

*Ethan Berger*

POLARIZABILITY MODELS  
FOR SIMULATING  
RAMAN SPECTRA WITH  
MOLECULAR DYNAMICS

UNIVERSITY OF OULU GRADUATE SCHOOL;  
UNIVERSITY OF OULU,  
FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING





ACTA UNIVERSITATIS OULUENSIS  
C Technica 966

*ETHAN BERGER*

**POLARIZABILITY MODELS FOR  
SIMULATING RAMAN SPECTRA  
WITH MOLECULAR DYNAMICS**

Academic dissertation to be presented with the assent of  
the Doctoral Programme Committee of Information  
Technology and Electrical Engineering of the University of  
Oulu for public defence in Auditorium Lo124, Linnanmaa,  
on 11 October 2024, at 12 noon

UNIVERSITY OF OULU, OULU 2024

Copyright © 2024  
Acta Univ. Oul. C 966, 2024

Supervised by  
Assistant Professor Hannu-Pekka Komsa  
Professor Krisztian Kordas

Reviewed by  
Professor David Egger  
Docent Johannes Niskanen

Opponent  
Professor Vincent Meunier

ISBN 978-952-62-4224-8 (Paperback)  
ISBN 978-952-62-4225-5 (PDF)

ISSN 0355-3213 (Printed)  
ISSN 1796-2226 (Online)

Cover Design  
Raimo Ahonen

PUNAMUSTA  
TAMPERE 2024

## **Berger, Ethan, Polarizability Models for Simulating Raman Spectra with Molecular Dynamics**

University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering

*Acta Univ. Oul. C 966, 2024*

University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

### ***Abstract***

Raman spectroscopy is a versatile and very popular method to study the vibrational properties of molecules and crystals. Simulations of Raman spectra can bring additional insight to experimental results, although such calculations are usually quite limited due to their large computational cost. In order to accelerate simulations, models can be used to predict properties of interest, namely forces and polarizabilities in the case of Raman spectroscopy. For the former, machine learning force fields and classical force fields are well established and commonly used in various applications. On the other hand, the development of polarizability models is still ongoing.

In this thesis, various polarizability models are presented and compared, with a particular interest in their application to Raman spectroscopy. The models range from simple empirical bond models (BPM) to state-of-the-art machine learning algorithms (SA-GPR and TNEP), and also include a newly developed method relying on the projection of vibrational modes (RGDOS). The accuracy of BPM, RGDOS and SA-GPR is first assessed using simple materials, namely boron arsenide (BAs) and molybdenum disulfide ( $\text{MoS}_2$ ). These two materials are also used to highlight the advantages and shortcomings of all models when used to obtain Raman spectra.

The best performing models are then applied to study complex materials and molecules, such as titanium carbide MXenes ( $\text{Ti}_3\text{C}_2\text{X}_2$ ), anharmonic halide perovskites ( $\text{CsPbBr}_3$  and  $\text{CsSnBr}_3$ ) and amino acids chains (known as peptides). By combining machine learning force fields in molecular dynamics with efficient polarizability models, we are able to study longer timescales and larger systems than before. As a result, we find good agreement with experimental measurements for all applications. Additionally, access to larger systems also allows for accurate simulations of defects in a  $\text{MoS}_2$  monolayer and the simulation of heterogeneous surfaces in the case of MXenes. Similarly, long MD trajectories enable simulations of Raman spectra at low frequencies, which are found to play an important role in the phase transition of halide perovskites. Finally, the transferability of machine learning models is tested using amino acids and peptide chains. These applications bring further insights into the vibrational properties of complex materials and show that all polarizability models presented here could be utilized depending on the investigated system.

**Keywords:** amino acids, density functional theory, halide perovskites, machine learning, MXenes, polarizability, Raman spectroscopy, transition metal dichalcogenides



## **Berger, Ethan, Polarisoituvuusmallit Raman-spektrien simulointiin molekyyliidynamiikalla**

Oulun yliopiston tutkijakoulu; Oulun yliopisto, tieto- ja sähkötekniikan tiedekunta

*Acta Univ. Oul. C 966, 2024*

Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

### ***Tiivistelmä***

Raman-spektroskopia on monipuolinen ja erittäin suosittu menetelmä molekyylien ja kiteiden värähtelyominaisuuksien tutkimiseen. Raman-spektrien simuloinnit voivat tuoda lisäymmärrystä kokeellisiin tuloksiin, vaikka tallaiset laskelmat ovat yleensä melko rajallisia niiden suurten laskentavaatimusten vuoksi. Simulointien nopeuttamiseksi voidaan käyttää malleja, joiden avulla voidaan ennustaa haluttuja ominaisuuksia, kuten voimia ja polarisaatioita Raman-spektroskopian tapauksessa. Edellisten osalta koneoppimiseen perustuvat voimakentät ja klassiset voimakentät ovat vakiintuneita ja yleisesti käytössä eri sovelluksissa. Toisaalta polarisaatiomallien kehittämiseen liittyy vielä paljon avoimia kysymyksiä.

Tässä väitöskirjassa esitellään ja vertaillaan erilaisia polarisaatiomalleja, ja erityistä huomiota kiinnitetään niiden soveltamiseen Raman-spektroskopiaan. Mallit vaihtelevat yksinkertaisista empiirisistä sidosmalleista (BPM) uusimpiin koneoppimisalgoritmeihin (SA-GPR ja TNEP), ja niihin sisältyy myös hiljattain kehitetty menetelmä, joka perustuu värähtelymoodien projektiioon (RGDOS). BPM:n, RGDOS:n ja SA-GPR:n tarkkuutta arvioidaan ensin yksinkertaisten materiaalien eli booriarsenidin (BAs) ja molybdeenidisulfidin ( $\text{MoS}_2$ ) avulla. Naita kahta materiaalia käytetään myös korostamaan kaikkien mallien etuja ja puutteita Raman-spektreihin sovellettaessa.

Parhaiten toimivia malleja sovelletaan sitten monimutkaisten materiaalien ja molekyylien, kuten titaanikarbidi-MXeenien ( $\text{Ti}_3\text{C}_2\text{X}_2$ ), anharmonisten halidiperovskiittien ( $\text{CsPbBr}_3$  ja  $\text{CsSnBr}_3$ ) ja aminohappoketjujen (niin sanottujen peptidien) tutkimiseen. Yhdistämällä koneoppimiseen perustuvien voimakenttien molekyyliidynamiikkaa tehokkaisiin polarisaatiomalleihin pystymme simuloimaan aiempää pidempää aikaskaalaa ja suurempia järjestelmiä. Tämän ansiosta lasketut spektrit ovat hyvin yhteneväisiä kokeellisten mittausten kanssa kaikissa sovelluksissa. Lisäksi pääsy suurempiin järjestelmiin mahdollistaa myös  $\text{MoS}_2$ :n kidevikojen tarkan simuloinnin ja heterogeenisten pintojen simuloinnin MXeenien tapauksessa. Vastaavasti pitkät MD-simulaatiot mahdollistavat matalille taajuuksille ulottuvien Raman-spektrien simuloinnin, joilla havaitaan olevan tärkeä rooli halidiperovskiittien faasimuutoksissa. Lopuksi testataan koneoppimismallien siirrettävyyttä aminohappojen ja peptidiketjujen avulla. Nämä sovellukset tuovat lisää tietoa monimutkaisten materiaalien värähtelyominaisuuksista ja osoittavat, että kaikki tässä väitöskirjassa esitellyt polarisaatiomallit voivat olla käyttökelpoisia tutkittavasta järjestelmästä riippuen.

*Asiasanat:* aminohapot, halidiperovskiitit, koneoppiminen, MXeenit, polarisaatio, Raman-spektroskopia, siirtymämetallidikalkogenidit, tiheysfunktionaaliteoria



*"Croyez ceux qui cherchent  
la vérité, doutez de ceux qui la trouvent."  
André Gide*



## Acknowledgements

First of all, I am most grateful to my supervisor Prof. Hannu-Pekka Komsa for his continuous trust and guidance throughout my doctorate studies. I also wish to thank my line manager Prof. Krisztian Kordas for his help, as well as Prof. Matti Alatalo, Prof. Wei Cao and Dr. Anssi Mäkynen for acting as my follow-up group. Additionally, I would like to acknowledge Prof. Jean-Philippe Ansermet and Prof. Alfredo Pasquarello from EPFL for their support in my earlier studies and encouraging me to pursue an academic career.

I am thankful to all the members of the Microelectronics Research Unit, and more specifically Prof. Heli Jantunen and Prof. Jari Juuti for making it a stimulating research environment. Thank you to all the MIC PhD students for the interesting discussions during our researchers' coffee meetings, as well as the invited speakers who agreed to share their academic journey with us. This thesis would not have been possible without the help of my collaborators Juha Niemelä, Outi Lampela, André Juffer and Zhong-Peng Lv, all of whom I wish to warmly thank.

I had the chance of making amazing friends who kept me sane when research was tough. Special thanks to Marcus for the numerous chess games, to Suhas for keeping me fit and to Mohammad for handling my grocery shopping (among many other things).

None of this would have been possible without the constant support of my parents, my siblings, my grandparents and the rest of my family.

Oulu, September 2024

Ethan Berger



## List of abbreviations and symbols

a.u.	Atomic units
BPM	Bond polarizability model
CEE	Committee error estimate
DFPT	Density functional perturbation theory
DFT	Density functional theory
GPR	Gaussian process regressor
IR	Infrared (spectroscopy)
MD	Molecular dynamics
ML	Machine learning
MLFF	Machine learning force field
NEP	Neuroevolutionary potential
RGDOS	Method to obtain Raman tensors or polarizabilities from the projection of phonon modes
RMSE	Root mean squared error
SA-GPR	Symmetry-adapted Gaussian process regressor
SERS	Surface-enhanced Raman spectroscopy
SNES	Separable natural evolution strategy
SOAP	Smooth overlap of atomic positions
TNEP	Tensorial NEP
VASP	Vienna ab initio software packages
BAs	Boron arsenide
CsPbBr <sub>3</sub>	Césium lead bromide perovskites
CsSnBr <sub>3</sub>	Césium tin bromide perovskites
MoS <sub>2</sub>	Molybdenum disulfide
Ti <sub>3</sub> C <sub>2</sub> T <sub>2</sub>	Titanium carbide MXenes
$\alpha, \chi$	Polarizability, dielectric susceptibility
$\varepsilon$	Orbitals energies
$\mu$	Electric dipole
$\xi_n$	Atomic displacements along eigenmode $n$
$\psi$	Wave function
$\omega$	(angular) Frequency
$E$	Electric field
$\mathbf{F}$	Force
$k, \mathbf{k}^\lambda$	Kernel and tensorial kernel used in (SA-)GPR

$m_k$	Mass of atom $k$
$n$	Electronic density
$\mathbf{P}, \mathbf{P}^\lambda$	Power spectra used in GPR and SA-GPR
$\mathbf{q}$	Wave vector in the reciprocal space
$R_n, R_{nm}$	First- and Second-order Raman tensors
$\mathbf{r}$	Position
$T$	Temperature
$t$	Time
$U$	Total energy
$X$	Atomic configuration

## List of original publications

This thesis is based on the following publications, which are referred throughout the text by their corresponding Roman numerals:

- I Berger, E., & Komsa, H.-P., (2024). Polarizability models for simulations of finite temperature Raman spectra from machine learning molecular dynamics. *Physical Review Materials*, 8, 043802. <https://doi.org/10.1103/PhysRevMaterials.8.043802>
- II Dash, A. K., Swaminathan, H., Berger, E., Mondal, M., Lehenkari, T., Prasad, P. R., Watanabe, K., Taniguchi, T., Komsa, H.-P., & Singh, A. (2023). Evidence of defect formation in monolayer MoS<sub>2</sub> at ultralow accelerating voltage electron irradiation. *2D Materials*, 10, 035002. <https://doi.org/10.1088/2053-1583/acc7b6>
- III Berger, E., Lv, Z.-P., & Komsa, H.-P. (2023). Raman spectra of 2D titanium carbide MXene from machine-learning force field molecular dynamics. *Journal of Materials Chemistry C*, 11, 1311–1319. <https://doi.org/10.1039/D2TC04374B>
- IV Berger, E., Niemelä, J., Lampela, O., Juffer, A. H., & Komsa, H.-P. (2024). Raman spectra of amino acids and peptides from machine learning polarizabilities. *Journal of Chemical Information and Modeling*, 64(12), 4601–4612. <https://doi.org/10.1021/acs.jcim.4c00077>

Publication I presents the polarizability models and their applications to BAs and halide perovskites. Publication II studies the impact of point defects in the MoS<sub>2</sub> monolayer on its Raman spectra. Publication III investigates the effect of surface terminations on the Raman spectra of titanium carbide MXenes. In Publication IV, machine learning models are compared using amino acids and peptides chains.

The author carried out all the simulations presented in these publications, except in Publication IV where the classical MD simulations were performed by coauthors. The experimental results in Publication I and III were also performed by coauthors.



# Contents

<b>Abstract</b>	
<b>Tiivistelmä</b>	
<b>Acknowledgements</b>	<b>9</b>
<b>List of abbreviations and symbols</b>	<b>11</b>
<b>List of original publications</b>	<b>13</b>
<b>Contents</b>	<b>15</b>
<b>1 Introduction</b>	<b>17</b>
<b>2 Theoretical background</b>	<b>21</b>
2.1 Vibrational properties .....	21
2.2 Raman spectroscopy .....	23
2.3 Density functional theory .....	28
2.3.1 Polarizability from DFT .....	29
2.4 Machine learning molecular dynamics .....	31
2.4.1 Gaussian process regressor .....	35
2.4.2 Neural network .....	36
<b>3 Polarizability models</b>	<b>39</b>
3.1 Thole model .....	39
3.2 Bond polarizability model .....	41
3.3 Phonon modes projection .....	42
3.4 Machine learning models .....	44
3.4.1 Symmetry-adapted Gaussian process regressor .....	44
3.4.2 Tensorial neuroevolution potential .....	46
<b>4 Applications</b>	<b>47</b>
4.1 Boron arsenide .....	47
4.2 Molybdenum disulfide .....	51
4.2.1 Resonant Raman spectra .....	52
4.2.2 Effect of defects .....	54
4.3 Titanium carbide MXenes .....	56
4.4 Inorganic halide perovskites .....	61
4.5 Amino acids and peptides .....	65
<b>5 Discussion and conclusions</b>	<b>75</b>
<b>References</b>	<b>81</b>
<b>Appendices</b>	<b>93</b>
<b>Original publications</b>	<b>97</b>



# 1 Introduction

Discovering and designing new materials is a tremendously complicated and lengthy task, yet it is necessary for the development of new technologies. Increasing the efficiency of solar cells [1, 2], improving the lifetime and storage capacity of batteries [3, 4], or reducing the size of electronic components [5, 6] are typical examples of current challenges. For potential candidate materials, a wide variety of properties have to be extensively investigated, such as optical and electronic properties, but also mechanical stability and toxicity in some cases. The study of atomic vibrations can also provide important information about materials. Typical examples would be the characterization of the atomic structure and the rich information on the material quality stemming from it, such as strain, doping, defects and grain sizes.

Vibrations represent oscillations of the atomic positions around their equilibrium positions. They are defined by an eigenvector which corresponds to the displacement of atoms and the vibration frequency. Certain atoms or bonds have specific frequencies, which makes it possible to obtain information about atomic structures from the vibrational frequencies. Experimentally, vibrational spectra can be obtained by exciting the system and in turn creating phonons. There are many types of spectroscopy based on various scattering processes. When using light (photons) as the source of excitation, there are two main kinds of spectroscopy. Infrared spectroscopy (IR) studies the elastic scattering of light, which corresponds to photons being absorbed to create phonons. On the other hand, Raman spectroscopy is based on the inelastic scattering of light. In this case, part of the photon energy is transferred into a phonon, and vibrational frequencies can be obtained by comparing the energy of incoming and outgoing photons. While closely related, IR and Raman give complementary insights and are both necessary to get a full picture of the vibrational properties of a material.

Raman scattering was first observed experimentally by C. V. Raman in 1928 [7]. Since then, Raman spectroscopy has become one of the most popular characterization tools in modern science [8]. It has the benefit of being non-destructive and being applicable to solids as well as liquids and gases. Due to its relatively small scattering cross section, Raman spectroscopy suffers from two issues. First, spectra suffer from low intensities, which lead to noisy spectra and challenging interpretations. Intensities can be increased by using, for example, resonant Raman spectroscopy or surface enhanced Raman spectroscopy [9, 10]. The second problem is the strong fluorescence background arising from electronic excitations. Various methods have been developed to reduce or

suppress this unwanted effect [11], ranging from algorithm-based baseline corrections [12, 13] to time-gated Raman spectroscopy [14, 15].

In addition to experiments, computer simulations at the atomic scale are commonly performed to further confirm experimental results or get additional insights. By simulating the motion of each individual atom, these kinds of calculations give access to additional information and can be compared to experimental results for verification. Additionally, having access to single atom resolution can also help in understanding the mechanism behind puzzling observations. Taking Raman spectroscopy as an example, assigning atomic vibrations to the experimentally observed peaks can prove challenging, while this becomes a trivial task with spectra obtained from first-principles calculations. There are two main methods to obtain Raman spectra from simulations: either by using the harmonic approximation and phonon modes, or by using molecular dynamics (MD). Both rely on calculating forces exerted on each atom and determining their polarizabilities. While harmonic approximation is less demanding in terms of computational efforts, it leads to less complete results. In some cases, it is necessary to use MD, which can prove lengthy and expensive. It then becomes crucial to have methods to compute both forces and polarizabilities efficiently.

At the atomic scale, all properties can be derived from the electronic structure, with energies, forces, dipole moment or polarizabilities being typical examples. First-principles (or *ab initio*) methods, which are methods relying only on physical laws and constants (as opposed to empirical methods which use observations and measurements) are usually used to obtain electronic structures. The Hartree-Fock method was one of the first *ab initio* method developed in the late 1920s [16, 17, 18]. While correctly accounting for the electron exchange, Hartree-Fock completely neglects electron correlations. More involved methods, usually referred to as post-Hartree-Fock methods, include correlation effects and allow for the computation of electronic structures more accurately. Examples of post-Hartree-Fock methods would be the coupled cluster expansion [19] or the Møller-Plesset perturbation theory [20]. However, these methods are extremely demanding and are usually limited to only a few atoms.

Density functional theory (DFT) represents a more affordable alternative and is currently the most popular electronic structure calculation method. Even though it was first introduced in 1960s by Kohn and Sham [21, 22], it only became popular in the 1990s due to the progress in functional approximations [23, 24, 25, 26]. Instead of calculating molecular orbitals like *ab initio* methods, DFT relies on computing the electronic density. The popularity of DFT comes mainly from its very good balance between accuracy and computational cost. With the continuous improvement of supercomputers, the number of atoms used in DFT calculations is also increasing, leading to simulations of hundreds

of atoms being performed routinely nowadays. However, computing the polarizabilities of large systems is still very demanding. Additionally, simulations of vibrational spectra require long MD simulations, and methods faster than DFT are often necessary in this case.

In the last decade, machine learning (ML) has invaded our everyday lives, and materials science is no exception. In the context of DFT calculations, ML makes it possible to directly obtain the desired properties without having to perform cumbersome calculations of the electronic density [27]. Properties such as total energies, forces, electric dipole or polarizabilities can be obtained directly from the atomic positions, making it much more efficient than the DFT or post-Hartree-Fock methods. Note, however, that ML algorithms still rely on training sets, which can be obtained using DFT calculations. It then becomes possible to obtain the accuracy of DFT calculations for only a fraction of its cost. In most applications, ML is used to predict total energies and forces, which can then be used to perform MD for a large number of atoms very efficiently. Systems containing millions of atoms can now be investigated, which makes it possible to access length scales and time scales that were previously unimaginable [28].

While the methods for obtaining forces from ML are well established, it is not clear which methods to predict polarizabilities are the best, with many models based on various physical concepts being suggested. There are numerous empirical models based on atomic positions and/or bonds [29, 30, 31, 32], while other models rely on projections of the atomic displacements [33, 34, 35]. Additionally, models based on ML algorithms have also been applied to predict polarizabilities [36, 37] and Raman spectra [38, 39]. However, it remains unclear how these models compare in term of accuracy and computational cost. Similarly, it is uncertain whether all models are suitable for the prediction of Raman spectra.

### *Content and outline of the thesis*

This thesis compiles four publications related to efficient simulation of Raman spectra. To this end, five different polarizability models are used and compared. Publication I presents three of these models and first benchmarks their performances using two fairly simple materials, namely boron arsenide and molybdenum disulfide. The models are further tested on inorganic halide perovskites, which represent more complex anharmonic materials. The other publications are related to applications of these models to scientifically relevant materials. Publication II studies the impact of point defects on the Raman spectra of MoS<sub>2</sub> monolayers. Another kind of 2D material called MXenes is

also investigated in Publication III. These materials possess inhomogeneous surfaces which can be characterized using Raman spectra. Finally, in Publication IV, two machine learning algorithms are trained for amino acids, which represent a set of 20 different molecules. The resulting models are then applied to study the Raman spectra of chains of amino acids (also known as peptides). All of these applications require simulations containing large numbers of atoms. They therefore greatly benefit from ML-based MD and efficient polarizability models.

The chapters are organized as follows. Chapter 2 introduces a theoretical background on Raman spectroscopy, DFT calculations and machine learning molecular dynamics. Polarizability models are then presented in Chapter 3. Chapter 4 contains the results, including the benchmarks of the models and their applications to various complex materials. Finally, the benefits and shortcomings of each model are discussed in Chapter 5, which also concludes the thesis.

## 2 Theoretical background

### 2.1 Vibrational properties

At the atomic scale, vibrations can be seen as oscillations of atoms around their optimal positions. Close to the equilibrium position, the potential surface along such a displacement can be approximated as harmonic. This means that for a single atom moving along one direction  $x$ , the total energy  $U$  takes the form

$$U(x) = \frac{1}{2}k_s x^2, \quad (1)$$

where  $k_s$  can be understood as the curvature of the potential. Using the definition of forces,  $F$ , they can be derived from this potential as

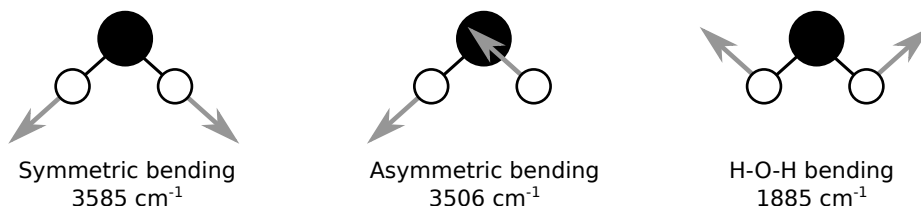
$$F = -\frac{\partial U}{\partial x} = -k_s x. \quad (2)$$

By combining these forces with Newton's second law of motion, we find the following equation of motion

$$\frac{\partial^2 x}{\partial t^2} = -\frac{k_s}{m}x, \quad (3)$$

where  $m$  is the mass of the atom. The solution to this differential equation is given by  $x(t) \sim \exp(-i\omega t)$ , which represents an oscillation at frequency  $\omega$  with  $\omega^2 = \frac{k_s}{m}$ .

Real atomic systems, however, contain many atoms which can all move in all three Cartesian directions. For a system of  $N$  atoms, this leads to  $3N$  degrees of freedom in total. For molecules, rotations and translations can be omitted, leading to a total of  $3N - 6$  vibrational modes. Each of them is defined by a displacement of atoms  $\xi$  and a frequency  $\omega$ . Taking the water molecule as an example, this leads to 3 vibrational modes which are represented in Fig. 1. For periodic systems, vibrational frequencies are usually represented using dispersion curves that plot the frequencies of each mode



**Fig. 1. Vibrational modes of a water molecule with their frequencies. Oxygen atoms are in black and hydrogen atoms in white.**

with respect to the wave vector  $\mathbf{q}$ . Fig. 2 shows the dispersion curves of monolayer molybdenum disulfide ( $\text{MoS}_2$ ) as an example.

In order to address all degrees of freedom, the quantity  $\frac{k_x}{m}$  appearing in Equation (3) must to be replaced with the dynamical matrix  $D$ , which can be defined as [40, 41]

$$D_{k\alpha,k'\beta}(a,b) = \frac{1}{\sqrt{m_k m_{k'}}} \frac{\partial U}{\partial r_{k\alpha}^a \partial r_{k'\beta}^b}, \quad (4)$$

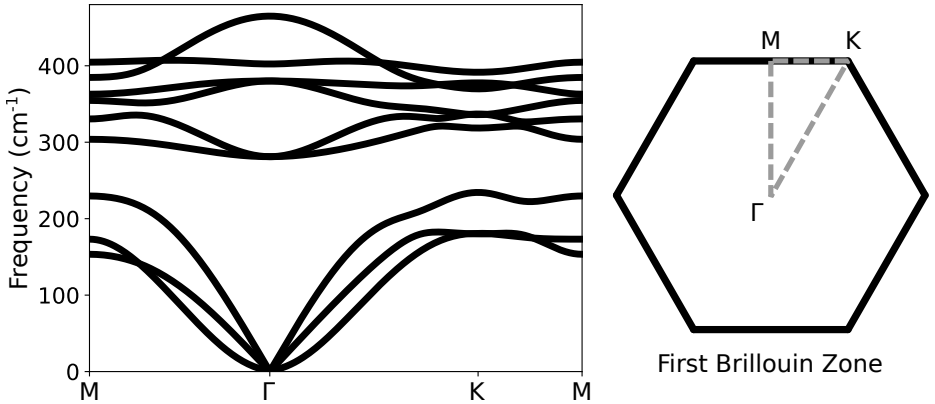
where  $r_{k\alpha}^a$  is the displacement of atom  $k$  in cell  $a$  along the Cartesian direction  $\alpha$  and  $m_k$  is the mass of the atom. The dynamical matrix in reciprocal space  $\tilde{D}$  is obtained through a Fourier transform and reads

$$\tilde{D}_{k\alpha,k'\beta}(\mathbf{q}) = \sum_b D_{k\alpha,k'\beta}(0,b) \cdot e^{i\mathbf{q}\mathbf{R}_b}, \quad (5)$$

where  $\mathbf{R}_b$  is the position of cell  $b$ . Using the equation of motion [Equation (3)], the relation between the dynamical matrix, the eigenmode  $\tilde{\xi}_n$  and the frequency  $\omega_n$  becomes

$$\tilde{D}(\mathbf{q})\tilde{\xi}_n(\mathbf{q}) = \omega_n^2(\mathbf{q})\tilde{\xi}_n(\mathbf{q}). \quad (6)$$

The eigenmodes  $\tilde{\xi}_n(\mathbf{q})$  and the squared frequencies  $\omega_n(\mathbf{q})$  are therefore obtained by diagonalization of the dynamical matrix  $\tilde{D}(\mathbf{q})$ . The vectors  $\tilde{\xi}_n(\mathbf{q})$  have a size of  $3N$  and form an orthogonal basis set of the atomic displacements. Dispersion curves like the one presented in Fig. 2 are obtained by plotting the frequencies  $\omega_n(\mathbf{q})$  along a path in the reciprocal space.



**Fig. 2. Left: Dispersion curves of a  $\text{MoS}_2$  monolayer obtained from DFT calculations. Right: First Brillouin zone of the  $\text{MoS}_2$ . The letters indicate the high-symmetry points and the dashed gray line shows the path used for the dispersion curves on the left.**

The atomic displacements  $\xi_{n,k}(\mathbf{q})$  associated with the mode  $\tilde{\xi}_{n,k}(\mathbf{q})$  is finally obtained as [42]

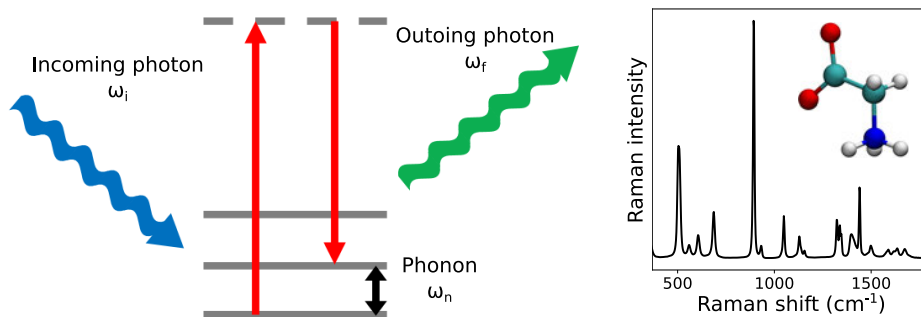
$$\xi_{n,k}(\mathbf{q}) = \frac{1}{\sqrt{m_k}} \tilde{\xi}_{n,k}(\mathbf{q}). \quad (7)$$

Due to the mass scaling, the basis formed by the displacements is no longer orthogonal. Displacements  $\xi_n$  are the principal ingredients used to obtain Raman spectra along with the frequencies  $\omega_n(\mathbf{q})$  and the Raman tensors, as will be presented in the following section.

