
UNIVERSITÉ DE LIÈGE

Machine-learning-driven atomistic modeling of molybdenum sulfides: anharmonic dynamics and amorphous structures

A thesis presented to obtain the academic grade of

Doctor in Science

by

SAMUEL LONGO



Academic year 2025–2026

JURY COMMITTEE

Pr. Matthieu Verstraete (Supervisor)
Pr. Françoise Remacle (President)
Dr. He Xu (Secretary)
Pr. Menno Bokdam
Dr. Hannah Johnson
Dr. Alberto Marmodoro

*Université de Liège
Université de Liège
Université de Liège
University of Twente
Toyota Motor Europe
Institute of Physics, Czech Academy of Sciences*

PhD Thesis:

**Machine-learning-driven atomistic modeling of
molybdenum sulfides: anharmonic dynamics and
amorphous structures**

Samuel Longo

2022-2026

Abstract

Molybdenum sulfides are critical for energy and catalytic applications, but modeling their finite-temperature dynamics and amorphous topologies is computationally prohibitive for standard *ab initio* methods. This thesis overcomes these limitations by integrating Machine-Learning Interatomic Potentials (MLIPs) with advanced configurational sampling. To automate and scale these workflows, two custom software packages were developed: MlipKit, a unified MLIP deployment API, and Piomlacs, an iterative active-learning framework. For crystalline 2H-MoS₂, MLIP-driven simulations were coupled with the Temperature Dependent Effective Potential (TDEP) method to calculate finite-temperature Raman spectra. The models successfully captured anharmonic lineshape broadenings and frequency shifts, demonstrating the necessity of incorporating quantum zero-point energy in low-temperature sampling, while exposing the boundaries of static perturbative methodologies. In parallel, large-scale MLIP-driven melt-and-quench simulations mapped the structural evolution of amorphous MoS_x ($x = 2, 4, 6$). The analysis revealed a profound topological transition: as sulfur content increases, the rigidly cross-linked metal-chalcogenide glass ($x = 2$) is fragmented into isolated molybdenum centers surrounded by a flexible polysulfide matrix ($x = 6$). Crucially, the realistic, slower cooling rates unlocked by the MLIPs proved that while local short-range coordination is thermodynamically locked, the extended mesostructure is sensitive to kinetic trapping, establishing a rigorous new baseline for modeling amorphous catalysts.

Résumé

Les sulfures de molybdène sont essentiels pour les applications énergétiques et catalytiques, mais la modélisation de leur dynamique à température finie et de leurs topologies amorphes est excessivement coûteuse en termes de calcul pour les méthodes *ab initio* standard. Cette thèse surmonte ces limites en intégrant des potentiels interatomiques basés sur l'apprentissage automatique (MLIP) à un échantillonnage configurationnel avancé. Pour automatiser et généraliser ces workflow, deux logiciels sur mesure ont été développés : MlipKit, une API unifiée de déploiement de MLIP, et Piomlacs, un cadre d'apprentissage actif itératif. Pour le 2H-MoS (cristallin) des simulations pilotées par les MLIP ont été couplées à la méthode du potentiel effectif dépendant de la température (TDEP) pour calculer les spectres Raman à température finie. Les modèles ont réussi à capturer les élargissements de raies anharmoniques et les décalages de fréquence, démontrant la nécessité d'intégrer l'énergie quantique de point zéro dans l'échantillonnage à basse température, tout en exposant les limites des méthodologies perturbatives statiques. En parallèle, des simulations de fusion-trempe à grande échelle pilotées par les MLIP ont cartographié l'évolution structurale du MoS_x amorphe ($x = 2, 4, 6$). L'analyse a révélé une profonde transition topologique : à mesure que la teneur en soufre augmente, le verre de métal-chalcogénure rigidement réticulé ($x = 2$) se fragmente en centres de molybdène isolés entourés d'une matrice polysulfure flexible ($x = 6$). Fait crucial, les vitesses de refroidissement réalistes et plus lentes permises par les MLIP ont prouvé que si la coordination locale à courte distance est thermodynamiquement verrouillée, la mésostructure étendue est sensible au piégeage cinétique, établissant ainsi une nouvelle référence rigoureuse pour la modélisation des catalyseurs amorphes.

Contents

Contents	vii
1 Theoretical background	3
1.1 Basics of quantum mechanics	3
1.1.1 Molecular systems	3
1.1.2 Periodic systems	9
1.1.3 Phonon Green's functions	21
1.2 Raman spectroscopy	24
1.2.1 Unpolarized experiments on isotropic samples . . .	29
1.3 Density Functional Theory	31
1.3.1 Density Functional Perturbation Theory (DFPT) . .	34
1.4 Molecular Dynamics	37
1.4.1 Thermostats	39
1.4.2 Barostats	40
1.4.3 Thermodynamic ensembles	40
1.4.4 The Ergodic Hypothesis	41
1.4.5 Equilibration of MD simulations	42
1.5 Machine Learning	43
1.5.1 Linear fit	44
1.5.2 Neural Networks (NN)	45
1.5.3 Backpropagation and Model Optimization	47
1.5.4 Activation functions	47
1.5.5 Training	48
1.5.6 Validation	49
1.5.7 Machine Learning Interatomic Potentials	50
1.5.8 Message Passing Neural Network with Atomic Cluster Expansion (MACE)	52
1.6 Temperature Dependent Effective Potential	56
1.6.1 Sampling for TDEP: Molecular Dynamics	59
1.6.2 Sampling for TDEP: Self Consistent Stochastic TDEP, sTDEP	59
1.7 Characterization of amorphous materials	61
1.7.1 Radial and Angular Distributions	61
1.7.2 Local Coordination Environments: ChemEnv Analysis	62
1.7.3 Intermediate-Range Order: Ring Statistics	62
1.7.4 Vibrational Properties: Vibrational Density of States (vibDOS)	63
2 Temperature-dependent Raman spectra of 2H-MoS₂	65
2.1 Introduction and motivation	65
2.2 2H-MoS ₂	66
2.3 Phonons at 0 K	68
2.4 Generation of a dataset and training of the MLIP (M1) . . .	70
2.5 Cell thermalization	72
2.6 Phonons at finite temperature	74
2.6.1 MD sampling	76

