.09)
median error	3.45 (±0.18)	32.9 (±0.5)	0.078 (±0.005)	0.31 (±0.03)
error rate	0.4% (±0.1%)			0.0% (±0.0%)

[a]Mean absolute errors and median errors of augmented and nonaugmented models are reported on the test set for energies, forces and dipoles. The polarizability tensor is calculated at the DFT equilibrium positions of the 343 molecules in the test set such that it corresponds to a different error rate. Results are averaged over four different models, and the value between parentheses is the standard deviation between those models.

representative randomly sampled subset of 5660 structures of the test set to reduce the computational time. It is immediately apparent that the MAE has increased multifold, while the median errors barely increase. This is due to the large tail of the error distribution. There will be a small amount of outliers, having errors orders of magnitude larger than the rest, dominating the MAE. Moreover, the standard deviation on the MAE has the same order of magnitude, indicating large fluctuations between different models. Both effects are closely linked with the error rate, also reported in Table 5. The error rate is the fraction of the number of systems for which the optimization of the electron pairs' positions fails. In those cases, the BFGS algorithm does not converge and yields solutions with nonzero electron pair forces. The particles of the model then also show erratic behavior: two or more particles are located at the same point or the particles are chaotically spread all across the molecule. For 2.4% of all configurations, the nonaugmented models cannot find a proper minimum. In rare cases, even for structures for which a solution for the electron pair positions is found, there might be large displacements compared to the Foster-Boys positions, giving rise to the outliers and the large MAEs. Nevertheless, in comparison to the static errors, the median errors remain almost unchanged and stable (small standard deviations). For the vast majority of molecules in the test set, energies and forces after the electron optimization are still predicted with the same accuracy as the static errors.

The eMLP was not explicitly trained to dipole moments but is able to reproduce it when the electron pair equilibrium positions stay close to their Foster-Boys targets, since the Foster-Boys locations exactly reproduce the dipole moment of the molecule. In Table 5, the MAE and median error on the norm of the dipole moment is reported for the structures in the test set. Again, the MAE suffers from exceptional outliers, making it less suitable to quantify the performance of the model. The median errors, however, show that accurate predictions are possible with errors as low as 0.034 D for the nonaugmented models.

The polarizability tensor characterizes the response of the molecular dipole moment to a homogenous external electric field. In the final column of Table 5, the MAE and median error on the components of the polarizability tensor are given, for each of the 343 molecules in the test. Only the polarizability tensors at the DFT equilibrium geometries are

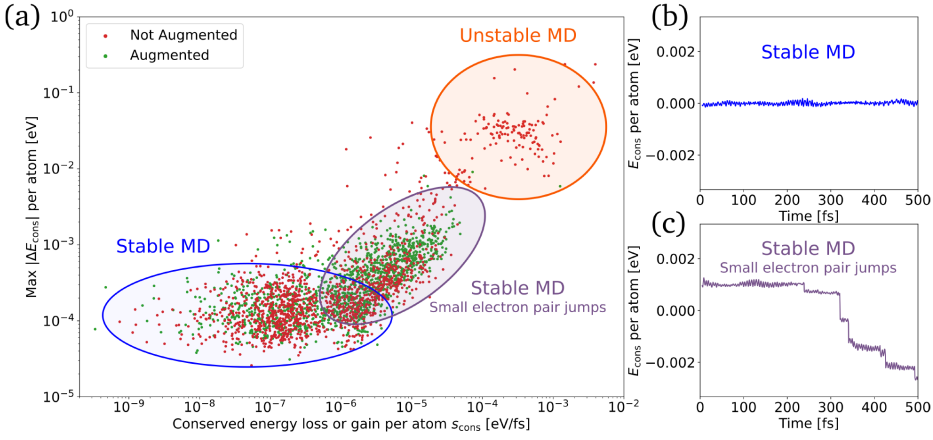


Figure 4. (a) Scatter plot of the stability of MD simulations. Each point represents one MD simulation of 500 fs of a single molecule, each for the four different augmented models (in green) and four nonaugmented models (in red). The conserved energy per atom is tracked throughout the simulation and the value of the slope of a linear function fitted to that curve is the rate of energy loss or increase. The maximum jump in conserved energy per atom between two MD steps is given on the *y*-axis. Three different regions are encircled: a stable MD region, a zone with small electron pair jumps and an unstable MD region. To illustrate the behavior in the first two regions, E_{cons} is tracked in panel b for stable MD and panel c for stable MD with small electron pair jumps.

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

considered. For the nonaugmented models, the median errors have a magnitude of about 0.30 bohr^3, while the MAEs are much larger. Furthermore, there is a small error rate of 0.2%. The small errors indicate that eMLP is not only capable of describing the ground-state configuration of the electron pairs, but also their response to an external field.

A comparison of the performance of eMLP for dipole moments and polarizabilities with other models from the literature is somewhat unfair: most state-of-the-art machine learning force fields train to these targets explicitly, whereas they are not explicitly included in our cost function. Still, the eMLP can be compared with models previously trained on the QM7b data set.[92,93] There, the errors are reported as a fraction of the intrinsic variability. Similarly, we can compute the ratio of the eMLP median errors with respect to the (median) intrinsic variabilities. A quick comparison shows that these ratios are similar (3−6%) for both the dipole moments and polarizabilities. Note that the data sets have some differences: QM7b contains only one data point per molecule (7122 versus 3 434 000 data points), and eQM7 in this work does not contain sulfur or chlorine.

3.1.4. Stabilizing MD Simulations with Data Augmentation. In this section, the applicability of eMLP to MD simulations is explored. We will primarily focus on the stability of the MD run by investigating the conserved quantity. This is a more challenging test compared to the static and dynamics errors of the previous subsections, because the nuclear motion explores a broader region in the coordinate space.

For every molecule in the test set and for each eMLP parametrization (four nonaugmented and four augmented models), a 500 fs NVT simulation at 300 K is performed with a time step of 0.5 fs, of which the initialization phase (first 10 steps) is discarded. Newton's equations are integrated in Yaff[94] with a Nosé−Hoover thermostat[95] with a chain length of 3. At every time step, the electron pair positions are optimized with the L-BFGS-B[96] algorithm. To speed up the convergence of L-BFGS-B, box constraints are imposed on the electron pair positions: their maximal displacement from the initial guess is limited to three times the largest nuclear displacement in the last MD step. Within these bounds a new local minimum is always found, except for some of the unstable MD runs discussed below. For each run, the conserved quantity divided by the number of atoms, E_{cons}, is characterized by two parameters. First, a line is fitted to E_{cons} as a function of time, the slope of which, s_{cons}, represents the rate of conserved energy loss or gain per atom. Second, discontinuities in the BO potential energy surface are quantified by the maximum jump of the conserved quantity per atom between two time steps, max $|\Delta E_{\mathrm{cons}}|$. Smaller values for both parameters correspond to a more stable MD run.

In Figure 4a, each MD simulation is represented by the two parameters (s_{cons}, max$|\Delta E_{\mathrm{cons}}|$). Every red point corresponds to a single MD run using one of the four nonaugmented models. Three different regions can be distinguished in this plot: a region of stable MD simulations, a transitional region, and a region with unstable MD simulations.

Stable MD simulations are characterized by small fluctuations in the conserved quantity without a noticeable increasing or decreasing trend. Formally, these trajectories are characterized by $\Delta E_{\mathrm{cons}} < 5 \times 10^{-4}$ eV. A typical example is shown in Figure 4b.

The transition region contains trajectories exhibiting sudden drops of the conserved energy per atom by 5×10^{-4} to at most 0.01 eV, resulting in a rate of energy loss of approximately 10^{-6} to 10^{-5} eV/fs. A representative example is shown in Figure 4c. Visualization of the trajectories reveals that the electron pairs also exhibit sudden displacements when the conserved quantity drops. These jumps occur most frequently in double bonds or lone pairs, and are small enough to preserve the chemical structure of the molecule. The drop in conserved quantity corresponds to the appearance of a lower local energy minimum for the electron pairs and the L-BFGS-B tends to find such lower minima whenever they are separated by only a negligible barrier from the current minimum. The resulting trajectories are still useful to sample the PES since the thermostat in MD simulations can compensate the loss or gain in conserved energy.

Unstable MD runs are characterized by max $|\Delta E_{\mathrm{cons}}| > 0.01$ eV, with corresponding dramatic and nonphysical rearrangements of the electronic centers. This behavior is seen for 9.8% of the MD runs with nonaugmented models. For example, electron centers begin to overlap (which is unexpected for centers of localized orbitals) or they are ejected out of the molecule. In these cases, the eMLP wrongly assigns a lower energy to unreasonable electron configurations for which no representative training data exists.

The results so far show that the nonaugmented models may result in stable MD simulations, except when electron centers move (even briefly) outside the region sampled in the training data. The data augmentation in this work is designed to teach eMLP that it should predict a high potential energy and restoring forces (pulling the electron pairs back) whenever they venture into uncharted territory. The green points in Figure 4a correspond to MD simulations with the augmented models, and these results confirm the effectiveness of the augmentation scheme. For all these MD runs, max $|\Delta E_{\mathrm{cons}}|$ stays below 0.01 eV, and nonphysical electron pair configurations were not observed.

The improvement of data augmentation on MD simulations comes at a minor cost. The augmented models are trained in part to "noisy" labels (energies and forces) because, when the electron pairs are displaced at random, only an educated guess of the correct label can be made. For that reason, the nonaugmented models have a slight edge of about 10% for all static errors, as shown in Table 4. On the other hand, Table 5 compares the augmented and nonaugmented models for the dynamic errors. The same behavior is again present. The median errors on energies, forces, dipoles, and polarizabilities are slightly higher for the augmented models. The opposite is true for the error rates. The augmented models have an almost negligible error rate of 0.4%, six times lower than the nonaugmented models, again providing evidence that data augmentation avoids erratic solutions for the electron centers. Moreover, the error rate of the polarizabilities vanishes (only the KS-DFT ground state geometries of the molecules are considered here). In conclusion, a small amount of accuracy is traded for an increase in stability, which improves the reliability of MD simulations.

3.1.5. Infrared Spectra. The computation of infrared (IR) spectra requires an excellent reproduction of both the PES of the molecule and the response of the dipole moment to changes in geometry. The static IR spectra at 0 K consist of peaks lying at the frequencies of the normal modes. The corresponding intensities are proportional to the square of the derivative of the dipole moment along those normal modes.[97] The DFT spectra are calculated with finite differences, while

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

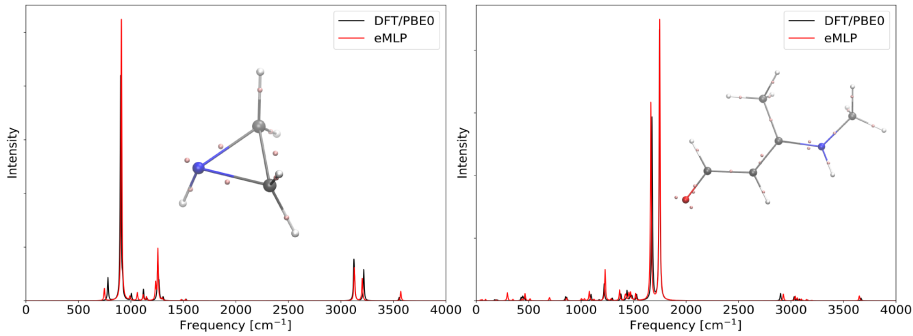


Figure 5. IR spectra of two molecules in the test set: aziridine (left) and 3-(methylamino)but-2-enal (right).

the spectra of the eMLP are calculated analytically by using automatic differentiation in TensorFlow.[98] Only the Hessian or second order derivatives of the energy with respect to the atomic cores, electron pairs, and the electric fields are necessary for the computation of the frequencies and intensities.

In Figure 5 we predict the IR spectra of two randomly selected molecules of the test set: azaridine in the left panel and 3-(methylamino)but-2-enal in the right panel. Similar results are obtained for all other molecules in the test set. Lorentzian line shape functions with a full width at half-maximum of 10 cm^{-1} are used to visualize the spectra. Both the intensities and frequencies correspond well with the DFT values for all vibrational modes, including the high frequencies belonging to the C−H and N−H stretches and the low frequencies belonging to more global modes of the molecule. Furthermore, the eMLP correctly identifies the peaks with the largest intensities. We emphasize that the eMLP is trained on neither molecules, showing the transferability of the eMLP for small molecules to predict IR spectra. For instance, the mean average error on all frequencies of all molecules in the validation set is approximately 15 cm^{-1}, which is about the same magnitude as the full width at half-maximum of the Lorentzian line shapes.