Similar to light and photons, the vibrational energy is quantized and takes discrete values. The corresponding quasiparticles are called phonons, which play an important role in thermal and electric conductivity. In the context of IR and Raman spectroscopies, phonons are created when photons interact with matter. Thus, representing vibrations as discrete energy levels becomes very important when studying Raman spectroscopy.

## 2.2 Raman spectroscopy

In its most general definition, Raman spectroscopy refers to the inelastic scattering of light caused by atomic vibrations. The mechanism behind Raman scattering is illustrated in Fig. 3. Incoming photons of frequency  $\omega_i$  excite the material to a virtual electronic state, which after returning to the ground state leads to the creation of a phonon of frequency  $\omega_n$  and the emission of a photon of frequency  $\omega_f$ . Information about the vibrational properties of the material is obtained by looking at the difference in frequency between incoming and emitted photons. The Raman spectrum of glycine is represented in Fig. 3 as an example. The difference  $\omega_f - \omega_i$  is usually used for the frequency and



**Fig. 3. Left: Schematic of non-resonant Raman scattering. The dashed gray line represents a virtual electronic excitation, while solid gray lines are the vibrational states. Right: Raman spectrum of the glycine molecule. Experimental data are taken from Ref. [43].**

it is labeled as the Raman shift. Peaks are clearly visible at some frequencies, which correspond to the frequencies  $\omega_n$  of the vibrational modes.

If the scattered frequency is lower than that of the incoming light, a phonon is created; this process is referred to as Stokes scattering. In this case the conservation of energy dictates that  $\omega_i = \omega_f + \omega_n$ . The annihilation of a phonon, on the other hand, is called anti-Stokes and  $\omega_i = \omega_f - \omega_n$ . In practice, both processes are observed and appear symmetrically on Raman spectra. It is also possible to create/annihilate more than one phonon of different frequencies  $\omega_n$  and  $\omega_{n'}$ . Typical examples are the creation of two phonons  $\omega_i = \omega_f + \omega_n + \omega_{n'}$ , annihilation of two phonons  $\omega_i = \omega_f - \omega_n - \omega_{n'}$  or creation and annihilation of two different phonons  $\omega_i = \omega_f + \omega_n - \omega_{n'}$  (referred to as difference scattering).

The absorption cross section of incoming photons greatly depends on their wavelengths, and this also impacts Raman scattering. For an arbitrary wavelength  $\omega_i$ , the absorption and emission proceed via a virtual state, leading to a low cross section and resulting in low Raman intensities. On the other hand, for certain values of  $\omega_i$  the system is excited to a real electronic state, which greatly increases absorption. This is referred to as resonant Raman scattering, which leads to spectra orders of magnitude more intense.

The following derivation presents the fundamental equations of Raman spectroscopy and how to obtain spectra from atomistic simulations [44]. For simplicity, we will focus only on Stokes processes, but the same treatment and results also apply to anti-Stokes. From the classical theory of scattering, the energy emitted by an electric dipole  $\mu$  vibrating at a frequency  $\omega_f$  is given by

$$\frac{\partial W_s}{\partial \Omega} = \frac{\omega_f^4}{(4\pi)^2 \epsilon_0 c^2} |\hat{e}_f \cdot \mu|^2, \quad (8)$$

where  $\Omega$  is the solid angle and  $\hat{e}_f$  is the polarization of the emitted light. In the case of Raman spectroscopy, such vibrating dipoles are created by incoming photons. The dipole is therefore related to the polarizability  $\alpha$  and the incoming electric field  $\mathbf{E} = E \cdot \hat{e}_i$  as

$$\mu = \alpha \cdot \hat{e}_i \cdot E, \quad (9)$$

with  $\hat{e}_i$  being the direction of the electric field. Instead of the emitted power, it is more convenient to use the scattering cross section  $\frac{\partial \sigma}{\partial \Omega}$ , which is the ratio between emitted and incoming power. By combining Equations (8) and (9), the cross section reads

$$\frac{\partial \sigma}{\partial \Omega} = \frac{\omega_f^4}{(4\pi\epsilon_0)^2 c^4} |\hat{e}_f \cdot \alpha \cdot \hat{e}_i|^2. \quad (10)$$

Let us now consider a phonon with frequency  $\omega_n$  and displacement  $\xi_n$ . The “electronic” frequencies  $\omega_i$  and  $\omega_f$  are much larger than the vibrational frequency  $\omega_n$ . This allows

consideration of the deformation along  $\xi_n$  as static. Therefore we can write the change of polarizability with time in terms of the displacement as a Taylor expansion

$$\alpha(\omega_f, t) = \alpha_0(\omega_f) + \frac{\partial \alpha(\omega_f)}{\partial \xi_n} \xi_n(t) + \frac{1}{2} \frac{\partial^2 \alpha(\omega_f)}{\partial \xi_n^2} \xi_n^2(t), \quad (11)$$

where  $\alpha_0$  denotes the polarizability in the vibrational ground state. The second term containing  $\xi_n(t) = \xi_n \cdot e^{-i\omega_n t}$  fluctuates with a frequency  $\omega_n$  and therefore corresponds to a single phonon. It is common to isolate this term to only study the first-order contribution. This is usually done by defining the first-order Raman tensor  $R_n$  as

$$R_n = \frac{\partial \alpha}{\partial \xi_n}. \quad (12)$$

Similarly, the second-order Raman tensor  $R_{nm} = \frac{\partial^2 \alpha}{\partial \xi_n^2}$  can be defined using the third term, and this approach can be extended to define higher-order terms as well. It has to be noted that for second- and higher order contributions, terms containing two different phonons  $n$  and  $m$  should also be included in Equation (11). In this case the Raman tensor takes the form  $R_{nm} = \frac{\partial^2 \alpha}{\partial \xi_n \partial \xi_m}$ .

The fact that  $\omega_f \approx \omega_i \gg \omega_n$  also implies that the scattered frequency is independent of the vibrational frequency and that the prefactor in Equation (10) can be approximated as a constant. By combining Equation (10) and the definition of the Raman tensor, we find that the Raman intensity  $I_n$  associated with phonon  $n$  is proportional to

$$I_n \propto \frac{n(\omega_n) + 1}{\omega_n} |\hat{e}_f \cdot R_n \cdot \hat{e}_i|^2, \quad (13)$$

where the prefactor comes from the thermal average of  $\xi_n$  and  $n(\omega_n)$  is the Bose-Einstein distribution [44, 45]. The Raman spectrum  $I(\omega)$  is then obtained by summing the Raman intensities of all phonon modes

$$I(\omega) \propto \sum_n \frac{n(\omega_n) + 1}{\omega_n} |\hat{e}_f \cdot R_n \cdot \hat{e}_i|^2 \cdot \delta(\omega - \omega_n). \quad (14)$$

In practice, the delta function is substituted with a broadening function, typically represented by a Lorentzian. Even though not explicitly written, the intensity still depends on  $\omega_f$  through the polarizability function  $\alpha(\omega_f)$  and the Raman tensors. This becomes relevant when considering resonant spectra.

It is important to note that the treatment is slightly different for periodic systems. Here the dielectric susceptibility  $\chi$  is used instead of the polarizability, but the treatment is similar. Additionally, in periodic systems, phonons also depend on the wave vector  $\mathbf{q}$ . The incoming and scattered photons also possess momentum described by wave vectors

$\mathbf{q}_i$  and  $\mathbf{q}_f$ , respectively. Due to the conservation of momentum, the following equation holds:  $\mathbf{q}_i = \mathbf{q}_f + \mathbf{q}$ . The typical wave vectors of photons used in Raman spectroscopy ( $\sim 10^5 \text{ cm}^{-1}$ ) are much smaller than the wavelength of phonons at the edge of the Brillouin zone ( $\sim 10^9 \text{ cm}^{-1}$ ), which implies that the approximation  $\mathbf{q} = \mathbf{q}_i - \mathbf{q}_f \approx 0$  can be used for one-phonon processes. Therefore, only modes located close to the center of the Brillouin zone (usually referred to as  $\Gamma$ -point) contribute to the first order Raman spectra. The definition of the Raman tensor can therefore be rewritten as

$$R_n = \frac{\partial \chi}{\partial \xi_n(\Gamma)}. \quad (15)$$

Raman spectra are then obtained in the exact same way as for molecules, using Equation (14). For second-order contributions, the conditions become such that the sum of the wave vectors is zero ( $\mathbf{q} + \mathbf{q}' \approx 0$ ). In this case, the second-order Raman tensor  $R_{nm}$  originates from the third term in Equation (11)

$$R_{nm}(\mathbf{q}) = \frac{\partial^2 \chi}{\partial \xi_n(\mathbf{q}) \partial \xi_m(-\mathbf{q})}. \quad (16)$$

An equation similar to Equation (13), including second-order Raman tensors, can be written and summed to obtain the second-order Raman spectra. Note that in this case, the sum has to be done over pairs of phonons  $n$  and  $m$  as well as over wave vectors  $\mathbf{q}$ . In terms of simulations, this requires more calculations since a large enough sampling of the reciprocal space is necessary, which in turn results in a greater number of polarizability calculations.

Including second and higher orders contributions is therefore challenging. Additionally, Raman spectra obtained from Equation (14) rely only on harmonic approximation and do not account for possible anharmonicity or the effect of temperature. The use of a Lorentzian with an arbitrary width for the artificial broadening is also quite restricting. In some cases, it is necessary to obtain realistic widths.

Another approach, based on fluctuations, makes use of time-correlation functions. It was first used by Green and Kubo to study transport phenomena [46, 47, 48], but can also be applied to obtain power spectra [49, 50] or infrared spectra [49, 51, 52] from the autocorrelation function of velocities or dipole moments, respectively. Similarly, the Raman intensity depends on the fluctuations of the polarizability [44, 53].

Without loss of generality, let us consider only one direction of polarizability. Additionally, instead of using phonon modes, let us now write the initial and final vibrational states  $|i\rangle$  and  $|f\rangle$ . From Equation (10), the Raman intensity can then be rewritten as

$$I(\omega) \propto \sum_{i,f} \rho_i |\langle i | \chi | f \rangle|^2 \delta(\omega_f - \omega_i - \omega), \quad (17)$$

where  $\rho_i$  is the probability of being in state  $|i\rangle$  initially. By using the definition of the delta function

$$\delta(\omega_f - \omega_i - \omega) = \frac{1}{2\pi} \int \exp [i(\omega_f - \omega_i - \omega)t] dt, \quad (18)$$

and inserting it in Equation (17), we obtain

$$I(\omega) \propto \int \sum_{i,f} \rho_i \langle i | \chi | f \rangle \langle f | e^{i\omega_f t} \chi e^{-i\omega_i t} | i \rangle e^{-i\omega t} dt. \quad (19)$$

One can then notice that  $e^{-i\omega_i t} | i \rangle = e^{-iHt/\hbar} | i \rangle$  (and similarly for  $\langle f |$ ), which allows us to rewrite  $e^{iHt/\hbar} \chi e^{-iHt/\hbar} = \chi(t)$ . Using the fact that  $\sum_f |f\rangle \langle f| = 1$ , we get

$$I(\omega) \propto \int \sum_i \rho_i \langle i | \chi(0) \chi(t) | i \rangle e^{-i\omega t} dt. \quad (20)$$

The sum over initial states can be replaced with an average over the configurations sampled by molecular dynamics (MD). In this case, the Raman intensity finally becomes

$$I(\omega) \propto \int \langle \chi(\tau) \chi(\tau+t) \rangle_{\tau} e^{-i\omega t} dt, \quad (21)$$

where  $\langle \chi(\tau) \chi(\tau+t) \rangle_{\tau} = \int \chi(\tau) \chi(\tau+t) d\tau$  denotes the polarizability autocorrelation function. Equation (21) therefore relies on obtaining the polarizability as a function of time  $\chi(t)$ , which can typically be obtained using MD. By changing the temperature used in MD, it becomes possible to study the impact of temperature on the Raman spectra or even study phase transitions. The resulting spectra also account for every phenomenon included in the MD simulations, including second (and higher) order Raman effects as well as possible anharmonicity. Additionally there is no need to artificially broaden the spectra and instead realistic widths are displayed. These advantages clearly show that spectra from MD are more complete than when just using the harmonic approximation.

Obtaining spectra from Equation (21), however, comes at a higher computational cost for two reasons. First, long MD trajectories are necessary to obtain converged spectra, typically requiring on the order of  $10^5$ – $10^6$  time steps. Such a number of steps can prove very challenging for large systems when using first-principles methods such as DFT. Nowadays machine learning (ML) can be used to accelerate these calculations and long MD simulations are not as challenging as they used to be. ML-based MD is presented in more details in Section 2.4. The second reason is the calculation of polarizability, which has to be performed at every time step of the MD (so  $10^5$ – $10^6$  times). Calculating polarizabilities can prove to be very costly, as will be shown in the next section. To reduce this cost, polarizability models can be used. These models represent the main topic of this thesis and are described in detail in Chapter 3.

## 2.3 Density functional theory

Many macroscopic quantities can be understood from atomic scale and electronic structure calculations. The quantum treatment of electrons in solids or molecules generally requires solving a many-body Schrödinger equation, which can prove challenging even for a small number of electrons. It becomes necessary to simplify the problem and, in this context, density functional theory (DFT) represents a great compromise between accuracy and computational cost, making it one of most popular methods for electronic structure calculations. The idea behind DFT is to compute the electronic density  $n(\mathbf{r})$  instead of the orbitals, making the whole calculation much more efficient. Its foundations date back to the 1960s, when Hohenberg and Kohn [21] showed that for a given external potential  $V_{\text{ext}}(\mathbf{r})$  there exists a unique electronic density  $n(\mathbf{r})$  which describes the ground state. In turn, a functional of the density exists such that the total energy  $U$  is minimized for the ground state density. The total energy  $U[n(\mathbf{r})]$  can then be written as a functional of the form

$$U[n(\mathbf{r})] = K[n(\mathbf{r})] + \int V_{\text{ext}}(\mathbf{r})n(\mathbf{r})d\mathbf{r} + \frac{1}{2} \iint \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}d\mathbf{r}' + U_{\text{xc}}[n(\mathbf{r})], \quad (22)$$

where the terms (from left to right) are the kinetic energy  $K[n(\mathbf{r})]$ , the external potential  $V_{\text{ext}}(\mathbf{r})$ , the Coulomb energy and the exchange-correlation energy  $U_{\text{xc}}[n(\mathbf{r})]$ , respectively. The exact form of this latter term is unknown, and approximations have to be used [24, 26, 25]. Minimizing the energy functional therefore leads to the ground state electronic density.

Kohn and Sham [22] later used this functional formulation to simplify the complex many-body problem into a system of one-electron equations. The variation of the energy functional with the density can be understood as the effective potential of a system of non-interacting electrons [54]. This effective potential is called the self-consistent field  $V_{\text{scf}}[n(\mathbf{r})]$  and from Equation (22) it can be written as

$$V_{\text{scf}}[n(\mathbf{r})] = \frac{\delta U[n(\mathbf{r})]}{\delta n(\mathbf{r})} = V_{\text{ext}}(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' + V_{\text{xc}}[n(\mathbf{r})], \quad (23)$$

with  $V_{\text{xc}}[n(\mathbf{r})] = \frac{\delta U_{\text{xc}}[n(\mathbf{r})]}{\delta n(\mathbf{r})}$  being the exchange-correlation potential. The self-consistent field can then be used to solve single-electron Schrödinger equations of the form

$$\left[ -\frac{1}{2}\nabla^2 + V_{\text{scf}}[n(\mathbf{r})] \right] \psi_i(\mathbf{r}) = \varepsilon_i \psi_i(\mathbf{r}), \quad (24)$$

where  $\psi_i(\mathbf{r})$  are called the Kohn-Sham orbitals and  $\varepsilon_i$  are the associated orbital energies. Since the electronic density depends on the orbitals, these represent a set of self-consistent equations known as Kohn-Sham equations. The electronic density  $n(\mathbf{r})$  then

takes the form

$$n(\mathbf{r}) = \sum_i^{N_e} |\psi_i(\mathbf{r})|^2, \quad (25)$$

where  $N_e$  is the number of electrons in the system. Equations (24) and (25) can be solved iteratively. Starting from an arbitrary density, Equation (24) is solved to obtain orbitals  $\psi_i(\mathbf{r})$ , which are then used in Equation (25) to obtain the new density. This process is repeated until reaching convergence.

Various properties can then be derived from the energy and the electronic density. In this work, we mainly focus on forces to perform MD and on evaluating polarizabilities to obtain the Raman spectrum. Forces  $\mathbf{F}_i$  on ion  $i$  are usually obtained using first-order perturbation theory and the Hellmann-Feynman theorem. For a perturbation  $\lambda$ , the first-order derivative of the energy can be written as

$$\frac{\partial U}{\partial \lambda} = \langle \psi | \frac{\partial H}{\partial \lambda} | \psi \rangle. \quad (26)$$

Using the fact that forces on nuclei can be written as  $\mathbf{F}_i = -\frac{\partial U}{\partial \mathbf{x}_i}$  (with  $\mathbf{x}_i$  being the position of nucleus  $i$ ) and explicitly writing the ionic Hamiltonian  $H$ , forces can finally be obtained as

$$\mathbf{F}_i = - \int n(\mathbf{r}) \frac{\partial V_{\text{ext}}(\mathbf{r})}{\partial \mathbf{x}_i} d\mathbf{r} - \frac{\partial U_N(\mathbf{x}_i)}{\partial \mathbf{x}_i}. \quad (27)$$

The terms correspond to the electron-ion and the ion-ion interactions, respectively. While the former was already written in Equation (22), the latter can be written as

$$U_N = \sum_{i < j} \frac{Z_i \cdot Z_j}{|\mathbf{x}_i - \mathbf{x}_j|}, \quad (28)$$

where  $Z_i$  denotes the charge of nucleus  $i$ . Derivatives of  $V_{\text{ext}}$  and  $U_N$  can be formulated analytically, making the computation of forces from Equation (27) straightforward once the electronic density  $n(\mathbf{r})$  is known.

### 2.3.1 Polarizability from DFT

As previously presented in Section 2.2, calculations of the polarizability are necessary to obtain Raman spectra. Although there exist many ways to compute the polarizability from the electronic density and DFT, obtaining it requires additional steps, making it much more demanding than energy calculations. Starting from Equation (9), the polarizability tensor  $\chi$  can be written as the second order derivatives of the energy with electric field  $\mathbf{E}$

$$\chi_{ij} = \frac{\partial \mu_i}{\partial E_j} = \frac{\partial^2 U}{\partial E_i \partial E_j}, \quad (29)$$

where  $i$  and  $j$  refer to Cartesian coordinates components. From this, the easiest way to obtain  $\chi$  is to compute the polarization  $\mu$  in presence of an electric field  $\mathbf{E}$  and use finite differences. In this case, four calculations are necessary: one without an electric field and three with the electric field applied along each of the Cartesian coordinates. It already becomes evident that obtaining polarizabilities is more demanding than just energy and force calculations. The application of this direct method is, however, fairly limited, mostly because polarization is poorly defined in periodic systems. Choosing the intensity of the electric field is another downside of this method.

Density functional perturbation theory (DFPT) is the most popular choice for the calculation of electronic polarizabilities. The polarizability tensor  $\chi$  is then obtained as [54]

$$\chi_{ij} = \frac{4}{\Omega} \sum_v \langle \phi_v^i | \Delta^{E_j} \psi_v \rangle, \quad (30)$$

where  $\Omega$  is the volume, indices  $v$  refer to the valence states,  $|\Delta^{E_j} \psi_v\rangle$  is the response of the wave function to the electric field and  $|\phi_v^i\rangle$  is an auxiliary wave function. The latter can be defined as  $|\phi_n^i\rangle = P_c r_i |\psi_n\rangle$  and could be understood as a ‘‘polarization vector’’ [55]. It is usually obtained by solving the equation [54]

$$(H_{\text{scf}} - \varepsilon_n) |\phi_n^i\rangle = P_c [H_{\text{scf}}, r_i] |\psi_n\rangle, \quad (31)$$

where  $H_{\text{scf}}(\mathbf{r}) = -\frac{1}{2}\nabla^2 + V_{\text{scf}}(\mathbf{r})$  is the single-electron Hamiltonian previously used in Equation (24),  $P_c$  is the projector onto empty bands and  $r_i$  is the position operator in the direction  $i$ . The functions  $|\phi_n^i\rangle$  are therefore directly obtained from the Kohn-Sham orbitals  $\psi_n$  and energies  $\varepsilon_n$ .

There exist a few ways to obtain the response  $|\Delta^{E_j} \psi_n\rangle$ , with typical examples including derivatives over wave vector  $\mathbf{q}$  [55] or variations in the overlap matrix [56]. However, they all rely on solving a set of self-consistent Sternheimer equations of the form

$$(H_{\text{scf}} - \varepsilon_n) |\Delta^{\mathbf{E}} \psi_n\rangle = -\sum_i |\phi_n^i\rangle - P_c \Delta V_{\text{scf}} |\psi_n\rangle, \quad (32)$$

where  $\Delta V_{\text{scf}}$  is the first-order correction to the self-consistent potential previously defined in Equation (23). The latter depends implicitly on the response of the density  $\Delta n(\mathbf{r})$ , which in turn depends on  $\Delta \psi_n(\mathbf{r})$ . Equation (32) therefore has to be solved iteratively. That is, once the density and wave functions have been solved from Equations (24) and (25), the Sternheimer equation has to be solved, further increasing the computational cost.

While not explicitly stated, we have only considered the static electronic dielectric constant  $\chi(\omega = 0)$  so far. To obtain resonant Raman spectra, it is important to obtain the

whole dielectric function. In this case, the dielectric function has a real and an imaginary part and can therefore be written as

$$\chi(\omega) = \chi'(\omega) + i\chi''(\omega). \quad (33)$$

The imaginary part  $\chi''(\omega)$  can be obtained as a sum over the overlap between valence and conduction bands, which reads [55]

$$\chi''_{ij}(\omega) = \frac{\pi}{\Omega} \lim_{\lambda \rightarrow 0} \frac{1}{\lambda^2} \sum_{c,v,\mathbf{q}} \delta(\varepsilon_{c,\mathbf{q}} - \varepsilon_{v,\mathbf{q}} - \omega) \cdot \langle \psi_{c,\mathbf{q}+\mathbf{e}_i\lambda} | \psi_{v,\mathbf{q}} \rangle \langle \psi_{v,\mathbf{q}} | \psi_{c,\mathbf{q}+\mathbf{e}_j\lambda} \rangle, \quad (34)$$

where  $\mathbf{q}$  denotes wave vectors and  $\mathbf{e}_i$  are unit vectors of the reciprocal space. Note that indices  $v$  and  $c$  relate to the valence and conduction states, respectively. The real part  $\chi'(\omega)$  is then simply obtained through the Kramers-Kronig transformation

$$\chi'_{ij}(\omega) = \frac{2}{\pi} P \int \frac{\chi''_{ij}(\nu) \cdot \nu}{\nu^2 - \omega^2 - i\eta} d\nu, \quad (35)$$

where  $P$  is the principal value and  $\eta$  is an infinitesimal number. While obtaining  $|\psi_{c,\mathbf{q}+\lambda}\rangle$  is not particularly time consuming, the sum in Equation (34) requires that a large number of conduction states are considered, leading to much heavier calculations. Additionally, commonly-used DFT algorithms do not optimize empty states, which sometimes results in inaccurate polarizabilities when using this summation method. Deriving electronic polarizabilities from Equations (34) and (35) is, however, much better suited for more accurate first principles methods, such as GW perturbation theory or post-Hartree-Fock algorithms.

Calculations of polarizabilities for the electronic structure therefore require additional expensive steps, regardless of which method is used. Repeating this process  $10^5$ – $10^6$  times to obtain high quality Raman spectra from molecular dynamics [and Equation (21)] becomes impossible. Similarly, obtaining forces for a large number of atoms would also prove too demanding. Instead, forces are obtained efficiently using a machine learning algorithm (presented in the next section), while various polarizability models are used (these are later presented in Chapter 3).

## 2.4 Machine learning molecular dynamics

The goal of molecular dynamics is to obtain a realistic evolution of atomic positions with time  $\mathbf{x}_i(t)$ , also referred to as trajectories, with index  $i$  referring to an atom. There exist many algorithms to integrate positions with time, one of the most popular being the

velocity-Verlet algorithm given as

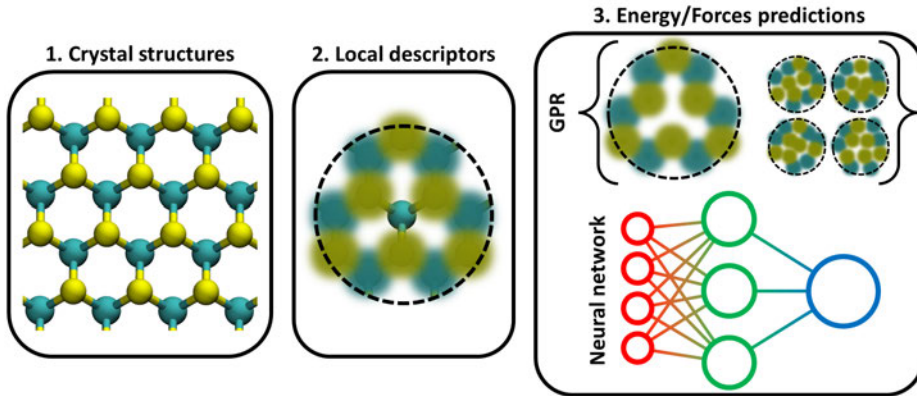
$$\begin{aligned}\mathbf{x}_i(t + \Delta t) &= \mathbf{x}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{\mathbf{F}_i(t)}{2m_i}\Delta t^2, \\ \mathbf{v}_i(t + \Delta t) &= \mathbf{v}_i(t) + \frac{\mathbf{F}_i(t) + \mathbf{F}_i(t + \Delta t)}{2m_i}\Delta t,\end{aligned}\tag{36}$$

where  $\mathbf{v}_i(t)$  denotes the velocity of atom  $i$  at time  $t$  and  $\Delta t$  is the integration time step. This algorithm works in three steps: the first equation is used to get new positions  $\mathbf{x}_i(t + \Delta t)$ , which are then used to compute new forces  $\mathbf{F}_i(t + \Delta t)$  and new velocities  $\mathbf{v}_i(t + \Delta t)$  are finally obtained from the second equation. By choosing a sufficiently small time step  $\Delta t$ , the discrete positions can be used as trajectories which are necessary in many applications. For example, these can be used in Equation (21) to obtain Raman spectra.

The velocity-Verlet algorithm can be used to run MD in the NVE ensemble, maintaining the number of particles, the volume and the energy constant. In numerous situations, however, it is useful to obtain trajectories at constant temperatures and use the NVT ensemble. Various thermostats have been suggested [57, 58, 59], with the most popular thermostat being the Nosé-Hoover thermostat [60, 61, 62]. The equations for the Nosé-Hoover chain thermostat [63] are presented in Appendix A. In addition to a thermostat, some simulations require a constant pressure (and changing volume), in which case the NPT ensemble is simulated using a combination of a thermostat and a barostat [58, 64, 65, 66].

Obtaining forces from atomic positions is the most time-consuming part of molecular dynamics. DFT calculations of the forces [as derived from Equation (27)] rely on obtaining the whole electronic density, which scales poorly with the number of atoms. This results in heavy calculations, which limit the length of the simulated trajectory. Instead, bypassing electronic structure calculations would significantly reduce calculation time. Machine learning (ML) can be used to this end. In atomistic simulations, ML is used to predict properties, usually energies and forces, directly from the atomic structures without relying on the electronic density. Other properties can also be predicted using ML [67], with examples being dipolar moments [68, 69], electronic structures [70, 71] or band gaps [72, 73, 74]. The prediction of polarizabilities is particularly relevant in the case of Raman spectroscopy and is discussed in detail later (see Section 3.4).