2.6.2	Stochastic sampling	80
2.6.3	Comparison and experimental reference	85
2.7	Conclusions and perspectives	92
3	Amorphous MoS_x	95
3.1	Introduction and Motivation	95
3.2	State of the Art: Computational generation of amorphous structures	97
3.3	MLIP, and training data	99
3.4	Amorphous structure generation	103
3.4.1	<i>a</i> -MoS ₂ structural exploration	103
3.4.2	<i>a</i> -MoS ₄ structural exploration	116
3.4.3	<i>a</i> -MoS ₆ structural exploration	123
3.4.4	Stoichiometry comparison	129
3.5	Conclusions and perspectives	133
4	MlipKit	137
4.1	Description of the MlipKit software package	137
4.2	Usage example	138
5	Python InterOperable MLACS (Piomlacs)	141
5.1	Decoupling the loop from the potential	141
5.2	Design	142
5.3	Usage example	143
6	Conclusions and outlook	145
	Bibliography	147
	Acknowledgments	159

Introduction

Molybdenum sulfides are versatile materials with applications ranging from nanoelectronics to energy storage and catalysis [1–12]. While crystalline 2H-MoS₂ is widely celebrated for its layer-dependent electronic properties and direct bandgap transition in monolayers, the reliable deployment of MoS₂-based devices under real operational conditions demands an accurate understanding of its thermal and dynamical response. Raman spectroscopy is among the favorite non-destructive techniques to benchmark these systems, serving as an exquisite probe of layer number, strain, and local temperature. However, interpreting finite-temperature Raman spectra requires an explicit treatment of lattice anharmonicity: thermal expansion and many-body phonon–phonon scattering inherently shift mode frequencies and govern finite phonon lifetimes. Traditional harmonic models and perturbative zero-temperature corrections completely break down in this regime, leaving lineshape broadenings and temperature-dependent softening unresolved.

In parallel, real-world catalytic and energy applications frequently depart from pristine crystal lattices, turning instead to defect-rich or fully amorphous molybdenum sulfides (MoS_x). The structural exploration of these amorphous phases is equally crucial: their highly tunable sulfur stoichiometries and abundant coordinatively unsaturated active sites make them exceptionally relevant for electrocatalytic hydrogen evolution and next-generation battery architectures [3, 13–15].

However, whether mapping the immensely complex structural landscape of amorphous MoS_x across varied compositions, or sampling the finite-temperature vibrational phase space needed to capture anharmonic phonon self-energies in crystalline MoS₂, computational modeling faces a severe bottleneck. Both problems require statistical sampling over supercell sizes and simulation timescales that far exceed the reach of standard *ab initio* Molecular Dynamics (AIMD).

This thesis overcomes these limitations by leveraging Machine Learning Interatomic Potentials (MLIPs). By training MLIPs to reproduce first-principles accuracy at a fraction of the computational cost, we enable the extensive molecular dynamics (MD) sampling required to rigorously explore these systems. This approach unlocks the ability to accurately predict complex

anharmonic lattice dynamics and simulate massive non-crystalline networks.

The remainder of this thesis is organized as follows:

Chapter 1 establishes the theoretical foundation, detailing density functional theory, lattice dynamics, the underlying principles of machine learning interatomic potentials, and the structural characterization methods used in this study.

Chapter 2 investigates the finite-temperature dynamics of crystalline 2H-MoS₂. By coupling MLIP-driven molecular dynamics with the Temperature Dependent Effective Potential (TDEP) method, this chapter calculates the finite-temperature Raman spectrum, explicitly capturing the anharmonic frequency shifts and intrinsic phonon broadenings that static harmonic models fail to describe.

Chapter 3 shifts focus to the crucial structural exploration of amorphous MoS_x. Utilizing large-scale MLIP-driven MD, we extensively sample the configurational phase space across varying sulfur stoichiometries ($x = 2, 4, 6$). This chapter provides critical insights into the local atomic coordination, structural motifs, and phase stability of these highly relevant disordered networks.

To facilitate the massive data generation and potential training required for these physical studies, tailored software frameworks were developed and are detailed in the final methodology chapters:

Chapter 4 presents **MlipKit**, a comprehensive software suite developed during this work to streamline the end-to-end training, validation, and deployment of machine-learned potentials for complex material simulations.

Chapter 5 introduces **Piomlacs**, a custom implementation of the Machine Learning Assisted Canonical Sampling (MLACS) algorithm. Designed to interface seamlessly with the Atomic Simulation Environment (ASE) and MlipKit, Piomlacs automates the active-learning generation of robust, highly representative training datasets.

Finally, the last chapter summarizes the conclusions of this thesis and outlines future perspectives for the predictive modeling of crystalline and amorphous materials.

1.1 Basics of quantum mechanics

1.1.1 Molecular systems

The Hamiltonian of a system of electrons and nuclei can be written as the sum of the kinetic and potential operators:

$$\hat{H}(\mathbf{r}, \mathbf{R}) = \hat{T}_e(\mathbf{r}) + \hat{T}_n(\mathbf{R}) + \hat{V}(\mathbf{r}, \mathbf{R}) \quad (1.1)$$

where $\hat{T}_e$ and $\hat{T}_n$ are the electronic and nuclear kinetic energy operators, respectively.

In the absence of external fields, the potential term can be expanded as:

$$\begin{aligned} \hat{V}(\mathbf{r}, \mathbf{R}) &= \hat{V}_{ee}(\mathbf{r}) + \hat{V}_{nn}(\mathbf{R}) + \hat{V}_{en}(\mathbf{r}, \mathbf{R}) \\ \hat{V}_{ee}(\mathbf{r}) &= \frac{1}{2} \sum_{ij, i \neq j}^n \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \\ \hat{V}_{nn}(\mathbf{R}) &= \frac{1}{2} \sum_{IJ, I \neq J}^M \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} \\ \hat{V}_{en}(\mathbf{r}, \mathbf{R}) &= - \sum_{iI}^{n, M} \frac{Z_I}{|\mathbf{r}_i - \mathbf{R}_I|} \end{aligned} \quad (1.2)$$

where n is the number of electrons and M is the number of nuclei. Note that the electron-nucleus sum couples both degrees of freedom. Moreover, electronic and nuclear variables can interact in any way in the functional form of the eigenstates. However, since electrons are $\sim 10^4$ times lighter than nuclei, the *adiabatic approximation* allows us to assume that the electrons are still in their ground state after any nuclear movement. The generic Hamiltonian's eigenstate can be written as:

$$\Psi(\mathbf{r}, \mathbf{R}) = \chi(\mathbf{R})\psi(\mathbf{r}; \mathbf{R}) \quad (1.3)$$

where $\chi(\mathbf{R})$ and $\psi(\mathbf{r}; \mathbf{R})$ are the nuclear and electronic terms, respectively.