3.2. β-Glycine. In this section, the eMLP will be utilized to model the response properties of β-glycine, with the main focus on the piezoelectric tensor. In essence, the piezoelectric tensor describes the coupling between mechanical properties (stress or strain) and electric properties (electric displacement or electric field). Multiple piezoelectric tensors exist, depending on the independent variables in the coupled equations. Here, we study the piezoelectric charge tensor $\boldsymbol{e}$:[99]

$$\boldsymbol{\sigma} = \mathbf{C}_{\mathcal{E}=0}\colon \boldsymbol{S} - \boldsymbol{e}^{\mathrm{T}}\cdot\boldsymbol{\mathcal{E}} \tag{33}$$

$$\boldsymbol{D} = \boldsymbol{e}\colon \boldsymbol{S} + \varepsilon_{\mathbf{S}=0}\cdot\boldsymbol{\mathcal{E}} \tag{34}$$

and the piezoelectric strain constant $\boldsymbol{d}$:

$$\boldsymbol{S} = \mathbf{C}^{-1}_{\mathcal{E}=0}\colon \boldsymbol{\sigma} + \boldsymbol{d}^{\mathrm{T}}\cdot\boldsymbol{\mathcal{E}} \tag{35}$$

$$\boldsymbol{D} = \boldsymbol{d}\colon \boldsymbol{\sigma} + \varepsilon_{\sigma=0}\cdot\boldsymbol{\mathcal{E}} \tag{36}$$

which both couple to the strain $\boldsymbol{S}$, stress $\boldsymbol{\sigma}$, electric field $\boldsymbol{\mathcal{E}}$ and electric displacement field $\boldsymbol{D} = \varepsilon_0\boldsymbol{\mathcal{E}} + \boldsymbol{P}$. In these equations, $\mathbf{C}$ and ε are the stiffness and dielectric tensor, respectively, while $\boldsymbol{P}$ is the induced dipole density. Hence,

$$e_{ijk} = \left(\frac{\partial D_i}{\partial S_{jk}}\right)^{\mathcal{E}} = -\left(\frac{\partial \sigma_{jk}}{\partial \mathcal{E}_i}\right)^{S} \tag{37}$$

$$d_{ijk} = \left(\frac{\partial D_i}{\partial \sigma_{jk}}\right)^{\mathcal{E}} = \left(\frac{\partial S_{jk}}{\partial \mathcal{E}_i}\right)^{\sigma} \tag{38}$$

The piezoelectric strain constant $\boldsymbol{d}$ cannot be calculated directly in Quantum ESPRESSO since it is impossible to compute stresses when there are electric fields being applied. Fortunately, there exists a relation[99] between the two piezoelectric constants:

$$\boldsymbol{d} = \boldsymbol{e}\colon \mathbf{C}^{-1}_{\mathcal{E}=0} \tag{39}$$

allowing us to calculate the piezoelectric strain constant by multiplying the charge tensor with the inverse of the stiffness tensor.

To show that the eMLP is able to reproduce these response properties, we train a single specialized model for β-glycine. A split of 90−5−5% is used for the training, validation, and test set. Stresses are included in the cost function of eq 31 and λ_E = 0.1/eV^2, λ_σ = 0.1/GPa^2. The initial learning rate starts at the same value 3×10^{-4} but decays exponentially with a factor of 2 every 1200 epochs. To calculate the periodic long-range interactions, Ewald-summation[100] has been used. Only models with data-augmentation are considered here, and the total training time is limited to 216 h (early stopping is not triggered in this time-period). All remaining hyperparameters of the neural network architecture or the training algorithm have the same value as for the data set of the small molecules. After training, the best performing model is selected, and this model is used for all results below. The static errors on the test set are 3.1 meV/atom, 35 meV/Å, and 0.063 GPa for the MAEs on the energies, forces, and stresses, respectively. These are lower than the ones for the small molecules because training and validation are done on the same material, which is an easier task. This is reflected in the almost perfect reproduction of the lattice constants and volume of the unit cell, reported in Table 6. The unit cell itself belongs to space group $P2_1$ and is visualized in Figure 6.

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

Table 6. DFT and Predicted Lattice Constants, Angles, and the Unit Cell Volume

	lattice lengths [Å]			angles [deg]	volume [Å^3]
	a	*b*	*c*	*β*	*V*
DFT/PBE	5.08	6.24	5.40	112.6	158.05
eMLP	5.08	6.25	5.40	112.4	158.41

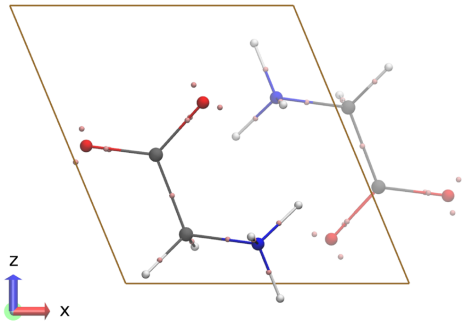


Figure 6. DFT optimized structure of *β*-glycine. The orientation of the unit cell and the electron pairs are visualized.

The DFT values of the stiffness and dielectric and piezoelectric tensors are calculated by finite differences. We were unable to reproduce the value of 195 pm/V for the d_{16} coefficient, as reported in Guerin et al.,[101] presumably because of differences in computational details and dispersion model. This is not surprising as the authors of the same paper report that the stiffness and piezoelectric strain constants are extremely sensitive to the dispersion corrections. The stiffness constant increases if dispersion corrections are included, and by making use of eq 39, it is obvious that the piezoelectric constants will decrease. Nevertheless, in this work we are mainly interested in the ability of the eMLP to reproduce our training data. The predicted response properties are calculated analytically by first constructing the Hessian, after which the equations of Appendix A are utilized.

The diagonal components of the (ion-clamped) dielectric tensor under constant strain are reported in Table 7. The relative error on the predicted results is 1% or less, which shows that the response of the electron pairs to an external field is almost perfectly replicated. This is a promising result, demonstrating that the eMLP should be able to describe various dielectrics. For the remaining response properties, we make use of the Voigt notation: $xx = 1$, $yy = 2$, $zz = 3$, $yz = 4$, $xz = 5$, and $xy = 6$. Note that in this notation, the off-diagonal components of the strain tensor are multiplied by two: $\boldsymbol{S} = (s_1, s_2, s_3, s_4, s_5, s_6) = (s_{xx}, s_{yy}, s_{zz}, 2s_{yz}, 2s_{xz}, 2s_{xy})$. The offdiagonal components of the piezoelectric strain tensor $\boldsymbol{d}$ should also be multiplied by two since they should be counted double in the contraction of eq 35 in Voigt notation. Table 7 shows that the diagonal components of the stiffness constants are accurately reproduced, except for the C_{22} coefficient for which the deviation from the DFT result is slightly higher. This corresponds to the direction in which the zwitterions are stacked on top of each other in Figure 6. In this direction, $\pi-\pi$ stacking between parallel molecules is expected to have a significant dispersion component, which is not yet explicitly included in eMLP. Furthermore, note that our cost function only tries to minimize the absolute (squared) errors and not the relative errors. For this reason, a minor absolute error of 1.9 GPa on the small C_{66} coefficient (which is comparable to errors of the other components), results in a more substantial relative error. In the same table, the DFT values of the piezoelectric charge tensor are compared to the ones predicted by the eMLP. Owing to the symmetry of the *β*-glycine crystal, only eight values of the piezoelectric tensor are nonzero. All the DFT values, other than the e_{22} coefficient, are accurately predicted. Finally, the piezoelectric strain constants are also given in Table 7. The deviations from the DFT values can be attributed to the errors on the stiffness constants. For instance, the d_{16} is dependent on the inverse of the C_{66} coefficient, for which there is a moderate relative error, such that non-negligible relative errors on the piezoelectric strain constants are unavoidable.

Table 7. PBE and eMLP Prediction of the Dielectric, Stiffness and Piezoelectric Constants in *β*-Glycine and the Deviation between Those Two Values

	PBE	eMLP	deviation
	Dielectric Constants		
ε_{11}	2.65	2.68	0.03
ε_{22}	2.19	2.21	0.02
ε_{33}	2.52	2.55	0.03
	Stiffness Constants [GPa]		
C_{11}	62.4	65.0	2.6
C_{22}	22.2	31.4	9.2
C_{33}	77.6	84.1	6.5
C_{44}	8.5	9.2	0.7
C_{55}	17.5	18.9	1.4
C_{66}	3.8	5.7	1.9
	Piezoelectric Charge Constant [C/m^2]		
e_{14}	−0.11	−0.11	0.00
e_{16}	0.24	0.24	0.00
e_{21}	0.04	0.04	0.00
e_{22}	−0.16	−0.03	0.13
e_{23}	0.12	0.10	0.02
e_{25}	−0.08	−0.05	0.03
e_{34}	0.01	0.02	0.01
e_{36}	−0.05	−0.05	0.00
	Piezoelectric Strain Constant [pm/V]		
d_{14}	−25.0	−19.8	5.2
d_{16}	71.6	46.5	25.1
d_{21}	1.0	0.3	0.7
d_{22}	−9.6	−1.8	7.8
d_{23}	1.9	1.2	0.7
d_{25}	−4.3	−1.3	3.0
d_{34}	3.2	3.9	0.7
d_{36}	−15.1	−9.9	5.2

3.3. Computational Efficiency. In the previous sections, several distinguishing features of eMLP were discussed, which can be attributed to its increase in model complexity compared to nonpolarizable MLPs or conventional explicit-electron models. The increase in complexity also comes with an extra computational burden. Compared to conventional force fields, eMLP is computationally more demanding for two reasons: (i) machine learning potentials are more expensive than simple pairwise interactions and (ii) the inclusion of explicit-electron particles increases the particle density. Both effects will be discussed qualitatively, because our implementation focused on

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

proof of concept and several aspects can still be made more efficient. For example, a regular Ewald sum was used for long-range electrostatics, which can be replaced by $N \log(N)$ scaling methods such as Particle-Mesh Ewald. One may also envision replacing SchNet by another machine-learning potential. Some illustrative numbers will be mentioned below, but bear in mind that these are subject to change with future hardware and implementation updates.

The computational cost of SchNet can be appreciated by comparing its wall time to that of the electrostatic term in eMLP. The ratio of both is a good indication of the overhead of SchNet compared to a conventional force field. For the molecules of the eQM7 data set, the computational cost of SchNet is five times as large as the electrostatic energy. However, for a β-glycine unit cell (20 atoms), SchNet is equally expensive as our GPU implementation of the Ewald sum.

The increase in particle density due to the explicit electrons has two effects. First, for a given system, the total number of particles increases (approximately doubles in this work), resulting in a linear increase of the cost of SchNet. In addition, also the number of particles within the cutoff doubles. Both effects combined result in an approximately 4-fold increase in cost due to the presence of explicit electrons, not yet taking into account the cost of relaxing the electron positions in a Born−Oppenheimer calculation.

Just like other machine-learning potentials, eMLP is still considerably more efficient than a DFT calculation. On a A100 GPU with 80GB of VRAM, the eMLP finishes a single energy and force evaluation in 3.1 ms for 1,3-dimethylazetidine [CC1CN(C1)C], a representative molecule of the eQM7 data set with 17 atomic cores and 18 electron pairs. A single Born−Oppenheimer evaluation requires on average 10 energy such evaluations to optimize the electron pair positions for that molecule. These timing are almost negligible compared to that of a single DFT evaluation, which takes 51 s on eight cores of an AMD Epyc 7H12 CPU. Hence, the eMLP is 3 to 4 orders of magnitude faster than DFT, for small molecules.

4. CONCLUSION AND OUTLOOK

In this work, we introduced the eMLP, a new explicit electron force field making use of machine learning for short-range interactions combined with classical electrostatics at longer ranges. Centers of localized Foster-Boys or Wannier orbitals serve as training data for the positions of electron pair particles in the eMLP, which has several advantages. These centers provide extensive reference data, they exactly reproduce molecular dipole moments, intuitively represent chemical features, and are well-defined. Two new data sets were created to showcase eMLP's capabilities and performance. The eQM7 data set consists of a variety of independently sampled configurations for each of the 6868 small molecules in the data set. The β-glycine data set uses a generalized version of normal mode sampling to sample configurations with different unit cells and fractional coordinates. It was shown that force errors under 0.05 eV/Å can be achieved, even after (re)optimizing the electron pairs with a trained eMLP. Furthermore, IR-spectra of unseen molecules are predicted accurately. For β-glycine, the eMLP is able to model elastic, dielectric, and piezoelectric responses, which are hard to accomplish with conventional force fields. These test cases demonstrate the potential of eMLP for the simulation of physical properties involving nontrivial electronic behavior. To run MD simulations, it is necessary to train eMLP with data augmentation, a technique in which an electron center is displaced over a larger distance with a large associated energy increase. Such data cannot be generated with electronic structure calculations but are needed in the training set to prevent extrapolation issues.

During the development of eMLP, new challenges arose for which future methodological advances are of interest. In this work, only relatively weak homogeneous electrical fields were used to perturb the centers of localized orbitals, which provides an incomplete picture on the electronic response function. Kohn−Sham DFT data for larger and more diverse displacements of the centers would be a valuable addition to the training set and may eventually replace the data augmentation in this work.

Several extensions or more complex use-cases of the eMLP can be realized in future developments. We do not expect any major difficulties extending the eMLP beyond the second row in the periodic table, provided that strongly bound electrons are lumped into the atomic cores. We expect configurations with a large density of valence electrons to be the most challenging. For example, for transition metals, the locations of the valence electron pairs will be sensitive to the geometry of the surrounding nuclei. The corresponding PES will have shallow minima which could hamper the BO approach. Also the simulation of chemical reactions with the eMLP should be investigated, since this could potentially enable a natural description of redox reactions and charge-transport phenomena. Up until now, the electronic degrees of freedom were modeled as electron pairs, but a subdivision in a separate class of spin-up and spin-down electrons, in analogy to LEWIS•, would enable the simulation of radicals and magnetic systems and could also be beneficial for chemical reactions. Another interesting improvement would be the explicit treatment of long-range dispersion interactions. Finally, it should be noted that the short-range interactions can be modeled by any state-of-the-art machine learning force field. For example, when more data-efficient models are proposed, they can be easily incorporated into eMLP.