Predictions of energies and/or forces are done using machine learning potentials (MLP) or machine learning force fields (MLFF). Fig. 4 shows the different steps of a typical MLP/MLFF. The first step is to get the crystal structures and the positions of ions. However, these positions cannot be directly used for predictions: hence one needs to encode them using descriptors. In materials science, one common practice is to



**Fig. 4. Schematic of ML predictions. Crystal structures are transformed into local descriptors, which are then used to make predictions using either GPR or NN.**

use local environments. Local descriptors are usually centered around atom  $i$  and only neighboring atoms within a predefined cutoff are considered. These are then used to predict local properties, typically the energy  $U_i$  associated with atom  $i$ . The total energy  $U$  is then simply obtained as the sum of local energies

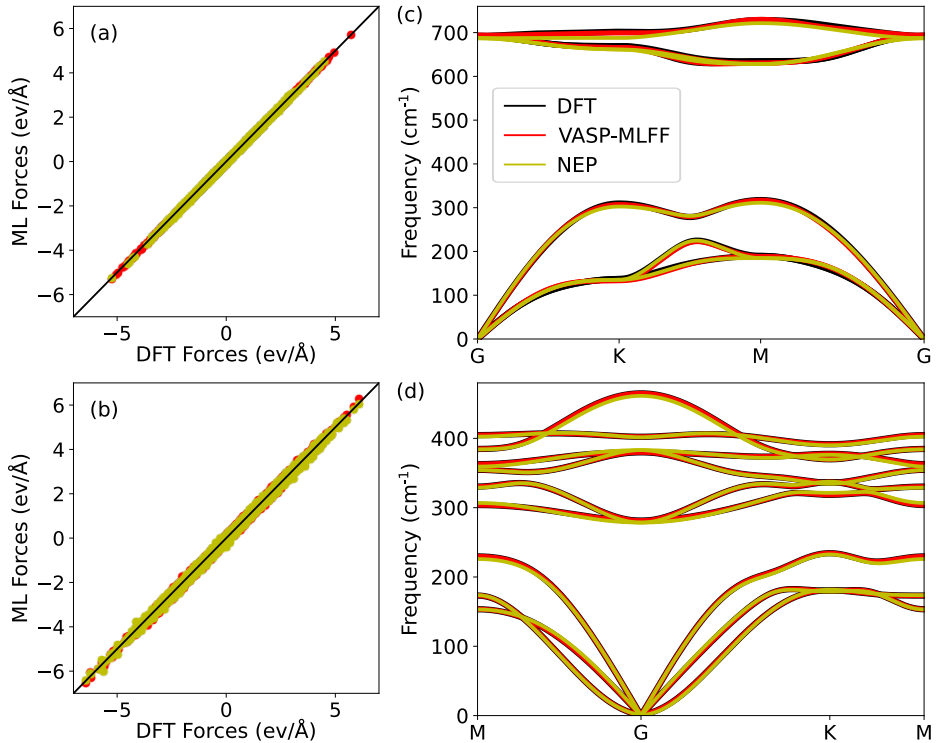
$$U = \sum_i U_i. \quad (37)$$

For periodic systems, using local environments has the benefit of making its application to different sizes of supercells very straightforward. Other global properties, such as the stress tensor or the polarizability, are also obtained from a similar sum. On the other hand, local properties such as forces are obtained directly.

Before they are applied, all ML-based models have to first be trained. In this training step, the desired properties are calculated for a large number of structures, called the training set, and the parameters of the ML model are optimized in order to minimize the differences between calculations and predictions for structure-property pairs in the training set. The quality of the resulting model depends greatly on the size and the accuracy of the training set. DFT calculations represent a good starting point for training sets, even though more accurate methods lead to more accurate ML predictions, which might be necessary depending on the predicted properties. Supercells included in the training set are usually quite small due to the poor scalability of first-principles calculations. Once the model is sufficiently trained, it can be applied to larger supercells for production.

After training the MLFF, it is common practice to assess its accuracy by comparing the predicted forces (and sometimes energies) with DFT calculations. For benchmarking

purposes, we trained two different ML models. The first model is based on the MLFF by Jinnouchi *et al.* [75, 76] and is labeled VASP-MLFF. For the second model, we trained a neuroevolution potential (NEP) first introduced by Fan *et al.* [77]. Note that the two MLFFs used here are presented in detail in the next two sections. Fig. 5(a) and (b) show a comparison between the forces from DFT with those predicted by both ML models for two materials, respectively boron arsenide (BAs) and molybdenum disulfide ( $\text{MoS}_2$ ). We find excellent agreement with DFT results for both MLFFs and both materials. Depending on the application, it is also necessary to test other properties. In the case of Raman spectroscopy, it can be useful to assess the accuracy of the phonon dispersion curves (previously introduced in Fig. 2). In Fig. 5(c) and (d), phonon bands from the MLFFs are compared with DFT calculations for BAs and  $\text{MoS}_2$ , respectively. Similar to forces, the agreement between MLFFs and DFT is excellent for both materials.



**Fig. 5.** (a–b) Comparison between DFT and ML forces sampled from MD trajectories for BAs and  $\text{MoS}_2$ , respectively. Forces from VASP-MLFF and NEP are both compared. (c–d) Phonon dispersion curves of BAs and  $\text{MoS}_2$ , respectively. Here again, bands from VASP-MLFF and NEP are compared to DFT calculations. The same colors are used in all four panels. [Panels (c–d) are adapted, with permission, from Publication I © 2024 American Physical Society].

While many forms of MLP/MLFF have been suggested [78, 79, 80], Gaussian process regressors (GPR) and neural networks (NN) represent the most popular choices. These two methods are introduced in the remainder of this section. Note that both GPR and NN have been extended to predict tensors such as polarizabilities. These extensions are discussed in detail in a later chapter (see Section 3.4).

### 2.4.1 Gaussian process regressor

The Gaussian process regressor (GPR) was first introduced by Bartók *et al.* [81]. It has since become a very popular ML method to create force fields [75, 82] as well as predict other properties such as electronic density [83, 84] or polarizabilities [85, 86]. GPR relies on kernel functions  $k(X, X')$ , which describe the similarity between two atomic configurations  $X$  and  $X'$ . The predicted value of some property, for example the local energy  $U_i$  of a local configuration  $X_i$  is given as a linear combination of the form

$$U_i(X_i) = \sum_n w_n k(X_i, X_n), \quad (38)$$

where  $X_n$  are the configurations in the training set and  $w_n$  are the associated weights. This equation can be recast as the matrix equation  $\mathbf{U} = \mathbf{w}\mathbf{k}$  with  $\mathbf{U}$  containing the energies of the training set. The training step simply consists of obtaining the weights  $\mathbf{w}$  by solving this matrix equation. Once the weights are known, Equation (38) can be used to make predictions.

While many descriptors have been suggested [87, 88, 89], one popular choice is the smooth overlap of atomic positions (SOAP) [90]. It is built using Gaussian smeared atomic densities  $\rho(\mathbf{r})$  and the resulting kernel reads

$$k(X, X') = \int d\hat{R} \left| \int d\mathbf{r} \rho(\mathbf{r}) \rho'(\hat{R}\mathbf{r}) \right|^2, \quad (39)$$

where the integral is performed over active rotations  $\hat{R}$ . This ensures that the kernel is invariant under rotation, which has to be true when the property of interest is also invariant. Calculations of kernels can be simplified by expanding atomic densities using a set of orthogonal radial functions  $g_n(r)$  and spherical harmonics  $Y_{lm}(\hat{r})$ , which reads

$$\rho(\mathbf{r}) = \sum_{nlm} c_{nlm} g_n(r) Y_{lm}(\hat{r}). \quad (40)$$

By using this form for the atomic densities in Equation (39), kernels can be rewritten in the simpler form

$$k(X, X') = \sum_{nn' ll'} P_{nn' ll'}(X) P_{nn' ll'}(X') = \mathbf{P}(X) \cdot \mathbf{P}(X'), \quad (41)$$

where  $P_{n'l'l'}(X) = \sum_m c_{nlm}(X) c_{n'l'm}(X)$ . Kernels are therefore simply obtained from the scalar product of two vectors  $\mathbf{P}$  (we later refer to these vectors as power spectra). Note that the choice of spherical harmonics is the most common for SOAP, but other spherical functions can be chosen and lead to different power spectra. Calculations of kernels for GPR with SOAP therefore rely on computing radial functions and spherical harmonics. The number of functions included in Equation (40) is not fixed. Including more functions leads to more accurate densities, which should result in more accurate predictions but at the cost of increased computational expense. Higher order expansions also lead to more parameters needing optimization, hereby requiring more training data.

### 2.4.2 Neural network

Feed-forward neural networks (sometimes referred to as multilayer perceptrons) have also been used to predict energies. It was first used by Behler and Parrinello [91], and many different neural networks have since then been successfully applied to obtain reliable and efficient force fields for MD simulations [92, 93, 94]. Fig. 6 shows a typical representation of a feed-forward neural network. It is composed of neurons (black circles) arranged in successive layers, which are connected by synapses (gray lines). Starting from the input layer, information is passed to the following layer through synapses using connectivity weights, bias vectors and activation functions until it reaches the output layer containing the predicted property. The layers between the input and

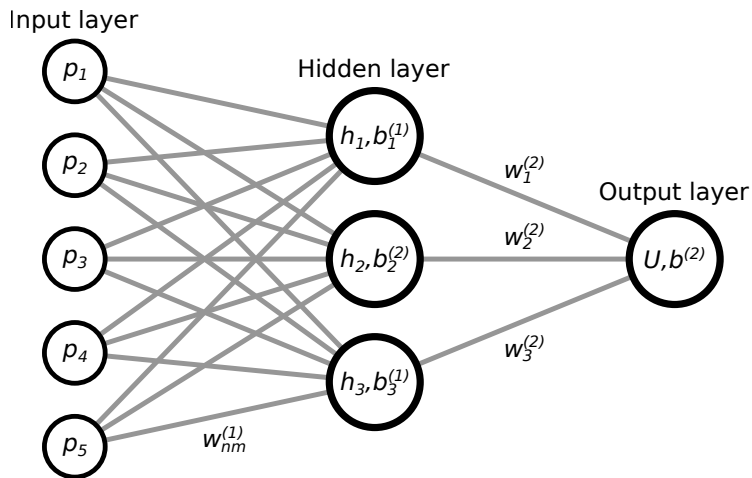


Fig. 6. Representation of a feed-forward neural network with its different layers. Neurons are represented by black circles and synapses by gray lines. Definitions and relations between the different quantities are given in Equations (42) and (43).

output layer are called hidden layers as they usually do not represent any meaningful quantity. In this work, we use the neuroevolution potential (NEP) first proposed by Fan *et al.* [77, 95, 96]. In the case of NEP, there is only one hidden layer containing only a few tens of neurons. More complex neural networks might possess many hidden layers composed of a much larger number of neurons [97, 98].

Similar to the power spectra of GPR, the local environment around atom  $i$  is transformed into descriptor vectors  $p^i$ , which are in turn used as the input layer. In NEP, radial functions are based on Chebyshev polynomials, and Legendre polynomials are used as spherical functions. Additionally, the hyperbolic tangent is used as activation function between the input and hidden layer. The hidden layer state vector  $h^i$  then takes the form

$$h_n^i = \tanh \left( \sum_m^{N_{\text{des}}} w_{mn}^{(1)} p_m^i + b_n^{(1)} \right), \quad (42)$$

where  $w_{mn}^{(1)}$  are the connectivity weights between the input  $p_m^i$  and the hidden layer,  $b_n^{(1)}$  are the biases of the hidden layer and  $N_{\text{des}}$  is the dimension of the descriptor vector. Also note that the dimensionality of the state vector  $h^i$  depends on the number of neurons in the hidden layer, denoted  $N_{\text{neu}}$ . Finally, in the NEP, the local energy  $U_i$  is obtained in the output layer as

$$U_i = \sum_n^{N_{\text{neu}}} w_n^{(2)} h_n^i + b^{(2)}, \quad (43)$$

where  $w_n^{(2)}$  are the connectivity weights between the hidden layer and the output and  $b^{(2)}$  is the output bias. Connectivity weights and biases are optimized during the training step. Additionally, parameters pertaining to the creation of the descriptors have to be optimized, which can represent thousands of parameters in total.

While the output layer could in general contain many output neurons, energy is the only output of NEP. Forces on atoms  $i$  are instead directly derived from local energies and read

$$\mathbf{F}_i = \sum_{j \neq i} \frac{\partial U_i}{\partial \mathbf{r}_{ij}} - \frac{\partial U_j}{\partial \mathbf{r}_{ji}}, \quad (44)$$

where  $\mathbf{r}_{ij}$  denotes the bond between atoms  $i$  and  $j$ . Similarly, stress tensors  $W$  are obtained as a sum over local tensors  $W_i$ , which reads

$$W_i = \sum_{j \neq i} \mathbf{r}_{ij} \otimes \frac{\partial U_j}{\partial \mathbf{r}_{ji}}, \quad (45)$$

where  $\otimes$  denotes the outer product between the two vectors. Note that derivatives  $\frac{\partial U_i}{\partial \mathbf{r}_{ij}}$  can be obtained analytically from Equations (42) and (43) and from the derivatives

of descriptors with respect to coordinates, making it more efficient than using finite differences.

We use values from DFT calculations as targets for total energies, forces and stress tensors. As previously mentioned, connectivities  $w$ , biases  $b$  and descriptors have to be optimized in order to minimize the difference between ML predictions and these target values. In other words, the goal is to minimize a loss function which, in the case of NEP, takes the following form

$$L = \lambda_U L_U + \lambda_F L_F + \lambda_W L_W + \lambda_1 L_1 + \lambda_2 L_2. \quad (46)$$

Each  $L_U$ ,  $L_F$  and  $L_W$  represents the RMSE of energy, forces and stress tensors, respectively, while  $\lambda$  are weights.  $L_1$  and  $L_2$  are regularization parameters taking the form

$$L_1 = \frac{1}{N_{\text{par}}} \sum_{n=1}^{N_{\text{par}}} |z_n| \quad (47)$$

and

$$L_2 = \left( \frac{1}{N_{\text{par}}} \sum_{n=1}^{N_{\text{par}}} z_n^2 \right)^{1/2}, \quad (48)$$

where  $z_n$  are the parameters of the neural network and  $N_{\text{par}}$  is the number of parameters. These parameters  $z_n$  are optimized by minimizing the loss function, which corresponds to minimizing the errors between training data at the DFT levels and predictions from NEP.

In total, NEP can contain thousands of parameters that need to be optimized, making this step particularly challenging. To simplify this problem, the Separable Natural Evolution Strategy (SNES) is used for optimization [77, 99, 100]. The SNES algorithm is presented in more detail in Appendix B.

### 3 Polarizability models

While ML has been established as a very powerful method to obtain MD trajectories, Raman spectra also rely on obtaining polarizabilities along these trajectories. As previously mentioned, computing polarizabilities with DFT would prove too expensive and computational gains from using ML would become irrelevant. To this end, various models have been suggested to predict polarizabilities efficiently and accurately. This chapter introduces some of these models, focusing on those used for applications later presented in Chapter 4. Each model relies on different physical properties. The Thole model and the bond polarizability models (BPM) are empirical models and rely only on the positions of ions and their bonds. Another model, called RGDOS, uses phonon projections to predict the polarizabilities of large supercells. Models based on the ML methods previously introduced in Chapter 2 also exist, although small modifications are necessary to correctly account for the fact that polarizabilities are tensors *i.e.*, properties with directions/orientations, unlike energies.

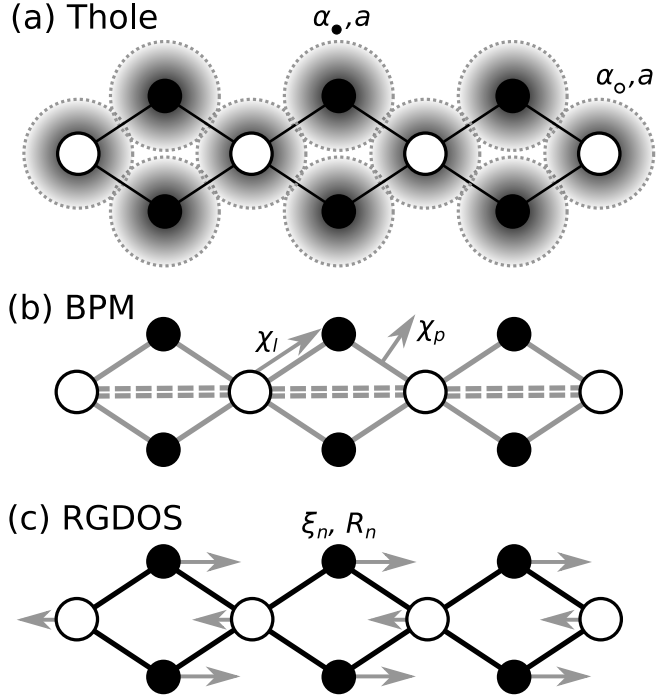
#### 3.1 Thole model

The Thole model is based on the dipole-dipole interaction [29, 30]. A representation of the Thole model can be found in Fig. 7(a). Each atom is attributed an atomic polarizability tensor  $\alpha$  of size  $3 \times 3$ , which only depends on the atomic species. The interaction between atoms and their polarizabilities is then accounted for by the dipole field tensor  $T$ . For a collection of atoms placed in a homogeneous electric field  $\mathbf{E}$ , the dipole moment  $\mu_p$  of atom  $p$  can be written as [101]

$$\mu_p = \alpha_p \left[ \mathbf{E}_p - \sum_{q \neq p} T_{qp} \mu_q \right], \quad (49)$$

where  $\alpha_p$  is the atomic polarizability of atom  $p$  and  $T_{qp}$  is the dipole field tensor between atoms  $p$  and  $q$ . This equation can be written as a system of equations of the form  $\mu = A\mathbf{E}$ , with the inverse of matrix  $A$  being defined as [30]

$$A^{-1} = \begin{bmatrix} \alpha_1^{-1} & 0 & \dots & 0 \\ 0 & \alpha_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_N^{-1} \end{bmatrix} + \begin{bmatrix} 0 & T_{12} & \dots & T_{1N} \\ T_{21} & 0 & \dots & T_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ T_{N1} & T_{N2} & \dots & 0 \end{bmatrix}. \quad (50)$$



**Fig. 7. Schematic of the polarizability models used in this thesis. The Thole model is built on the atomic polarizabilities  $\alpha$  and the screening length  $a$ , while BPM relies on the longitudinal and perpendicular polarizabilities  $\chi_l$  and  $\chi_p$ , and RGDOS uses the eigenmodes  $\xi_n$  and Raman tensors  $R_n$ .**

In this form, the equation clearly corresponds to the usual relation between dipolar moment and (effective) polarizability. The molecular polarizability is then retrieved by summing the elements of  $A$

$$\chi_{ij} = \sum_{qp} (A_{ij})_{qp}. \quad (51)$$

It can be also noted that for  $T = 0$  (*i.e.* the dipole-dipole interaction is neglected)  $A$  is simply given by the atomic polarizabilities. In this case, the polarizability would be unaffected by changes in the atomic positions and therefore would be unable to describe phonon oscillations or obtain Raman spectra. For a bond vector  $\mathbf{r}$ , the general form of the dipole field tensor  $T$  is

$$T_{ij} = \frac{\delta_{ij}}{r^3} f_e - \frac{3r_i r_j}{r^5} f_t, \quad (52)$$

where  $f_e$  and  $f_t$  are distance-dependent screening functions. Applequist [101] initially proposed this form with  $f_e = f_t = 1$ , which leads to infinite polarizabilities at a certain distance  $r$ , thereby necessitating improvement. To fix this issue, Thole proposed various

forms for the screening functions, all including a screening range  $a$  as a parameter [29]. Here we use the exponential shape, for which  $f_e$  and  $f_i$  take the form

$$f_e = 1 - \left( \frac{a^2 r^2}{2} + ar + 1 \right) e^{-ar} \quad (53)$$

and

$$f_i = 1 - \left( \frac{a^3 r^3}{6} + \frac{a^2 r^2}{2} + ar + 1 \right) e^{-ar}. \quad (54)$$

The screening length  $a$  along with atomic polarizabilities of H, C, N and O atoms were originally optimized using the polarizabilities of 16 molecules [29]. The model was later extended to more atoms (mainly halogens) using a larger training set, making the resulting model particularly suitable for organic molecules [30]. The Thole model is utilized only in Publication IV, where it is used to compute the Raman spectra of peptides. The results are then compared with ML-based polarizability models. In this case, we use the screening range  $a$  and the atomic polarizability previously optimized for amino acids [102].

Note, however, that the Thole model used here is trained using only optimized molecular configurations. It is therefore unclear if the model correctly accounts for vibrations and reproduces the positions and intensity of Raman peaks. This was investigated by comparing Raman spectra from the Thole model with other polarizability models.

### 3.2 Bond polarizability model

The bond polarizability model (BPM) is a fairly simple yet powerful model [103, 104, 105]. It has already been applied to many systems, including  $\alpha$ -quartz [31], aluminosilicates [106], germanium telluride [107] and molecules such as alkanes [108]. In this model, each bond contributes to the total polarizability and is attributed a longitudinal and a perpendicular component, respectively written  $\chi_l$  and  $\chi_p$ . For each bond  $\mathbf{r}$  selected into the model, the contribution to the polarizability tensor  $\chi_{ij}$  can be written as

$$\chi_{ij} = \frac{r_i r_j}{r^2} \chi_l(r) + \left( \delta_{ij} - \frac{r_i r_j}{r^2} \right) \chi_p(r). \quad (55)$$

Both components  $\chi_l$  and  $\chi_p$  are illustrated in Fig. 7(b). The fact that only one perpendicular component is used implies that the bonds are considered to be cylindrical, that is, the perpendicular contributions are the same in all directions. While reasonable in crystal, this approximation has to be modified for special symmetries. For example, when

applying BPM to boron nitride sheets and nanotubes, the perpendicular contribution  $\chi_p$  has been decomposed into one in-plane and one out-of-plane component [109]. For the materials studied in this work however, the cylindrical approximation holds and is used.

Different forms of  $\chi(r)$  have been suggested, with the most common being polynomials [31, 106, 107]. In this case, both the longitudinal  $\chi_l$  and perpendicular  $\chi_p$  terms simply read

$$\chi_l(r) = a_0 + a_1 r + a_2 r^2 + a_3 r^3 + \mathcal{O}(r^4). \quad (56)$$

The coefficients  $a_n$  are then parameters that need to be optimized. Recently, a more complex shape of  $\chi(r)$  involving Lorentzian functions has been suggested [32] and applied to oxide perovskites, resulting in very accurate models. In this work, we use the polynomial parametrization and optimize the coefficients  $a_n$  by minimizing the root-mean-square error (RMSE), which can be defined as

$$\text{RMSE} = \sqrt{\frac{1}{9N} \sum_{n,i,j} \left( \chi_{n,i,j}^{\text{BPM}} - \chi_{n,i,j}^{\text{DFT}} \right)^2}, \quad (57)$$

where  $N$  is the number of configurations in the training set, index  $n$  runs over the configurations in the training set and  $i$  and  $j$  are Cartesian directions. Given the small number of parameters to optimize, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [110, 111, 112, 113] is used to optimize the polynomial coefficients.

The accuracy of BPM will depend on the chosen expansion order in Equation (56). A higher order should lead to more accurate models, at the cost of more parameters to optimize. Additionally, one has to carefully choose which bonds to include in BPM. Using MoS<sub>2</sub> [represented in Fig. 7(b)] as an example, Mo-S bonds (solid lines) have to be included, but it remains unclear if other bonds, such as Mo-Mo bonds (double dashed lines) also have to be included. Similar to the expansion order, including more bonds increases the number of parameters, which in turn makes the optimization more challenging.

### 3.3 Phonon modes projection

Compared to the previous models, which relied on the atomic positions or bonds, the RGDOS model is based on an expansion around the projection of phonon modes  $\xi_n$  and their Raman tensors  $R_n$ . A schematic of RGDOS is shown in Fig. 7(c). This model was first proposed by Hashemi *et al.* [33], and was designed to study Raman spectra of large supercells, which could be alloys or contain defects. In this context, it has already been successfully applied to transition metal dichalcogenides alloys [33, 114], molybdenum disulfide defects [115] and tin sulfide multilayer films [116]. Using the mass-scaled

vibrational modes of a smaller unit cell  $\xi_n$  as a basis set, Raman tensors of a larger cell  $R_M$  can be written as an expansion over those of the unit cell  $R_n$ . Since  $\xi_n$  do not form an orthogonal basis, the projection coefficients  $Q_{Mn}$  are found by solving a system of linear equations

$$\sum_i \langle \xi_n | \xi_i \rangle Q_{Mi} = \langle \xi_n | \xi_M \rangle, \quad (58)$$

where  $\xi_M$  corresponds to an eigenmode of the large supercell. The Raman tensors of the supercell then read

$$R_M = \sum_n Q_{Mn} R_n. \quad (59)$$

Since both  $\xi_M$  and  $R_M$  are known, the Raman spectra can then be obtained using Equation (14). RGDOS therefore only requires Raman tensors of the unit cell, thus making expensive polarizability calculations for the larger cell unnecessary. This “harmonic” version of RGDOS was used in Publication II in combination with an ML force field to obtain the Raman spectra of MoS<sub>2</sub> with different types of defects at a very low concentration.

RGDOS can also be used in combination with MD and Equation (21), in which case it is used to predict polarizabilities. Instead of projecting eigenvectors of the supercell  $\xi_M$ , displacements from the equilibrium position  $\mathbf{u}$  are projected. The coefficients  $P_n(\mathbf{u})$  associated with  $\mathbf{u}$  are then found by solving

$$\sum_i \langle \xi_n | \xi_i \rangle Q_i(\mathbf{u}) = \langle \xi_n | \mathbf{u} \rangle. \quad (60)$$

These can in turn be used to obtain polarizabilities for any displacement  $\mathbf{u}$  using a Taylor expansion similar to Equation (11) but with derivatives over  $\mathbf{u}$ . The polarizability then takes the form

$$\chi(t) = \chi_0 + \sum_n R_n Q_n(\mathbf{u}(t)) + \frac{1}{2} \sum_{n,\mathbf{q}} R_{nn}(\mathbf{q}) Q_n^2(\mathbf{u}(t)), \quad (61)$$

where  $R_{nn}(\mathbf{q})$  are the second-order Raman tensors for wave vector  $\mathbf{q}$  previously defined in Equation (16). While not written explicitly, first order Raman tensors  $R_n$  are only taken at the center of the Brillouin zone ( $\mathbf{q} = 0$ ). Similar to the “harmonic” RGDOS, this “MD” version only requires Raman tensors of a small unit cell to predict polarizabilities for a larger cell at any point during MD, making it very efficient.

Note that Equation (61) contains only the same-phonon terms  $R_{nn}$ , while a complete expression would also contain mixed phonon terms  $R_{nm}$ . Obtaining these latter terms would, however, prove very challenging and the expansion is therefore simplified. Additionally, every wave vector  $\mathbf{q}$  should be included in the sum. Sampling more wave

vectors requires additional polarizability calculations of larger cells, quickly making the whole method too demanding. For these reasons, the expansion is limited only to the same-phonon terms  $R_{nn}$  and a few high symmetry  $\mathbf{q}$ . These inconveniences do not affect the first-order term, making RGDOS particularly suitable when only interested in first-order spectra.

Any basis set of displacement could be used for the expansion in Equation (61) [and the corresponding Equation (11)]. For example, one could use the non-mass-scaled eigenmodes  $\tilde{\xi}_n$  which form an orthogonal basis. In this case, the coefficients  $Q_n(\mathbf{u})$  would simply be given by  $\langle \tilde{\xi}_n | \mathbf{u} \rangle$  and Raman tensors are replaced by  $\frac{\partial \chi}{\partial \tilde{\xi}_n}$ . Alternatively, one could take the displacements of atoms along the Cartesian directions as a basis set. This has been chosen in Ref. [34], showing good agreement with RGDOS results. Here again the Raman tensors have to be replaced and now take the form  $\frac{\partial \chi}{\partial r_{k\alpha}}$ , where  $r_{k\alpha}$  is the displacement of atom  $k$  along the Cartesian direction  $\alpha$ .