The nuclear wavefunction does not depend on $\mathbf{r}$ because of the adiabatic approximation: because any variation in $\mathbf{r}$ happens too quickly for the nuclei to be influenced by it. Note that the electronic wavefunction has

a parametric dependence on $\mathbf{R}$, because nuclei are much slower, hence, the electronic dynamics can be described considering the nuclei as frozen. Furthermore, we neglect the effect of the nuclear kinetic operator on the electronic wavefunction.

We can then solve the electronic and nuclear problems separately. For a given nuclear geometry, the electronic Hamiltonian reads:

$$\hat{H}_e(\mathbf{r}, \mathbf{R}) = \hat{T}_e(\mathbf{r}) + \hat{V}(\mathbf{r}, \mathbf{R}) \quad (1.4)$$

where we included the constant nucleus-nucleus interaction terms. Electronic states can be found as eigenstates of $\hat{H}_e$:

$$\hat{H}_e \psi_i(\mathbf{r}; \mathbf{R}) = E_i^e \psi_i(\mathbf{r}; \mathbf{R}) \quad (1.5)$$

Note that $\psi_i(\mathbf{r}; \mathbf{R})$ describes how the electronic probability density behaves as the nuclei move, assuming the electrons remain in the state i .

The nuclear problem

Assuming we can solve the electronic Hamiltonian (eq. 1.5) for any nuclear geometry, the eigenvalue $E_i^e \equiv U(\mathbf{R})$ is the potential of the field the nuclei move through as a function of their position, also known as Potential Energy Surface (PES) (we will drop any electronic index in this section for clarity).

The nuclear Hamiltonian then reads:

$$[\hat{T}_n(\mathbf{R}) + U(\mathbf{R})] \chi_k(\mathbf{R}) = E_k^{TOT} \chi_k(\mathbf{R}) \quad (1.6)$$

where the index k runs over the nuclear Hamiltonian spectrum. While the kinetic term is analytically known, the PES is arbitrarily complex, and we need to simplify it in order to solve the equation.

We expand the PES as a Taylor series of the nuclear coordinates around an

equilibrium geometry (where the PES reaches a minimum):

$$\begin{aligned}
 U(\mathbf{R}) &= U(\mathbf{R}^{eq}) + \sum_i \left. \frac{\partial U}{\partial R_i} \right|_{eq} (R_i - R_i^{eq}) \\
 &+ \frac{1}{2} \sum_{ij} \left. \frac{\partial^2 U}{\partial R_i \partial R_j} \right|_{eq} (R_i - R_i^{eq}) (R_j - R_j^{eq}) \\
 &+ \dots \\
 \mathbf{R}^{eq} &= \{R_1^{eq}, R_2^{eq}, \dots, R_{3M}^{eq}\}
 \end{aligned} \tag{1.7}$$

where the indices encode the nucleus and the Cartesian direction (each displacement is an independent degree of freedom). We can make a change of variables $R_i \rightarrow u_i = R_i - R_i^{eq}$ and rewrite the PES as follows:

$$U(\mathbf{u}) = U(0) + \sum_i \left. \frac{\partial U}{\partial u_i} \right|_0 u_i + \frac{1}{2} \sum_{ij} \left. \frac{\partial^2 U}{\partial u_i \partial u_j} \right|_0 u_i u_j + \dots \tag{1.8}$$

with $\mathbf{u} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{3M}\}$. Note that the derivatives in the first-order terms are zero by construction.

Now we introduce mass-weighted coordinates $Q_i = \sqrt{m_i} u_i$, and the conjugated momenta $P_i = p_i / \sqrt{m_i}$ (p_i is the conjugate momentum of u_i , which preserves the commutation of the respective operators $[\hat{u}_i, \hat{p}_j] = [\hat{Q}_i, \hat{P}_j] = i\hbar \delta_{ij}$).

Now we can rewrite the PES expansion as:

$$U(\mathbf{Q}) = U(0) + \frac{1}{2} \sum_{ij} \Phi_{ij}^{(2)} Q_i Q_j + \dots \tag{1.9}$$

where we define the mass-weighted second-order Interatomic Force Constant (IFC):

$$\Phi_{ij}^{(2)} = \left. \frac{\partial^2 U}{\partial u_i \partial u_j} \right|_0 \frac{1}{\sqrt{m_i m_j}} \tag{1.10}$$

The harmonic approximation

If we truncate the PES expansion at the second order, it describes a set of $3M$ coupled harmonic oscillators.

The IFC matrix lives in the space of the relative atomic displacements, and is written in the canonical basis representing displacements of one atom along one Cartesian direction by a fixed arbitrary unit. By diagonalizing the IFC matrix, we find a basis called *normal modes of vibrations*.

We can then rewrite the PES as a function of some new displacements corresponding to the normal mode basis. The operators associated with these displacements now are:

$$\{\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_{3M}\} \rightarrow \{\hat{\eta}_1, \dots, \hat{\eta}_{3M}\} \quad (1.11)$$

$$\hat{\eta}_i = \sum_j C_{ji} \hat{Q}_j \quad (1.12)$$

where C_{ji} is the ji element of the coefficient matrix from the diagonalization of the IFC, matrix $\mathbf{C}$, and $\hat{Q}_j$ is the j -th displacement operator.