A. ANALYTICAL EXPRESSIONS OF EMLP RESPONSE PROPERTIES

Here, we give a brief overview of the necessary equations to calculate response properties analytically with eMLP for solids and molecules. The evaluation of derivatives appearing in these expressions is implemented with automatic differentiation. The starting point is the full extended Hessian with respect to the fractional coordinates of the atomic cores and electron pairs, the elements of the unit cell, and the electric field:

$$\mathrm{H}_{\mathrm{extended}} = \begin{pmatrix} \mathrm{H}_{\mathrm{ff}} & \mathrm{H}_{\mathrm{fa}} & \mathrm{H}_{\mathrm{fe}} \\ \mathrm{H}_{\mathrm{af}} & \mathrm{H}_{\mathrm{aa}} & \mathrm{H}_{\mathrm{ae}} \\ \mathrm{H}_{\mathrm{ef}} & \mathrm{H}_{\mathrm{ea}} & \mathrm{H}_{\mathrm{ee}} \end{pmatrix} \tag{40}$$

The subscripts denote the respective groups of derivatives of the Hessian:

f: Toward fractional coordinates of the N atomic cores and C electron pairs ($3N + 3C$ elements in total).

a: Toward elements of the unit cell (nine elements in total). Usually two indices are used to identify rows (for cell vectors) and columns (for components of cell vectors).

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

e: Toward the electric field (three elements).

For instance, using this notation, $H_{fa} = H_{af}^{T} \in \mathbb{R}^{(3N+3C)\times 9}$ is the off-diagonal block of the Hessian with the first index being the fractional coordinates and the second index being the unit cell elements. In the following equations, the unit cell matrix is A_{ij} for which the *rows* are the lattice vectors. All the following response properties are calculated at 0 K after an optimization of the structure such that the Hessian has no negative eigenvalues and is invertible. For the stiffness tensor, we find

$$C_{ijkl} = \frac{1}{V}\sum_{mn} A_{mi}A_{nk}(H_{aa} - H_{af}H_{ff}^{-1}H_{fa})_{mjnl} \tag{41}$$

where the second term is a consequence of the relaxation of the fractional coordinates, calculated using vibrational subsystem analysis (VSA).[102] Furthermore, the term in parentheses is a 9 × 9 matrix of which the elements are reordered in a 3 × 3 × 3 × 3 tensor to perform the contraction over m and n. The total polarizability of a molecule is simply

$$p_{ij} = (H_{ef}H_{ff}^{-1}H_{fe})_{ij} \tag{42}$$

This formula also includes atomic core relaxations. These can be frozen as well, by only taking the submatrices corresponding to the electron pairs into account. The latter are reported in the main text. The (relative) dielectric constant is

$$\varepsilon_{ij} = \delta_{ij} + \frac{1}{V\varepsilon_0}(H_{ef}H_{ff}^{-1}H_{fe})_{ij} \tag{43}$$

where V is the volume of the unit cell. Again, by excluding the atomic cores in the sum matrices, one can compute the so-called ion-clamped static dielectric tensor, which is reported in the main text. Next, the piezoelectric charge constant is

$$\widetilde{e_{ijk}} = \frac{1}{V}\sum_{m} A_{mj}(H_{ef}H_{ff}^{-1}H_{fa})_{imk} \tag{44}$$

where again the matrices are reshaped at the end. Note that this is the *proper* piezoelectric tensor $\widetilde{e_{ijk}}$, which can be measured experimentally.[103] Furthermore, the proper piezoelectric tensor has the correct symmetry and is invariant under equivalent displacements of the particles. Indeed, displacing a particle over an integer multiple of the lattice vectors (putting it in a neighboring cell), should not affect any observable quantities. The dipole vector itself is not invariant under such a displacement, which is not an issue as it cannot be measured experimentally for periodic systems. Only the changes of the dipole vector, due to internal relaxations of the particles, can be measured experimentally. The proper and improper piezoelectric tensor e_{ijk} are related by[103]

$$\widetilde{e_{ijk}} = e_{ijk} + \delta_{jk}P_i - \delta_{ik}P_j \tag{45}$$

where P_i is the dipole density. The improper piezoelectric tensor e_{ijk} is not symmetric in general and depends on the unit cell under consideration. Only the proper piezoelectric tensors are reported in the main text.

B. DERIVATION OF AUGMENTED FORCES

In the augmentation procedure, the electron pair i is displaced over a distance $\Delta\boldsymbol{r}_i$, and as a result, it receives an additional force $\Delta\boldsymbol{f}_i$, which wants to push the particle back to its original position:

$$\Delta\boldsymbol{f}_i = -k\Delta\boldsymbol{r}_i \tag{46}$$

In the following discussion, the index j is used for all particles, atomic cores, and electron pairs. Since the only external force on the system is a constant electric field $\mathcal{E}$, the net force on a electrically neutral system must be zero:

$$\sum_j \boldsymbol{f}'_j = 0 \tag{47}$$

and the total torque must be equal to

$$\sum_j \boldsymbol{r}'_j \times \boldsymbol{f}'_j = \boldsymbol{d}' \times \mathcal{E} \tag{48}$$

where $\boldsymbol{d}'$ is the dipole vector of the system, $\boldsymbol{f}'_j$ and $\boldsymbol{r}'_j$ are the forces and positions of the particles *after* the electron pair has been moved. The goal is to find the extra forces $\Delta\boldsymbol{f}_j$ for $j \neq i$ such that eqs 47 and 48 are fulfilled. This results in the following equations for $\Delta\boldsymbol{f}_j$:

$$\begin{cases} \sum_{j\neq i} \Delta\boldsymbol{f}_j = -\Delta\boldsymbol{f}_i \\ \sum_{j\neq i} \boldsymbol{r}_j \times \Delta\boldsymbol{f}_j = q_i\Delta\boldsymbol{r}_i \times \mathcal{E} - \Delta\boldsymbol{r}_i \times (\boldsymbol{f}_i + \Delta\boldsymbol{f}_i) - \boldsymbol{r}_i \times \Delta\boldsymbol{f}_i \end{cases} \tag{49}$$

Note that the equations are invariant to a global translation. This property will be used later on. If one also demands that the extra forces are as small as possible due to the disruption caused by the augmentation, one should find the stationary point of the following Lagrangian:

$$C(\{\Delta\boldsymbol{f}_j\}) = \frac{1}{2}\sum_{j\neq i}\phi(r_{ji})\|\Delta\boldsymbol{f}_j\|^2 + \boldsymbol{\lambda}\cdot\left(-\Delta\boldsymbol{f}_i - \sum_{j\neq i}\Delta\boldsymbol{f}_j\right) + \boldsymbol{\mu}\cdot\left(\boldsymbol{M} - \sum_{j\neq i}\boldsymbol{r}_j \times \Delta\boldsymbol{f}_j\right) \tag{50}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are two vectorial Langrange multipliers. The vector $\boldsymbol{M} = q_i\Delta\boldsymbol{r}_i \times \mathcal{E} - \Delta\boldsymbol{r}_i \times (\boldsymbol{f}_i + \Delta\boldsymbol{f}_i) - \boldsymbol{r}_i \times \Delta\boldsymbol{f}_i$ is introduced to simplify the notational burden and $\phi(r_{ji})$ is a possible weight factor depending on the distance between the particles i and j. Minimization with respect to $\Delta\boldsymbol{f}_j$ yields

$$\phi(r_{ji})\boldsymbol{f}_j - \boldsymbol{\lambda} - \boldsymbol{\mu} \times \boldsymbol{r}_j = 0 \tag{51}$$

or

$$\boldsymbol{f}_j = \frac{\boldsymbol{\lambda} + \boldsymbol{\mu} \times \boldsymbol{r}_j}{\phi(r_{ji})} \tag{52}$$

Maximizing with respect to the langrangian multipliers and substituting this result, yields

$$\begin{cases} \lambda Q_0 + \boldsymbol{r} \times \boldsymbol{Q}_1 = -\Delta\boldsymbol{f}_i \\ \boldsymbol{Q}_1 \times \boldsymbol{\lambda} + \overline{\overline{Q}}_2\boldsymbol{\mu} = \boldsymbol{M} \end{cases} \tag{53}$$

with $Q_0 = \sum_{j\neq i}\frac{1}{\phi(r_{ji})}$, $\boldsymbol{Q}_1 = \sum_{j\neq i}\frac{\boldsymbol{r}_j}{\phi(r_{ji})}$ and the 3-by-3 tensor $\overline{\overline{Q}}_2 = \sum_{j\neq i}\frac{\|\boldsymbol{r}_j\|^2\,\mathrm{I} - \boldsymbol{r}_j\boldsymbol{r}_j}{\phi(r_{ji})}$. Next, consider following translation: $\boldsymbol{r}_j \rightarrow \boldsymbol{r}_j + \boldsymbol{t}$. For the following choice of the translation vector,

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

$$t = -\frac{\mathbf{Q}_1}{Q_0} \tag{54}$$

$\mathbf{Q}_1$ in the new coordinate system will be zero, simplifying the equations. Furthermore, the vector $\boldsymbol{M}$ should now also be calculated in this coordinate system, which is valid since eq 49 has the same form after a translation. Hence,

$$\begin{cases} \lambda = -\dfrac{\Delta f_i}{Q_0} \\ \boldsymbol{\mu} = \overline{\overline{Q}}_2^{-1}\boldsymbol{M} \end{cases} \tag{55}$$

Finally, substituting these in eq 52 gives us the sought for answer. In this work, the following weight function is being used, yielding larger forces for particles closer to the displaced electron pair:

$$\phi(r_{ji}) = \frac{r_{ji}^2}{f_{\text{cutoff}}(r_{ji})} \tag{56}$$

where $f_{\text{cutoff}}(r_{ji})$ is the cutoff function of eq 16. (The inverse of ϕ appears in the solution, such that forces on particles outside the cutoff sphere of electron pair i become zero.)

AUTHOR INFORMATION

Corresponding Author

Toon Verstraelen − *Center for Molecular Modeling (CMM), Ghent University, B-9052 Gent, Belgium;* orcid.org/0000-0001-9288-5608; Email: toon.verstraelen@ugent.be

Authors

Maarten Cools-Ceuppens − *Center for Molecular Modeling (CMM), Ghent University, B-9052 Gent, Belgium*

Joni Dambre − *IDLab, Electronics and Information Systems Department, Ghent University-imec, B-9052 Gent, Belgium*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c00978

Notes

The authors declare no competing financial interest.

Data and software availability: The eQM7 data set is available on the Materials Cloud Archive (10.24435/materialscloud:66-9j). The β-glycine data set is also available on the Materials Cloud Archive (10.24435/materialscloud:jn-44). A reference implementation of the eMLP is available on github at https://github.com/mcoolsce/eMLP and a release is archived on Zenodo (10.5281/zenodo.5526796).

ACKNOWLEDGMENTS

This work is supported by the Fund for Scientific Research Flanders (FWO, Grant No. 11D0420N). The work is furthermore supported by the Research Board of Ghent University (BOF). The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO, and the Flemish Government−department EWI.

REFERENCES

(1) Warshel, A.; Kato, M.; Pisliakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **2007**, *3*, 2034−2045.

(2) Harrison, J. A.; Schall, J. D.; Maskey, S.; Mikulski, P. T.; Knippenberg, M. T.; Morrow, B. H. Review of force fields and intermolecular potentials used in atomistic computational materials research. *Appl. Phys. Rev.* **2018**, *5*, 031104.

(3) Dykstra, C. E. Electrostatic interaction potentials in molecular force fields. *Chem. Rev.* **1993**, *93*, 2339−2353.

(4) Neves-Petersen, M. T.; Petersen, S. B. Protein electrostatics: A review of the equations and methods used to model electrostatic equations in biomolecules − Applications in biotechnology. *Biotechnology Annual Review; Elsevier* **2003**, *9*, 315−395.

(5) Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J. Polarization effects in molecular mechanical force fields. *J. Phys. Cond. Mater.* **2009**, *21*, 333102.

(6) Jing, Z.; Liu, C.; Cheng, S. Y.; Qi, R.; Walker, B. D.; Piquemal, J.-P.; Ren, P. Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annu. Rev. Biophys.* **2019**, *48*, 371−394.

(7) Rappe, A. K.; Goddard, W. A. Charge equilibration for molecular dynamics simulations. *J. Phys. Chem.* **1991**, *95*, 3358−3363.

(8) Resta, R.; Vanderbilt, D. *Physics of Ferroelectrics: A Modern Perspective*; Springer: Berlin, Heidelberg, 2007; pp 31−68.

(9) Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell, A. D. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **2016**, *116*, 4983−5013.

(10) Thole, B. Molecular polarizabilities calculated with a modified dipole interaction. *Chem. Phys.* **1981**, *59*, 341−350.

(11) Caldwell, J.; Dang, L. X.; Kollman, P. A. Implementation of nonadditive intermolecular potentials by use of molecular dynamics: development of a water-water potential and water-ion cluster interactions. *J. Am. Chem. Soc.* **1990**, *112*, 9144−9147.

(12) Mortier, W. J.; Ghosh, S. K.; Shankar, S. Electronegativity-equalization method for the calculation of atomic charges in molecules. *J. Am. Chem. Soc.* **1986**, *108*, 4315−4320.

(13) Patel, S.; Brooks, C. L., III CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.* **2004**, *25*, 1−16.

(14) Metcalf, D. P.; Jiang, A.; Spronk, S. A.; Cheney, D. L.; Sherrill, C. D. Electron-Passing Neural Networks for Atomic Charge Prediction in Systems with Arbitrary Molecular Charge. *J. Chem. Inf. Mod.* **2021**, *61*, 115−122.

(15) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **2021**, *12*, 398.