While “harmonic” RGDOS has already found successful applications, the “MD” version of RGDOS is a new model. It is therefore unclear how this model compares to others for predictions of the polarizability. It is also important to carefully benchmark how the limited sampling of the reciprocal space in Equation (61) impacts predictions of polarizability and the resulting Raman spectra.

The accuracy of “MD” RGDOS is first tested on simple materials and compared to other models in Publication I, while Publication III shows a more demanding application of this method to obtain the first-order Raman spectra of 2D MXenes. The results of these applications are presented in detail in Chapter 4.

### 3.4 Machine learning models

Machine learning can also be used to predict polarizabilities. Various methods have been proposed. For example, GPR can be used to predict the electronic density and the effect of an external field on it, which in turn can be used to predict polarizabilities [71, 83]. Another example uses the Thole model (presented in Section 3.1) with the dipole interaction field being given by a neural network [117]. However, we focus here on the methods already presented in Section 2.4 and their extensions to predict tensorial properties, which are presented in the remainder of this section.

#### 3.4.1 Symmetry-adapted Gaussian process regressor

Symmetry-adapted GPR (SA-GPR) is an extension of GPR to predict tensorial properties; it was first proposed by Grisafi *et al.* [36]. While applicable to tensors of any rank,

it has mostly been used for dipole moments or polarizabilities. Given the context of Raman spectroscopy, we will focus on the latter. SA-GPR has already been successfully applied to obtain polarizabilities of a large set of molecules [85] as well as paracetamol molecules and crystals [38]. For polarizabilities, this is done by replacing the kernel in Equation (38) by a tensorial kernel  $\mathbf{k}^\lambda(X, X')$ . Instead of learning every component of the tensor separately, SA-GPR uses irreducible spherical tensor representation and  $\lambda$  identifies subspaces of size  $2\lambda + 1$ . For a rank-2 polarizability tensor, both subspaces  $\lambda = 0$  (corresponding to the trace) and  $\lambda = 2$  (corresponding to traceless symmetric matrix elements) have to be considered. Spherical tensor components of the polarizability can then be predicted using GPR similarly to Equation (38), which reads

$$\chi_\mu(X) = \sum_{I, J, i, v} w_i^{\lambda v} k_{\mu v}^\lambda(X^I, X_i^J), \quad (62)$$

where indices  $\mu$  and  $v$  refer to the dimensions of the spherical subspace. SOAP [90] can also be adapted from Equation (39). In this case, Wigner D-matrices  $\mathbf{D}^\lambda$  have to be included and the tensorial kernel becomes [36]

$$\mathbf{k}^\lambda(X, X') = \int d\hat{R} \mathbf{D}^\lambda(\hat{R}) \left| \int d\mathbf{r} \rho(\mathbf{r}) \rho'(\hat{R}\mathbf{r}) \right|^2. \quad (63)$$

Again, similar to GPR, atomic densities can be expanded using radial and spherical functions. Kernels can therefore be reformulated using power spectra  $\mathbf{P}^{\lambda\mu}(X)$ . Note that there is now one power spectrum for each element  $\mu$  of the spherical representation. The tensorial kernel then reads

$$k_{\mu v}^\lambda(X, X') = \sum_{n, n', l, l'} P_{nn' ll'}^{\lambda\mu}(X) P_{nn' ll'}^{\lambda v*}(X'), \quad (64)$$

and by inserting this form of the kernel in Equation (62), the polarizability can be written as

$$\begin{aligned} \chi_\mu(t) &= \left( \frac{1}{N_I} \sum_I \mathbf{P}^{\lambda\mu}(X^I(t)) \right) \cdot \left( \frac{1}{N_J} \sum_{J, i, v} w_i^{\lambda v} \mathbf{P}^{\lambda v \dagger}(X_i^J) \right) \\ &= \mathbf{P}^{\lambda\mu}(t) \cdot \mathbf{P}_T^{\lambda \dagger}. \end{aligned} \quad (65)$$

In this form, each spherical tensor component is obtained by simply projecting its power spectra  $\mathbf{P}^{\lambda\mu}$  onto the power spectra of the training set  $\mathbf{P}_T^{\lambda \dagger}$ . When applied to MD, the latter is constant and therefore does not need to be computed at every time step, making it more efficient. For polarizabilities, there are two of these training power spectra (one for  $\lambda = 0$  and one for  $\lambda = 2$ ). Similarly, there are two sets of weights  $w^{\lambda v}$  which have

to be trained separately. Once spherical tensorial components  $\chi_\mu$  are obtained, they can be transformed back into Cartesian coordinates  $\chi_{ij}$ .

While SA-GPR has been successfully applied to predict the polarizabilities of a wide variety of molecules, it remains unclear how this would transfer to the prediction of Raman spectra of solids. In general, ML methods require large amount of training data. In Publication I we compared SA-GPR with RGDOS and BPM in order to better understand the advantages of each model.

### 3.4.2 Tensorial neuroevolution potential

The neuroevolution potential (NEP) has recently been extended by Xu *et al.* to predict tensorial properties [37], in which case the method is referred to as TNEP. The main motivation was to predict polarization and polarizability in order to compute infrared and Raman spectra, respectively. It has already successfully been applied to predict the Raman spectra of water molecules as well as BaZrO<sub>3</sub> crystals [37]. TNEP uses the fact that stress tensors and polarizabilities obey similar transformation under rotation. Starting from the stress tensor of Equation (45), local polarizabilities  $\chi_i$  are obtained as

$$\chi_i = O_i \cdot 1 - \sum_{j \neq i} \mathbf{r}_{ij} \otimes \frac{\partial O_j}{\partial \mathbf{r}_{ji}}, \quad (66)$$

where  $O_j$  are the output of the neural network and 1 is the unit tensor of rank 2. Note that the first term has been added to account for the fact that polarizability tensors usually have large diagonal values. With regard to polarizability,  $O_i$  has a unit of polarizability and no energy anymore. It therefore cannot be used to obtain energies, forces and stress tensors as was previously possible for interatomic potential. This means that the loss function from Equation (46) used for the training of TNEP has to be modified and contains only one polarizability term in addition to the regularization terms. The neural network is then optimized by minimizing this new loss function using the same SNES algorithm as for NEP.

TNEP was introduced only recently and still lacks extensive testing and applications. In particular, it remains unclear how this model compares to other ML methods such as SA-GPR. One of the remaining challenges of ML is its transferability, that is, its applicability to systems outside training data. In Publication IV, the transferability of NEP and SA-GPR were compared using biomolecules, namely amino acids and peptides.

## 4 Applications

This chapter presents applications of the polarizability models to various materials. Table 1 summarizes the different applications and the choice of polarizability models and interatomic potentials. Boron arsenide (BAs) is first used as a simple material to compare models and investigate their possible limits. Models are then applied to more complex materials, namely single-layer molybdenum disulfide ( $\text{MoS}_2$ ), 2D titanium carbide MXenes ( $\text{Ti}_3\text{C}_2\text{T}_2$ ), inorganic halide perovskites ( $\text{CsPbBr}_3$  and  $\text{CsSnBr}_3$ ) and peptide chains. In these cases, only one of the polarizability models is used depending on the studied systems and the desired properties. The choice of interatomic potential also changes for each system.

**Table 1. Summary of the materials presented here as applications, including the interatomic potentials and polarizability models used as well as the properties of interest.**

Material	Interatomic potentials	Polarizability models	Properties of interest	Publication
BAs	NEP	RGDOS, BPM & SA-GPR	Models benchmark	I
$\text{MoS}_2$	NEP	BPM & SA-GPR	Resonant spectra	I
$\text{MoS}_2$	VASP MLFF	RGDOS (harmonic)	Point defects	II
MXenes	VASP MLFF	RGDOS (MD)	Surface terminations	III
$\text{CsXBr}_3$	NEP*	SA-GPR	Central peak	I
Peptides	CHARMM27	SA-GPR & TNEP	Models transferability	IV

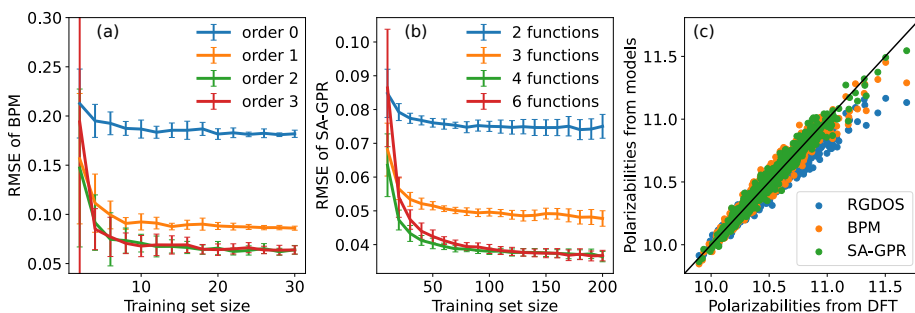
\*NEP potentials previously trained by Fransson *et al.* [118].

### 4.1 Boron arsenide

Boron arsenide (BAs) is a semiconductor crystal with a particularly high thermal conductivity, making it interesting for applications in electronics and heat management. Its crystal structure is similar to that of a diamond, and the material only contains two different chemical elements, making it a fairly simple material to model. The Raman spectra of BAs has been extensively studied experimentally. It shows only one first-order peak around  $700\text{ cm}^{-1}$  and a broad second-order peak between  $1200$  and  $1500\text{ cm}^{-1}$ . Given the simple nature of BAs and its Raman spectra, this material represents an ideal test subject for the polarizability models. In Publication I, RGDOS, BPM and SA-GPR are all trained and their accuracy compared. Models are then used to obtain first- and second-order Raman spectra, which are again compared. Additionally, the

spectra of BAs are sensitive to the concentration of boron isotope. Spectra at various concentrations are therefore also computed and reported.

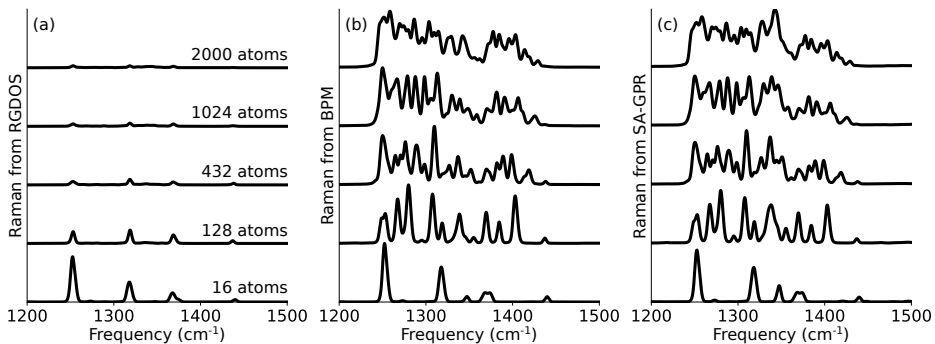
All models are trained using  $2 \times 2 \times 2$  supercells containing 16 atoms. For RGDOS all 48 modes of the supercell are considered and polarizabilities are expanded up to the second order [similar to what is written in Equation (61)], leading to a total of 97 calculations. RGDOS is the only model using a fixed number of modes and calculations. For BPM and SA-GPR, the quality of the model depends on the number of structures used in the training set, which requires careful testing. Such benchmarks typically include looking at the root mean squared error (RMSE) decreasing as the training set size increases. In Fig. 8, the RMSE of the validation set (containing 100 structures independent from the training set) is shown as a function of the training set size. For BPM, polynomial orders are compared in Fig. 8(a). While zero-th and first order BPM are clearly less accurate, the RMSE is converged for orders higher than 2. Also note that for BPM, only a few structures (roughly 15) are necessary to reach convergence. On the other hand, SA-GPR requires more training data, as can be observed in Fig. 8(b). In this case, models only converged for training sets larger than 150 structures, but the resulting RMSE is lower than in the case of BPM. From Fig. 8(b), one can also conclude that using more than 4 radial and angular functions does not lead to significant improvements in model accuracy. The resulting converged models are lastly compared to DFT results in Fig. 8(c). All models accurately reproduce DFT polarizabilities, with RMSE of 0.086, 0.058 and 0.043 for RGDOS, BPM and SA-GPR, respectively. While RGDOS shows the highest error, it must be noted that this model does not take into account some effects, such as the higher than second order terms and mixed phonon terms [due to the finite Taylor expansion written in Equation (61)]. Average polarizability



**Fig. 8. (a) Convergence of BPM root mean-squared error (RMSE) with training set size for different expansion orders. (b) Convergence of SA-GPR with training set size for different numbers of radial and angular functions. (c) Comparison of DFT polarizabilities with those obtained from the three final models. (Reprinted, with permission, from Publication I © 2024 American Physical Society).**

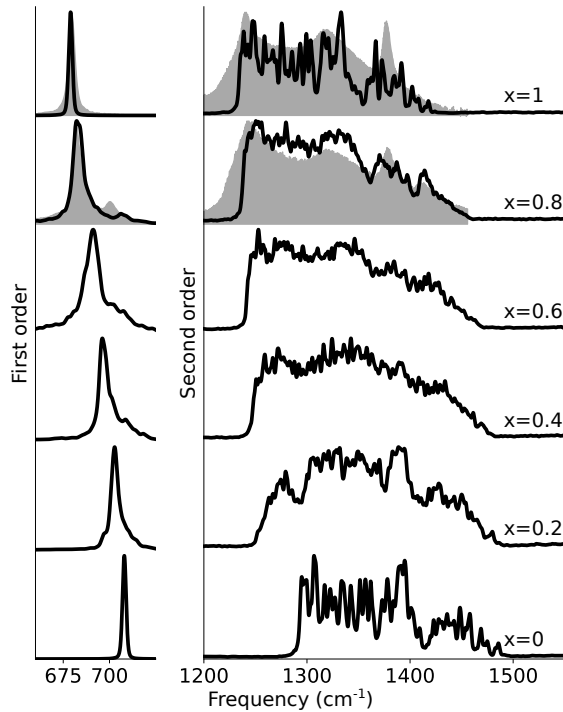
is therefore less accurate, even though first and second order Raman spectra are still correctly reproduced. Also note that the RGDOS expansion can easily be restricted to particular modes or parts of the Brillouin zone, allowing for decomposition of the different contributions to the Raman spectra. While this is not really useful in the case of BAs (with only one active mode), this feature will be used later when investigating MXenes. Such decomposition is impossible with BPM and SA-GPR.

With the polarizability models being sufficiently trained, one can now use them in combination with MD to produce Raman spectra. In the case of BAs, NEP is used as interatomic potential, and 1 ns long trajectories are performed at 300 K for  $10 \times 10 \times 10$  supercells (containing 2000 atoms). The resulting Raman spectra from different polarizability models all show a very similar first-order peak around  $700 \text{ cm}^{-1}$  in good agreement with expectations from experimental results. However, the second-order spectra depend greatly on the method as well as on the supercell size, as is reported in Fig. 9 for all three models. For small  $2 \times 2 \times 2$  supercells (16 atoms), all models lead to similar second-order spectra, exhibiting clear peaks. Each of these correspond to points in the reciprocal space directly sampled with a  $2 \times 2 \times 2$  supercell. In the case of RGDOS [Fig. 9(a)] increasing the supercell size does not change the resulting spectra, as the points sampled are limited by the modes included in the expansion. In the case of SA-GPR (as well as BPM), increasing the supercell size leads to a better sampling of the reciprocal space and additional peaks are observed. For a large  $10 \times 10 \times 10$  supercell (2000 atoms), second-order spectrum from both BPM and SA-GPR shows a large band between  $1250$  and  $1450 \text{ cm}^{-1}$ , which agrees well with experiments.



**Fig. 9. Second-order Raman of BAs for different sizes of supercells using (a) RGDOS, (b) BPM and (c) SA-GPR polarizability models. (Adapted, with permission, from Publication I © 2024 American Physical Society).**

Finally, we investigate the effect of B isotope content in  $B_x^{11}B_{1-x}^{10}As$ . The results from SA-GPR are shown in Fig. 10, where the first-order spectra are represented on the left panel and the second-order on the right one. Experimental results from Ref. [119] are also represented for pure  $B^{11}As$  and  $B_{0.8}^{11}B_{0.2}^{10}As$  (corresponding to the natural isotope content of BAs). Experimental spectra are redshifted by  $19\text{ cm}^{-1}$  to account for the error of DFT/NEP and obtain better visual comparison. The first-order spectra peak shifts from  $679\text{ cm}^{-1}$  to  $708\text{ cm}^{-1}$  when going from pure  $B^{11}$  to pure  $B^{10}$ . This shift of  $30\text{ cm}^{-1}$  is in good agreement with previous calculations as well as experimental observations [119, 120, 121]. At natural isotope content ( $x = 0.8$ ), a small shift of  $4\text{ cm}^{-1}$  can be observed as well as the appearance of a smaller peak close to the  $B^{10}$  frequency. These two features also agree well with previous experimental measurements. For the second-order spectra (right panel), our results compare well with the experimental spectra for pure  $B^{11}As$ , including the peak around  $1320\text{ cm}^{-1}$ . This peak comes from perpendicular contributions (off-diagonal components of the Raman tensors). For



**Fig. 10.** First (left) and second (right) order Raman spectra of  $B_x^{11}B_{1-x}^{10}As$  with different isotope content  $x$ . The areas shaded in gray show experimental measurements from Ref. [119]. The spectra are obtained using SA-GPR. (Adapted, with permission, from Publication I © 2024 American Physical Society).

$B_{0.8}^{11}B_{0.2}^{10}As$ , an additional peak around  $1400\text{ cm}^{-1}$  appears in both our simulated and the experimental spectra. For higher concentration of  $B^{10}$ , the spectra shift to higher frequencies, which is also in good agreement with experimental results [121].

RGDOS, BPM and SA-GPR were successfully tested and applied to BAs. All three models are capable of accurately predicting polarizabilities. While RGDOS is found to be the least accurate model, it also possesses some advantages, mainly flexibility in the expansion and a fixed number of calculations. While BPM was found to require only tens of structures to correctly predict polarizabilities, SA-GPR requires a larger training set but leads to more accurate predictions. It is also important to note that SA-GPR was found to be the slowest method, mainly due to the computation of descriptors. RGDOS and BPM rely only on displacements/bonds, making them simpler to implement and generally faster. Overall, all the models have advantages and disadvantages and can find applications depending on the investigated system.

## 4.2 Molybdenum disulfide

Molybdenum disulfide ( $MoS_2$ ) is a 2D semiconductor with potential applications in electronics. The nature of 2D materials makes them highly anisotropic, with in-plane modes and out-of-plane modes behaving independently and at different frequencies. Single-layer  $MoS_2$  has two Raman active modes, which are represented in Fig. 11. A double peak is found around  $380\text{ cm}^{-1}$  due to the asymmetric motion of sulfur atoms in-plane, which are labeled  $E_g$ . The second peak originates from out-of-plane vibrations of sulfur atoms and is referred to as  $A_{1g}$ . A representation of these two modes at the edge of the Brillouin zone is also shown in Fig. 11. Contrary to BAs, second-order

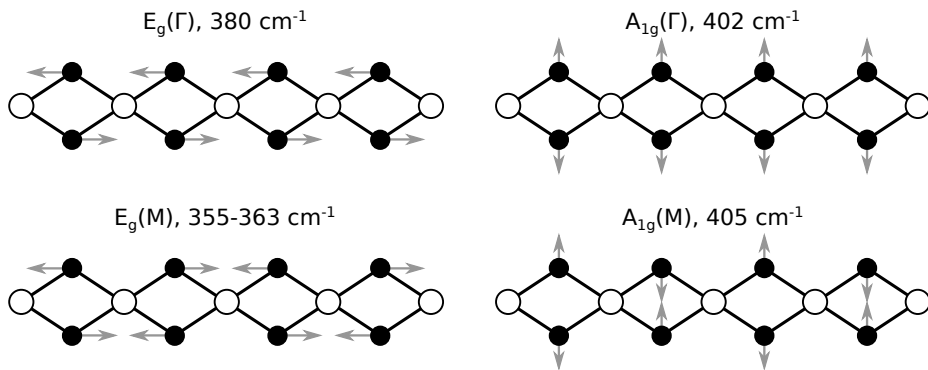


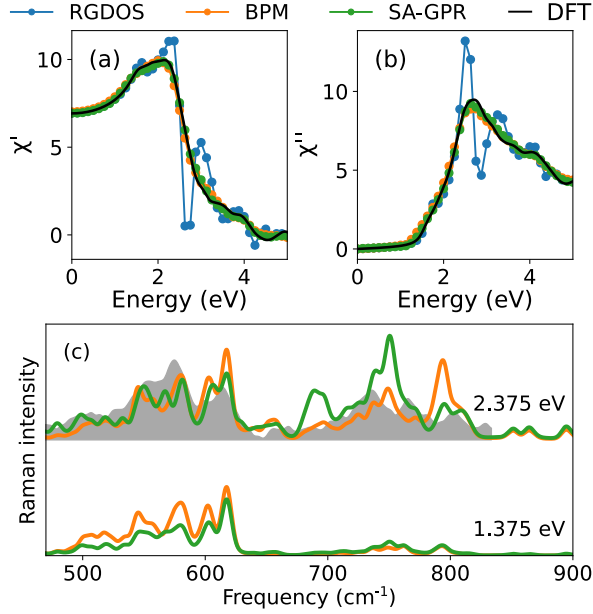
Fig. 11. Representation of the main vibrational modes of  $MoS_2$  with their frequency calculated at the DFT level. Mo are in white and S in black.

peaks of MoS<sub>2</sub> are observed experimentally only in the resonant case [122, 123]. This material is therefore a good subject to study resonant spectra and for testing how it can be reproduced with each model. Additionally, the formation of point defects are very common in 2D materials due to the high surface area in contact with ambient air. In the case of MoS<sub>2</sub>, such defects affect the electronic properties and could be used to tune its properties. Simulations of defects require large supercells, which rapidly makes Raman calculations at the DFT level impossible. Investigations of defects could therefore greatly benefit from the efficient polarizability models presented so far. This section presents results for both resonant Raman spectra of MoS<sub>2</sub> and the impact of point defects on the spectra.

#### **4.2.1 Resonant Raman spectra**

Similar to BAs, RGDOS, BPM and SA-GPR were all trained for MoS<sub>2</sub>. 2×2 supercells (containing 12 atoms) were used for the training of the polarizability models. For RGDOS, the expansion contains all 36 modes up to the second order and mixed phonon terms are omitted. Second order polynomials were used for BPM and radial and angular functions up to the fourth order are included for SA-GPR. These correspond to the optimal parameters previously found for BAs. Note that for MoS<sub>2</sub>, Mo-Mo bonds are also considered in addition to the first nearest neighbor Mo-S bonds. In total, 500 structures were included in the training set, which should be sufficient considering previous values found for BAs. While not explicitly stated previously, only the real part of the dielectric function was previously used in the non-resonant case since the imaginary part is always zero. This is no longer true for resonant spectra and one has to take into account the imaginary part. This means that both parts have to be trained separately. Similarly to BAs, RGDOS, BPM and SA-GPR are all trained in order to compare their accuracy. To reproduce the whole dielectric function, 40 models were trained at fixed frequencies equally distributed between 0 and 4 eV. In total, 80 sets of parameters have to be trained independently.

The resulting real and imaginary parts are compared with DFT in Fig. 12(a) and (b), respectively. For low excitation energies, all three models lead to great accuracies, with respective RMSEs of 0.079, 0.070 and 0.022 for RGDOS, BPM and SA-GPR. These values compare well to the ones previously obtained for BAs, indicating again the good accuracy of all the models. At higher energies, RGDOS is found to lead to much worse predictions, while BPM and SA-GPR both perform well for both parts of the dielectric functions at all frequencies. RGDOS is therefore discarded and only BPM and SA-GPR are used to obtain Raman spectra.



**Fig. 12. (a) Real and (b) imaginary parts of the dielectric function of a displaced structure. RGDOS, BPM and SA-GPR are compared with DFT results (in black). (c) Resonant second-order Raman spectra of MoS<sub>2</sub> at different excitation energies (1.375 eV and 2.375 eV). Results from BPM and SA-GPR are compared with experimental results (in gray) from Ref. [122]. (Adapted, with permission, from Publication I © 2024 American Physical Society).**

Second-order Raman spectra from BPM and SA-GPR are shown in Fig. 12(c), where they are compared with experimental results from Ref. [122]. The experimental results are obtained using a laser of 532 nm, corresponding to an excitation energy of 2.33 eV (close to the value of 2.375 we used). Also note that the experimental spectrum is redshifted by 14  $\text{cm}^{-1}$  to account for the error from DFT calculations. At excitation energy of 1.375 eV, a few peaks are visible between 500 and 600  $\text{cm}^{-1}$ . One peak is clearly visible at 619  $\text{cm}^{-1}$ , which agrees with other experimental observations [122, 123]. Peaks in this region are usually associated with second-order scattering between optical and acoustic modes at the border of the Brillouin zone. For higher excitation energy, additional peaks are visible between 700 and 800  $\text{cm}^{-1}$ , which correspond to second-order peaks from two optical modes. Although both models show great agreement between 500–600  $\text{cm}^{-1}$ , they agree less at higher frequencies. Experimental measurements also report a very intense peak around 750  $\text{cm}^{-1}$  when using a laser of 354 nm (3.5 eV), which are not reproduced in our results. Such discrepancies might come from the inaccuracy of the dielectric constant at the DFT level. Dielectric functions from more accurate methods (typically many-body theory

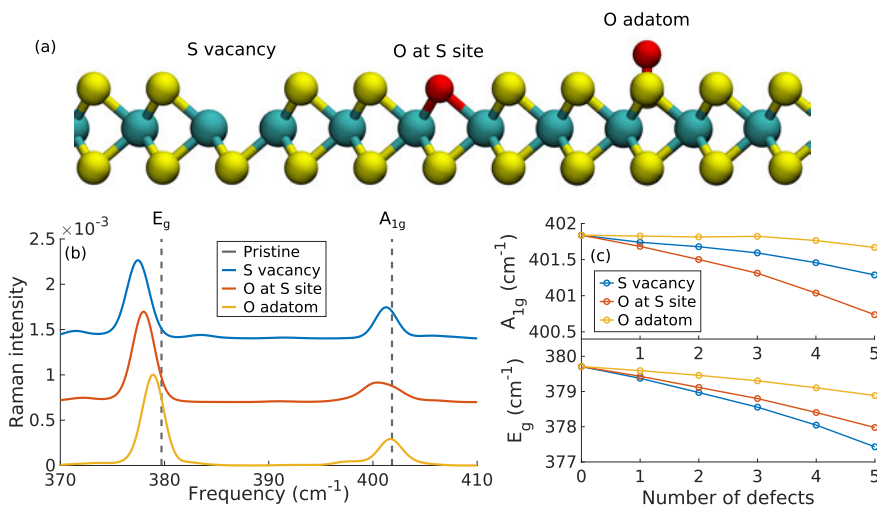
or time-dependent DFT) might improve the agreement with the experimental results [124, 125]. For example, SA-GPR has already been trained using calculations more accurate than DFT, resulting in models more accurate than those trained on DFT [85]. This shows that ML cannot be more accurate than the data used for training, in our case DFT polarizabilities.

While RGDOS, BPM and SA-GPR can all be successfully applied in the non-resonant case for both BAs and MoS<sub>2</sub>, we find that only the two latter lead to accurate dielectric functions in the resonant case. Both models also lead to reasonable resonant Raman spectra in good agreement with dielectric functions from DFT. This also made it possible to reproduce experimental observations of the second-order Raman spectra of MoS<sub>2</sub>, although agreement could be improved by using higher level simulation methods.

#### **4.2.2 Effect of defects**

2D materials can contain a high concentration of defects arising from their large surface area in contact with ambient air. These defects can either be unintentionally created during synthesis or artificially added to serve a particular purpose. This latter case can be particularly useful in the case of MoS<sub>2</sub>, where defect engineering can be used to tune optoelectronic properties. To this end, electron (or ion) irradiation can be used to produce point defects in the MoS<sub>2</sub> monolayer and this is what was studied in Publication II.