Note that $\boldsymbol{\eta}_i = (C_{1i}, C_{2i}, \dots, C_{3Mi})$ is the i -th eigenvector written in the canonical basis of the space of relative displacements. The invariance of the system's total energy under a uniform rigid translation dictates the existence of exactly 3 zero-frequency macroscopic translational modes (or 3 zero-frequency acoustic branches at the Brillouin zone center for extended periodic crystals). This physical constraint is mathematically enforced on the force constant matrix via the Acoustic Sum Rule (ASR). For isolated molecules, energy is additionally invariant under rigid spatial rotations, yielding a total of 6 (or 5 for a single linear molecule) zero-frequency normal modes.

The harmonic (second order) PES written as a function of the normal mode coordinates reads (written in operator formalism):

$$\hat{U}(\boldsymbol{\eta}) \approx U(0) + \frac{1}{2} \sum_i \omega_i^2 \hat{\eta}_i^2 \quad (1.13)$$

where ω_i^2 is the i -th eigenvalue, and has units of a squared frequency.

Applying the same transformation to the momentum operators as in eq. 1.12:

$$\hat{\pi}_i = \sum_j C_{ji} \hat{P}_j \quad (1.14)$$

we can rewrite the nuclear Hamiltonian (eq. 1.6) as a sum of uncoupled harmonic oscillators:

$$\hat{H}_n = \sum_i \frac{\hat{\pi}_i^2}{2} + \frac{1}{2} \omega_i^2 \hat{\eta}_i^2 \quad (1.15)$$

where we write the kinetic operator as $\hat{T}_i = \hat{\pi}_i^2/2$. Also note that:

$$[\hat{\eta}_i, \hat{\pi}_j] = i\hbar \delta_{i,j} \quad (1.16)$$

Each harmonic oscillator has a set of eigenstates $\{\theta_\lambda(\eta)\}$ which can be found by solving the differential equation:

$$\left(\frac{\hat{\pi}^2}{2} + \frac{1}{2}\omega^2\hat{\eta}^2\right)\theta_\lambda(\eta) = E_\lambda\theta_\lambda(\eta) \quad (1.17)$$

This differential equation admits the following solutions:

$$\begin{aligned} \theta_\lambda(\eta) &= N_\lambda \cdot H_\lambda\left(\sqrt{\frac{\omega}{\hbar}}\eta\right) \cdot e^{-\frac{\omega}{2\hbar}\eta^2} \\ N_\lambda &= \frac{1}{\sqrt{2^\lambda\lambda!}}\left(\frac{\omega}{\pi\hbar}\right)^{1/4} \end{aligned} \quad (1.18)$$

with λ is a non-negative integer, and H_λ is the λ -th-order Hermite polynomial.

The kinetic and total energy of state θ_λ are

$$E_\lambda^K = \frac{1}{2}\hbar\omega\left(n_\lambda + \frac{1}{2}\right) \quad (1.19)$$

$$E_\lambda = \hbar\omega\left(n_\lambda + \frac{1}{2}\right). \quad (1.20)$$

Note that the lowest state has finite kinetic energy (and potential) $E_0 = \hbar\omega/2$ due to the Heisenberg uncertainty principle; this is called *Zero Point Energy* (ZPE), associated with the *Zero Point Motion* (ZPM), and is a purely quantum result.

The eigenstates of $\hat{H}_n$ can be built by choosing one solution for every harmonic oscillator and multiplying them:

$$\chi_k(\boldsymbol{\eta}) = \prod_i \theta_{\lambda_k}^i(\eta_i) \quad (1.21)$$

and the energy is given by the sum of the individual energies of the oscillators. The eigenstates can be written in Cartesian space by applying eq. 1.12:

$$\tilde{\theta}_\lambda^i(\mathbf{Q}) \equiv \theta_\lambda^i(\eta_i) = \theta_\lambda^i\left(\sum_j C_{ji}Q_j\right) \quad (1.22)$$

and similarly for χ_k .

The n_λ number counts how many quanta of vibration are in each oscillator. In the second quantization picture, n_λ is the *occupation* of the phonon λ , and the zero-occupation state represents the ZPM.

For a system at thermal equilibrium at temperature T , it can be proved that n_λ follows the Bose-Einstein statistics:

$$n(\omega, T) = \frac{1}{e^{(\hbar\omega - \mu)/k_B T} - 1} \quad (1.23)$$

where μ is any bias, like the chemical potential.

Nuclear motion in the classical limit

We just derived the quantum description of the nuclear states; however, the nuclear motion is often framed in a classical context, in order to describe more intuitively some concepts, such as polarization and angular momentum of vibrational modes. While at low temperature, the ZPM, which is purely quantum, represents a large contribution to the nuclear motion, at high temperature the atoms oscillate with larger amplitude, and the contribution of the ZPM is much less important. However, very light elements, like H and He, strongly deviate from the classical treatment up to very high temperatures. We will now describe the analysis of the nuclear motion in the classical case, under the assumption of harmonic interactions.

For the nuclear displacement i , the equation of motion reads:

$$\frac{d^2 Q_i(t)}{dt^2} = - \sum_j \Phi_{ij} Q_j \quad (1.24)$$

where Φ_{ij} is the mass-weighted force constant that couples displacements i and j . Note that under the harmonic approximation, we take $\Phi_{ij} = \Phi_{ij}^{(2)}$ (see section 1.1.1).

Since eq. 1.24 is hard to solve due to the coupled degrees of freedom, we make the same change of basis we performed for the quantum case, obtaining a set of uncoupled differential equations of the form:

$$\frac{d^2 \eta_i(t)}{dt^2} = -\omega_i^2 \eta_i(t) \quad (1.25)$$

where again ω_i^2 is the respective eigenvalue of the weighted IFC matrix. As before, the eigenvectors are $\boldsymbol{\eta}_i = (C_{1i}, C_{2i}, \dots, C_{3Mi})$, according to eq. 1.12.

The equations admit solutions of the form:

$$\eta_i(t) = A_i^+ e^{i(\omega_i t + \phi_i^+)} + A_i^- e^{-i(\omega_i t + \phi_i^-)} \quad (1.26)$$

where A_i^+ and A_i^- are the real amplitudes, and ϕ_i^+ and ϕ_i^- are arbitrary phase factors. By imposing real displacement, we obtain that $A_i^+ = A_i^- \equiv \frac{1}{2}A_i$ and $\phi_i^+ = \phi_i^- \equiv \phi_i$, leading to:

$$\eta_i(t) = A_i \cos(\omega_i t + \phi_i). \quad (1.27)$$

The nuclear displacement is obtained through the eigenvectors:

$$Q_i(t) = \sum_j C_{ij} \eta_j(t) \quad (1.28)$$

Note that nuclear displacements move with the same phases $\{\phi_i\}$.