(16) Bai, C.; Kale, S.; Herzfeld, J. Chemistry with semi-classical electrons: reaction trajectories auto-generated by sub-atomistic force fields. *Chem. Sci.* **2017**, *8*, 4203−4210.

(17) Kale, S.; Herzfeld, J. Natural polarizability and flexibility via explicit valency: The case of water. *J. Chem. Phys.* **2012**, *136*, 084109.

(18) Kale, S.; Herzfeld, J.; Dai, S.; Blank, M. Lewis-inspired representation of dissociable water in clusters and Grotthuss chains. *J. Biol. Phys.* **2012**, *38*, 49−59.

(19) Ekesan, S.; Kale, S.; Herzfeld, J. Transferable pseudoclassical electrons for aufbau of atomic ions. *J. Comput. Chem.* **2014**, *35*, 1159−1164.

(20) Ekesan, S.; Herzfeld, J. Pointillist rendering of electron charge and spin density suffices to replicate trends in atomic properties. *P. R. Soc. A Math. Phys.* **2015**, *471*, 20150370.

(21) Ekesan, S.; Lin, D. Y.; Herzfeld, J. Magnetism and Bond Order in Diatomic Molecules Described by Semiclassical Electrons. *J. Phys. Chem. B* **2016**, *120*, 6264−6269.

(22) Nistor, R. A.; Polihronov, J. G.; Müser, M. H.; Mosey, N. J. A generalization of the charge equilibration method for nonmetallic materials. *J. Chem. Phys.* **2006**, *125*, 094108.

(23) Lee Warren, G.; Davis, J. E.; Patel, S. Origin and control of superlinear polarizability scaling in chemical potential equalization methods. *J. Chem. Phys.* **2008**, *128*, 144110.

(24) Nistor, R. A.; Müser, M. H. Dielectric properties of solids in the regular and split-charge equilibration formalisms. *Phys. Rev. B* **2009**, *79*, 104303.
(25) Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. *Phys. Rev. B* **2015**, *92*, 045131.
(26) Faraji, S.; Ghasemi, S. A.; Rostami, S.; Rasoulkhani, R.; Schaefer, B.; Goedecker, S.; Amsler, M. High accuracy and transferability of a neural network potential through charge equilibration for calcium fluoride. *Phys. Rev. B* **2017**, *95*, 104105.
(27) Xie, X.; Persson, K. A.; Small, D. W. Incorporating Electronic Information into Machine Learning Potential Energy Surfaces via Approaching the Ground-State Electronic Energy as a Function of Atom-Based Electronic Populations. *J. Chem. Theory Comput.* **2020**, *16*, 4256−4270.
(28) Spaldin, N. A. A beginner's guide to the modern theory of polarization. *J. Solid. State Chem.* **2012**, *195*, 2−10.
(29) Su, J. T.; Goddard, W. A. Excited Electron Dynamics Modeling of Warm Dense Matter. *Phys. Rev. Lett.* **2007**, *99*, 185003.
(30) Su, J. T.; Goddard, W. A. The dynamics of highly excited electronic systems: Applications of the electron force field. *J. Chem. Phys.* **2009**, *131*, 244501.
(31) Xiao, H.; Jaramillo-Botero, A.; Theofanis, P. L.; Goddard, W. A. Non-adiabatic dynamics modeling framework for materials in extreme conditions. *Mech. Mater.* **2015**, *90*, 243−252. Proceedings of the IUTAM Symposium on Micromechanics of Defects in Solids.
(32) Theofanis, P. L.; Jaramillo-Botero, A.; Goddard, W. A.; Xiao, H. Nonadiabatic Study of Dynamic Electronic Effects during Brittle Fracture of Silicon. *Phys. Rev. Lett.* **2012**, *108*, 045501.
(33) Islam, M. M.; Kolesov, G.; Verstraelen, T.; Kaxiras, E.; van Duin, A. C. T. eReaxFF: A Pseudoclassical Treatment of Explicit Electrons within Reactive Force Field Simulations. *J. Chem. Theory Comput.* **2016**, *12*, 3463−3472.
(34) Leven, I.; Hao, H.; Das, A. K.; Head-Gordon, T. A Reactive Force Field with Coarse-Grained Electrons for Liquid Water. *J. Phys. Chem. Lett.* **2020**, *11*, 9240−9247.
(35) Leven, I.; Head-Gordon, T. C-GeM: Coarse-Grained Electron Model for Predicting the Electrostatic Potential in Molecules. *J. Phys. Chem. Lett.* **2019**, *10*, 6820−6826.
(36) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133−A1138.
(37) Herzfeld, J.; Ekesan, S. Exchange potentials for semi-classical electrons. *Phys. Chem. Chem. Phys.* **2016**, *18*, 30748−30753.
(38) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem. Rev.* **2021**, *121*, 10142−10186.
(39) Huang, B.; von Lilienfeld, O. A. Ab Initio Machine Learning in Chemical Compound Space. *Chem. Rev.* **2021**, *121*, 10001−10036.
(40) Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **2021**, *121*, 10037−10072.
(41) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(42) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.
(43) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(44) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proceedings of the 34th International Conference on Machine Learning* **2017**, 1263−1272.
(45) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet − A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(46) Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **2019**, *5*, No. 6490, DOI: 10.1126/sciadv.aav6490.
(47) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **2018**, *148*, 241715.
(48) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678−3693.
(49) Klicpera, J.; Giri, S.; Margraf, J. T.; Günnemann, S. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. *arXiv* (Computer Science, Machine Learning) 2011.14115 ver. 2, November 28, 2020; https://arxiv.org/abs/2011.14115 (accessed 2022-02-02).
(50) Schütt, K.; Unke, O.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *Proceedings of the 38th International Conference on Machine Learning* **2021**, 9377−9388.
(51) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887.
(52) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(53) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.
(54) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **2020**, *152*, 044107.
(55) Christensen, A. S.; Faber, F. A.; von Lilienfeld, O. A. Operators in quantum machine learning: Response properties in chemical space. *J. Chem. Phys.* **2019**, *150*, 064105.
(56) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K.-R. SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15*, 448−455.
(57) Vassilev-Galindo, V.; Fonseca, G.; Poltavsky, I.; Tkatchenko, A. Challenges for machine learning force fields in reproducing potential energy surfaces of flexible molecules. *J. Chem. Phys.* **2021**, *154*, 094119.
(58) Zaverkin, V.; Kästner, J. Gaussian Moments as Physically Inspired Molecular Descriptors for Accurate and Scalable Machine Learning Potentials. *J. Chem. Theory Comput.* **2020**, *16*, 5410−5421.
(59) Zubatyuk, R.; Smith, J. S.; Nebgen, B. T.; Tretiak, S.; Isayev, O. Teaching a neural network to attach and detach electrons from molecules. *Nat. Commun.* **2021**, *12*, 4870.
(60) Unke, O. T.; Chmiela, S.; Gastegger, M.; Schütt, K. T.; Sauceda, H. E.; Müller, K.-R. SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun.* **2021**, *12*, 7273.
(61) Gastegger, M.; Schütt, K. T.; Müller, K.-R. Machine learning of solvent effects on molecular spectra and reactions. *Chem. Sci.* **2021**, *12*, 11473−11483.
(62) Grisafi, A.; Nigam, J.; Ceriotti, M. Multi-scale approach for the prediction of atomic scale properties. *Chem. Sci.* **2021**, *12*, 2078−2090.
(63) Guerin, S.; Tofail, S. A. M.; Thompson, D. Organic piezoelectric materials: milestones and potential. *NPG Asia Mater.* **2019**, *11*, 10.
(64) Politzer, P.; Murray, J. S.; Clark, T. Mathematical modeling and physical reality in noncovalent interactions. *J. Mol. Model.* **2015**, *21*, 52.
(65) Foster, J. M.; Boys, S. F. Canonical Configurational Interaction Procedure. *Rev. Mod. Phys.* **1960**, *32*, 300−302.
(66) Marzari, N.; Mostofi, A. A.; Yates, J. R.; Souza, I.; Vanderbilt, D. Maximally localized Wannier functions: Theory and applications. *Rev. Mod. Phys.* **2012**, *84*, 1419−1475.
(67) Marzari, N.; Vanderbilt, D. Maximally localized generalized Wannier functions for composite energy bands. *Phys. Rev. B* **1997**, *56*, 12847−12865.

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

(68) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(69) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.
(70) Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew−Burke−Ernzerhof exchange-correlation functional. *J. Chem. Phys.* **1999**, *110*, 5029−5036.
(71) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.
(72) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796−6806.
(73) Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(74) Parrish, R. M.; Burns, L. A.; Smith, D. G. A.; Simmonett, A. C.; DePrince, A. E.; Hohenstein, E. G.; Bozkaya, U.; Sokolov, A. Y.; Di Remigio, R.; Richard, R. M.; Gonthier, J. F.; James, A. M.; McAlexander, H. R.; Kumar, A.; Saitow, M.; Wang, X.; Pritchard, B. P.; Verma, P.; Schaefer, H. F.; Patkowski, K.; King, R. A.; Valeev, E. F.; Evangelista, F. A.; Turney, J. M.; Crawford, T. D.; Sherrill, C. D. Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.* **2017**, *13*, 3185−3197.
(75) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, 134.
(76) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **2017**, *4*, 170193.
(77) Behler, J.; Csányi, G. Machine learning potentials for extended systems: a perspective. *Eur. Phys. J. B* **2021**, *94*, 142.
(78) Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L; Cococcioni, M.; Dabo, I.; Dal Corso, A.; de Gironcoli, S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Kokalj, A.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Paolini, S.; Pasquarello, A.; Paulatto, L.; Sbraccia, C.; Scandolo, S.; Sclauzero, G.; Seitsonen, A. P; Smogunov, A.; Umari, P.; Wentzcovitch, R. M QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Cond. Mater.* **2009**, *21*, 395502.
(79) Giannozzi, P; Andreussi, O; Brumme, T; Bunau, O; Buongiorno Nardelli, M; Calandra, M; Car, R; Cavazzoni, C; Ceresoli, D; Cococcioni, M; Colonna, N; Carnimeo, I; Dal Corso, A; de Gironcoli, S; Delugas, P; DiStasio, R A; Ferretti, A; Floris, A; Fratesi, G; Fugallo, G; Gebauer, R; Gerstmann, U; Giustino, F; Gorni, T; Jia, J; Kawamura, M; Ko, H-Y; Kokalj, A; Kucukbenli, E; Lazzeri, M; Marsili, M; Marzari, N; Mauri, F; Nguyen, N L; Nguyen, H-V; Otero-de-la-Roza, A; Paulatto, L; Ponce, S; Rocca, D; Sabatini, R; Santra, B; Schlipf, M; Seitsonen, A P; Smogunov, A; Timrov, I; Thonhauser, T; Umari, P; Vast, N; Wu, X; Baroni, S Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys. Cond. Mater.* **2017**, *29*, 465901.
(80) Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **1996**, *105*, 9982−9985.
(81) Vanderbilt, D. Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Phys. Rev. B* **1990**, *41*, 7892−7895.
(82) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(83) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.
(84) Monkhorst, H. J.; Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **1976**, *13*, 5188−5192.
(85) Bernasconi, M.; Chiarotti, G.; Focher, P.; Scandolo, S.; Tosatti, E.; Parrinello, M. First-principle-constant pressure molecular dynamics. *J. Phys. Chem. Solids* **1995**, *56*, 501−505. Proceedings of the sixth International Conference on High Pressure Semiconductor Physics.
(86) Pizzi, G.; Vitale, V.; Arita, R.; Blügel, S.; Freimuth, F.; Géranton, G.; Gibertini, M.; Gresch, D.; Johnson, C.; Koretsune, T.; Ibañez-Azpiroz, J.; Lee, H.; Lihm, J.-M.; Marchand, D.; Marrazzo, A.; Mokrousov, Y.; Mustafa, J. I.; Nohara, Y.; Nomura, Y.; Paulatto, L.; Poncé, S.; Ponweiser, T.; Qiao, J.; Thöle, F.; Tsirkin, S. S.; Wierzbowska, M.; Marzari, N.; Vanderbilt, D.; Souza, I.; Mostofi, A. A.; Yates, J. R. Wannier90 as a community code: new features and applications. *J. Phys. Cond. Mater.* **2020**, *32*, 165902.
(87) Shorten, C.; Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60.
(88) Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations, Calgary, Canada, April 14, 2014.
(89) Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Red Hook: NY, USA, 2017; pp 992−1002.
(90) *Numerical Optimization*; Springer: New York, NY, 2006; pp 135−163.
(91) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; et al. SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261−272.
(92) Veit, M.; Wilkins, D. M.; Yang, Y.; DiStasio, R. A.; Ceriotti, M. Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles. *J. Chem. Phys.* **2020**, *153*, 024113.
(93) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 3401−3406.
(94) Verstraelen, T.; Vanduyfhuys, L.; Vandenbrande, S.; Rogge, S. M. J. Y. Yet another force field. https://github.com/molmod/yaff (accessed 2022-02-02).
(95) Martyna, G. J.; Klein, M. L.; Tuckerman, M. Nosé−Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* **1992**, *97*, 2635−2643.
(96) Zhu, C.; Byrd, R. H.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Trans. Math. Softw.* **1997**, *23*, 550−560.
(97) Dovesi, R.; Kirtman, B.; Maschio, L.; Maul, J.; Pascale, F.; Rérat, M. Calculation of the Infrared Intensity of Crystalline Systems. A Comparison of Three Strategies Based on Berry Phase, Wannier Function, and Coupled-Perturbed Kohn−Sham Methods. *J. Phys. Chem. C* **2019**, *123*, 8336−8346.
(98) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015; https://www.tensorflow.org, Software available from tensorflow.org (accessed 2022-02-02).