The formation of defects also impacts the vibrational properties of materials and in turn the Raman spectrum, which can be used to study the type and density of the defects. Experimentally, it is very challenging to understand the impact of defects on the Raman spectra, as the exact type of defect and their concentration is usually unknown. In such situation, atomic scale calculations can prove very useful. In Publication II, three types of defects, namely sulfur vacancies, oxygen at the sulfur site and oxygen adatom [represented in Fig. 13(a)], are studied and their impact on Raman spectra is compared to experimental results. Spectra are obtained using the harmonic version of RGDOS and Equation (59). Vibrational modes of large supercells are obtained using on-the-fly machine learning force fields (as implemented in VASP). The MLFF model is first trained using  $4 \times 4$  pristine MoS<sub>2</sub> structures, before being further trained for all three kinds of defects. The final model is used to obtain phonon modes of large  $10 \times 10$  defective (and pristine) supercells. Note that such large supercells are necessary to study the low density of defects and possibly random distribution. For comparison, using  $4 \times 4$  supercells only enable reaching a defect concentration of  $\sim 6\%$ . With  $10 \times 10$



**Fig. 13.** (a) Balls-and-sticks representation of the three kinds of defects, with Mo in cyan, S in yellow and O in red. (b) Raman spectra of MoS<sub>2</sub> for all three kinds of defects at 5% concentration. Peak positions of the pristine structure are represented with dashed black lines. (c) Evolution of peak positions of  $A_{1g}$  and  $E_g$  modes with defect concentrations. [Panels (b-c) are adapted, with permission, from Publication II © 2023 IOP Publishing Ltd].

supercells, a concentration of up to 1% can be studied. The simulation of such large supercells at the DFT level would prove impossible, hence the necessary use of ML.

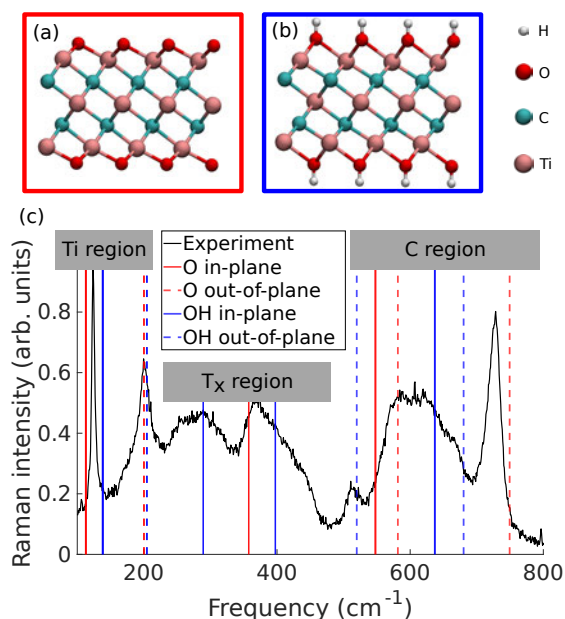
Fig. 13(b) shows the Raman spectra with 5% defects. Peak position is compared with frequencies for the pristine supercell (represented using a dashed vertical line). While all three kinds of defects lead to a redshift of the  $E_g$  mode, only sulfur vacancies and oxygen at the sulfur sites also lead to a redshift of the  $A_{1g}$  mode, while the frequency of this latter mode remains unchanged in the presence of oxygen adatom. Note that the shift is larger for the  $E_g$  mode than for the  $A_{1g}$  mode. Fig. 13(c) shows the position of peaks with respect to the number of defects.

In Publication II, respective shifts of  $\sim 1.2 \text{ cm}^{-1}$  and  $\sim 0.3 \text{ cm}^{-1}$  are experimentally observed for the  $E_g$  and  $A_{1g}$  peaks when irradiating the MoS<sub>2</sub> monolayer with an electron beam. According to Fig. 13(c), such a shift would correspond to a high density of oxygen adatoms, discarding this kind of defect. For oxygen at the sulfur site, no single density can simultaneously satisfy both shifts. This leaves only sulfur vacancies, for which a concentration of 3–4% shows good agreement with both shifts. Harmonic RGDOS was therefore successfully used to understand the Raman spectra of defective MoS<sub>2</sub>. The results were used to support experimental observations and to help identify which kinds of defects were created.

### 4.3 Titanium carbide MXenes

MXenes represent one of the largest families of 2D materials. They take the stoichiometric form  $M_{n+1}X_nT_2$ , where M is an early transition metal, X is either carbon or nitrogen, T are the surface terminations and  $n = 1-4$  defines the thickness of the material. MXenes are synthesised from the MAX parent phase, where A is group A element (typically aluminium). After etching of the A-layers, dangling bonds at the surfaces are filled with random surface terminations. The most common terminations are  $-O$ ,  $-OH$  and  $-F$ , but their exact ratio cannot be controlled during synthesis, leading to generally unknown compositions. We focus here on the most studied MXene  $Ti_3C_2T_2$ . Ball-and-sticks representations of pure  $Ti_3C_2O_2$  and  $Ti_3C_2(OH)_2$  are shown in Fig. 14(a) and (b), respectively. Raman spectroscopy could be used to study terminations ratios. Experimental spectra exhibit clear sharp peaks around  $100$  and  $200\text{ cm}^{-1}$ , as well as wide peaks between  $300-400\text{ cm}^{-1}$  and around  $600\text{ cm}^{-1}$ . Most computational studies focus on homogeneous surfaces, in which case the calculated frequencies of Raman peaks do not agree well with experimental measurements. A comparison between experimental spectra and calculated frequencies for homogeneous surfaces is shown in Fig. 14(c). Mixed surfaces are therefore required in order to obtain more realistic spectra. Similar to  $MoS_2$ , a simulation of heterogeneous surfaces calls for large supercells in order to study small changes in surface concentrations, which in turn requires the use of ML methods. Also note that wide peaks cannot be obtained when using harmonic approximation. MD is therefore required to obtain realistic peak widths.

In Publication III, Raman spectra of large supercells are computed to study the impact of surface terminations on the spectra. For this particular application, we decided to use RGDOS to model the polarizabilities. This choice is motivated by the fact that mixed surface MXenes contain many different atomic species (C, Ti, O, H and possibly F), making the application of BPM impossible. On the other hand, vibrational modes are found to be very similar for every type of termination. This makes RGDOS easily applicable, as only one set of phonons and Raman tensors is necessary. Additionally, the expected Raman spectra contain only first-order terms, making RGDOS application even more suitable. SA-GPR and NEP would probably also be suitable in this case, but unnecessarily complex. For  $Ti_3C_2X_2$  MXenes, there are only 21 Raman tensors to compute, hence only 43 polarizability calculations to get RGDOS. This number is far lower than what would be necessary to obtain good quality ML models for systems containing four different atomic species. Note that this number would be far greater if the second-order terms had to be included in the expansion.

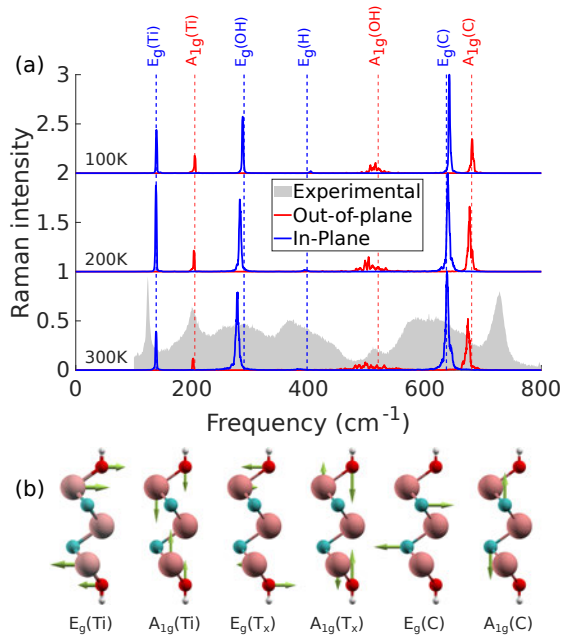


**Fig. 14.** (a)-(b) Balls-and-sticks representation of MXene with pure  $-O$  and pure  $-OH$  termination, respectively. (c) Frequencies of the Raman-active peaks for pure  $-O$  (in red) and pure  $-OH$  (in blue) MXene. Experimental spectrum (in black) is added for comparison. (Reprinted under CC BY 3.0 license from Publication III © 2023 Authors).

MD trajectories are obtained using VASP MLFF. The final training set contains 1364 structures of  $4 \times 4$  supercells with various concentrations of surface terminations. Raman spectra are calculated from 100 ps long trajectories of  $8 \times 8$  supercells. For each concentration, 5 different runs are performed using different distributions of the terminations and the average Raman spectra is taken<sup>1</sup>. This process is repeated for various concentrations ranging from pure  $-O$  to pure  $-OH$  structures.

The vibrational modes of MXenes are similar to those of  $MoS_2$ , with each pair of atomic layers (other than the central titanium) having one  $E_g$  and one  $A_{1g}$  mode. Note that the hydrogen modes, labeled  $E_g(H)$  and  $A_{1g}(H)$  can be neglected. While  $A_{1g}(H)$  has a frequency around  $3600 \text{ cm}^{-1}$  and is irrelevant to the current application,  $E_g(H)$  shows negligible intensity as can be seen from Fig. 15(a). There are therefore 6 Raman active modes in total, which are represented on Fig. 15(b). Before investigating the effect of surface terminations, we first look at the effect of temperature on the Raman spectra. The results are presented in Fig. 15 for a pure  $-OH$  surface. Raman spectra

<sup>1</sup>We note that the distribution of surface termination does not seem to impact the resulting Raman spectra. More information on this topic can be found directly in Publication III.



**Fig. 15. (a) Raman spectra of  $\text{Ti}_3\text{C}_2(\text{OH})_2$  at different temperatures. The blue lines represent the in-plane modes while the red ones represent the out-of-plane modes. The gray areas show the experimental results and the dashed vertical lines show the frequencies from phonon calculations using the MLFF. (b) Raman-active eigenmodes. (Adapted under CC BY 3.0 license from Publication III © 2023 Authors).**

at various temperature are compared with experimental measurements (in gray) and with the frequencies obtained from DFT calculations (vertical dashed lines). For some peaks, mainly  $E_g(\text{OH})$  and  $A_{1g}(\text{C})$ , increasing temperatures lead to a small shift to lower frequencies. Moreover, some peaks show slightly wider peaks at higher temperatures. At room temperature (300 K), these widths are, however, still underestimated compared to experimental observations. Additionally, the positions of sharp peaks around  $120\text{ cm}^{-1}$  and  $700\text{ cm}^{-1}$  are very poorly reproduced. These discrepancies hint that something is missing and that inhomogeneous surfaces might be necessary to correctly reproduce experimental results.

Raman spectra for different concentrations of surface termination are presented in Fig. 16, where in-plane and out-of plane contributions are separated, and the results are compared with experimental spectra (in gray). The  $A_{1g}$  mode of titanium at  $200\text{ cm}^{-1}$  is largely unaffected by changes of the surface. The other two sharp peaks, namely  $E_g(\text{Ti})$  at  $120\text{ cm}^{-1}$  and  $A_{1g}(\text{C})$  at  $700\text{ cm}^{-1}$ , both show a linear shift with the  $-\text{O}/-\text{OH}$  ratio. The position of these three peaks could therefore be used to obtain an idea of the

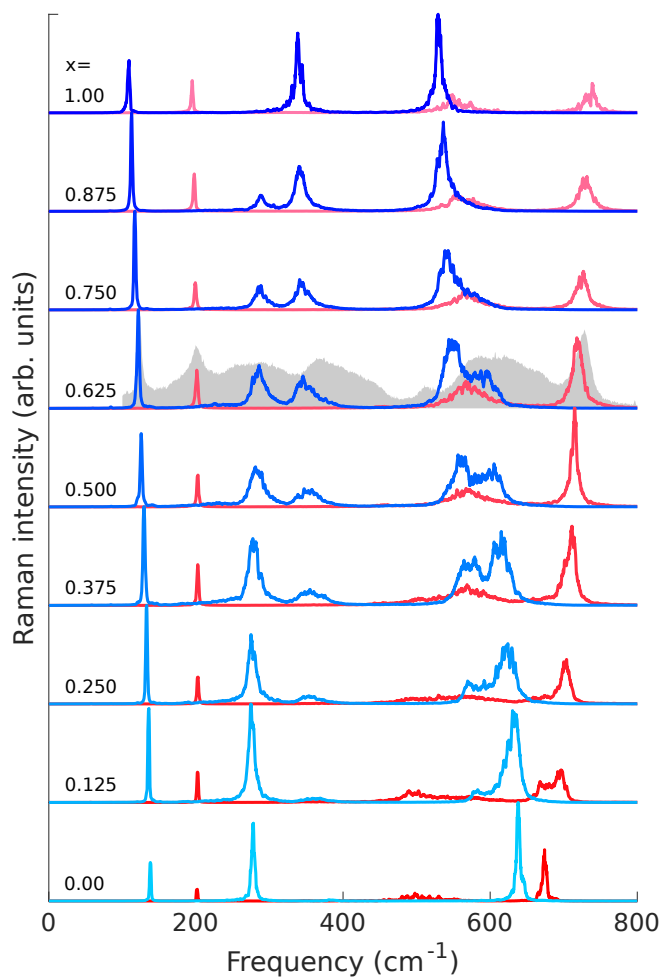


Fig. 16. Raman spectra of  $\text{Ti}_3\text{C}_2(\text{O}_x\text{OH}_{1-x})_2$  for different surface group compositions. Labels on the left show the concentration of oxygen at the surface, e.g. the top and bottom lines represent pure  $-\text{O}$  and  $-\text{OH}$  surfaces, respectively. In-plane modes are represented by the blue lines while out-of-plane modes are in red. The gray area shows an experimental spectrum for comparison. (Adapted under CC BY 3.0 license from Publication III © 2023 Authors).

surface composition from experiment. Table 2 compares the peak positions at different concentrations with experimental positions from Ref. [126]. The best agreement is found for concentrations of 0.5–0.625, which agree well with other estimations from calculations [127]. For these concentrations, wide peaks at 300–400  $\text{cm}^{-1}$  and 600  $\text{cm}^{-1}$  also show good agreement. The  $E_g$  modes of the terminations show two distinct relatively wide peaks, which are not correctly reproduced when only considering pure surfaces. The  $A_{1g}$  mode of the terminations and  $E_g(\text{C})$  also exhibit a wider peak around 600  $\text{cm}^{-1}$ , leading to an overall much better agreement with experiments.

**Table 2. Frequencies (in  $\text{cm}^{-1}$ ) of the  $E_g(\text{Ti})$ ,  $A_{1g}(\text{Ti})$ , and  $A_{1g}(\text{C})$  modes for different ratios of O/OH surface terminations, and comparison to experimental values. (Adapted under CC BY 3.0 license from Publication III © 2023 Authors).**

Mode	Simulated							Experiments
	1.000	0.875	0.750	0.625	0.500	0.250	0.000	Ref. [126]
$E_g(\text{Ti})$	108.4	112.5	116.8	121.6	125.7	132.9	138.1	119.8–124.5
$A_{1g}(\text{Ti})$	194.9	197.7	199.2	200.3	202.3	202.5	201.5	201.0–206.7
$A_{1g}(\text{C})$	737.4	729.4	725.2	718.9	714.5	702.0	673.9	737.7–719.6

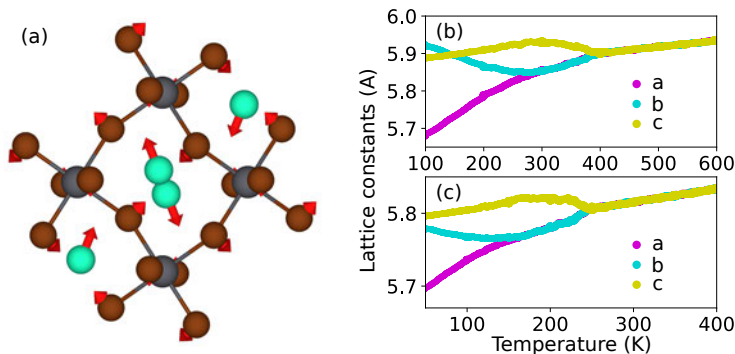
Note that the results presented so far only include modes at the center of the Brillouin zone ( $\Gamma$ -point) since only these contribute to the first-order Raman spectra. However, if the translational symmetry of the crystal is broken, modes outside of  $\Gamma$  will start contributing. In MXenes, such a reduction of symmetry can happen because of the finite size of the flake, the presence of defects or simply because of their inhomogeneous surfaces. To get an idea of how these modes would contribute to the Raman spectra, we investigated the phonon density of states in Publication III. We found that these additional modes lead to further broadening some peaks, especially in between 300–400  $\text{cm}^{-1}$  and around 600  $\text{cm}^{-1}$ , as well as around the peak at 200  $\text{cm}^{-1}$ . Although this would lead to better agreement with experimental measurements, computing the Raman tensor of these modes would require calculations of polarizabilities for very large supercells, which we deemed impossible. Similarly, explicitly including defects into the simulations would prove very challenging, as training of the model would be more complex and expensive calculations of large supercells would be necessary.

Overall, the “MD” version of RGDOS was successfully applied to MXenes monolayers with inhomogeneous surfaces. The high efficiency of the model made it possible to easily obtain Raman spectra of large supercells with mixed surface terminations, improving the quality of previous calculations for this material and providing a valuable benchmark for future experiments.

## 4.4 Inorganic halide perovskites

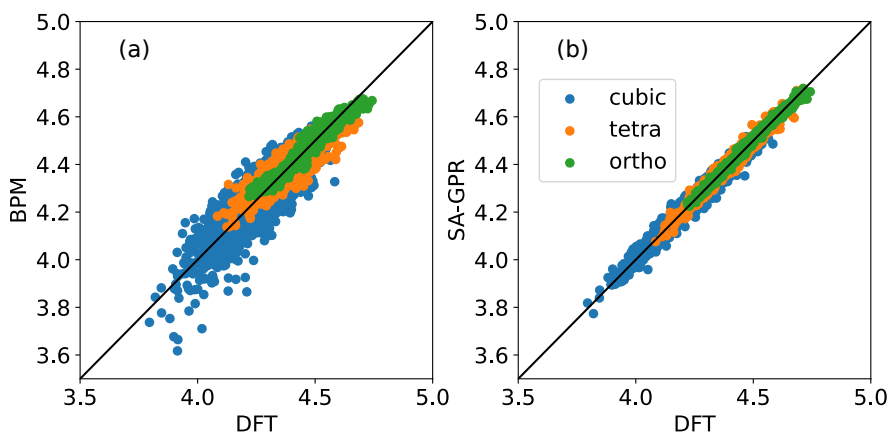
Halide perovskites represent some of the most promising materials for the next generation of solar panels, with a record efficiency of more than 25% [128, 129]. They adopt the stoichiometric form  $ABX_3$ , with A being a monovalent cation, B being a metal and X being a halide. Most efficient solar cells use alloys containing many species, including organic molecules as cations [129, 130]. For simplicity, we only consider cesium-based perovskites that take the form  $CsXBr_3$ , with X being either lead or tin. In this case, the cesium cation is located at the center of a cage made of eight  $XBr_6$  octahedra. The orientation of octahedra changes with temperature, leading to three different phases. At low temperatures, octahedra are tilted in fixed positions and the phase is orthorhombic. Fig. 17 shows a representation of the orthorhombic phase. With increased temperature, perovskites become tetragonal. In this phase, octahedra can rotate in one direction while still being fixed in others. Finally, at high temperatures, octahedra are completely free to rotate and their average structure is cubic. Phase transition can be observed by monitoring the changes in lattice constants during heating (or cooling), which are represented in Fig. 17(b) and (c) for  $CsPbBr_3$  and  $CsSnBr_3$  respectively. Phase transitions from orthorhombic to tetragonal are found to happen around 250 K and 150 K, while tetragonal to cubic happens at 400 K and 250 K for  $CsPbBr_3$  and  $CsSnBr_3$ , respectively.

Given the complex anharmonic behavior of perovskites and their three different phases, the modeling of such materials is challenging. In Publication I, the main goal was to obtain Raman spectra at finite temperatures for all phases and at the phase



**Fig. 17. (a) Atomic structure of inorganic perovskites, with cesium in cyan, bromide in brown and lead/tin in dark gray. Red arrows show the phonon mode pertaining to octahedral tilting. (b-c) Lattice constants with temperature during heating for  $CsPbBr_3$  and  $CsSnBr_3$ , respectively. The different colors correspond to the different lattice constants. (Adapted, with permission, from Publication I © 2024 American Physical Society).**

transition temperatures. This therefore requires that both the interatomic potential and the polarizability model are properly trained for all phases. Fransson *et al.* recently developed NEP for a wide range of cesium-based perovskites [118]. These potentials contain all three phases and are reused here to perform MD. For the polarizability model, a method based on phonons like RGDOS is clearly not applicable given the large differences in atomic structures between the three phases. Both BPM and SA-GPR models were trained for CsPbBr<sub>3</sub> using  $\sqrt{2} \times \sqrt{2} \times 2$  supercells (containing 20 atoms) at various temperatures. The training set contains 300 orthorhombic structures, 300 tetragonal and 600 cubic, leading to a total of 1200 structures. Note that this number is much larger than the training sets previously used for BAs and MoS<sub>2</sub> due to the additional complexity of perovskites (three phases and anharmonicity at high temperature). A comparison of the resulting models with DFT can be found in Fig. 18.

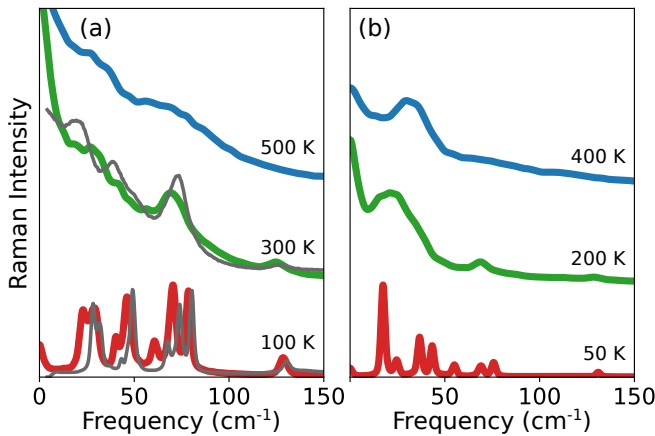


**Fig. 18. Comparison between polarizabilities from (a) BPM and (b) SA-GPR with DFT for CsPbBr<sub>3</sub>. Three temperatures are compared to show all phases. (Adapted ,with permission, from Publication I © 2024 American Physical Society).**

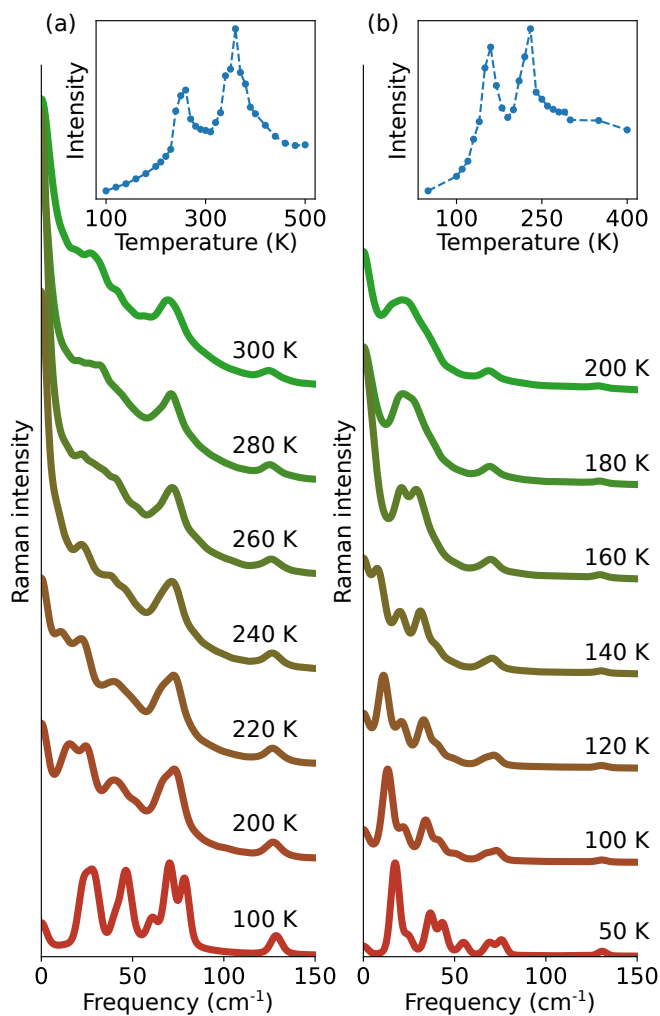
While BPM shows accurate predictions for the orthorhombic phase, it is clearly not suitable for the cubic phase. This is likely due to BPM only accounting for bond length and neglecting three-body (and higher terms) which might play an important role in correctly accounting for octahedra rotations at high temperatures. On the other hand, SA-GPR is found to have similar accuracy for all three phases. For comparison, SA-GPR leads to RMSE values of 0.011, 0.019 and 0.023 for the orthorhombic, tetragonal and cubic phase, while the RMSE values of BPM are 0.025, 0.048 and 0.071 for the same phases. It is pretty clear that SA-GPR is well trained and can accurately predict polarizabilities in all three phases. SA-GPR was therefore used to obtain the Raman spectra of CsPbBr<sub>3</sub> and CsSnBr<sub>3</sub>.

Raman spectra were obtained from 4 ns long trajectories using  $6\sqrt{2} \times 6\sqrt{2} \times 8$  supercells (containing 2880 atoms). For  $\text{CsPbBr}_3$  and  $\text{CsSnBr}_3$ , MDs are run at different temperatures between 100–500 K and 50–400 K, respectively. The resulting spectra for each phase are shown in Fig. 19, where spectra of  $\text{CsPbBr}_3$  are also compared with experiments from Ref. [131]. In the low temperature orthorhombic phase, Raman spectra show clear sharp peaks. For  $\text{CsPbBr}_3$  these peaks are in great agreement with experimental measurements, showing only a small redshift for frequencies below  $100 \text{ cm}^{-1}$ . Above phase transition temperatures, a wide central peak appears at  $\omega = 0$ , and most other peaks at low frequencies are not visible anymore. Only one peak around  $75 \text{ cm}^{-1}$  remains for  $\text{CsPbBr}_3$ . Both this peak and the shape of the central peak agree well with experiments. Finally, in the cubic phase (at 500 K), the central peak becomes dominant, and no other peaks are visible. For  $\text{CsSnBr}_3$ , the central peak is less intense and one broad peak between  $30\text{--}50 \text{ cm}^{-1}$  remains in the tetragonal and cubic phases. Surprisingly, the intensity of the central peak is found to be larger in the tetragonal phase for both materials.

To further study the behavior of the central peak, we look at the Raman spectra at the phase transition temperatures. Results are shown in Fig. 20(a) and (b) for  $\text{CsPbBr}_3$  and  $\text{CsSnBr}_3$ , respectively. From this figure, it becomes clear that the peak remaining in the tetragonal phase of  $\text{CsPbBr}_3$  comes from the peak observed around  $75 \text{ cm}^{-1}$  in the orthorhombic phase. Similarly, in  $\text{CsSnBr}_3$ , the peak that is still visible at 200 K arises from the peak around  $40 \text{ cm}^{-1}$ . For both materials, the lowest frequency peaks



**Fig. 19.** Raman spectra at different temperatures for (a)  $\text{CsPbBr}_3$  and (b)  $\text{CsSnBr}_3$ . Temperatures are selected so that all three phases are included. For  $\text{CsPbBr}_3$ , experimental results from Ref. [131] are represented with gray lines. (Adapted, with permission, from Publication I © 2024 American Physical Society).