For a single classical oscillator the energy is given by:

$$E_i = \frac{1}{2} \dot{\eta}_i^2 + \frac{1}{2} \omega_i^2 \eta_i^2 \quad (1.29)$$

and by plugging the solution for η_i , we obtain:

$$E_i = \frac{1}{2} \omega_i^2 A_i^2 \quad (1.30)$$

By the equipartition theorem, at thermal equilibrium at temperature T , the energy of a single oscillator is equal to $k_B T$, meaning that the thermal average of the squared amplitude is related to the temperature by:

$$\langle A_i^2 \rangle = \frac{2k_B T}{\omega_i^2} \quad (1.31)$$

Note that the classical case fails to include the ZPM, as $\langle A_i^2 \rangle = 0$ at $T = 0$. However, it recovers the high-temperature limit of the quantum occupation.

1.1.2 Periodic systems

Electrons

We start by defining the *lattice vectors*:

$$\begin{aligned} \mathbf{a}_1 &= a_{x1} \mathbf{e}_x + a_{y1} \mathbf{e}_y + a_{z1} \mathbf{e}_z \\ \mathbf{a}_2 &= a_{x2} \mathbf{e}_x + a_{y2} \mathbf{e}_y + a_{z2} \mathbf{e}_z \\ \mathbf{a}_3 &= a_{x3} \mathbf{e}_x + a_{y3} \mathbf{e}_y + a_{z3} \mathbf{e}_z \end{aligned} \quad (1.32)$$

where $\mathbf{e}_x$, $\mathbf{e}_y$ and $\mathbf{e}_z$ are the Cartesian unit vectors.

The *lattice* is the set of points $\{\mathbf{P}\}$ in space of the form:

$$\mathbf{P} = n_1 \mathbf{a}_1 + n_2 \mathbf{a}_2 + n_3 \mathbf{a}_3$$

with $n_1, n_2, n_3 \in \mathbb{Z}$.

A periodic system of atoms can be described as a unit of atoms repeating with a certain periodicity. The repeating unit is called *unit cell* and can be uniquely identified with a lattice and a *basis*. The basis specifies, for each atom in the repeating unit, its species and its position in the unit cell. Conventionally, the unit cell is chosen as the parallelepiped generated by the lattice parameters.

For a given lattice, its *reciprocal lattice* is given by the *reciprocal lattice parameters*:

$$\begin{aligned} \mathbf{b}_1 &= \frac{2\pi}{V} (\mathbf{a}_2 \times \mathbf{a}_3) \\ \mathbf{b}_2 &= \frac{2\pi}{V} (\mathbf{a}_3 \times \mathbf{a}_1) \\ \mathbf{b}_3 &= \frac{2\pi}{V} (\mathbf{a}_1 \times \mathbf{a}_2) \end{aligned} \quad (1.33)$$

where V is the volume of the conventional unit cell $\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)$. The reciprocal lattice vectors are of the form:

$$\mathbf{G} = m_1 \mathbf{b}_1 + m_2 \mathbf{b}_2 + m_3 \mathbf{b}_3$$

with $m_1, m_2, m_3 \in \mathbb{Z}$. Note that $\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi \delta_{ij}$.

We define the first *Brillouin zone* (BZ) as the Wigner-Seitz cell of the reciprocal lattice centered at the origin.

According to *Bloch's theorem*, in the presence of a periodic potential (such as nuclear field in crystals), the solutions of the Schrödinger equation must be a periodic function modulated by a phase factor. For one particle with position $\mathbf{x}$:

$$\psi(\mathbf{x}) = e^{i\mathbf{k} \cdot \mathbf{x}} u(\mathbf{x}) \quad (1.34)$$

where $\mathbf{k}$ is the *wavevector*, and $u(\mathbf{x})$ is a periodic function with the periodicity of the lattice. Each solution has its own periodic function and wavevector. Note that we can always take $\mathbf{k}$ within the first BZ (*folding*) and incorporate any additional phase factor $e^{i\mathbf{G} \cdot \mathbf{x}}$ into $u(\mathbf{x})$ without changing its periodicity:

$$e^{i(\mathbf{k}+\mathbf{G}) \cdot \mathbf{x}} u(\mathbf{x}) = e^{i\mathbf{k} \cdot \mathbf{x}} [e^{i\mathbf{G} \cdot \mathbf{x}} u(\mathbf{x})] = e^{i\mathbf{k} \cdot \mathbf{x}} \tilde{u}(\mathbf{x})$$

For n particles, Bloch's theorem states:

$$\psi(\mathbf{x}_1, \mathbf{x}_2, \dots) = e^{i \sum_l^n \mathbf{k}_l \cdot \mathbf{x}_l} u(\mathbf{x}_1, \mathbf{x}_2, \dots) \quad (1.35)$$

where $\sum_l \mathbf{k}_l$ is the *total crystal momentum*.

Let us keep the discussion restricted to the one-particle cases, for the sake of clarity.

We approximate finite crystals having N_1, N_2, N_3 repetitions of the unit cell along the three lattice vector directions, by applying the Born-von Karman periodic boundary conditions:

$$\psi(\mathbf{x} + \mathbf{N}) = \psi(\mathbf{x})$$

with $\mathbf{N} = \lambda_1 \mathbf{a}_1 + \lambda_2 \mathbf{a}_2 + \lambda_3 \mathbf{a}_3$ and $\lambda_i \in \{0, N_i\}$. Due to translational symmetry, the components of $\mathbf{k}$ can only assume some discrete values. It can be proved that the set of relevant wavevectors is:

$$\left\{ \frac{g_1}{N_1} \mathbf{b}_1 + \frac{g_2}{N_2} \mathbf{b}_2 + \frac{g_3}{N_3} \mathbf{b}_3 \right\} \quad (1.36)$$

with g_i integer in $\left[-\frac{N_i}{2}, \frac{N_i}{2}\right)$.