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

(99) Qin, Q.-H. *Advanced Mechanics of Piezoelectricity*; Springer: Berlin, Heidelberg, 2013; pp 1−19.
(100) Frenkel, D.; Smit, B. In *Understanding Molecular Simulation*, 2nd ed.; Frenkel, D., Smit, B., Eds.; Academic Press: San Diego, 2002; pp 291−320.
(101) Guerin, S.; Stapleton, A.; Chovan, D.; Mouras, R.; Gleeson, M.; McKeown, C.; Noor, M. R.; Silien, C.; Rhen, F. M. F.; Kholkin, A. L.; Liu, N.; Soulimane, T.; Tofail, S. A. M.; Thompson, D. Control of piezoelectricity in amino acids by supramolecular packing. *Nat. Mater.* **2018**, *17*, 180−186.
(102) Woodcock, H. L.; Zheng, W.; Ghysels, A.; Shao, Y.; Kong, J.; Brooks, B. R. Vibrational subsystem analysis: A method for probing free energies and correlations in the harmonic limit. *J. Chem. Phys.* **2008**, *129*, 214109.
(103) Vanderbilt, D. Berry-phase theory of proper piezoelectric response. *J. Phys. Chem. Solids* **2000**, *61*, 147−151.

https://doi.org/10.1021/acs.jctc.1c00978
J. Chem. Theory Comput. 2022, 18, 1672−1691

B

List of Publications

Publications in international peer–reviewed journals

1. **Quantum free energy profiles for molecular proton transfers**
   Aran Lamaire, Maarten Cools-Ceuppens, Massimo Bocus, Toon Verstraelen, and Veronique Van Speybroeck
   J. Chem. Theory Comput., **2022**, *in revision*

2. **On the impact of nuclear quantum effects on zeolite proton hopping kinetics through machine learning potentials and path integral molecular dynamics simulations**
   Massimo Bocus, Ruben Goeminne, Aran Lamaire, Maarten Cools-Ceuppens, Toon Verstraelen, and Veronique Van Speybroeck
   Nat. Commun., **2022**, *in revision*

3. **Machine Learning Potentials for Metal-Organic Frameworks with Thermodynamic Transferability**
   Sander Vandenhaute, Maarten Cools-Ceuppens, Toon Verstraelen, and Veronique Van Speybroeck
   NPJ Comput. Mater., **2022**, *in revision*

4. **Modeling Electronic Response Properties with an Explicit-Electron Machine Learning Potential**
   Maarten Cools-Ceuppens, Joni Dambre, and Toon Verstraelen
   J. Chem. Theory Comput., **2022**, *18*, 1672–1691
   IF: 6.006. Number of citations: 3

5. **IOData: A python library for reading, writing, and converting computational chemistry file formats and generating input files**
   Toon Verstraelen, William Adams, Leila Pujal, Alireza Tehrani, Braden D Kelly, Luis Macaya, Fanwang Meng, Michael Richer, Raymundo Hernández-Esparza, Xiaotian Derrick Yang, Matthew Chan, Taewon David Kim, Maarten Cools-Ceuppens, Valerii Chuiko, Esteban Vöhringer-Martinez, Paul W Ayers, and Farnaz Heidar-Zadeh
   J. Comput. Chem., **2021**, *42*, 458-464
   IF: 3.376. Number of citations: 11

Conference contributions

Oral presentations

1. **The influence of nuclear quantum effects on proton hopping kinetics in the H-SSZ-13 zeolite through ab initio derived machine learning potentials**
   Massimo Bocus, Ruben Goeminne, Aran Lamaire, Maarten Cools-Ceuppens, Toon Verstraelen, and Veronique Van Speybroeck
   NCCC XXIII, Noordwijkerhout, The Netherlands, May 9 – May 11, 2022

2. **The eMLP: a novel machine learning potential to model electronic properties with explicit-electrons**
   Maarten Cools-Ceuppens, Joni Dambre, and Toon Verstraelen
   AutoCheMo International Reactive Force Field Workshop, Ghent, Belgium, December 8 – December 9, 2021

3. **Evaluating a linear machine learning force field for aluminium**
   Maarten Cools-Ceuppens, and Toon Verstraelen
   L'intelligence artificielle pour la chimie des matériaux, Paris, France, September 25, 2018

Poster presentations

1. **Comparing Different Machine Learning Force Fields: A Case Study of Aluminium**
   Maarten Cools-Ceuppens, Joni Dambre, and Toon Verstraelen
   MOFSIM 2019, Ghent, Belgium, April 10–12, 2019

2. **Comparing Different Machine Learning Force Fields: A Case Study of Aluminium**
   Maarten Cools-Ceuppens, Joni Dambre, and Toon Verstraelen
   The 34th Winter School in Theoretical Chemistry: Machine Learning, Helsinki, Finland, December 10–13, 2018

3. **Machine learning and materials science: ab initio screening to microstructure analysis**
   Michiel Larmuseau, Maarten Cools-Ceuppens, Michael Sluydts, Toon Verstraelen, and Stefaan Cottenier
   The 34th Winter School in Theoretical Chemistry: Machine Learning, Helsinki, Finland, December 10–13, 2018

Master's thesis

Uncertainty prediction in molecular simulations using ab initio derived force fields

Maarten Cools-Ceuppens

Master's thesis performed at the Center for Molecular Modeling (CMM), Ghent University, 2016–2017

Supervisor: prof. dr. ir. Toon Verstraelen

Counsellors: prof. dr. ir. Louis Vanduyfhuys and dr. ir. Steven Vandenbrande

C

List of Software Packages

In the context of this Ph.D. dissertation, various software packages have been used extensively. Below, an overview of these software packages is provided.

CP2K

CP2K is a quantum chemistry and solid state program. It allows for DFT calculations using a mixed plane wave and atomic basis set. CP2K was adopted in this PhD to perform first-principles calculations on MOFs. More information about this software is available in Ref. 228.

eMLP

The eMLP is Python library developed at the Center for Molecular Modeling and built on top of TensorFlow, to train and run MD simulations with eMLP models. It was extensively used in this PhD dissertation to train and validate all the different eMLP models. The eMLP is available online at https://github.com/mcoolsce/eMLP.

Psi4

Psi4 is a quantum chemistry code in which various methods such as HF, DFT or CCSD(T) are implemented with atomic basis sets. Psi4 was adopted in this PhD to construct the eQM7 data set. More information about this software is available in Ref. 229.

Quantum ESPRESSO

Quantum ESPRESSO is a solid-state DFT code, making use of a plane wave basis set and pseudopotentials. It was extensively used in this PhD dissertation to construct the β-glycine data set. More information about this software is available in Ref. 230.

TensorFlow

Tensorflow is primarily a Python library to train and perform inference of deep neural networks. It can run on multiple GPUs and supports automatic differentiation. Tensorflow was adopted in this PhD dissertation to implement the eMLP. More information about this software is available in Ref. 195.

Yaff

Yaff is a Python library developed at the Center for Molecular Modeling to run MD simulations, geometry optimizations and other tools with conventional force fields. In this PhD dissertation, it was interfaced with the eMLP to perform MD simulations. More information about this software is available at https://github.com/molmod/yaff.

Bibliography

[1] S. Harmand, J. E. Lewis, C. S. Feibel, C. J. Lepre, S. Prat, A. Lenoble, X. Boës, R. L. Quinn, M. Brenet, A. Arroyo, N. Taylor, S. Clément, G. Daver, J.-P. Brugal, L. Leakey, R. A. Mortlock, J. D. Wright, S. Lokorodi, C. Kirwa, D. V. Kent, and H. Roche, "3.3-million-year-old stone tools from lomekwi 3, west turkana, kenya," *Nature*, vol. 521, pp. 310–315, May 2015.

[2] M. Park, J. Ryu, W. Wang, and J. Cho, "Material design and engineering of next-generation flow-battery technologies," *Nature Reviews Materials*, vol. 2, p. 16080, Nov 2016.

[3] D. Stork and S. Zinkle, "Introduction to the special issue on the technical status of materials for a fusion reactor," *Nuclear Fusion*, vol. 57, p. 092001, jun 2017.

[4] S. Berardi, S. Drouet, L. Francàs, C. Gimbert-Suriñach, M. Guttentag, C. Richmond, T. Stoll, and A. Llobet, "Molecular artificial photosynthesis," *Chem. Soc. Rev.*, vol. 43, pp. 7501–7519, 2014.

[5] J. J. de Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, J. Analytis, I. Dabo, D. M. DeLongchamp, G. A. Fiete, G. M. Grason, G. Hautier, Y. Mo, K. Rajan, E. J. Reed, E. Rodriguez, V. Stevanovic, J. Suntivich, K. Thornton, and J.-C. Zhao, "New frontiers for the materials genome initiative," *npj Computational Materials*, vol. 5, p. 41, Apr 2019.

[6] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011002, 2013.

[7] C. Draxl and M. Scheffler, "The NOMAD laboratory: from data sharing to artificial intelligence," *Journal of Physics: Materials*, vol. 2, p. 036001, may 2019.

[8] E. Schrödinger, "An undulatory theory of the mechanics of atoms and molecules," *Phys. Rev.*, vol. 28, pp. 1049–1070, Dec 1926.

[9] P. K. Weiner and P. A. Kollman, "Amber: Assisted model building with energy refinement. a general program for modeling molecules and their interactions," *Journal of Computational Chemistry*, vol. 2, no. 3, pp. 287–303, 1981.

[10] M. Cools-Ceuppens, J. Dambre, and T. Verstraelen, "Modeling electronic response properties with an explicit-electron machine learning potential," *Journal of Chemical Theory and Computation*, vol. 18, no. 3, pp. 1672–1691, 2022. PMID: 35171606.

[11] C. Adamo and V. Barone, "Toward reliable density functional methods without adjustable parameters: The pbe0 model," *The Journal of Chemical Physics*, vol. 110, no. 13, pp. 6158–6170, 1999.

[12] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, "A fifth-order perturbation comparison of electron correlation theories," *Chemical Physics Letters*, vol. 157, no. 6, pp. 479–483, 1989.

[13] "AMBER 16 GPU ACCELERATION SUPPORT." https://ambermd.org/gpus16/benchmarks.htm. Accessed: 10-12-2021.

[14] D. E. Shaw, J. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L.-S. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y.-H. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. B. Schafer, N. Siddique, C. W. Snyder, J. Spengler, P. T. P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang, and C. Young, "Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer," in *SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 41–53, 2014.

[15] S. Vandenhaute, S. M. J. Rogge, and V. Van Speybroeck, "Large-scale molecular dynamics simulations reveal new insights into the phase

transition mechanisms in mil-53(al)," *Frontiers in Chemistry*, vol. 9, p. 699, 2021.

[16] M. I. Zimmerman, J. R. Porter, M. D. Ward, S. Singh, N. Vithani, A. Meller, U. L. Mallimadugula, C. E. Kuhn, J. H. Borowsky, R. P. Wiewiora, M. F. D. Hurley, A. M. Harbison, C. A. Fogarty, J. E. Coffland, E. Fadda, V. A. Voelz, J. D. Chodera, and G. R. Bowman, "Sars-cov-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome," *Nature Chemistry*, vol. 13, pp. 651–659, Jul 2021.

[17] L. Casalino, A. C. Dommer, Z. Gaieb, E. P. Barros, T. Sztain, S.-H. Ahn, A. Trifan, A. Brace, A. T. Bogetti, A. Clyde, H. Ma, H. Lee, M. Turilli, S. Khalid, L. T. Chong, C. Simmerling, D. J. Hardy, J. D. Maia, J. C. Phillips, T. Kurth, A. C. Stern, L. Huang, J. D. McCalpin, M. Tatineni, T. Gibbs, J. E. Stone, S. Jha, A. Ramanathan, and R. E. Amaro, "Ai-driven multiscale simulations illuminate mechanisms of sars-cov-2 spike dynamics," *The International Journal of High Performance Computing Applications*, vol. 35, no. 5, pp. 432–451, 2021.

[18] G. Zhao, J. R. Perilla, E. L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A. M. Gronenborn, K. Schulten, C. Aiken, and P. Zhang, "Mature hiv-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics," *Nature*, vol. 497, pp. 643–646, May 2013.

[19] F. Vitalini, A. S. J. S. Mey, F. Noé, and B. G. Keller, "Dynamic properties of force fields," *The Journal of Chemical Physics*, vol. 142, no. 8, p. 084101, 2015.

[20] K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller, *Machine Learning meets Quantum Physics*. Springer International Publishing, 2020.

[21] J. Behler, "Perspective: Machine learning potentials for atomistic simulations," *The Journal of Chemical Physics*, vol. 145, no. 17, p. 170901, 2016.

[22] B. Huang and O. A. von Lilienfeld, "Ab initio machine learning in chemical compound space," *Chemical Reviews*, vol. 121, no. 16, pp. 10001–10036, 2021. PMID: 34387476.

[23] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine learning force fields," *Chemical Reviews*, vol. 121, no. 16, pp. 10142–10186, 2021. PMID: 33705118.

[24] P. Friederich, F. Häse, J. Proppe, and A. Aspuru-Guzik, "Machine-learned potentials for next-generation matter simulations," *Nature Materials*, vol. 20, pp. 750–761, Jun 2021.

[25] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[26] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science," *APL Materials*, vol. 4, no. 5, p. 053208, 2016.

[27] N. M. BALL and R. J. BRUNNER, "Data mining and machine learning in astronomy," *International Journal of Modern Physics D*, vol. 19, no. 07, pp. 1049–1106, 2010.