**Fig. 20. Raman spectra at different temperatures during the orthorhombic to tetragonal phase transition for (a) CsPbBr<sub>3</sub> and (b) CsSnBr<sub>3</sub>. Insets show the intensity of the central peak ( $\omega = 0$ ) with temperature for the respective materials. (Reprinted, with permission, from Publication I © 2024 American Physical Society).**

shift to lower frequencies until forming the central peak after the phase transition. The insets of Fig. 20 show the intensity of the central peak with temperature. Naturally, at a low temperature, there is no central peak and the intensity is close to zero. The intensity increases with temperature and reaches a first maximum at 260 K and 130 K for CsPbBr<sub>3</sub> and CsSnBr<sub>3</sub>, respectively. These temperatures correspond to the phase transition temperatures previously obtained from heating curves [see Fig. 17(b) and (c)]. A second maximum of the central peak intensity is also observed at higher temperatures, more precisely at 360 K and 260 K, which again correspond to the phase transitions from tetragonal to cubic for CsPbBr<sub>3</sub> and CsSnBr<sub>3</sub>, respectively.

While the central peak is usually attributed to local polar fluctuations and anharmonicity [132, 133], our results show that it is also linked to both phase transitions in perovskites. An explanation for the maximal intensities of the central peak at a phase transition might come from the potential energy of octahedra tilts, which takes a double well shape [134, 94]. In the orthorhombic phase, octahedra oscillate close to the relaxed positions and remain at one of the potential energy minima, leading to one clear Raman peak. Above the transition temperature, octahedra start tilting, and hopping between the two wells start happening. This hopping between wells directly causes disorder. At temperatures slightly above the phase transition, the hopping rarely occurs, leading to a sharp and intense central peak. As the temperature increases, the hopping rate also increases [135] and the central peak becomes less intense and broader. Such a mechanism is applicable to both phase transitions and would explain the higher intensity of the central peak at these temperatures.

Overall, SA-GPR is found to correctly reproduce polarizabilities from DFT in all three phases of halide perovskites, while other methods (such as BPM) fail at high temperatures. The complex features of the Raman spectrum of these materials, in particular the central peak due to disorder, are all correctly reproduced and Raman spectra in the different phases agree well with experiments. Given the efficiency of this model, it becomes possible to obtain many Raman spectra at various temperatures and therefore precisely investigate Raman spectra during phase transition. The maximal intensities of the central peaks are observed and linked to the phase transitions.

## 4.5 Amino acids and peptides

Amino acids are organic molecules formed by an amino group NH<sub>3</sub><sup>+</sup>, a carboxylic acid group COO<sup>-</sup> and a side chain. While there could exist a wide variety of side chains, only 20 are commonly observed in biological macromolecules. Some of them (which are most frequently used in this thesis) are represented in Fig. 21. Glycine

(gly) is the simplest amino acid with only one hydrogen atom as its side chain. Others possess longer hydrocarbon side chains, such as alanine (ala) and leucine (leu). Larger amino acids might contain aromatic rings, with examples including phenylalanine (phe), tyrosine (tyr) or tryptophan (trp). Asparagine (asn) and glutamine (gln) have amide (O–C–N bonds) in their side chains. Finally, two amino acids, cysteine (cys) and methionine (met), contain sulfur atoms. This wide variety of atomic structures makes simulations of amino acids relatively challenging. Additionally, side chains are not static and can assume a large number of positions (usually called conformations) during MD simulations. Long MD trajectories are therefore required to correctly account for the different conformations.

Amino acids can also bond in chains to form macromolecules. Small chains of 2 to 50 amino acids are called peptides, which can in turn form larger biomolecules known as proteins. Figure 22 shows how a peptide bond between two amino acids is formed. The amine and carboxylic acid merge to form an amide O–C–N (referred to as a peptide bond) and liberate a water molecule. Such bonding can repeat to form chains. Similar to the side chains of amino acids, peptide bonds are flexible and can lead to various conformations. Additionally, the amides within peptide bonds lead to identifiable Raman peaks. The position and intensities of these peaks can give information about the conformation of peptides or the folding of proteins. Additionally, this wide variety

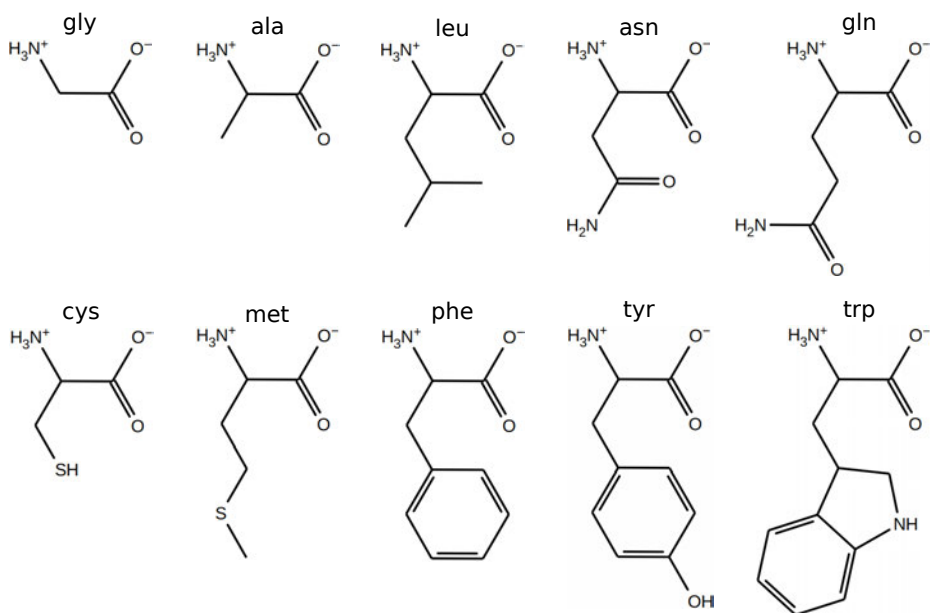
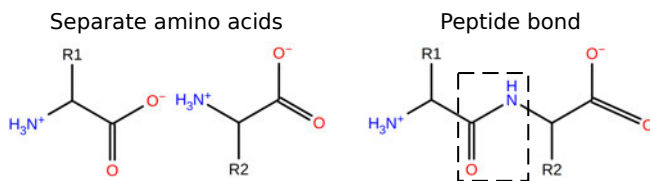


Fig. 21. Representation of a few peptides and their abbreviations.



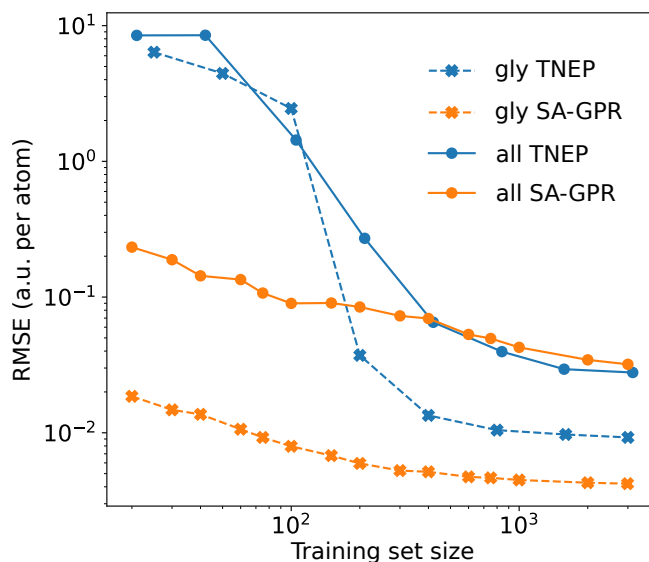
**Fig. 22. Representation of the peptide bond formation. (Adapted, with permission, from Publication IV © 2024 American Chemical Society).**

of side chains lead to very different Raman spectra, which could be used to identify amino acids and obtain the sequence of a given peptide or protein. In this context, surface-enhanced Raman spectroscopy (SERS) represents a promising candidate for amino acid sequencing [136, 137]. This shows that Raman spectra can be used in many ways to study the complex structures of such macromolecules.

While simulations of peptides and proteins are challenging, accurate classical force fields have been developed that enable the generation of long MD trajectories for large biomolecules. Some methods have also been developed to predict polarizabilities, such as the Thole model. However, these are usually trained using only optimized atomic positions and therefore do not account for vibrations, which are crucial to obtain accurate Raman spectra. For this purpose, another option could be to use BPM, which has already been applied to molecules, particularly long chains of carbons and hydrogens (called alkanes) [108]. However, in the case of amino acids and peptides, the training of BPM would prove very challenging due to the large number of bonds present. Similarly, it is impossible to use phonons as a basis set given the large number of conformations, making RGDOS inapplicable. This leaves only ML models, which are the most suitable for challenging systems (as was previously observed in the case of perovskites).

One limitation of ML models is transferability, that is, application of the model to structures outside of the training set. For periodic systems, transferability is less of an issue (obviously models trained on BAs are not applicable to MoS<sub>2</sub>). For molecules however, transferable models could be beneficial, with peptides being a perfect example of that. In principle, a model trained on all amino acids should be reasonably applicable to peptides and proteins. Careful testing is, however, necessary to make sure that predictions for larger peptides are accurate enough. This is what is done in Publication IV, where both SA-GPR and TNEP are trained for all amino acids and further tested on small peptides. The transferability of the resulting models is then compared and the best performing one is used to study the Raman spectra of peptides.

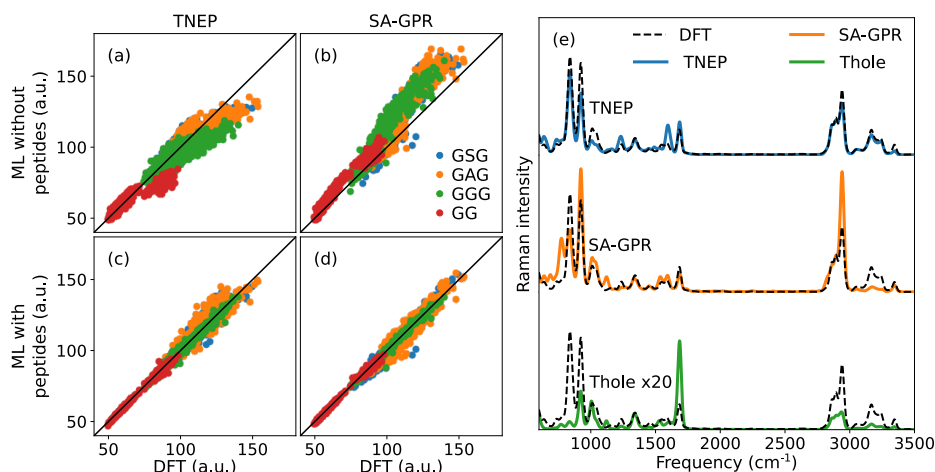
The models are first trained only for glycine, which is the simplest amino acid. Fig. 23 shows the evolution of the RMSE per atom with increasing training set size, and the results from TNEP and SA-GPR are compared. RMSE is calculated from a validation



**Fig. 23. Convergence of the RMSE of polarizabilities with the training set size. Results for models trained on all amino acids (dots with solid lines) and only on glycine (crosses with dashed lines) are both represented. (Adapted, with permission, from Publication IV © 2024 American Chemical Society).**

set kept constant even when increasing the size of the training set. For a single molecule, SA-GPR is found to be more accurate than TNEP, with RMSE values of  $4.21 \cdot 10^{-3}$  and  $9.24 \cdot 10^{-3}$  atomic units (a.u.) per atom, respectively, when using 3000 structures for training. For smaller training set sizes, the accuracy of SA-GPR is much better than TNEP. The models are then extended to all 20 amino acids and the results are also presented in Fig. 23. Similar to glycine models, SA-GPR demonstrates greater accuracy than TNEP for small training sets, even though the addition of different kinds of molecules clearly lowered the accuracy. For larger training sets containing more than 500 structures, SA-GPR and TNEP lead to similar accuracy, with RMSE values of  $3.20 \cdot 10^{-2}$  and  $2.78 \cdot 10^{-2}$  a.u. per atom, respectively.

The models are then tested on small peptides, namely glycine-glycine (GG), glycine-glycine-glycine (GGG), glycine-alanine-glycine (GAG) and glycine-serine-glycine (GSG). Fig. 24(a) and (b) compare polarizabilities from DFT with those from TNEP and SA-GPR models, respectively. The models used here have been trained exclusively on amino acids. Both models lead to inaccurate polarizabilities. While SA-GPR overestimates polarizabilities, TNEP shows large errors for all four peptides. Such differences might appear because the training set does not contain any structures with peptide bonds. New models containing di- and tri-glycine are therefore trained, and their



**Fig. 24. Comparison of polarizabilities between DFT and ML models for glycine-based peptides: (a) TNEP and (b) SA-GPR without peptides in the training set, and (c) TNEP and (d) SA-GPR including GG and GGG during training. (e) Raman spectra of GAG from DFT polarizabilities compared with TNEP, SA-GPR and the Thole model. (Adapted, with permission, from Publication IV © 2024 American Chemical Society).**

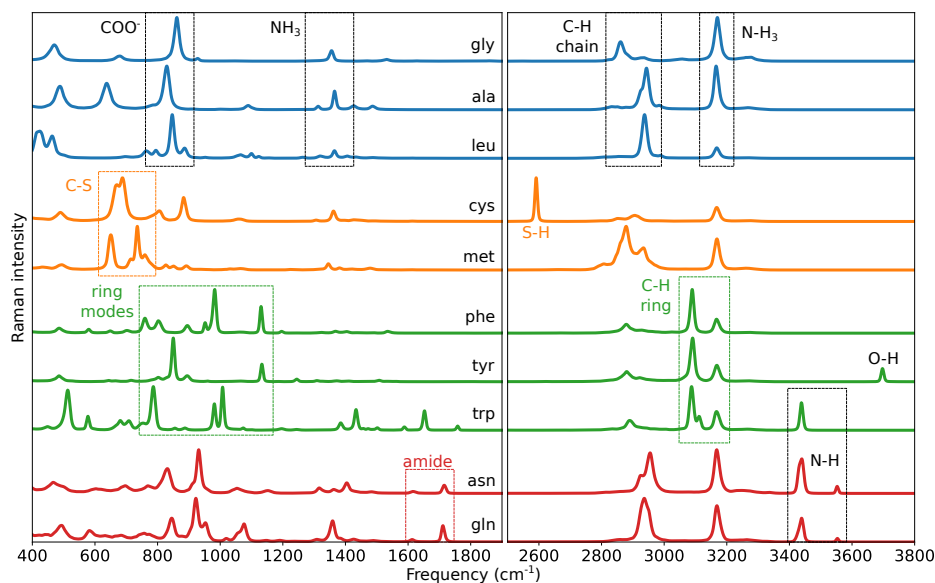
resulting predictions are presented in Fig. 24(c) and (d). As expected, both models now yield very accurate polarizabilities for GG and GGG. The RMSE values are  $5.79 \cdot 10^{-2}$  and  $6.96 \cdot 10^{-2}$  a.u. per atom for TNEP and  $5.34 \cdot 10^{-2}$  and  $4.79 \cdot 10^{-2}$  a.u. per atom for SA-GPR, respectively. These values are similar to those previously found when training amino acids. Both models also show satisfactory accuracy for other peptides, GAG and GSG, with RMSE values ranging from  $1.4$ – $1.6 \cdot 10^{-1}$  a.u. per atom. Note that for SA-GPR, we found that increasing the radial cutoff greatly improves the accuracy for peptides outside the training set. For TNEP, a cutoff of  $4 \text{ \AA}$  is used while a cutoff of  $6 \text{ \AA}$  has to be used for SA-GPR. Details regarding the choice of cutoff can be found in the supplementary information of Publication IV.

The transferability of TNEP and SA-GPR is further tested for the Raman spectra of GAG, which is compared with DFT and the Thole model in Fig. 24(e). Note that only the first 6.5 ps of the MD trajectory are used since DFT calculation of the polarizability is too expensive. Although this represents a short trajectory and definitely does not account for every conformation, it still allows for a comparison with spectra from DFT. From Fig. 24, we find that all models correctly reproduce most of the peak positions, although the intensities show large discrepancies for SA-GPR and the Thole model. On the other hand, spectra from TNEP show very good agreement with DFT. Although both TNEP and SA-GPR showed similar RMSE for the polarizability in Fig. 24(c) and (d),

we find that TNEP reproduces Raman spectra more accurately. This model is therefore used to investigate the Raman spectra of amino acids and peptides.

The Raman spectra of amino acids can be divided into two distinct regions. The region below  $1800\text{ cm}^{-1}$  contains peaks due to the movement of carbon, oxygen and nitrogen, including aromatic modes around  $1000\text{ cm}^{-1}$  and very-low frequency modes due to the change of conformations. This region contains most of the Raman peaks of amino acids, which can be used to identify amino acids in peptides [138], although attributing specific peaks can be quite challenging due to their large number. Hydrogen modes are found at higher frequencies between  $2600$  and  $3800\text{ cm}^{-1}$ . For simplicity, we focus here on the Raman spectra of the ten amino acids represented in Fig. 21. The results are presented in Fig. 25. The Raman spectrum of glycine shows two sharp peaks at  $860$  and  $1350\text{ cm}^{-1}$ , which agree well with experimental measurements [139]. These peaks are attributed to the  $\text{COO}^-$  and the amine group, respectively. For other amino acids with hydrocarbon side chains, such as alanine and leucine, the same two peaks are observed. Note that the  $\text{COO}^-$  peak shifts to  $830\text{ cm}^{-1}$  for alanine, in good agreement with experimental measurements [140, 141]. Longer side chains also lead to additional peaks, with, for example, peaks at  $1430$  and  $1490\text{ cm}^{-1}$  for alanine, which correspond to rocking of the  $-\text{CH}_3$  chain. The same  $\text{COO}^-$  and amine peaks are observed for cysteine. In this case, an additional peak due to the C–S bond appears at  $680\text{ cm}^{-1}$ , in good agreement with experiments [142, 143]. Methionine, the other amino acid with a sulfur side chain, possesses similar peaks at  $650$  and  $730\text{ cm}^{-1}$ , which can be attributed to the two different C–S bonds. For aromatic side chains, high intensity is observed for the breathing mode of benzene at  $990\text{ cm}^{-1}$ , which is correctly reproduced for phenylalanine. This peak shifts to  $850\text{ cm}^{-1}$  in tyrosine due to the additional OH added to the ring. For these two amino acids, another intense peak is found at  $1130\text{ cm}^{-1}$  and corresponds to another mode within the benzene ring (denoted 9a in Ref. [144]). For tryptophan, which contains both a pyrrole and a benzene ring, breathing modes are observed at  $790$  and around  $1000\text{ cm}^{-1}$  respectively, in good agreement with experiments [145]. Asparagine and glutamine, two amino acids with amide side chains, possess similar Raman spectra. In particular, they both show a unique peak at  $1710\text{ cm}^{-1}$ , which has been observed experimentally and assigned to the amide [146, 147].

The high-frequency region of the Raman spectra is also shown in Fig. 25. Cysteine is the only molecule showing a very clear peak around  $2600\text{ cm}^{-1}$ , which can be attributed to the S–H bond. Broad peaks between  $2800$ – $3100\text{ cm}^{-1}$  are due to the C–H vibrations. For amino acids with aromatic side chains, sharp peaks around  $3100\text{ cm}^{-1}$  can be attributed to C–H modes inside the carbon ring. Most amino acids show a clear peak around  $3200\text{ cm}^{-1}$  which correspond to the hydrogen motion inside the amino group



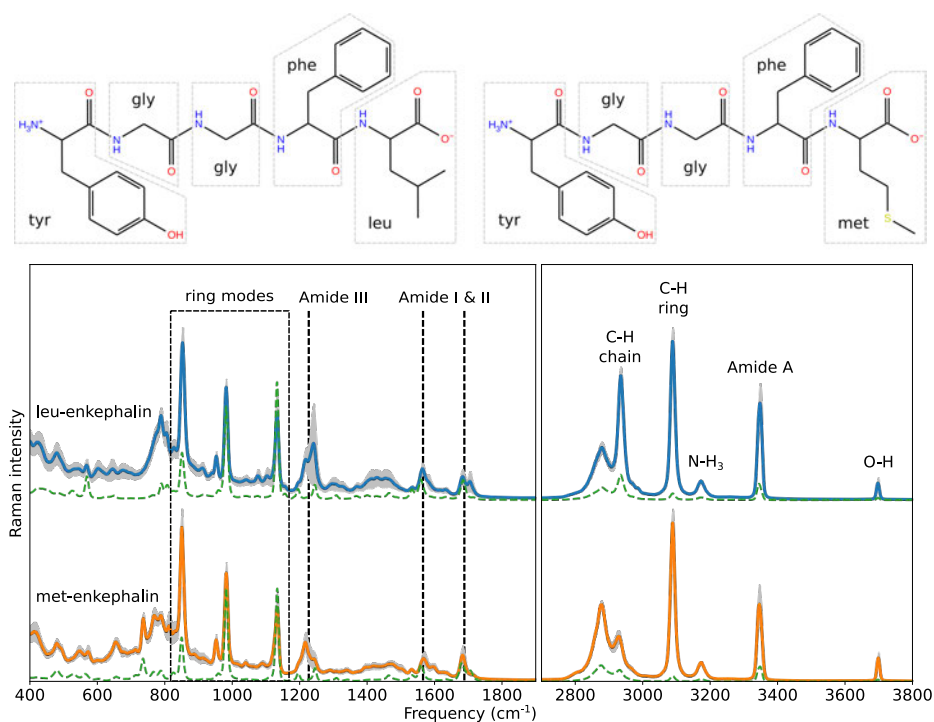
**Fig. 25.** Low- and high-frequency region of the Raman spectra for some amino acids in aqueous solution. Labels show peak assignment, where black labels correspond to general mode and colored labels refer to modes specific to the side chain: hydrocarbons in blue, sulfur in orange, aromatic in green, and amide in red. (Reprinted, with permission, from Publication IV © 2024 American Chemical Society).

(NH<sub>3</sub>). Proline is the only amino acid without such a group and does not show this peak (see Fig. S3 in Supporting Information of Publication IV). Instead, a peak at a higher frequency is observed and attributed to the N–H bond. Other peaks between 3400 and 3600 cm<sup>-1</sup> are found for tryptophan, asparagine and glutamine, which are also attributed to N–H bonds. Finally, O–H bonds, which are present in tyrosine, yield one peak around 3700 cm<sup>-1</sup>. This peak attribution is in great agreement with previous experimental results from [148].

Finally, we investigate Raman spectra of larger pentapeptides not included in the training step, namely met- and leu-enkephalins. They take the form of tyr-gly-gly-phe-met and tyr-gly-gly-phe-leu for met- and leu-enkephalin, respectively, as illustrated in Fig. 26, along with the resulting spectra. For larger peptides, polarizabilities at the DFT level are too expensive to be computed for many structures and it becomes crucial to obtain error estimates from the ML model. In this work, committee error estimates (CEE) are used to obtain such errors (see Refs. [149, 150] for details regarding CEE). 10 TNEP models are trained using the same training set (containing amino acids and peptides). Given the stochastic nature of the TNEP training, each model will have

different optimized parameters and therefore result in slightly different predictions. The quality of the model predictions can be assessed by comparing the model predictions, which is here quantified by evaluating the standard error of the predicted spectra. By looking at the frequency regions with larger errors, it becomes possible to get an idea of which vibrations are not sufficiently included in the training set. Additionally, Raman spectra obtained using the Thole model are also used as references.

The resulting spectra of leu- and met-enkephalins are presented in Fig. 26, along with CEE (in gray) and the results from the Thole model (in green). Large errors are observed in the very low frequencies region below  $400\text{ cm}^{-1}$  related to conformational changes. The low-frequency region is mainly dominated by peaks pertaining to the aromatic rings, with peaks at  $850\text{ cm}^{-1}$ ,  $980\text{ cm}^{-1}$  and  $1130\text{ cm}^{-1}$ , which respectively correspond to the breathing mode of tyrosine and phenylalanine and the 9a mode already observed for isolated amino acids. Errors for these three modes are relatively small. Additionally,



**Fig. 26. Top: representation of leu-enkephalin (left) and met-enkephalin (right). Bottom: Raman spectra of leu- and met-enkephalin. Committee error estimates are represented with gray area and spectra from the Thole model (multiplied by a factor 5000) in dashed green. (Reprinted, with permission, from Publication IV © 2024 American Chemical Society).**

new peaks pertaining to the peptide bonds appear at  $1230\text{ cm}^{-1}$ ,  $1560\text{ cm}^{-1}$  and  $1680\text{ cm}^{-1}$ . These peaks are quite similar to those previously obtained for amino acids with amide side chains (see asn and gln in Fig. 25) and are respectively labeled amide III, II and I [151, 152]. Amide peaks exhibit larger errors than modes coming from amino acids, hinting at the fact that the model is less trained for peptide bonds.

In the high-frequency region, the observed peaks are in great agreement with what was previously obtained for amino acids, with one additional peak at  $3350\text{ cm}^{-1}$  corresponding to the N–H vibration within the amide (labeled amide A) [151, 152]. In this region, peaks from aromatic rings and amide are the most intense. Also note that modes from C–H vibrations show wide peaks due to the many different environments existing in such large peptides. Peaks from  $\text{NH}_3$  and OH are also observed with smaller intensities. Also, note that the CEE are very small in the high-frequency region, indicating excellent training to capture hydrogen motions.

Raman spectra from the Thole model are also shown in Fig. 26 to assess its accuracy. Most of the peaks are correctly picked out, including both ring breathing modes, the 9a mode at  $1130\text{ cm}^{-1}$  and the amide I and II modes. In the high frequency region, all modes compare well, even though the Thole model clearly underestimates all intensities.

Based on the overall low committee error estimates, the Raman spectra of enkephalins should be fairly accurate even though these peptides were not included in the training. Committee error estimates can be useful for assessing which peaks are likely to be less accurate and can guide possible further training. In the case of enkephalins, amide/peptide bonds and conformational changes could benefit from further training. The TNEP model presented here represents a great starting point to obtain the Raman spectra of large peptides, or even proteins, as it can already accurately predict polarizabilities for all 20 amino acids. Extending the model to larger macromolecules would only require the training of a few conformations, which could be guided using CEE. Similarly, it could be extended to study amino acids in other conditions, with the investigation of amino acids on surfaces and SERS being particularly promising directions [136, 137].



## 5 Discussion and conclusions

Machine learning force fields are now routinely used to produce MD trajectories, enabling system sizes and simulation time scales that were previously impossible. It is sometimes necessary to obtain additional properties, such as polarizabilities in the case of Raman spectroscopy. It then becomes crucial to accurately predict polarizabilities with efficiencies similar to MLFFs. Various examples of such polarizability models and their applications to diverse materials were presented throughout this thesis. From these results, a detailed comparison of these methods is possible. While all of them show potential for specific applications, each of them clearly possess different advantages that should be taken into account when selecting the model that is most suitable for a given system. In the remainder of this section, the five polarizability models used in this work are discussed and compared in order to highlight their benefits and limitations.

### *Thole model*

The Thole model is a simple empirical model based on the dipole-dipole interaction. Since it relies only on atomic positions and bonds, this is a really fast method. This model is, however, restricted to the few atomic species present in organic molecules, which greatly restricts its field of application. Additionally, the training set contains only relaxed structures and vibrations are not really accounted for, making this model less suitable to Raman spectroscopy. Despite these concerns, the application of the Thole model in Publication IV led to reasonable Raman spectra of amino acids, peptides and enkephalins, demonstrating that the predicted polarizabilities are indeed reliable. One of the current limitations of the Thole model is that it greatly underestimates Raman intensities. This issue likely stems from the use of a training set that does not include displaced structures, as well as from the shape of the dipole field tensor. By extending the training set, it could in principle be possible to improve the accuracy of Raman spectra. The model could also be extended to other atoms and possibly applied to periodic systems as well. However, it remains uncertain whether such a dipole-interaction model correctly captures the effects of atomic vibrations. Although promising, such extension of the model would require careful testing.

### *Bond polarizability model*

BPM is another relatively simple model, where each bond contributes to the total polarizability. Similar to the Thole model, BPM relies only on bonds between atoms, making it highly efficient. It has the advantage of correctly accounting for the symmetry of the system. BPM also requires less training data to reach a converged model than its ML counterparts, but also comes with the price of being less accurate. Nonetheless BPM yields accurate Raman spectra for simple systems as shown in Publication I. It also has the benefit of being able to account for the second-order effects (unlike RGDOS). However, BPM fails when applied to complex systems, with the anharmonicity in perovskites being a good example. Overall, the simplicity of this model makes it most suitable for simpler materials. It is also important to note that recent developments in BPM allow it to accurately determine polarizabilities in anharmonic materials [32].