Now let us apply the Hamiltonian to the generic Bloch state:

$$\left[-\frac{\hbar^2}{2m} \hat{\nabla}^2 + U(\mathbf{x}) \right] e^{i\mathbf{k} \cdot \mathbf{x}} u(\mathbf{x}) \quad (1.37)$$

with some basic algebra, we realize that the phase can cancel out, and we are left with an equation for $u(\mathbf{x})$ which depends on $\mathbf{k}$:

$$\left[\frac{\hbar^2}{2m} \left(-\hat{\nabla}^2 - 2i\mathbf{k} \cdot \hat{\nabla} + |\mathbf{k}|^2 \right) + U(\mathbf{x}) \right] u(\mathbf{x}) = E u(\mathbf{x}) \quad (1.38)$$

In other words, in order to obtain the eigenstates, we can simply find the solutions to eq. 1.38 for every possible $\mathbf{k}$. The eigenstates can then be grouped according to their wavevector and ordered by energy. The ordering is noted with the *band index*, i . The states can then be written as:

$$\psi_{\mathbf{k},i}(\mathbf{x}) = e^{i\mathbf{k} \cdot \mathbf{x}} u_{\mathbf{k},i}(\mathbf{x}). \quad (1.39)$$

Nuclei

The treatment of nuclei in a periodic system is similar to that in the non-periodic case. We apply the same periodic boundary conditions as we did for electrons. Unlike for electrons, though, the nuclear potential is not periodic, so we cannot apply Bloch's theorem to the nuclear wavefunction. However, the interaction is periodic.

Let us denote the position of the nucleus i in the cell replica l with:

$$\mathbf{R}_{iL} = \mathbf{P}_L + \mathbf{x}_i \quad (1.40)$$

where $\mathbf{P}_L$ is the origin of the cell the atom is located in, and $\mathbf{x}_i$ is its position in that cell.

We can write the harmonic potential energy surface as a function of the mass-weighted coordinates:

$$U(\mathbf{Q}) \approx U(0) + \frac{1}{2} \sum_{LKij\alpha\beta} \Phi_{Li\alpha,Kj\beta} Q_{Li\alpha} Q_{Kj\beta} \quad (1.41)$$

The pairwise (quadratic) interaction is translationally invariant:

$$\Phi_{Li\alpha,Kj\beta} = \Phi_{(L-L')i\alpha,(K-L')j\beta} \quad (1.42)$$

where α and β run over Cartesian directions, and $L' \in \mathbb{Z}$.

Note that the matrix is now written on the canonical basis $\{|L, i\rangle\}$ where $|L, i\rangle$ corresponds to displacement i in the cell L by one unit.

For the sake of clarity, let us order the matrix elements in this way:

$$\Phi = \begin{pmatrix} [\Phi_{1,1}] & [\Phi_{1,2}] & \cdots & [\Phi_{1,N}] \\ [\Phi_{2,1}] & [\Phi_{2,2}] & \cdots & [\Phi_{2,N}] \\ \vdots & \vdots & \ddots & \vdots \\ [\Phi_{N,1}] & [\Phi_{N,2}] & \cdots & [\Phi_{N,N}] \end{pmatrix} \quad (1.43)$$

where N is the total number of repetitions of the unit cell in the box ($N = N_1 \cdot N_2 \cdot N_3$) $[\Phi_{L,K}]$ is a $3M \times 3M$ matrix that contains the IFCs

between only atoms in the cells L and K .

$$[\Phi_{L,K}] = \begin{pmatrix} \Phi_{L1x,K1x} & \Phi_{L1x,K1y} & \Phi_{L1x,K1z} & \Phi_{L1x,K2x} & \cdots & \Phi_{L1x,KMz} \\ \Phi_{L1y,K1x} & \Phi_{L1y,K1y} & \Phi_{L1y,K1z} & \Phi_{L1y,K2x} & \cdots & \Phi_{L1y,KMz} \\ \Phi_{L1z,K1x} & \Phi_{L1z,K1y} & \Phi_{L1z,K1z} & \Phi_{L1z,K2x} & \cdots & \Phi_{L1z,KMz} \\ \Phi_{L2x,K1x} & \Phi_{L2x,K1y} & \Phi_{L2x,K1z} & \Phi_{L2x,K2x} & \cdots & \Phi_{L2x,KMz} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \Phi_{LMz,K1x} & \Phi_{LMz,K1y} & \Phi_{LMz,K1z} & \Phi_{LMz,K2x} & \cdots & \Phi_{LMz,KMz} \end{pmatrix} \quad (1.44)$$

From now on, we will incorporate the Cartesian index α within i .

Let us define a new basis set, $\{|\mathbf{q}, i\rangle\}$, where $|\mathbf{q}, i\rangle$ represents the displacement i in *every* cell, by a fraction of unit specific for its cell, and proportional to $e^{i\mathbf{q}\cdot\mathbf{P}_L}$:

$$|\mathbf{q}, i\rangle = \frac{1}{\sqrt{N}} \sum_{L=1}^N e^{i\mathbf{q}\cdot\mathbf{P}_L} |L, i\rangle \quad (1.45)$$

for a set of N vectors¹ $\mathbf{q}$.

If we apply the same basis transformation to Φ we obtain the dynamical matrix:

$$\mathbf{D} = \mathbf{U}^\dagger \Phi \mathbf{U} \quad (1.46)$$

where

$$U_{(Li),(\mathbf{q}j)} = \frac{1}{\sqrt{N}} e^{i\mathbf{q}\cdot\mathbf{P}_L} \delta_{ij} \quad (1.47)$$

If we choose the set of $\mathbf{q}$ vectors to be the points in eq. 1.36, because of translational symmetry and Born von Karman boundary conditions, $\mathbf{D}$ is block-diagonal, and each block is associated with a $\mathbf{q}$ vector.