[28] N. Shah, S. Engineer, N. Bhagat, H. Chauhan, and M. Shah, "Research trends on the usage of machine learning and artificial intelligence in advertising," *Augmented Human Research*, vol. 5, p. 19, Nov 2020.

[29] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, "Machine learning and the physical sciences," *Rev. Mod. Phys.*, vol. 91, p. 045002, Dec 2019.

[30] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," *Information Fusion*, vol. 50, pp. 71–91, 2019.

[31] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," *Annual Review of Physical Chemistry*, vol. 71, no. 1, pp. 361–390, 2020. PMID: 32092281.

[32] OpenAI, :, C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. d. O. Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang, "Dota 2 with large scale deep reinforcement learning," 2019.

[33] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[34] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[35] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[36] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Phys. Rev. Lett.*, vol. 98, p. 146401, Apr 2007.

[37] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," *The Journal of Chemical Physics*, vol. 134, no. 7, p. 074106, 2011.

[38] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," *Phys. Rev. Lett.*, vol. 104, p. 136403, Apr 2010.

[39] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Phys. Rev. B*, vol. 87, p. 184115, May 2013.

[40] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1: an extensible neural network potential with dft accuracy at force field computational cost," *Chem. Sci.*, vol. 8, pp. 3192–3203, 2017.

[41] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," *Science Advances*, vol. 3, no. 5, p. e1603015, 2017.

[42] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, PMLR, 06–11 Aug 2017.

[43] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," *Nature Communications*, vol. 8, p. 13890, Jan 2017.

[44] V. L. Deringer and G. Csányi, "Machine learning based interatomic potential for amorphous carbon," *Phys. Rev. B*, vol. 95, p. 094203, Mar 2017.

[45] M. R. G. Marques, J. Wolff, C. Steigemann, and M. A. L. Marques, "Neural network force fields for simple metals and semiconductors: construction and application to the calculation of phonons and melting temperatures," *Phys. Chem. Chem. Phys.*, vol. 21, pp. 6506–6516, 2019.

[46] L. Bonati and M. Parrinello, "Silicon liquid structure and crystal nucleation from ab initio deep metadynamics," *Phys. Rev. Lett.*, vol. 121, p. 265701, Dec 2018.

[47] M. Eckhoff and J. Behler, "From molecular fragments to the bulk: Development of a neural network potential for mof-5," *Journal of Chemical Theory and Computation*, vol. 15, no. 6, pp. 3793–3809, 2019. PMID: 31091097.

[48] Y. Yu, W. Zhang, and D. Mei, "Artificial neural network potential for encapsulated platinum clusters in mof-808," *The Journal of Physical Chemistry C*, vol. 126, no. 2, pp. 1204–1214, 2022.

[49] B. C. Sweeny, H. Pan, A. Kassem, J. C. Sawyer, S. G. Ard, N. S. Shuman, A. A. Viggiano, S. Brickel, O. T. Unke, M. Upadhyay, and M. Meuwly, "Thermal activation of methane by mgo+: temperature dependent kinetics, reactive molecular dynamics simulations and statistical modeling," *Phys. Chem. Chem. Phys.*, vol. 22, pp. 8913–8923, 2020.

[50] D. Lu, J. Behler, and J. Li, "Accurate global potential energy surfaces for the h + ch3oh reaction by neural network fitting with permutation invariance," *The Journal of Physical Chemistry A*, vol. 124, no. 28, pp. 5737–5745, 2020. PMID: 32530628.

[51] S. Käser, O. T. Unke, and M. Meuwly, "Isomerization and decomposition reactions of acetaldehyde relevant to atmospheric processes from dynamics simulations on neural network-based potential energy surfaces," *The Journal of Chemical Physics*, vol. 152, no. 21, p. 214304, 2020.

[52] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, "General-purpose machine learning potentials capturing nonlocal charge transfer," *Accounts of Chemical Research*, vol. 54, no. 4, pp. 808–817, 2021. PMID: 33513012.

[53] C. G. Staacke, H. H. Heenen, C. Scheurer, G. Csányi, K. Reuter, and J. T. Margraf, "On the role of long-range electrostatics in machine-learned interatomic potentials for complex battery materials," *ACS Applied Energy Materials*, vol. 4, no. 11, pp. 12562–12569, 2021.

[54] H. Furukawa, K. E. Cordova, M. O'Keeffe, and O. M. Yaghi, "The chemistry and applications of metal-organic frameworks," *Science*, vol. 341, no. 6149, p. 1230444, 2013.

[55] A. G. Slater and A. I. Cooper, "Function-led design of new porous materials," *Science*, vol. 348, no. 6238, p. aaa8075, 2015.

[56] S. M. J. Rogge, A. Bavykina, J. Hajek, H. Garcia, A. I. Olivos-Suarez, A. Sepúlveda-Escribano, A. Vimont, G. Clet, P. Bazin, F. Kapteijn, M. Daturi, E. V. Ramos-Fernandez, F. X. Llabrés i Xamena, V. Van Speybroeck, and J. Gascon, "Metal–organic and covalent organic frameworks as single-site catalysts," *Chem. Soc. Rev.*, vol. 46, pp. 3134–3184, 2017.

[57] C. A. Trickett, A. Helal, B. A. Al-Maythalony, Z. H. Yamani, K. E. Cordova, and O. M. Yaghi, "The chemistry of metal–organic frameworks for co2 capture, regeneration and conversion," *Nature Reviews Materials*, vol. 2, p. 17045, Jul 2017.

[58] R. Zhao, Z. Liang, R. Zou, and Q. Xu, "Metal-organic frameworks for batteries," *Joule*, vol. 2, no. 11, pp. 2235–2259, 2018.

[59] L. E. Kreno, K. Leong, O. K. Farha, M. Allendorf, R. P. Van Duyne, and J. T. Hupp, "Metal–organic framework materials as chemical sensors," *Chemical Reviews*, vol. 112, no. 2, pp. 1105–1125, 2012. PMID: 22070233.

[60] M. Born and R. Oppenheimer, "Zur quantentheorie der molekeln," *Annalen der Physik*, vol. 389, no. 20, pp. 457–484, 1927.

[61] *Quasi-Newton Methods*, pp. 135–163. New York, NY: Springer New York, 2006.

[62] D. Frenkel and B. Smit, "Chapter 4 - molecular dynamics simulations," in *Understanding Molecular Simulation (Second Edition)* (D. Frenkel and B. Smit, eds.), pp. 63–107, San Diego: Academic Press, second edition ed., 2002.

[63] T. E. Markland and M. Ceriotti, "Nuclear quantum effects enter the mainstream," *Nature Reviews Chemistry*, vol. 2, p. 0109, Feb 2018.

[64] A. Lamaire, J. Wieme, S. M. J. Rogge, M. Waroquier, and V. Van Speybroeck, "On the importance of anharmonicities and nuclear quantum effects in modelling the structural properties and thermal expansion of mof-5," *The Journal of Chemical Physics*, vol. 150, no. 9, p. 094503, 2019.

[65] R. G. Parr, D. P. Craig, and I. G. Ross, "Molecular orbital calculations of the lower excited electronic levels of benzene, configuration interaction included," *The Journal of Chemical Physics*, vol. 18, no. 12, pp. 1561–1563, 1950.

[66] D. R. Hartree and W. Hartree, "Self-consistent field, with exchange, for beryllium," *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences*, vol. 150, no. 869, pp. 9–33, 1935.

[67] J. Thijssen, *Computational Physics*. Cambridge University Press, 2 ed., 2007.

[68] R. Ditchfield, W. J. Hehre, and J. A. Pople, "Self-consistent molecular-orbital methods. ix. an extended gaussian-type basis for molecular-orbital studies of organic molecules," *The Journal of Chemical Physics*, vol. 54, no. 2, pp. 724–728, 1971.

[69] 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*, vol. 90, no. 2, pp. 1007–1023, 1989.

[70] R. P. Feynman, "Forces in molecules," *Phys. Rev.*, vol. 56, pp. 340–343, Aug 1939.

[71] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Phys. Rev.*, vol. 136, pp. B864–B871, Nov 1964.

[72] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Phys. Rev.*, vol. 140, pp. A1133–A1138, Nov 1965.

[73] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko, "Density functional theory is straying from the path toward the exact functional," *Science*, vol. 355, no. 6320, pp. 49–52, 2017.

[74] J. P. Perdew and K. Schmidt, "Jacob's ladder of density functional approximations for the exchange-correlation energy," *AIP Conference Proceedings*, vol. 577, no. 1, pp. 1–20, 2001.

[75] J. P. Perdew, M. Ernzerhof, and K. Burke, "Rationale for mixing exact exchange with density functional approximations," *The Journal of Chemical Physics*, vol. 105, no. 22, pp. 9982–9985, 1996.

[76] M. Ernzerhof and G. E. Scuseria, "Assessment of the perdew–burke–ernzerhof exchange-correlation functional," *The Journal of Chemical Physics*, vol. 110, no. 11, pp. 5029–5036, 1999.

[77] F. London, "Zur theorie und systematik der molekularkräfte," *Zeitschrift für Physik*, vol. 63, pp. 245–279, Mar 1930.

[78] J. Hermann, R. A. DiStasio, and A. Tkatchenko, "First-principles models for van der waals interactions in molecules and materials: Concepts, theory, and applications," *Chemical Reviews*, vol. 117, no. 6, pp. 4714–4758, 2017. PMID: 28272886.

[79] S. Grimme, J. Antony, S. Ehrlich, and 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*, vol. 132, no. 15, p. 154104, 2010.

[80] E. Artacho, D. Sánchez-Portal, P. Ordejón, A. García, and J. M. Soler, "Linear-scaling ab-initio calculations for large and complex systems," *physica status solidi (b)*, vol. 215, no. 1, pp. 809–817, 1999.

[81] J. C. A. Prentice, J. Aarons, J. C. Womack, A. E. A. Allen, L. Andrinopoulos, L. Anton, R. A. Bell, A. Bhandari, G. A. Bramley, R. J. Charlton, R. J. Clements, D. J. Cole, G. Constantinescu, F. Corsetti, S. M.-M. Dubois, K. K. B. Duff, J. M. Escartín, A. Greco, Q. Hill, L. P. Lee, E. Linscott, D. D. O'Regan, M. J. S. Phipps, L. E. Ratcliff, A. R. Serrano, E. W. Tait, G. Teobaldi, V. Vitale, N. Yeung, T. J. Zuehlsdorff, J. Dziedzic, P. D. Haynes, N. D. M. Hine, A. A. Mostofi, M. C. Payne, and C.-K. Skylaris, "The onetep linear-scaling density functional theory program," *The Journal of Chemical Physics*, vol. 152, no. 17, p. 174111, 2020.

[82] C. Møller and M. S. Plesset, "Note on an approximation treatment for many-electron systems," *Phys. Rev.*, vol. 46, pp. 618–622, Oct 1934.

[83] C. David Sherrill and H. F. Schaefer, "The configuration interaction method: Advances in highly correlated approaches," vol. 34 of *Advances in Quantum Chemistry*, pp. 143–269, Academic Press, 1999.

[84] J. A. Harrison, J. D. Schall, S. Maskey, P. T. Mikulski, M. T. Knippenberg, and B. H. Morrow, "Review of force fields and intermolecular potentials used in atomistic computational materials research," *Applied Physics Reviews*, vol. 5, no. 3, p. 031104, 2018.

[85] A. D. Mackerell Jr., "Empirical force fields for biological macromolecules: Overview and issues," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1584–1604, 2004.

[86] R. S. Mulliken, "Electronic population analysis on lcao–mo molecular wave functions. ii. overlap populations, bond orders, and covalent bond energies," *The Journal of Chemical Physics*, vol. 23, no. 10, pp. 1841–1846, 1955.

[87] C. I. Bayly, P. Cieplak, W. Cornell, and 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*, vol. 97, no. 40, pp. 10269–10280, 1993.

[88] F. L. Hirshfeld, "Bonded-atom fragments for describing molecular charge densities," *Theoretica chimica acta*, vol. 44, pp. 129–138, Jun 1977.

[89] T. Verstraelen, S. Vandenbrande, F. Heidar-Zadeh, L. Vanduyfhuys, V. Van Speybroeck, M. Waroquier, and P. W. Ayers, "Minimal basis iterative stockholder: Atoms in molecules for force-field development," *Journal of Chemical Theory and Computation*, vol. 12, no. 8, pp. 3894–3912, 2016. PMID: 27385073.

[90] F. J. Dyson, "Ground-state energy of a finite system of charged particles," *Journal of Mathematical Physics*, vol. 8, no. 8, pp. 1538–1545, 1967.

[91] R. A. Buckingham and J. E. Lennard-Jones, "The classical equation of state of gaseous helium, neon and argon," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 168, no. 933, pp. 264–283, 1938.

[92] S. Grimme, "A general quantum mechanically derived force field (qmdff) for molecules and condensed phase simulations," *Journal of Chemical Theory and Computation*, vol. 10, no. 10, pp. 4497–4514, 2014. PMID: 26588146.

[93] S. Bureekaew, S. Amirjalayer, M. Tafipolsky, C. Spickermann, T. K. Roy, and R. Schmid, "Mof-ff – a flexible first-principles derived force field for metal-organic frameworks," *physica status solidi (b)*, vol. 250, no. 6, pp. 1128–1141, 2013.

[94] L. Vanduyfhuys, S. Vandenbrande, T. Verstraelen, R. Schmid, M. Waroquier, and V. Van Speybroeck, "Quickff: A program for a quick and easy derivation of force fields for metal-organic frameworks from ab initio input," *Journal of Computational Chemistry*, vol. 36, no. 13, pp. 1015–1027, 2015.