### *RGDOS*

RGDOS is another highly efficient method that relies on projecting phonon modes. Since it relies only on atomic positions, its speed is similar to that of BPM and faster than ML models, which rely on descriptors. Additionally, RGDOS has the advantage of requiring a set number of “training” data, usually lower than what is necessary for other models. RGDOS also makes it possible to separate the Raman spectra by selecting only certain modes. This can be useful to track the shift of specific Raman modes when changing the temperature or the chemistry of the system, such as the surface of MXenes in Publication III. Finally, RGDOS benefits from good transferability to systems with similar vibrational modes, typically alloys or point defects, as shown in Publications II and III. On the other hand, RGDOS cannot be applied to systems containing various phases and very different phonon modes, such as perovskites. It also struggles to correctly reproduce the second-order contributions, which makes it necessary to include many reciprocal points in the expansion. RGDOS is therefore perfectly suited for harmonic materials where only first-order spectra is of interest, as well as systems containing defects and alloys.

### *Symmetry-adapted GPR*

SA-GPR is an ML-based model that relies on obtaining the similarity between structures through kernel functions. While this method is much more efficient than DFT, it remains slower than BPM or RGDOS, mainly due to the calculation of descriptors. It also

requires larger training sets than BPM, as observed in Publication I. Even though SA-GPR is more demanding than BPM and RGDOS, it leads to more accurate predictions. GPR also benefits from a straightforward training step, in which a simple matrix equation has to be solved. Additionally, this model is applicable to the more challenging systems where other simpler methods would fail. Examples include anharmonic materials such as perovskites in Publication I or different molecules such as amino acids in Publication IV. Overall, SA-GPR is most suitable for complex materials where RGDOS and BPM would fail.

### *TNEP*

Polarizabilities can also be obtained using a neural network, as shown by the TNEP method. The prediction efficiency of TNEP is expected to be similar to that of SA-GPR since both models need to compute local descriptors.<sup>2</sup> TNEP is also applicable to the most challenging systems and correctly accounts for any effects. Overall, it has the same benefits and shortcomings as SA-GPR when compared to BPM or RGDOS. The main differences between TNEP and SA-GPR are highlighted in Publication IV. While reaching similar accuracies for large training sets, TNEP is found to be less accurate for smaller training sets. In other words, while SA-GPR might be applicable with a limited training set, TNEP requires more training data in general. Also note that the training step of a neural network is very demanding, as there is no straightforward method to obtain the optimized parameters. These additional costs come at the benefit of better transferability, that is, TNEP leads to more accurate Raman spectra for systems not included in the training set. This was shown in Publication IV for amino acids and peptides, but should also hold true for other groups of molecules or crystals.

### *Conclusions*

Besides the choice of polarizability model, the quality and size of the training set also play an important role on the resulting Raman spectra. The accuracy of models is limited by the accuracy of the data in the training set. In the case of the resonant spectra of MoS<sub>2</sub>, polarizabilities from methods more accurate than DFT could improve the agreement between simulation and experiments. As for the size of the training set, we observed that complex systems, such as perovskites and amino acids, require large

---

<sup>2</sup>The comparison of efficiency here is slightly unfair since TNEP predictions are performed using an implementation on GPU while other methods are simple Python implementations. For similar implementations, TNEP and SA-GPR should have similar efficiencies.

amount of training data. On the other hand, smaller training sets can be used for simpler materials such as BAs and MoS<sub>2</sub>.

The application of the models to various materials showed good overall agreement with experimental measurements, further proving the quality of the predictions. In the case of BAs, the effect of isotopes on both first- and second-order Raman spectra was correctly reproduced in Publication I. While obtaining spectra for alloys is usually very demanding, the integration of ML MD and the polarizability model made it more efficient and easier. Additionally, this combination enables the simulation of large supercells, which is necessary to investigate low concentration alloys. Similarly large simulation cells are required when studying point defects. Publication II is a perfect example of this, where the changes in the spectra were compared upon the introduction of three kinds of defects. The resulting shifts in the Raman spectra were in quantitative agreement with experimental measurements, making it possible to reliably identify the presence of defects in the MoS<sub>2</sub> monolayer from Raman spectra. Good agreement with experimental results is also found for the resonant Raman of the MoS<sub>2</sub> monolayer (Fig. 12), MXenes with inhomogeneous surfaces (Fig. 16) and all three phases of cesium halide perovskites (Fig. 19).

In addition to providing strong comparisons with experimental observations, some of the applications also give further insights into the chemical composition of materials. Simulations of Raman spectra for mixed-surface MXenes allow for a better understanding of the impact of the terminations on spectra. That kind of information would be very challenging to obtain from experiments. It is important to keep in mind that the simulation of heterogeneous surfaces, similar to alloys, requires large supercells. Using MLFF MD makes it possible to efficiently obtain trajectories and repeat the process for many distributions and concentrations of surface terminations, while RGDOS allows for the easy computation of polarizabilities and first-order Raman spectra. Results from our calculations and comparison to experiments suggest a O/OH ratio between 0.5 and 0.625 and other concentrations can also be compared with future experimental spectra to get an idea of the surface compositions.

Similarly, simulations of perovskites are usually very demanding since long trajectories are required to obtain spectra at low frequencies. The complex anharmonic phases of these materials make the simulations even more complicated. Here again MLFF allows for the easy acquisition of forces and trajectories, but the treatment of polarizabilities still requires attention. Simple models such as RGDOS or BPM fail to reproduce the complex behavior of these materials at high temperatures and only ML models lead to accurate polarizabilities. Although we only tested SA-GPR, we expect TNEP to also perform well. In the end, the combination of NEP with SA-GPR

made it possible to obtain Raman spectra that are in great agreement with experimental observations. Thanks to the efficiency of this scheme, it becomes possible to obtain many Raman spectra at various temperatures and therefore precisely investigate Raman spectra during phase transitions, with maximal intensities of the central peaks observed at the phase transition temperatures. Our simulations of halide perovskites therefore enable the investigation of their phase transitions in detail and deepen our understanding of their vibrational properties.

Another challenge of ML methods is their transferability to systems outside of the training set. The transferability of SA-GPR and TNEP was tested on amino acids and peptide chains. We found that TNEP has better transferability, which also comes at the cost of requiring a larger amount of training data. Raman spectra obtained by combining classical force fields MD with TNEP polarizabilities reproduce many features observed experimentally in amino acids well. Additionally, the scheme extends well to enkephalins, showing good agreement with experimental spectra as well as with the Thole model.

In conclusion, all the models presented in this work are capable of correctly predicting polarizabilities for a fraction of the cost of first-principles calculations. This in turn allows for the efficient computation of Raman spectra from long MD trajectories containing thousands of atoms. While each model comes with its own strengths and weaknesses, all of them can find applications depending on the system being investigated. Simpler models do not require large training sets but in turn have restricted application, making them particularly suitable for simpler materials. More complicated systems require more complicated models such as SA-GPR or TNEP. These models come with the downside of needing much larger training sets and less efficient predictions due to the calculations of descriptors. In the end, a compromise between accuracy/applicability and computational cost has to be found, reminding us of the same dilemma when dealing with *ab initio* methods.



## References

- [1] M. Kokkonen, P. Talebi, J. Zhou, S. Asgari, S. A. Soomro, F. Elsehrawy, J. Halme, S. Ahmad, A. Hagfeldt, and S. G. Hashmi, "Advanced research trends in dye-sensitized solar cells," *J. Mater. Chem. A*, vol. 9, pp. 10 527–10 545, 2021.
- [2] E. Berger, M. Bagheri, S. Asgari, J. Zhou, M. Kokkonen, P. Talebi, J. Luo, A. F. Nogueira, T. Watson, and S. G. Hashmi, "Recent developments in perovskite-based precursor inks for scalable architectures of perovskite solar cell technology," *Sustainable Energy Fuels*, vol. 6, pp. 2879–2900, 2022.
- [3] N. Yabuuchi, K. Kubota, M. Dahbi, and S. Komaba, "Research development on sodium-ion batteries," *Chemical Reviews*, vol. 114, no. 23, pp. 11 636–11 682, 2014.
- [4] T. Hosaka, K. Kubota, A. S. Hameed, and S. Komaba, "Research development on k-ion batteries," *Chemical Reviews*, vol. 120, no. 14, pp. 6358–6466, 2020.
- [5] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, and A. Kis, "Single-layer MoS<sub>2</sub> transistors," *Nature Nanotechnology*, vol. 6, no. 3, pp. 147–150, Mar. 2011.
- [6] K. Liu, B. Jin, W. Han, X. Chen, P. Gong, L. Huang, Y. Zhao, L. Li, S. Yang, X. Hu, J. Duan, L. Liu, F. Wang, F. Zhuge, and T. Zhai, "A wafer-scale van der Waals dielectric made from an inorganic molecular crystal film," *Nature Electronics*, vol. 4, no. 12, pp. 906–913, Dec. 2021.
- [7] C. V. Raman and K. S. Krishnan, "A new type of secondary radiation," *Nature*, vol. 121, no. 3048, pp. 501–502, Mar. 1928.
- [8] R. S. Das and Y. Agrawal, "Raman spectroscopy: Recent advancements, techniques and applications," *Vib. Spectrosc.*, vol. 57, no. 2, pp. 163–176, 2011.
- [9] J. R. Lombardi and R. L. Birke, "A unified approach to surface-enhanced raman spectroscopy," *J. Phys. Chem. C*, vol. 112, no. 14, pp. 5605–5617, 2008.
- [10] P. L. Stiles, J. A. Dieringer, N. C. Shah, and R. P. Van Duyne, "Surface-enhanced raman spectroscopy," *Annu. Rev. Anal. Chem.*, vol. 1, pp. 601–626, 2008.
- [11] D. Wei, S. Chen, and Q. Liu, "Review of fluorescence suppression techniques in Raman spectroscopy," *Appl. Spectrosc. Rev.*, vol. 50, no. 5, pp. 387–406, 2015.
- [12] T. Hasegawa, J. Nishijo, and J. Umemura, "Separation of raman spectra from fluorescence emission background by principal component analysis," *Chemical Physics Letters*, vol. 317, no. 6, pp. 642–646, 2000. [Online]. Available: [https://doi.org/10.1016/S0009-2614\(99\)01427-X](https://doi.org/10.1016/S0009-2614(99)01427-X)
- [13] P. J. Cadusch, M. M. Hlaing, S. A. Wade, S. L. McArthur, and P. R. Stoddart, "Improved methods for fluorescence background subtraction from raman spectra," *Journal of Raman Spectroscopy*, vol. 44, no. 11, pp. 1587–1595, 2013.

- [14] J. Kostamovaara, J. Tenhunen, M. Kögler, I. Nissinen, J. Nissinen, and P. Keränen, “Fluorescence suppression in raman spectroscopy using a time-gated cmos spad,” *Opt. Express*, vol. 21, no. 25, pp. 31 632–31 645, Dec. 2013. [Online]. Available: <https://doi.org/10.1364/OE.21.031632>
- [15] M. Kögler and B. Heilala, “Time-gated Raman spectroscopy – a review,” *Measurement Science and Technology*, vol. 32, no. 1, p. 012002, Oct. 2020. [Online]. Available: <https://dx.doi.org/10.1088/1361-6501/abb044>
- [16] D. R. Hartree, “The wave mechanics of an atom with a non-coulomb central field. Part II. Some results and discussion,” *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, no. 1, p. 111–132, 1928.
- [17] V. Fock, “Näherungsmethode zur lösung des quantenmechanischen mehrkörperproblems,” *Zeitschrift für Physik*, vol. 61, no. 1, pp. 126–148, Jan. 1930.
- [18] J. C. Slater, “A simplification of the hartree-fock method,” *Phys. Rev.*, vol. 81, pp. 385–390, Feb. 1951.
- [19] H. J. Monkhorst, “Calculation of properties with the coupled-cluster method,” *International Journal of Quantum Chemistry*, vol. 12, no. S11, pp. 421–432, 1977.
- [20] C. Møller and M. S. Plesset, “Note on an approximation treatment for many-electron systems,” *Phys. Rev.*, vol. 46, pp. 618–622, Oct. 1934.
- [21] P. Hohenberg and W. Kohn, “Inhomogeneous electron gas,” *Phys. Rev.*, vol. 136, pp. B864–B871, Nov. 1964.
- [22] W. Kohn and L. J. Sham, “Self-consistent equations including exchange and correlation effects,” *Phys. Rev.*, vol. 140, pp. A1133–A1138, Nov. 1965.
- [23] D. M. Ceperley and B. J. Alder, “Ground state of the electron gas by a stochastic method,” *Phys. Rev. Lett.*, vol. 45, pp. 566–569, Aug. 1980. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.45.566>
- [24] J. P. Perdew and Y. Wang, “Accurate and simple analytic representation of the electron-gas correlation energy,” *Phys. Rev. B*, vol. 45, pp. 13 244–13 249, Jun. 1992.
- [25] A. D. Becke, “A new mixing of Hartree–Fock and local density-functional theories,” *The Journal of Chemical Physics*, vol. 98, no. 2, pp. 1372–1377, Jan. 1993.
- [26] J. P. Perdew, K. Burke, and M. Ernzerhof, “Generalized gradient approximation made simple,” *Phys. Rev. Lett.*, vol. 77, pp. 3865–3868, Oct. 1996. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.77.3865>
- [27] H. J. Kulik, T. Hammerschmidt, J. Schmidt, S. Botti, M. A. L. Marques, M. Boley, M. Scheffler, M. Todorović, P. Rinke, C. Oses, A. Smolyanyuk, S. Curtarolo, A. Tkatchenko, A. P. Bartók, S. Manzhos, M. Ihara, T. Carrington, J. Behler, O. Isayev, M. Veit, A. Grisafi, J. Nigam, M. Ceriotti, K. T. Schütt, J. Westermayr, M. Gastegger, R. J. Maurer, B. Kalita, K. Burke, R. Nagai, R. Akashi, O. Sugino, J. Hermann, F. Noé, S. Pilati, C. Draxl, M. Kuban, S. Rigamonti, M. Scheidgen, M. Esters, D. Hicks, C. Toher, P. V. Balachandran, I. Tamblin, S. Whitelam, C. Bellinger, and L. M. Ghiringhelli, “Roadmap on machine learning in electronic structure,” *Electronic Structure*, vol. 4, no. 2, p. 023004, Aug. 2022.

- [28] D. Lu, H. Wang, M. Chen, L. Lin, R. Car, W. E. W. Jia, and L. Zhang, “86 pflops deep potential molecular dynamics simulation of 100 million atoms with ab initio accuracy,” *Computer Physics Communications*, vol. 259, p. 107624, 2021. [Online]. Available: <https://doi.org/10.1016/j.cpc.2020.107624>
- [29] B. Thole, “Molecular polarizabilities calculated with a modified dipole interaction,” *Chemical Physics*, vol. 59, no. 3, pp. 341–350, 1981.
- [30] P. T. van Duijnen and M. Swart, “Molecular and atomic polarizabilities: Thole’s model revisited,” *The Journal of Physical Chemistry A*, vol. 102, no. 14, pp. 2399–2407, 1998.
- [31] P. Umari, A. Pasquarello, and A. Dal Corso, “Raman scattering intensities in  $\alpha$ -quartz: A first-principles investigation,” *Phys. Rev. B*, vol. 63, p. 094305, Feb. 2001. [Online]. Available: <https://doi.org/10.1103/PhysRevB.63.094305>
- [32] A. Paul, A. Ruffino, S. Masiuk, J. Spanier, and I. Grinberg, “An atomistic model of electronic polarizability for calculation of Raman scattering from large-scale md simulations,” Apr. 2023, arXiv preprint arXiv:2304.07536. [Online]. Available: <https://doi.org/10.48550/arXiv.2304.07536>
- [33] A. Hashemi, A. V. Krasheninnikov, M. Puska, and H.-P. Komsa, “Efficient method for calculating Raman spectra of solids with impurities and alloys and its application to two-dimensional transition metal dichalcogenides,” *Phys. Rev. Materials*, vol. 3, p. 023806, Feb. 2019. [Online]. Available: <https://doi.org/10.1103/PhysRevMaterials.3.023806>
- [34] E. Guerrero and D. A. Strubbe, “Structure, thermodynamics, and Raman spectroscopy of Rhenium-doped bulk MoS<sub>2</sub> from first principles,” *The Journal of Physical Chemistry C*, vol. 126, no. 43, pp. 18 393–18 403, 2022.
- [35] M. Grumet, C. von Scarpatetti, T. Bučko, and D. A. Egger, “Delta machine learning for predicting dielectric properties and raman spectra,” *The Journal of Physical Chemistry C*, vol. 128, no. 15, pp. 6464–6470, 2024.
- [36] 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. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.120.036002>
- [37] N. Xu, P. Rosander, C. Schäfer, E. Lindgren, N. Österbacka, M. Fang, W. Chen, Y. He, Z. Fan, and P. Erhart, “Tensorial properties via the neuroevolution potential framework: Fast simulation of infrared and raman spectra,” *Journal of Chemical Theory and Computation*, vol. 20, no. 8, pp. 3273–3284, 2024.
- [38] N. Raimbault, A. Grisafi, M. Ceriotti, and M. Rossi, “Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals,” *New Journal of Physics*, vol. 21, no. 10, p. 105001, Oct. 2019. [Online]. Available: <https://doi.org/10.1088/1367-2630/ab4509>
- [39] G. M. Sommers, M. F. Calegari Andrade, L. Zhang, H. Wang, and R. Car, “Raman spectrum and polarizability of liquid water from deep neural networks,” *Phys. Chem. Chem. Phys.*, vol. 22, pp. 10 592–10 602, 2020.

- [40] X. Gonze and C. Lee, “Dynamical matrices, born effective charges, dielectric permittivity tensors, and interatomic force constants from density-functional perturbation theory,” *Phys. Rev. B*, vol. 55, pp. 10 355–10 368, Apr. 1997.
- [41] M. Bagheri and H.-P. Komsa, “High-throughput computation of raman spectra from first principles,” *Sci. Data*, vol. 10, no. 1, p. 80, Feb. 2023.
- [42] P. Umari and A. Pasquarello, “Infrared and raman spectra of disordered materials from first principles,” *Diamond and Related Materials*, vol. 14, no. 8, pp. 1255–1261, 2005, sMAC ’04 Conference Proceeding S.I.
- [43] J. De Gelder, K. De Gussem, P. Vandenabeele, and L. Moens, “Reference database of raman spectra of biological molecules,” *Journal of Raman Spectroscopy*, vol. 38, no. 9, pp. 1133–1147, 2007.
- [44] M. Cardona, *Resonance phenomena*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1982, pp. 19–178. [Online]. Available: <https://doi.org/10.1007/3-540-11380-0>
- [45] H. M. J. Smith and M. Born, “The theory of the vibrations and the Raman spectrum of the diamond lattice,” *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 241, no. 829, pp. 105–145, 1948.
- [46] M. S. Green, “Markoff random processes and the statistical mechanics of time-dependent phenomena. II. Irreversible processes in fluids,” *The Journal of Chemical Physics*, vol. 22, no. 3, pp. 398–413, Mar. 1954.
- [47] R. Kubo, “Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems,” *Journal of the Physical Society of Japan*, vol. 12, no. 6, pp. 570–586, 1957.
- [48] D. A. McQuarrie, *Statistical Mechanics*. New York: Harper Collins, 1976.
- [49] 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. [Online]. Available: <http://dx.doi.org/10.1039/C3CP44302G>
- [50] J. Lahnsteiner and M. Bokdam, “Anharmonic lattice dynamics in large thermodynamic ensembles with machine-learning force fields: CsPbbr<sub>3</sub>, a phonon liquid with Cs rattlers,” *Phys. Rev. B*, vol. 105, p. 024302, Jan. 2022.
- [51] V. Dubois, P. Umari, and A. Pasquarello, “Dielectric susceptibility of dipolar molecular liquids by ab initio molecular dynamics: application to liquid hcl,” *Chemical Physics Letters*, vol. 390, no. 1, pp. 193–198, 2004.
- [52] E. Berger, J. Wiktor, and A. Pasquarello, “Low-frequency dielectric response of tetragonal perovskite ch<sub>3</sub>nh<sub>3</sub>pb<sub>3</sub>i<sub>3</sub>,” *The Journal of Physical Chemistry Letters*, vol. 11, no. 15, pp. 6279–6285, 2020.
- [53] A. Putrino and M. Parrinello, “Anharmonic raman spectra in high-pressure ice from ab initio simulations,” *Phys. Rev. Lett.*, vol. 88, p. 176401, Apr. 2002. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.88.176401>

- [54] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, “Phonons and related crystal properties from density-functional perturbation theory,” *Rev. Mod. Phys.*, vol. 73, pp. 515–562, Jul. 2001. [Online]. Available: <https://doi.org/10.1103/RevModPhys.73.515>
- [55] M. Gajdoš, K. Hummer, G. Kresse, J. Furthmüller, and F. Bechstedt, “Linear optical properties in the projector-augmented wave methodology,” *Phys. Rev. B*, vol. 73, p. 045112, Jan. 2006. [Online]. Available: <https://doi.org/10.1103/PhysRevB.73.045112>
- [56] S. Lubber, M. Iannuzzi, and J. Hutter, “Raman spectra from ab initio molecular dynamics and its application to liquid s-methyloxirane,” *The Journal of Chemical Physics*, vol. 141, no. 9, p. 094503, 2014. [Online]. Available: <https://doi.org/10.1063/1.4894425>
- [57] H. C. Andersen, “Molecular dynamics simulations at constant pressure and/or temperature,” *The Journal of Chemical Physics*, vol. 72, no. 4, pp. 2384–2393, Feb. 1980.
- [58] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, “Molecular dynamics with coupling to an external bath,” *The Journal of Chemical Physics*, vol. 81, no. 8, pp. 3684–3690, Oct. 1984.
- [59] G. Bussi and M. Parrinello, “Canonical resampling through velocity rescaling,” *J. Chem. Phys.*, vol. 126, no. 1, p. 014101, 2007.
- [60] S. Nosé, “A unified formulation of the constant temperature molecular dynamics methods,” *The Journal of Chemical Physics*, vol. 81, no. 1, pp. 511–519, 1984.
- [61] W. G. Hoover, “Canonical dynamics: Equilibrium phase-space distributions,” *Phys. Rev. A*, vol. 31, pp. 1695–1697, Mar. 1985.
- [62] N. Shuichi, “Constant Temperature Molecular Dynamics Methods,” *Progress of Theoretical Physics Supplement*, vol. 103, pp. 1–46, Jan. 1991.
- [63] G. J. Martyna, M. L. Klein, and M. Tuckerman, “Nosé–Hoover chains: The canonical ensemble via continuous dynamics,” *The Journal of Chemical Physics*, vol. 97, no. 4, pp. 2635–2643, Aug. 1992.
- [64] M. Parrinello and A. Rahman, “Polymorphic transitions in single crystals: A new molecular dynamics method,” *J. Appl. Phys.*, vol. 52, pp. 7182–7190, 1981.
- [65] G. J. Martyna, D. J. Tobias, and M. L. Klein, “Constant pressure molecular dynamics algorithms,” *The Journal of Chemical Physics*, vol. 101, no. 5, pp. 4177–4189, Sep. 1994.
- [66] M. Bernetti and G. Bussi, “Pressure control using stochastic cell rescaling,” *The Journal of Chemical Physics*, vol. 153, no. 11, p. 114107, Sep. 2020.
- [67] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, “Accelerating materials property predictions using machine learning,” *Scientific Reports*, vol. 3, no. 1, p. 2810, Sep. 2013.
- [68] M. Gastegger, J. Behler, and P. Marquetand, “Machine learning molecular dynamics for the simulation of infrared spectra,” *Chem. Sci.*, vol. 8, pp. 6924–6935, 2017.
- [69] M. Veit, D. M. Wilkins, Y. Yang, J. DiStasio, Robert A., and M. Ceriotti, “Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles,” *The Journal of Chemical Physics*, vol. 153, no. 2, p. 024113, Jul. 2020.