Transforming the mass-weighted displacements accordingly, we obtain an intermediate basis:

$$\{Q_{1,1}, Q_{1,2}, \dots, Q_{N,3M}\} \rightarrow \{\tilde{Q}_1^{\mathbf{q}_1}, \dots, \tilde{Q}_{3M}^{\mathbf{q}_1}, \dots, \tilde{Q}_1^{\mathbf{q}_N}, \dots, \tilde{Q}_{3M}^{\mathbf{q}_N}\} \quad (1.48)$$

$$\begin{aligned} \tilde{Q}_j^{\mathbf{q}_n} &= \sum_{L=1, i=1}^{N, 3M} U_{(Li),(\mathbf{q}_n, j)}^* Q_{Li} \\ &= \frac{1}{\sqrt{N}} \sum_{L=1}^N e^{-i\mathbf{q}_n \cdot \mathbf{P}_L} Q_{Lj} \end{aligned} \quad (1.49)$$

Finally, we can diagonalize the blocks of $\mathbf{D}$ individually, obtaining a new basis set (the eigenvectors), namely the *phonon modes*, and the squared vibrational frequencies (eigenvalues). Let us denote by $\mathbf{D}(\mathbf{q})$ the block of

1: We already defined N as the number of repetition of the unit cell in the box.

dynamical matrix associated with $\mathbf{q}$. We diagonalize $\mathbf{D}(\mathbf{q})$:

$$\{\tilde{Q}_1^{\mathbf{q}}, \dots, \tilde{Q}_{3M}^{\mathbf{q}}\} \rightarrow \{\eta_1^{\mathbf{q}}, \dots, \eta_{3M}^{\mathbf{q}}\} \quad (1.50)$$

$$\eta_i^{\mathbf{q}} = \sum_{j=1}^{3M} (X_{ji}^{\mathbf{q}})^* \tilde{Q}_j^{\mathbf{q}} \quad (1.51)$$

$$= \frac{1}{\sqrt{N}} \sum_{j=1}^{3M} \left[(X_{ji}^{\mathbf{q}})^* \sum_{L=1}^N e^{-i\mathbf{q} \cdot \mathbf{P}_L} Q_{Lj} \right] \quad (1.52)$$

where $\mathbf{X}^\dagger(\mathbf{q}) \mathbf{D}(\mathbf{q}) \mathbf{X}(\mathbf{q}) = \mathbf{\Lambda}(\mathbf{q})$, and $\mathbf{\Lambda}(\mathbf{q})$ is the matrix of eigenvalues. Note that the i -th column of $\mathbf{X}(\mathbf{q})$, namely the i -th phonon eigenvector, represents the polarization of the nuclear motion for the i -th phonon mode. We denote the phonon eigenvector as $\eta_i^{\mathbf{q}} = (X_{1,i}^{\mathbf{q}}, \dots, X_{3M,i}^{\mathbf{q}})$, for convenience.

By adopting the new coordinates, the Hamiltonian becomes easy to attack. For each $\mathbf{q}$ wavevector, there are $3M$ uncoupled harmonic oscillators. They can be solved analogously to the non-periodic case, with both quantum and classical operators, yielding wavefunctions and trajectories, respectively. For the mode i with wavevector $\mathbf{q}$, the quantum solutions are identical to eq. 1.18, except for some indices:

$$\begin{aligned} \theta_{im}^{\mathbf{q}}(\eta_i^{\mathbf{q}}) &= N_{im}^{\mathbf{q}} \cdot H_m \left(\sqrt{\frac{\omega_i(\mathbf{q})}{\hbar}} \eta_i^{\mathbf{q}} \right) \cdot e^{-\frac{\omega_i(\mathbf{q})}{2\hbar} (\eta_i^{\mathbf{q}})^2} \\ N_{im}^{\mathbf{q}} &= \frac{1}{\sqrt{2^m m!}} \left(\frac{\omega_i(\mathbf{q})}{\pi \hbar} \right)^{1/4} \end{aligned} \quad (1.53)$$

The considerations on the statistical occupation factor seen for the non-periodic case are valid here as well.

The classical solutions are similar to eq. 1.27:

$$\eta_i^{\mathbf{q}}(t) = A_i(\mathbf{q}) e^{-i(\omega_i(\mathbf{q})t + \phi_i^{\mathbf{q}})} \quad (1.54)$$

where we preserve a single exponential solution to preserve the meaning of $\mathbf{q}$ as wavevector, and the nuclear displacement is obtained as:

$$Q_{Li}(t) = \frac{1}{\sqrt{N}} \sum_{\mathbf{q}_n} \sum_{j=1}^{3M} X_{ij}(\mathbf{q}_n) e^{i\mathbf{q}_n \cdot \mathbf{P}_L} \eta_j^{\mathbf{q}_n}(t) \quad (1.55)$$

Presuming time-reversal symmetry of the crystal lattice, the eigenvectors

and frequencies for opposite wavevectors satisfy $\mathbf{X}(-\mathbf{q}_n) = \mathbf{X}^*(\mathbf{q}_n)$ and $\omega_j(-\mathbf{q}_n) = \omega_j(\mathbf{q}_n)$. By grouping the $+\mathbf{q}_n$ and $-\mathbf{q}_n$ terms together, the imaginary components perfectly cancel, guaranteeing that the physical displacement is strictly real:

$$Q_{Li}(t) = \frac{2}{\sqrt{N}} \sum_{\mathbf{q}_n \in \text{HBZ}^+} \sum_{j=1}^{3M} A_j(\mathbf{q}_n) |X_{ij}(\mathbf{q}_n)| \cos(\mathbf{q}_n \cdot \mathbf{P}_L - \omega_j(\mathbf{q}_n)t - \phi_j^{\mathbf{q}_n} + \phi_{ij}^{\mathbf{q}_n}) \quad (1.56)$$

where we wrote the complex eigenvector in its polar form $X_{ij}(\mathbf{q}_n) = |X_{ij}(\mathbf{q}_n)| e^{i\phi_{ij}^{\mathbf{q}_n}}$. Again, $A_j(\mathbf{q}_n)$ is the amplitude of the oscillation, which is related to the occupation of the phonon mode, and $\phi_j^{\mathbf{q}_n}$ is an arbitrary initial phase. Note that the sum over $\mathbf{q}$ is only performed in one half of the Brillouin Zone.

From the reality of the atomic displacements, $Q_{Li} = (Q_{Li})^*$, it can be proved through eq. 1.55 that $(\eta_j^{\mathbf{q}_n})^* = \eta_j^{-\mathbf{q}_n}$.