[95] L. Vanduyfhuys, S. Vandenbrande, J. Wieme, M. Waroquier, T. Verstraelen, and V. Van Speybroeck, "Extension of the quickff force field protocol for an improved accuracy of structural, vibrational, mechanical and thermal properties of metal–organic frameworks," *Journal of Computational Chemistry*, vol. 39, no. 16, pp. 999–1011, 2018.

[96] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, "Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations," *Journal of the American Chemical Society*, vol. 114, no. 25, pp. 10024–10035, 1992.

[97] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," *Journal of Computational Chemistry*, vol. 25, no. 9, pp. 1157–1174, 2004.

[98] D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, M. K. Gilson, and P. K. Eastman, "Escaping atom types in force fields using direct chemical perception," *Journal of Chemical Theory and Computation*, vol. 14, no. 11, pp. 6076–6092, 2018. PMID: 30351006.

[99] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell Jr., "Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields," *Journal of Computational Chemistry*, vol. 31, no. 4, pp. 671–690, 2010.

[100] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "Charmm: The biomolecular simulation program," *Journal of Computational Chemistry*, vol. 30, no. 10, pp. 1545–1614, 2009.

[101] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "Charmm: A program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, no. 2, pp. 187–217, 1983.

[102] L. Vanduyfhuys, T. Verstraelen, M. Vandichel, M. Waroquier, and V. Van Speybroeck, "Ab initio parametrized force field for the flexible metal–organic framework mil-53(al)," *Journal of Chemical Theory and Computation*, vol. 8, no. 9, pp. 3217–3231, 2012. PMID: 26605731.

[103] M. S. Daw and M. I. Baskes, "Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals," *Phys. Rev. B*, vol. 29, pp. 6443–6453, Jun 1984.

[104] A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, "Reaxff: A reactive force field for hydrocarbons," *The Journal of Physical Chemistry A*, vol. 105, no. 41, pp. 9396–9409, 2001.

[105] K. ichi Nomura, R. K. Kalia, A. Nakano, and P. Vashishta, "A scalable parallel algorithm for large-scale reactive force-field molecular dynamics simulations," *Computer Physics Communications*, vol. 178, no. 2, pp. 73–87, 2008.

[106] P. Cieplak, F.-Y. Dupradeau, Y. Duan, and J. Wang, "Polarization effects in molecular mechanical force fields," *Journal of Physics: Condensed Matter*, vol. 21, p. 333102, jul 2009.

[107] Z. Jing, C. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J.-P. Piquemal, and P. Ren, "Polarizable force fields for biomolecular simulations: Recent advances and applications," *Annual Review of Biophysics*, vol. 48, no. 1, pp. 371–394, 2019. PMID: 30916997.

[108] A. Warshel, M. Kato, and A. V. Pisliakov, "Polarizable force fields: History, test cases, and prospects," *Journal of Chemical Theory and Computation*, vol. 3, no. 6, pp. 2034–2045, 2007. PMID: 26636199.

[109] P. Cieplak, J. Caldwell, and P. Kollman, "Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and n-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases," *Journal of Computational Chemistry*, vol. 22, no. 10, pp. 1048–1057, 2001.

[110] S. Patel and C. L. Brooks III, "Charmm fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations," *Journal of Computational Chemistry*, vol. 25, no. 1, pp. 1–16, 2004.

[111] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon, "Current status of the amoeba polarizable force field," *The Journal of Physical Chemistry B*, vol. 114, no. 8, pp. 2549–2564, 2010. PMID: 20136072.

[112] S. W. Rick and S. J. Stuart, *Potentials and Algorithms for Incorporating Polarizability in Computer Simulations*, ch. 3, pp. 89–146. John Wiley & Sons, Ltd, 2002.

[113] J. A. Lemkul, J. Huang, B. Roux, and A. D. MacKerell, "An empirical polarizable force field based on the classical drude oscillator model: Development history and recent applications," *Chemical Reviews*, vol. 116, no. 9, pp. 4983–5013, 2016. PMID: 26815602.

[114] B. Thole, "Molecular polarizabilities calculated with a modified dipole interaction," *Chemical Physics*, vol. 59, no. 3, pp. 341–350, 1981.

[115] J. Caldwell, L. X. Dang, and P. A. Kollman, "Implementation of nonadditive intermolecular potentials by use of molecular dynamics: development of a water-water potential and water-ion cluster interactions," *Journal of the American Chemical Society*, vol. 112, no. 25, pp. 9144–9147, 1990.

[116] A. C. Simmonett, F. C. Pickard, J. W. Ponder, and B. R. Brooks, "An empirical extrapolation scheme for efficient treatment of induced dipoles," *The Journal of Chemical Physics*, vol. 145, no. 16, p. 164101, 2016.

[117] F. Aviat, L. Lagardère, and J.-P. Piquemal, "The truncated conjugate gradient (tcg), a non-iterative/fixed-cost strategy for computing polarization in molecular dynamics: Fast evaluation of analytical forces," *The Journal of Chemical Physics*, vol. 147, no. 16, p. 161724, 2017.

[118] W. J. Mortier, S. K. Ghosh, and S. Shankar, "Electronegativity-equalization method for the calculation of atomic charges in molecules," *Journal of the American Chemical Society*, vol. 108, no. 15, pp. 4315–4320, 1986.

[119] A. K. Rappe and W. A. Goddard, "Charge equilibration for molecular dynamics simulations," *The Journal of Physical Chemistry*, vol. 95, no. 8, pp. 3358–3363, 1991.

[120] R. A. Nistor, J. G. Polihronov, M. H. Müser, and N. J. Mosey, "A generalization of the charge equilibration method for nonmetallic materials," *The Journal of Chemical Physics*, vol. 125, no. 9, p. 094108, 2006.

[121] O. T. Unke and M. Meuwly, "Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges," *Journal of Chemical Theory and Computation*, vol. 15, no. 6, pp. 3678–3693, 2019. PMID: 31042390.

[122] X. Xie, K. A. Persson, and D. W. Small, "Incorporating electronic information into machine learning potential energy surfaces via approaching the ground-state electronic energy as a function of atom-based electronic populations," *Journal of Chemical Theory and Computation*, vol. 16, no. 7, pp. 4256–4270, 2020. PMID: 32502350.

[123] D. P. Metcalf, A. Jiang, S. A. Spronk, D. L. Cheney, and C. D. Sherrill, "Electron-passing neural networks for atomic charge prediction in systems with arbitrary molecular charge," *Journal of Chemical Information and Modeling*, vol. 61, no. 1, pp. 115–122, 2021. PMID: 33326247.

[124] R. Zubatyuk, J. S. Smith, B. T. Nebgen, S. Tretiak, and O. Isayev, "Teaching a neural network to attach and detach electrons from molecules," *Nature Communications*, vol. 12, p. 4870, Aug 2021.

[125] S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, "Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network," *Phys. Rev. B*, vol. 92, p. 045131, Jul 2015.

[126] S. Faraji, S. A. Ghasemi, S. Rostami, R. Rasoulkhani, B. Schaefer, S. Goedecker, and M. Amsler, "High accuracy and transferability of a neural network potential through charge equilibration for calcium fluoride," *Phys. Rev. B*, vol. 95, p. 104105, Mar 2017.

[127] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, "A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer," *Nature Communications*, vol. 12, p. 398, Jan 2021.

[128] C. Staacke, S. Wengert, C. Kunkel, G. Csányi, K. Reuter, and J. T. Margraf, "Kernel charge equilibration: Efficient and accurate prediction of molecular dipole moments with a machine-learning enhanced electron density model," *ChemRxiv*, 2021.

[129] J. Behler, "Four generations of high-dimensional neural network potentials," *Chemical Reviews*, vol. 121, no. 16, pp. 10037–10072, 2021. PMID: 33779150.

[130] R. Resta and D. Vanderbilt, *Theory of Polarization: A Modern Approach*, pp. 31–68. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[131] N. A. Spaldin, "A beginner's guide to the modern theory of polarization," *Journal of Solid State Chemistry*, vol. 195, pp. 2–10, 2012. Polar Inorganic Materials: Design Strategies and Functional Properties.

[132] G. Lee Warren, J. E. Davis, and S. Patel, "Origin and control of superlinear polarizability scaling in chemical potential equalization methods," *The Journal of Chemical Physics*, vol. 128, no. 14, p. 144110, 2008.

[133] C. Bai, S. Kale, and J. Herzfeld, "Chemistry with semi-classical electrons: reaction trajectories auto-generated by sub-atomistic force fields," *Chem. Sci.*, vol. 8, pp. 4203–4210, 2017.

[134] S. Ekesan, D. Y. Lin, and J. Herzfeld, "Magnetism and bond order in diatomic molecules described by semiclassical electrons," *The Journal*

of Physical Chemistry B, vol. 120, no. 26, pp. 6264–6269, 2016. PMID: 27197811.

[135] M. M. Islam, G. Kolesov, T. Verstraelen, E. Kaxiras, and A. C. T. van Duin, "ereaxff: A pseudoclassical treatment of explicit electrons within reactive force field simulations," *Journal of Chemical Theory and Computation*, vol. 12, no. 8, pp. 3463–3472, 2016. PMID: 27399177.

[136] M. M. Islam and A. C. T. van Duin, "Reductive decomposition reactions of ethylene carbonate by explicit electron transfer from lithium: An ereaxff molecular dynamics study," *The Journal of Physical Chemistry C*, vol. 120, no. 48, pp. 27128–27134, 2016.

[137] I. Leven and T. Head-Gordon, "C-gem: Coarse-grained electron model for predicting the electrostatic potential in molecules," *The Journal of Physical Chemistry Letters*, vol. 10, no. 21, pp. 6820–6826, 2019. PMID: 31613629.

[138] G. N. Lewis, "The atom and the molecule.," *Journal of the American Chemical Society*, vol. 38, no. 4, pp. 762–785, 1916.

[139] J. Herzfeld and S. Ekesan, "Exchange potentials for semi-classical electrons," *Phys. Chem. Chem. Phys.*, vol. 18, pp. 30748–30753, 2016.

[140] R. Car and M. Parrinello, "Unified approach for molecular dynamics and density-functional theory," *Phys. Rev. Lett.*, vol. 55, pp. 2471–2474, Nov 1985.

[141] J. T. Su and W. A. Goddard, "Excited electron dynamics modeling of warm dense matter," *Phys. Rev. Lett.*, vol. 99, p. 185003, Nov 2007.

[142] J. T. Su and W. A. Goddard, "The dynamics of highly excited electronic systems: Applications of the electron force field," *The Journal of Chemical Physics*, vol. 131, no. 24, p. 244501, 2009.

[143] H. Xiao, A. Jaramillo-Botero, P. L. Theofanis, and W. A. Goddard, "Nonadiabatic dynamics modeling framework for materials in extreme conditions," *Mechanics of Materials*, vol. 90, pp. 243–252, 2015. Proceedings of the IUTAM Symposium on Micromechanics of Defects in Solids.

[144] P. L. Theofanis, A. Jaramillo-Botero, W. A. Goddard, and H. Xiao, "Nonadiabatic study of dynamic electronic effects during brittle fracture of silicon," *Phys. Rev. Lett.*, vol. 108, p. 045501, Jan 2012.

[145] S. Kale and J. Herzfeld, "Natural polarizability and flexibility via explicit valency: The case of water," *The Journal of Chemical Physics*, vol. 136, no. 8, p. 084109, 2012.

[146] S. Kale, J. Herzfeld, S. Dai, and M. Blank, "Lewis-inspired representation of dissociable water in clusters and grotthuss chains," *Journal of Biological Physics*, vol. 38, pp. 49–59, Jan 2012.

[147] S. Ekesan, S. Kale, and J. Herzfeld, "Transferable pseudoclassical electrons for aufbau of atomic ions," *Journal of Computational Chemistry*, vol. 35, no. 15, pp. 1159–1164, 2014.

[148] S. Ekesan and J. Herzfeld, "Pointillist rendering of electron charge and spin density suffices to replicate trends in atomic properties," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 471, no. 2181, p. 20150370, 2015.

[149] I. Leven, H. Hao, A. K. Das, and T. Head-Gordon, "A reactive force field with coarse-grained electrons for liquid water," *The Journal of Physical Chemistry Letters*, vol. 11, no. 21, pp. 9240–9247, 2020. PMID: 33073998.

[150] A. G. Donchev, V. D. Ozrin, M. V. Subbotin, O. V. Tarasov, and V. I. Tarasov, "A quantum mechanical polarizable force field for biomolecular interactions," *Proceedings of the National Academy of Sciences*, vol. 102, no. 22, pp. 7829–7834, 2005.

[151] W. Kohn, "Density functional and density matrix method scaling linearly with the number of atoms," *Phys. Rev. Lett.*, vol. 76, pp. 3168–3171, Apr 1996.

[152] E. Prodan and W. Kohn, "Nearsightedness of electronic matter," *Proceedings of the National Academy of Sciences*, vol. 102, no. 33, pp. 11635–11638, 2005.

[153] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "Schnet – a deep learning architecture for molecules and materials," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241722, 2018.

[154] J. Behler and G. Csányi, "Machine learning potentials for extended systems: a perspective," *The European Physical Journal B*, vol. 94, p. 142, Jul 2021.

[155] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao, "Enhanced sampling in molecular dynamics," *The Journal of Chemical Physics*, vol. 151, no. 7, p. 070902, 2019.