- [70] A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, and C. Corminboeuf, "Electron density learning of non-covalent systems," *Chem. Sci.*, vol. 10, pp. 9424–9432, 2019. [Online]. Available: <http://doi.org/10.1039/C9SC02696G>
- [71] A. M. Lewis, A. Grisafi, M. Ceriotti, and M. Rossi, "Learning electron densities in the condensed phase," *Journal of Chemical Theory and Computation*, vol. 17, no. 11, pp. 7203–7214, 2021.
- [72] Y. Zhuo, A. Mansouri Tehrani, and J. Brgoch, "Predicting the band gaps of inorganic solids by machine learning," *The Journal of Physical Chemistry Letters*, vol. 9, no. 7, pp. 1668–1673, 2018.
- [73] Y. Huang, C. Yu, W. Chen, Y. Liu, C. Li, C. Niu, F. Wang, and Y. Jia, "Band gap and band alignment prediction of nitride-based semiconductors using machine learning," *J. Mater. Chem. C*, vol. 7, pp. 3238–3245, 2019.
- [74] J. R. Moreno, J. Flick, and A. Georges, "Machine learning band gaps from the electron density," *Phys. Rev. Mater.*, vol. 5, p. 083802, Aug. 2021.
- [75] R. Jinnouchi, J. Lahnsteiner, F. Karsai, G. Kresse, and M. Bokdam, "Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with bayesian inference," *Phys. Rev. Lett.*, vol. 122, p. 225701, Jun. 2019. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.122.225701>
- [76] 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. [Online]. Available: <https://doi.org/10.1103/PhysRevB.100.014105>
- [77] Z. Fan, Z. Zeng, C. Zhang, Y. Wang, K. Song, H. Dong, Y. Chen, and T. Ala-Nissila, "Neuroevolution machine learning potentials: Combining high accuracy and low cost in atomistic simulations and application to heat transport," *Phys. Rev. B*, vol. 104, p. 104309, Sep. 2021. [Online]. Available: <https://doi.org/10.1103/PhysRevB.104.104309>
- [78] 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.
- [79] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," *Multiscale Modeling & Simulation*, vol. 14, no. 3, pp. 1153–1173, 2016.
- [80] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, and S. P. Ong, "Performance and cost assessment of machine learning interatomic potentials," *The Journal of Physical Chemistry A*, vol. 124, no. 4, pp. 731–745, 2020.
- [81] 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. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.104.136403>
- [82] 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. 10 073–10 141, 2021, PMID: 34398616. [Online]. Available: <https://doi.org/10.1021/acs.chemrev.1c00022>

- [83] A. M. Lewis, P. Lazzaroni, and M. Rossi, “Predicting the electronic density response of condensed-phase systems to electric field perturbations,” *The Journal of Chemical Physics*, vol. 159, no. 1, p. 014103, Jul. 2023.
- [84] A. Grisafi, A. M. Lewis, M. Rossi, and M. Ceriotti, “Electronic-structure properties from atom-centered predictions of the electron density,” *Journal of Chemical Theory and Computation*, vol. 19, no. 14, pp. 4451–4460, 2023.
- [85] D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio, and M. Ceriotti, “Accurate molecular polarizabilities with coupled cluster theory and machine learning,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 9, pp. 3401–3406, 2019.
- [86] M. G. Zauchner, S. D. Forno, G. Csányi, A. Horsfield, and J. Lischner, “Predicting polarizabilities of silicon clusters using local chemical environments,” *Mach. learn.: sci. technol.*, vol. 2, no. 4, p. 045029, Oct. 2021. [Online]. Available: <https://dx.doi.org/10.1088/2632-2153/ac2cfe>
- [87] J. Behler, “Atom-centered symmetry functions for constructing high-dimensional neural network potentials,” *The Journal of Chemical Physics*, vol. 134, no. 7, p. 074106, Feb. 2011.
- [88] R. Drautz, “Atomic cluster expansion for accurate and transferable interatomic potentials,” *Phys. Rev. B*, vol. 99, p. 014104, Jan. 2019. [Online]. Available: <https://doi.org/10.1103/PhysRevB.99.014104>
- [89] H. Huo and M. Rupp, “Unified representation of molecules and crystals for machine learning,” *Machine Learning: Science and Technology*, vol. 3, no. 4, p. 045017, Nov. 2022. [Online]. Available: <https://dx.doi.org/10.1088/2632-2153/aca005>
- [90] A. P. Bartók, R. Kondor, and G. Csányi, “On representing chemical environments,” *Phys. Rev. B*, vol. 87, p. 184115, May 2013. [Online]. Available: <https://doi.org/10.1103/PhysRevB.87.184115>
- [91] J. Behler and M. Parrinello, “Generalized neural-network representation of high-dimensional potential-energy surfaces,” *Phys. Rev. Lett.*, vol. 98, p. 146401, Apr. 2007. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.98.146401>
- [92] E. Kocer, T. W. Ko, and J. Behler, “Neural network potentials: A concise overview of methods,” *Annual Review of Physical Chemistry*, vol. 73, no. 1, pp. 163–186, 2022.
- [93] S. K. Natarajan and J. Behler, “Neural network molecular dynamics simulations of solid–liquid interfaces: water at low-index copper surfaces,” *Phys. Chem. Chem. Phys.*, vol. 18, pp. 28 704–28 725, 2016.
- [94] E. Fransson, J. M. Rahm, J. Wiktor, and P. Erhart, “Revealing the free energy landscape of halide perovskites: Metastability and transition characters in cspbbr3 and mapbi3,” *Chemistry of Materials*, vol. 35, no. 19, pp. 8229–8238, 2023.
- [95] Z. Fan, “Improving the accuracy of the neuroevolution machine learning potential for multi-component systems,” *Journal of Physics: Condensed Matter*, vol. 34, no. 12, p. 125902, Jan. 2022. [Online]. Available: <https://doi.org/10.1088/1361-648X/ac462b>

- [96] Z. Fan, Y. Wang, P. Ying, K. Song, J. Wang, Y. Wang, Z. Zeng, K. Xu, E. Lindgren, J. M. Rahm, A. J. Gabourie, J. Liu, H. Dong, J. Wu, Y. Chen, Z. Zhong, J. Sun, P. Erhart, Y. Su, and T. Ala-Nissila, “Gpumd: A package for constructing accurate machine-learned potentials and performing highly efficient atomistic simulations,” *The Journal of Chemical Physics*, vol. 157, no. 11, p. 114801, 2022.
- [97] 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, no. 1, p. 398, Jan. 2021.
- [98] J. Behler, “Four generations of high-dimensional neural network potentials,” *Chemical Reviews*, vol. 121, no. 16, pp. 10 037–10 072, 2021.
- [99] T. Schaul, T. Glasmachers, and J. Schmidhuber, “High dimensions and heavy tails for natural evolution strategies,” in *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO ’11. New York, NY, USA: Association for Computing Machinery, 2011, p. 845–852.
- [100] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber, “Natural evolution strategies,” *Journal of Machine Learning Research*, vol. 15, no. 27, pp. 949–980, 2014. [Online]. Available: <http://jmlr.org/papers/v15/wierstra14a.html>
- [101] J. Applequist, J. R. Carl, and K.-K. Fung, “Atom dipole interaction model for molecular polarizability. application to polyatomic molecules and determination of atom polarizabilities,” *Journal of the American Chemical Society*, vol. 94, no. 9, pp. 2952–2960, 1972.
- [102] M. Swart, J. G. Snijders, and P. T. van Duijnen, “Polarizabilities of amino acid residues,” *Journal of Computational Methods in Sciences and Engineering*, vol. 4, pp. 419–425, 2004, 3.
- [103] M. V. Wolkenstein, “Intensities of vibrational spectra of molecules,” *Compt. Rend. Acad. Sci. URSS*, vol. 30, pp. 791–794, 1941.
- [104] ———, “Polarisability of molecules and intramolecular forces,” *Compt. Rend. Acad. Sci. URSS*, vol. 32, pp. 185–188, 1941.
- [105] S. Go, H. Bilz, and M. Cardona, “Bond charge, bond polarizability, and phonon spectra in semiconductors,” *Phys. Rev. Lett.*, vol. 34, pp. 580–583, Mar. 1975.
- [106] K. S. Smirnov, D. Bougeard, and P. Tandon, “Electro-optical parameters of bond polarizability model for aluminosilicates,” *The Journal of Physical Chemistry A*, vol. 110, no. 13, pp. 4516–4523, 2006, pMID: 16571058. [Online]. Available: <https://doi.org/10.1021/jp060151+>
- [107] R. Mazzarello, S. Caravati, S. Angioletti-Uberti, M. Bernasconi, and M. Parrinello, “Signature of tetrahedral Ge in the Raman spectrum of amorphous phase-change materials,” *Phys. Rev. Lett.*, vol. 104, p. 085503, Feb. 2010. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.104.085503>
- [108] M. Fang, S. Tang, Z. Fan, Y. Shi, N. Xu, and Y. He, “Transferability of machine learning models for predicting Raman spectra,” *The Journal of Physical Chemistry A*, vol. 128, no. 12, pp. 2286–2294, 2024.

- [109] L. Wirtz, M. Lazzeri, F. Mauri, and A. Rubio, “Raman spectra of bn nanotubes: Ab initio and bond-polarizability model calculations,” *Phys. Rev. B*, vol. 71, p. 241402, Jun. 2005. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevB.71.241402>
- [110] C. G. Broyden, “The convergence of a class of double-rank minimization algorithms 1. General considerations,” *IMA Journal of Applied Mathematics*, vol. 6, no. 1, pp. 76–90, Mar. 1970.
- [111] R. Fletcher, “A new approach to variable metric algorithms,” *The Computer Journal*, vol. 13, no. 3, pp. 317–322, Jan. 1970.
- [112] D. Goldfarb, “A family of variable-metric methods derived by variational means,” *Math. Comput.*, vol. 24, no. 109, pp. 23–26, 1970.
- [113] D. F. Shanno, “Conditioning of quasi-Newton methods for function minimization,” *Math. Comput.*, vol. 24, no. 111, pp. 647–656, 1970.
- [114] S. M. Oliver, J. J. Fox, A. Hashemi, A. Singh, R. L. Cavalero, S. Yee, D. W. Snyder, R. Jaramillo, H.-P. Komsa, and P. M. Vora, “Phonons and excitons in ZrSe<sub>2</sub>-ZrS<sub>2</sub> alloys,” *J. Mater. Chem. C*, vol. 8, pp. 5732–5743, 2020. [Online]. Available: <http://dx.doi.org/10.1039/D0TC00731E>
- [115] Z. Kou, A. Hashemi, M. J. Puska, A. V. Krasheninnikov, and H.-P. Komsa, “Simulating raman spectra by combining first-principles and empirical potential approaches with application to defective MoS<sub>2</sub>,” *npj Comput Mater*, vol. 6, p. 59, 2020.
- [116] P. Sutter, H. Komsa, H. Lu, A. Gruverman, and E. Sutter, “Few-layer tin sulfide (sns): Controlled synthesis, thickness dependent vibrational properties, and ferroelectricity,” *Nano Today*, vol. 37, p. 101082, 2021. [Online]. Available: <https://doi.org/10.1016/j.nantod.2021.101082>
- [117] C. Feng, J. Xi, Y. Zhang, B. Jiang, and Y. Zhou, “Accurate and interpretable dipole interaction model-based machine learning for molecular polarizability,” *Journal of Chemical Theory and Computation*, vol. 19, no. 4, pp. 1207–1217, 2023.
- [118] E. Fransson, J. Wiktor, and P. Erhart, “Phase transitions in inorganic halide perovskites from machine-learned potentials,” *The Journal of Physical Chemistry C*, vol. 127, no. 28, pp. 13 773–13 781, 2023.
- [119] V. G. Hadjiev, M. N. Iliev, B. Lv, Z. F. Ren, and C. W. Chu, “Anomalous vibrational properties of cubic boron arsenide,” *Phys. Rev. B*, vol. 89, p. 024308, Jan. 2014. [Online]. Available: <https://doi.org/10.1103/PhysRevB.89.024308>
- [120] H. Sun, K. Chen, G. Gamage, H. Ziyace, F. Wang, Y. Wang, V. Hadjiev, F. Tian, G. Chen, and Z. Ren, “Boron isotope effect on the thermal conductivity of boron arsenide single crystals,” *Materials Today Physics*, vol. 11, p. 100169, 2019. [Online]. Available: <https://doi.org/10.1016/j.mtphys.2019.100169>
- [121] A. Rai, S. Li, H. Wu, B. Lv, and D. G. Cahill, “Effect of isotope disorder on the raman spectra of cubic boron arsenide,” *Phys. Rev. Mater.*, vol. 5, p. 013603, Jan. 2021. [Online]. Available: <https://doi.org/10.1103/PhysRevMaterials.5.013603>

- [122] H.-L. Liu, H. Guo, T. Yang, Z. Zhang, Y. Kumamoto, C.-C. Shen, Y.-T. Hsu, L.-J. Li, R. Saito, and S. Kawata, “Anomalous lattice vibrations of monolayer MoS<sub>2</sub> probed by ultraviolet Raman scattering,” *Phys. Chem. Chem. Phys.*, vol. 17, pp. 14 561–14 568, 2015.
- [123] X. Zhang, X.-F. Qiao, W. Shi, J.-B. Wu, D.-S. Jiang, and P.-H. Tan, “Phonon and raman scattering of two-dimensional transition metal dichalcogenides from monolayer, multilayer to bulk material,” *Chem. Soc. Rev.*, vol. 44, pp. 2757–2785, 2015.
- [124] Y. Gillet, M. Giantomassi, and X. Gonze, “First-principles study of excitonic effects in Raman intensities,” *Phys. Rev. B*, vol. 88, p. 094305, Sep. 2013. [Online]. Available: <https://doi.org/10.1103/PhysRevB.88.094305>
- [125] Y. Gillet, S. Kontur, M. Giantomassi, C. Draxl, and X. Gonze, “Ab initio approach to second-order resonant Raman scattering including exciton-phonon interaction,” *Scientific Reports*, vol. 7, no. 1, p. 7344, 2017.
- [126] A. Sarycheva and Y. Gogotsi, “Raman spectroscopy analysis of the structure and surface chemistry of Ti<sub>3</sub>C<sub>2</sub>Tx MXene,” *Chemistry of Materials*, vol. 32, no. 8, pp. 3480–3488, 2020.
- [127] R. Ibragimova, M. J. Puska, and H.-P. Komsa, “ph-dependent distribution of functional groups on Titanium-based MXenes,” *ACS Nano*, vol. 13, no. 8, pp. 9171–9181, 2019. [Online]. Available: <https://doi.org/10.1021/acsnano.9b03511>
- [128] H. Min, D. Y. Lee, J. Kim, G. Kim, K. S. Lee, J. Kim, M. J. Paik, Y. K. Kim, K. S. Kim, M. G. Kim, T. J. Shin, and S. Il Seok, “Perovskite solar cells with atomically coherent interlayers on SnO<sub>2</sub> electrodes,” *Nature*, vol. 598, no. 7881, pp. 444–450, Oct. 2021.
- [129] M. Kim, J. Jeong, H. Lu, T. K. Lee, F. T. Eickemeyer, Y. Liu, I. W. Choi, S. J. Choi, Y. Jo, H.-B. Kim, S.-I. Mo, Y.-K. Kim, H. Lee, N. G. An, S. Cho, W. R. Tress, S. M. Zakeeruddin, A. Hagfeldt, J. Y. Kim, M. Grätzel, and D. S. Kim, “Conformal quantum dot–SnO<sub>2</sub> layers as electron transporters for efficient perovskite solar cells,” *Science*, vol. 375, no. 6578, pp. 302–306, 2022.
- [130] N. J. Jeon, J. H. Noh, W. S. Yang, Y. C. Kim, S. Ryu, J. Seo, and S. I. Seok, “Compositional engineering of perovskite materials for high-performance solar cells,” *Nature*, vol. 517, no. 7535, pp. 476–480, Jan. 2015.
- [131] V. G. Hadjiev, C. Wang, Y. Wang, X. Su, H. A. Calderon, F. R. Hernandez, Z. M. Wang, and J. M. Bao, “Phonon fingerprints of CsPb<sub>2</sub>Br<sub>5</sub>,” *Journal of Physics: Condensed Matter*, vol. 30, no. 40, p. 405703, Sep. 2018. [Online]. Available: <https://dx.doi.org/10.1088/1361-648X/aadeb4>
- [132] O. Yaffe, Y. Guo, L. Z. Tan, D. A. Egger, T. Hull, C. C. Stoumpos, F. Zheng, T. F. Heinz, L. Kronik, M. G. Kanatzidis, J. S. Owen, A. M. Rappe, M. A. Pimenta, and L. E. Brus, “Local polar fluctuations in lead halide perovskite crystals,” *Phys. Rev. Lett.*, vol. 118, p. 136001, Mar. 2017. [Online]. Available: <https://doi.org/10.1103/PhysRevLett.118.136001>
- [133] M. Menahem, N. Benshalom, M. Asher, S. Aharon, R. Korobko, O. Hellman, and O. Yaffe, “Disorder origin of Raman scattering in perovskite single crystals,” *Phys. Rev. Mater.*, vol. 7, p. 044602, Apr. 2023. [Online]. Available: <https://doi.org/10.1103/PhysRevMaterials.7.044602>

- [134] R. X. Yang, J. M. Skelton, E. L. da Silva, J. M. Frost, and A. Walsh, "Assessment of dynamic structural instabilities across 24 cubic inorganic halide perovskites," *The Journal of Chemical Physics*, vol. 152, no. 2, p. 024703, Jan. 2020. [Online]. Available: <https://doi.org/10.1063/1.5131575>
- [135] —, "Spontaneous octahedral tilting in the cubic inorganic cesium halide perovskites CsSnX<sub>3</sub> and CsPbX<sub>3</sub> (X = F, Cl, Br, I)," *The Journal of Physical Chemistry Letters*, vol. 8, no. 19, pp. 4720–4726, 2017, pMID: 28903562. [Online]. Available: <https://doi.org/10.1021/acs.jpcclett.7b02423>
- [136] J.-A. Huang, M. Z. Mousavi, G. Giovannini, Y. Zhao, A. Hubarevich, M. A. Soler, W. Rocchia, D. Garoli, and F. De Angelis, "Multiplexed discrimination of single amino acid residues in polypeptides in a single sers hot spot," *Angew. Chem., Int. Ed.*, vol. 59, no. 28, pp. 11 423–11 431, 2020.
- [137] Y. Zhao, M. Iarossi, A. F. De Fazio, J.-A. Huang, and F. De Angelis, "Label-free optical analysis of biomolecules in solid-state nanopores: Toward single-molecule protein sequencing," *ACS Photonics*, vol. 9, no. 3, pp. 730–742, 2022.
- [138] P. Candeloro, E. Grande, R. Raimondo, D. Di Mascolo, F. Gentile, M. L. Coluccio, G. Perozziello, N. Malara, M. Francardi, and E. Di Fabrizio, "Raman database of amino acids solutions: a critical study of extended multiplicative signal correction," *Analyst*, vol. 138, pp. 7331–7340, 2013.
- [139] I. V. Krauklis, A. V. Tulub, A. V. Golovin, and V. P. Chelibanov, "Raman spectra of glycine and their modeling in terms of the discrete–continuum model of their water solvation shell," *Optics and Spectroscopy*, vol. 128, no. 10, pp. 1598–1601, Oct. 2020.
- [140] B. Hernández, F. Pflüger, M. Nsangou, and M. Ghomi, "Vibrational analysis of amino acids and short peptides in hydrated media. IV. Amino acids with hydrophobic side chains: l-alanine, l-valine, and l-isoleucine," *The Journal of Physical Chemistry B*, vol. 113, no. 10, pp. 3169–3178, 2009.
- [141] G. Zhu, X. Zhu, Q. Fan, and X. Wan, "Raman spectra of amino acids and their aqueous solutions," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 78, no. 3, pp. 1187–1195, Mar. 2011.
- [142] B. Hernández, F. Pflüger, A. Adenier, S. G. Kruglik, and M. Ghomi, "Side chain flexibility and protonation states of sulfur atom containing amino acids," *Phys. Chem. Chem. Phys.*, vol. 13, pp. 17 284–17 294, 2011. [Online]. Available: <http://dx.doi.org/10.1039/C1CP21054H>
- [143] D. Świąch, N. Piergies, G. Palumbo, and C. Paluszkiwicz, "In situ and ex situ Raman studies of cysteine's behavior on a titanium surface in buffer solution," *Coatings*, vol. 13, no. 1, 2023. [Online]. Available: <https://doi.org/10.3390/coatings13010175>
- [144] E. B. Wilson, "The normal modes and frequencies of vibration of the regular plane hexagon model of the benzene molecule," *Phys. Rev.*, vol. 45, pp. 706–714, May 1934. [Online]. Available: <https://doi.org/10.1103/PhysRev.45.706>
- [145] R. P. Rava and T. G. Spiro, "Resonance enhancement in the ultraviolet Raman spectra of aromatic amino acids," *The Journal of Physical Chemistry*, vol. 89, no. 10, pp. 1856–1861, 1985. [Online]. Available: <https://doi.org/10.1021/j100256a007>

- [146] E. J. Baran, I. Viera, and M. H. Torre, "Vibrational spectra of the Cu(II) complexes of l-asparagine and l-glutamine," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 66, no. 1, pp. 114–117, 2007. [Online]. Available: <https://doi.org/10.1016/j.saa.2006.01.052>
- [147] S. Sylvestre, S. Sebastian, S. Edwin, M. Amalanathan, S. Ayyapan, T. Jayavarthanam, K. Oudayakumar, and S. Solomon, "Vibrational spectra (FT-IR and FT-Raman), molecular structure, natural bond orbital, and TD-DFT analysis of l-asparagine monohydrate by density functional theory approach," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 133, pp. 190–200, 2014. [Online]. Available: <https://doi.org/10.1016/j.saa.2014.05.040>
- [148] A. L. Jenkins, R. A. Larsen, and T. B. Williams, "Characterization of amino acids using Raman spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 61, no. 7, pp. 1585–1594, 2005.
- [149] 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, Sep. 2020.
- [150] J. Carrete, H. Montes-Campos, R. Wanzenböck, E. Heid, and G. K. H. Madsen, "Deep ensembles vs committees for uncertainty estimation in neural-network force fields: Comparison and application to active learning," *The Journal of Chemical Physics*, vol. 158, no. 20, p. 204801, May 2023.
- [151] T. Hayashi and S. Mukamel, "Vibrational-exciton couplings for the amide I, II, III, and A modes of peptides," *J. Phys. Chem. B*, vol. 111, no. 37, pp. 11 032–11 046, 2007.
- [152] Y. Ji, X. Yang, Z. Ji, L. Zhu, N. Ma, D. Chen, X. Jia, J. Tang, and Y. Cao, "DFT-calculated IR spectrum amide I, II, and III band contributions of *N*-methylacetamide fine components," *ACS Omega*, vol. 5, no. 15, pp. 8572–8578, 2020.

## Appendices

Appendix A Nosé-Hoover thermostat

Appendix B Separable natural evolution strategy

## Appendix A : Nosé-Hoover thermostat

The thermostat is used to keep the temperature constant during MD simulations. In this work we use the Nosé-Hoover thermostat as already implemented in the GPUMD software [96]. It follows the Nosé-Hoover chains first proposed by Martyna *et al.* [63].

According to Hamiltonian mechanics, the evolution of positions  $\mathbf{x}_i$  and momenta  $\mathbf{p}_i$  of a system of atoms can be written as

$$\begin{aligned}\frac{d\mathbf{x}_i}{dt} &= \frac{\mathbf{p}_i}{m_i}, \\ \frac{d\mathbf{p}_i}{dt} &= \mathbf{F}_i.\end{aligned}\tag{A1}$$

The initial idea of Nosé [60] was to add an additional degree of freedom  $\eta$  and its corresponding momentum  $p_\eta$ . By interacting with the system, this additional degree of freedom acts as a heat bath to control the temperature. Equations (A1) then take the form

$$\begin{aligned}\frac{d\mathbf{x}_i}{dt} &= \frac{\mathbf{p}_i}{m_i}, \\ \frac{d\mathbf{p}_i}{dt} &= \mathbf{F}_i - \mathbf{p}_i \frac{p_\eta}{Q}, \\ \frac{dp_\eta}{dt} &= \sum_i \frac{\mathbf{p}_i^2}{m_i} - Nk_B T,\end{aligned}\tag{A2}$$

where  $Q$  is a constant representing the "mass" of the additional degree of freedom  $\eta$ . While trajectories from Equations (A2) are now in the NVT ensemble, they are not necessarily ergodic. In fact, the Nosé-Hoover thermostat usually suffers from poor ergodicity as the system does not correctly sample the whole phase-space. To solve this issue, a chain of thermostats can be used [63]. This means that the additional degree of freedom  $\eta$  is itself interacting with another  $\eta_2$  which can also interact with  $\eta_3$ , in turn creating a chain of any size. For a chain of size  $M$ , the evolution of  $p_{\eta_j}$  then reads

$$\begin{aligned}\frac{dp_\eta}{dt} &= \left[ \sum_i \frac{\mathbf{p}_i^2}{m_i} - Nk_B T \right] - p_\eta \frac{p_{\eta_2}}{Q_2}, \\ \frac{dp_{\eta_j}}{dt} &= \left[ \frac{p_{\eta_{j-1}}^2}{Q_{j-1}} - k_B T \right] - p_{\eta_j} \frac{p_{\eta_{j+1}}}{Q_{j+1}}, \\ \frac{dp_{\eta_M}}{dt} &= \left[ \frac{p_{\eta_{M-1}}^2}{Q_{M-1}} - k_B T \right].\end{aligned}\tag{A3}$$

In GPUMD, the length of the chain is fixed at  $M = 4$ . The parameter  $Q$  is chosen so that  $Q = Nk_B T \tau^2$  and  $\tau/\Delta t = 100$ , where  $\tau$  is a time parameter and  $\Delta t$  is the time step.

## Appendix B : Separable natural evolution strategy

Training the many parameters of a neural networks represents a very challenging task. The connectivities, biases and descriptor parameters of NEP can be combined into one vector  $\mathbf{z}$ . In separable natural evolution strategy (SNES), these values are attributed a mean  $\mathbf{m}$  and a variance  $\mathbf{s}$ , which are repeatedly evolved until convergence. Initial values can be chosen such that means are random numbers between  $-\frac{1}{2}$  and  $\frac{1}{2}$  and all variances are set to 1. They are then evolved according to the following algorithm:

- (1) A population  $\mathbf{z}_k$  is created according to

$$z_{k,i} \leftarrow m_i + s_i \cdot r_{k,i}, \quad (\text{B1})$$

where  $\mathbf{r}_k$  are random numbers taken from a Gaussian distribution of mean 0 and variance 1.

- (2) Loss functions of the populations  $L(\mathbf{z}_k)$  are calculated from Equation (46). The solutions  $\mathbf{z}_k$  are then ranked and attributed a utility rank  $u_k$  (individuals with a low loss function  $L(\mathbf{z}_k)$  are attributed a larger rank  $u_k$ ). Note that the shape of  $u_k$  has to be monotonous (see Ref. [100] for additional details).

- (3) Gradients for the mean  $\mathbf{J}^m$  and variances  $\mathbf{J}^s$  are obtained as

$$\begin{aligned} J_i^m &\leftarrow \sum_k u_k \cdot r_{k,i}, \\ J_i^s &\leftarrow \sum_k u_k \cdot (r_{k,i}^2 - 1). \end{aligned} \quad (\text{B2})$$

- (4) Using these gradients, the updated means  $\mathbf{m}$  and variances  $\mathbf{s}$  are given by

$$\begin{aligned} m_i &\leftarrow m_i + \eta_m \cdot s_i \cdot J_i^m, \\ s_i &\leftarrow s_i \cdot \exp\left(\frac{\eta_s}{2} \cdot J_i^s\right), \end{aligned} \quad (\text{B3})$$

where  $\eta_m$  and  $\eta_s$  are the mean and variance learning rates, respectively.

This loop is repeated until the convergence criterion is reached. In the GPUMD implementation, the number of iterations is fixed at the beginning of the training. Other convergence criterion based on the variance  $\mathbf{s}$  could also be used.



## Original publications

- I Berger, E., & Komsa, H.-P., (2024). Polarizability models for simulations of finite temperature Raman spectra from machine learning molecular dynamics. *Physical Review Materials*, 8, 043802. <https://doi.org/10.1103/PhysRevMaterials.8.043802>
- II Dash, A. K., Swaminathan, H., Berger, E., Mondal, M., Lehenkari, T., Prasad, P. R., Watanabe, K., Taniguchi, T., Komsa, H.-P., & Singh, A. (2023). Evidence of defect formation in monolayer MoS<sub>2</sub> at ultralow accelerating voltage electron irradiation. *2D Materials*, 10, 035002. <https://doi.org/10.1088/2053-1583/acc7b6>
- III Berger, E., Lv, Z.-P., & Komsa, H.-P. (2023). Raman spectra of 2D titanium carbide MXene from machine-learning force field molecular dynamics. *Journal of Materials Chemistry C*, 11, 1311–1319. <https://doi.org/10.1039/D2TC04374B>
- IV Berger, E., Niemelä, J., Lampela, O., Juffer, A. H., & Komsa, H.-P. (2024). Raman spectra of amino acids and peptides from machine learning polarizabilities. *Journal of Chemical Information and Modeling*, 64(12), 4601–4612. <https://doi.org/10.1021/acs.jcim.4c00077>

Reprinted with permissions from APS (Publication I © 2024 American Physical Society), IOP Publishing (Publication II © 2023 IOP Publishing Ltd) and ACS (Publication IV © 2024 American Chemical Society), or under Creative Commons CC BY 3.0 licence<sup>3</sup> (Publication III © 2023 Authors).

Original publications are not included in the electronic version of the dissertation.

---

<sup>3</sup><https://creativecommons.org/licenses/by/3.0/>



952. Tahir, Muhammad Naeem (2024) Advanced vehicular communications through cellular and short-range networks exploiting road weather and traffic observation data
953. Ren, Zhongfei (2024) Adsorption and advanced oxidation/reduction process for elimination of per- and polyfluoroalkyl substances and pharmaceutical pollutants in water
954. Beddiar, Djamilia Romaiassa (2024) Deep learning-based automatic captioning for medical imaging
955. Zhao, He (2024) Bismuth halide perovskites as photocatalysts for hydrogen production
956. Kaksonen, Rauli (2024) Transparent and tool-driven security assessment for sustainable IoT cybersecurity
957. Yastrebova-Castillo, Anastasia (2024) LEO satellite constellations for autonomous system operation in the Arctic region : situational awareness and communication aspects
958. Airaksinen, Susanna (2024) Oxide scale formation of stainless steel in transition of annealing and reheating towards fossil-free methods
959. Bahador, Nooshin (2024) Assessment of neurological function with multimodal and multichannel physiological signal analysis using machine and deep learning techniques
960. Akbar, Rehman (2024) CMOS integrated wideband, flexible and scalable beamforming architecture
961. Asad Ullah, Muhammad (2024) Machine-type direct-to-satellite communications : modeling and performance analysis
962. Yue, Xin (2024) Lignin processing by deep eutectic solvents : from structural chemistry to its advanced valorization
963. Lyons, Kevin (2024) Potable water from shallow wells : a multi-method study of factors influencing physicochemical and microbiological quality
964. Kinnunen, Juho (2024) On-site wastewater treatment in Northern conditions : real world performance, pollutant load and fate
965. Bagheri, Mohammad (2024) High-throughput computation of Raman spectra by atomistic first-principles methods

S E R I E S E D I T O R S

**A**  
**SCIENTIAE RERUM NATURALIUM**  
*University Lecturer Mahmoud Filali*

**B**  
**HUMANIORA**  
*University Lecturer Santeri Palviainen*

**C**  
**TECHNICA**  
*Senior Research Fellow Antti Kajjalainen*

**D**  
**MEDICA**  
*University Lecturer Pirjo Kaakinen*

**E**  
**SCIENTIAE RERUM SOCIALIUM**  
*University Lecturer Henri Pettersson*

**E**  
**SCRIPTA ACADEMICA**  
*Strategy Officer Mari Katvala*

**G**  
**OECONOMICA**  
*University Researcher Marko Korhonen*

**H**  
**ARCHITECTONICA**  
*Associate Professor Anu Soikkeli*

**EDITOR IN CHIEF**  
*University Lecturer Santeri Palviainen*