[156] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules," *Scientific Data*, vol. 4, p. 170193, Dec 2017.

[157] J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev, and S. Tretiak, "The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules," *Scientific Data*, vol. 7, p. 134, May 2020.

[158] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241733, 2018.

[159] C. Schran, K. Brezina, and O. Marsalek, "Committee neural network potentials control generalization errors and enable active learning," *The Journal of Chemical Physics*, vol. 153, no. 10, p. 104105, 2020.

[160] G. Sivaraman, A. N. Krishnamoorthy, M. Baur, C. Holm, M. Stan, G. Csányi, C. Benmore, and Á. Vázquez-Mayagoitia, "Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide," *npj Computational Materials*, vol. 6, p. 104, Jul 2020.

[161] K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev, "Machine learning of molecular properties: Locality and active learning," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241727, 2018.

[162] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, (New York, NY, USA), p. 287–294, Association for Computing Machinery, 1992.

[163] Z. Li, J. R. Kermode, and A. De Vita, "Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces," *Phys. Rev. Lett.*, vol. 114, p. 096405, Mar 2015.

[164] R. Jinnouchi, F. Karsai, and G. Kresse, "On-the-fly machine learning force field generation: Application to melting points," *Phys. Rev. B*, vol. 100, p. 014105, Jul 2019.

[165] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Big data meets quantum chemistry approximations: The Δ-machine learning approach," *Journal of Chemical Theory and Computation*, vol. 11, no. 5, pp. 2087–2096, 2015. PMID: 26574412.

[166] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Phys. Rev. Lett.*, vol. 108, p. 058301, Jan 2012.

[167] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," *The Journal of Physical Chemistry Letters*, vol. 6, no. 12, pp. 2326–2331, 2015. PMID: 26113956.

[168] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld, "Fchl revisited: Faster and more accurate quantum machine learning," *The Journal of Chemical Physics*, vol. 152, no. 4, p. 044107, 2020.

[169] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, "wacsf—weighted atom-centered symmetry functions as descriptors in machine learning potentials," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241709, 2018.

[170] A. S. Christensen and O. A. von Lilienfeld, "On the role of gradients for machine learning of molecular energies and forces," *Machine Learning: Science and Technology*, vol. 1, p. 045018, oct 2020.

[171] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, "Towards exact molecular dynamics simulations with machine-learned force fields," *Nature Communications*, vol. 9, p. 3887, Sep 2018.

[172] A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, "Symmetry-adapted machine learning for tensorial properties of atomistic systems," *Phys. Rev. Lett.*, vol. 120, p. 036002, Jan 2018.

[173] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth, and B. Kozinsky, "Learning local equivariant representations for large-scale atomistic dynamics," 2022.

[174] L. Zhang, J. Han, H. Wang, R. Car, and W. E, "Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics," *Phys. Rev. Lett.*, vol. 120, p. 143001, Apr 2018.

[175] V. Zaverkin and J. Kästner, "Gaussian moments as physically inspired molecular descriptors for accurate and scalable machine learning potentials," *Journal of Chemical Theory and Computation*, vol. 16, no. 8, pp. 5410–5421, 2020. PMID: 32672968.

[176] J. Klicpera, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," 2020.

[177] M. Gastegger, K. T. Schütt, and K.-R. Müller, "Machine learning of solvent effects on molecular spectra and reactions," *Chem. Sci.*, vol. 12, pp. 11473–11483, 2021.

[178] J. Klicpera, F. Becker, and S. Günnemann, "Gemnet: Universal directional graph neural networks for molecules," 2021.

[179] N. Lubbers, J. S. Smith, and K. Barros, "Hierarchical modeling of molecular energies using a deep neural network," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241715, 2018.

[180] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," *Nature Communications*, vol. 13, p. 2453, May 2022.

[181] M. Haghighatlari, J. Li, X. Guan, O. Zhang, A. Das, C. J. Stein, F. Heidar-Zadeh, M. Liu, M. Head-Gordon, L. Bertels, H. Hao, I. Leven, and T. Head-Gordon, "Newtonnet: A newtonian message passing network for deep learning of interatomic potentials and forces," 2021.

[182] K. T. Schütt, O. T. Unke, and M. Gastegger, "Equivariant message passing for the prediction of tensorial properties and molecular spectra," 2021.

[183] Y. Shao, M. Hellström, P. D. Mitev, L. Knijff, and C. Zhang, "Pinn: A python library for building atomic neural networks of molecules and materials," *Journal of Chemical Information and Modeling*, vol. 60, no. 3, pp. 1184–1193, 2020. PMID: 31935100.

[184] O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda, and K.-R. Müller, "Spookynet: Learning force fields with electronic degrees of freedom and nonlocal effects," *Nature Communications*, vol. 12, p. 7273, Dec 2021.

[185] Z. Qiao, A. S. Christensen, M. Welborn, F. R. Manby, A. Anandkumar, and T. F. M. I. au2, "Unite: Unitary n-body tensor equivariant network with applications to quantum chemistry," 2021.

[186] R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," *Phys. Rev. B*, vol. 99, p. 014104, Jan 2019.

[187] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," *Multiscale Modeling & Simulation*, vol. 14, no. 3, pp. 1153–1173, 2016.

[188] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials," *Journal of Computational Physics*, vol. 285, pp. 316–330, 2015.

[189] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[190] S. Theodoridis and K. Koutroumbas, "Chapter 4 - nonlinear classifiers," in *Pattern Recognition (Fourth Edition)* (S. Theodoridis and K. Koutroumbas, eds.), pp. 151–260, Boston: Academic Press, fourth edition ed., 2009.

[191] A. S. Christensen, F. A. Faber, and O. A. von Lilienfeld, "Operators in quantum machine learning: Response properties in chemical space," *The Journal of Chemical Physics*, vol. 150, no. 6, p. 064105, 2019.

[192] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, "Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241730, 2018.

[193] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules," *Chemical Reviews*, vol. 121, no. 16, pp. 10073–10141, 2021. PMID: 34398616.

[194] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct 1986.

[195] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org, accessed April 13, 2022.

[196] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner,

L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.

[197] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[198] L. Prechelt, *Early Stopping — But When?*, pp. 53–67. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[199] V. Vassilev-Galindo, G. Fonseca, I. Poltavsky, and A. Tkatchenko, "Challenges for machine learning force fields in reproducing potential energy surfaces of flexible molecules," *The Journal of Chemical Physics*, vol. 154, no. 9, p. 094119, 2021.

[200] W. Pronobis, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Capturing intensive and extensive dft/tddft molecular properties with machine learning," *The European Physical Journal B*, vol. 91, p. 178, Aug 2018.

[201] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, "Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds," 2018.

[202] E. P. Wigner, *On the Matrices Which Reduce the Kronecker Products of Representations of S. R. Groups*, pp. 608–654. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993.

[203] A. Gao and R. C. Remsing, "Self-consistent determination of long-range electrostatics in neural network potentials," *Nature Communications*, vol. 13, p. 1572, Mar 2022.

[204] J. H. Cavka, S. Jakobsen, U. Olsbye, N. Guillou, C. Lamberti, S. Bordiga, and K. P. Lillerud, "A new zirconium inorganic building brick forming metal organic frameworks with exceptional stability," *Journal of the American Chemical Society*, vol. 130, no. 42, pp. 13850–13851, 2008. PMID: 18817383.

[205] T. Loiseau, C. Serre, C. Huguenard, G. Fink, F. Taulelle, M. Henry, T. Bataille, and G. Férey, "A rationale for the large breathing of the porous aluminum terephthalate (mil-53) upon hydration," *Chemistry – A European Journal*, vol. 10, no. 6, pp. 1373–1382, 2004.

[206] S. M. J. Rogge, P. G. Yot, J. Jacobsen, F. Muniz-Miranda, S. Vandenbrande, J. Gosch, V. Ortiz, I. E. Collings, S. Devautour-Vinot, G. Maurin, N. Stock, and V. Van Speybroeck, "Charting the metal-dependent high-pressure stability of bimetallic uio-66 materials," *ACS Materials Letters*, vol. 2, no. 4, pp. 438–445, 2020. PMID: 32296781.

[207] L. Vanduyfhuys, S. M. J. Rogge, J. Wieme, S. Vandenbrande, G. Maurin, M. Waroquier, and V. Van Speybroeck, "Thermodynamic insight into stimuli-responsive behaviour of soft porous crystals," *Nature Communications*, vol. 9, p. 204, Jan 2018.

[208] J. M. Foster and S. F. Boys, "Canonical configurational interaction procedure," *Rev. Mod. Phys.*, vol. 32, pp. 300–302, Apr 1960.

[209] N. Marzari and D. Vanderbilt, "Maximally localized generalized wannier functions for composite energy bands," *Phys. Rev. B*, vol. 56, pp. 12847–12865, Nov 1997.

[210] N. Marzari, A. A. Mostofi, J. R. Yates, I. Souza, and D. Vanderbilt, "Maximally localized wannier functions: Theory and applications," *Rev. Mod. Phys.*, vol. 84, pp. 1419–1475, Oct 2012.

[211] L. Zhang, H. Wang, M. C. Muniz, A. Z. Panagiotopoulos, R. Car, and W. E, "A deep potential model with long-range electrostatic interactions," *The Journal of Chemical Physics*, vol. 156, no. 12, p. 124107, 2022.

[212] T. Verstraelen, L. Vanduyfhuys, S. Vandenbrande, and S. M. J. Rogge, "Yaff, yet another force field." accessed May 25, 2022.

[213] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[214] D. Frenkel and B. Smit, "Chapter 12 - long-range interactions," in *Understanding Molecular Simulation (Second Edition)* (D. Frenkel and B. Smit, eds.), pp. 291–320, San Diego: Academic Press, second edition ed., 2002.

[215] C. Edmiston and K. Ruedenberg, "Localized atomic and molecular orbitals," *Rev. Mod. Phys.*, vol. 35, pp. 457–464, Jul 1963.

[216] J. Pipek and P. G. Mezey, "A fast intrinsic localization procedure applicable for ab initio and semiempirical linear combination of atomic orbital wave functions," *The Journal of Chemical Physics*, vol. 90, no. 9, pp. 4916–4926, 1989.

[217] I.-M. Høyvik, B. Jansik, and P. Jørgensen, "Orbital localization using fourth central moment minimization," *The Journal of Chemical Physics*, vol. 137, no. 22, p. 224114, 2012.

[218] M. Cools-Ceuppens, J. Dambre, and T. Verstraelen, "eQM7: a dataset for small molecules with Foster-Boys centers," *Materials Cloud Archive*, 2021.

[219] S. Guerin, A. Stapleton, D. Chovan, R. Mouras, M. Gleeson, C. McKeown, M. R. Noor, C. Silien, F. M. F. Rhen, A. Kholkin, N. Liu, T. Soulimane, S. A. M. Tofail, and D. Thompson, "Control of piezoelectricity in amino acids by supramolecular packing," *Nature Materials*, vol. 17, pp. 180–186, Feb 2018.

[220] S. Guerin, S. A. M. Tofail, and D. Thompson, "Organic piezoelectric materials: milestones and potential," *NPG Asia Materials*, vol. 11, p. 10, Mar 2019.

[221] L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi, and N. Marzari, "Materials cloud, a platform for open computational science," *Scientific Data*, vol. 7, p. 299, Sep 2020.

[222] K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "Schnet: A continuous-filter convolutional neural network for modeling quantum interactions," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[223] M. Thomas, M. Brehm, R. Fligg, P. Vöhringer, and B. Kirchner, "Computing vibrational spectra from ab initio molecular dynamics," *Phys. Chem. Chem. Phys.*, vol. 15, pp. 6608–6622, 2013.

[224] M. Cools-Ceuppens, J. Dambre, and T. Verstraelen, "A dataset for beta-glycine with Wannier centers," *Materials Cloud Archive*, 2021.

[225] Q.-H. Qin, *Introduction to Piezoelectricity*, pp. 1–19. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[226] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, p. 60, Jul 2019.

[227] C. Zhu, R. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on Mathematical Software*, vol. 23, pp. 550–560, Dec. 1997.

[228] T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borštnik, M. Taillefumier, A. S. Jakobovits, A. Lazzaro, H. Pabst, T. Müller, R. Schade, M. Guidon, S. Andermatt, N. Holmberg, G. K. Schenter, A. Hehn, A. Bussy, F. Belleflamme, G. Tabacchi, A. Glöß, M. Lass, I. Bethune, C. J. Mundy, C. Plessl, M. Watkins, J. VandeVondele, M. Krack, and J. Hutter, "Cp2k: An electronic structure and molecular dynamics software package - quickstep: Efficient and accurate electronic structure calculations," *The Journal of Chemical Physics*, vol. 152, no. 19, p. 194103, 2020.

[229] D. G. A. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O'Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer, A. Y. Sokolov, K. Patkowski, A. E. DePrince, U. Bozkaya, R. A. King, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill, "Psi4 1.4: Open-source software for high-throughput quantum chemistry," *The Journal of Chemical Physics*, vol. 152, no. 18, p. 184108, 2020.

[230] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. D. Corso, S. de Gironcoli, P. Delugas, R. A. DiStasio, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Küçükbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H.-V. Nguyen, A. O. de-la Roza, L. Paulatto, S. Poncé, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu, and S. Baroni, "Advanced capabilities for materials modelling with quantum ESPRESSO," *Journal of Physics: Condensed Matter*, vol. 29, p. 465901, oct 2017.

This research was supported by the Research Foundation Flanders (FWO) through a personal mandate (Grant No. 11D0418N and 11D0420N).

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, and FWO.