Non-analyticity at Γ (LO-TO splitting)

In polar crystals, the relative displacement of oppositely charged atoms during an optical vibration induces a macroscopic dipole moment [16]. For phonons near the center of the Brillouin zone ($\mathbf{q} \rightarrow 0$), this local dipole density acts as a macroscopic polarization $\mathbf{P}$. The interaction between this polarization and the macroscopic electric field $\mathbf{E}$ it generates is responsible for lifting the degeneracy of the optical modes, a phenomenon known as *Longitudinal-Transverse Optical (LO-TO) splitting*.

Because we are in the electrostatic limit (the phonon phase velocity is much slower than the speed of light), Maxwell's equations dictate that the curl of the macroscopic electric field is zero ($\nabla \times \mathbf{F} = 0$), meaning the field must be entirely longitudinal.

Consequently, the induced electric field depends strictly on the angle between the polarization vector $\mathbf{P}$ and the phonon wavevector $\mathbf{q}$:

- **Transverse Optical (TO) Modes:** The atomic displacements, and hence the macroscopic polarization $\mathbf{P}$, are perpendicular to the propagation direction $\mathbf{q}$ ($\mathbf{q} \cdot \mathbf{P} = 0$): no macroscopic restoring electric field is generated ($\mathbf{F} = 0$).
- **Longitudinal Optical (LO) Modes:** The atomic displacements are parallel to the propagation direction $\mathbf{q}$ ($\mathbf{q} \parallel \mathbf{P}$). This creates a divergence

of the polarization field, generating a macroscopic electric field $\mathbf{F}$ that opposes the atomic motion.

This induced electric field exerts an additional restoring force on the ions. Because the restoring force is stronger for the longitudinal motion, the LO phonon requires more energy to excite, shifting its frequency higher than that of the TO phonon ($\omega_{LO} > \omega_{TO}$).

Mathematically, this long-range Coulomb interaction introduces a non-analytic term to the dynamical matrix near Γ . The total dynamical matrix is written as the sum of the standard short-range analytical part and the macroscopic non-analytical correction:

$$\mathbf{D}_{ij}(\mathbf{q} \rightarrow 0) = \mathbf{D}_{ij}(\mathbf{q}) + \mathbf{D}_{ij}^{\text{NA}}(\mathbf{q}) \quad (1.57)$$

The non-analytical correction depends strictly on the direction of $\mathbf{q}$, not its magnitude, and is governed by the Born effective charge tensor Z^* (which couples displacements to the polarization) and the high-frequency dielectric tensor ε^∞ (which screens the electric field). We temporarily restore explicitly $i \rightarrow i\alpha$ and $j \rightarrow j\beta$, and write the correction as:

$$\mathbf{D}_{i\alpha,j\beta}^{\text{NA}}(\mathbf{q}) = \frac{1}{\Omega \varepsilon_0 \sqrt{M_i M_j}} \frac{\left(\sum_\gamma q_\gamma Z_{i,\gamma\alpha}^* \right) \left(\sum_\delta q_\delta Z_{j,\delta\beta}^* \right)}{\sum_{\mu\nu} q_\mu \varepsilon_{\mu\nu}^\infty q_\nu} \quad (1.58)$$

where Ω is the volume of the unit cell, ε_0 is the vacuum permittivity, and $Z_{i,\alpha\beta}^* = \left. \frac{\partial^2 U}{\partial F_\alpha \partial Q_{i\beta}} \right|_0$. Because of the $q_\gamma q_\delta$ terms in the numerator, the eigenvalue solutions to this matrix will change depending on the exact direction from which $\mathbf{q}$ approaches zero, and LO and TO phonon branches will no longer be degenerate.

Second quantization

In order to simplify the description of vibrational states, and leave behind cumbersome explicit Hermite polynomials, we can resort to second quantization. Before starting, we note that the squares of quantum operators are actually a particular case of products of adjoints, which is the case of Hermitian operators, namely operators associated with (real) observables. In order to properly derive what follows, we need to adopt the rigorous notation $\hat{A}^\dagger \hat{A}$ instead of $\hat{A}^2$.

We start by pairing terms of the Hamiltonian with opposite wavevector:

$$\hat{H} = \sum_{\lambda}^{BZN^+} \hat{H}_{\text{pair}}^{\lambda}$$

$$H_{\text{pair}}^{\lambda} = \left(\frac{1}{2} \hat{\pi}_{\lambda}^{\dagger} \hat{\pi}_{\lambda} + \frac{1}{2} \omega_{\lambda}^2 \hat{\eta}_{\lambda}^{\dagger} \hat{\eta}_{\lambda} \right) + \left(\frac{1}{2} \hat{\pi}_{-\lambda}^{\dagger} \hat{\pi}_{-\lambda} + \frac{1}{2} \omega_{-\lambda}^2 \hat{\eta}_{-\lambda}^{\dagger} \hat{\eta}_{-\lambda} \right) \quad (1.59)$$

where we recall λ being an index which identifies the phonon mode, including both the wavevector and the branch index.

Analogously to the classical case, it can be proven that

$$\hat{\eta}_{\lambda}^{\dagger} = \hat{\eta}_{-\lambda} \quad (1.60)$$

$$\hat{\pi}_{\lambda}^{\dagger} = \hat{\pi}_{-\lambda} \quad (1.61)$$

$$\omega_{\lambda} = \omega_{-\lambda} \quad (1.62)$$

$$[\hat{\eta}_{\lambda}, \hat{\pi}_{\lambda'}] = i\hbar \delta_{\lambda, -\lambda'} \quad (1.63)$$

Then, the Hamiltonian simplifies:

$$H_{\text{pair}}^{\lambda} = \hat{\pi}_{-\lambda} \hat{\pi}_{\lambda} + \omega_{\lambda}^2 \hat{\eta}_{-\lambda} \hat{\eta}_{\lambda} \quad (1.64)$$

where we also used the commutation rule $[\hat{\eta}_{\lambda}^{\dagger}, \hat{\eta}_{\lambda}] = 0$.

Now we can define the “annihilation” and “creation” operators, also known as “ladder operators”:

$$\hat{a}_{\lambda} = \sqrt{\frac{\omega_{\lambda}}{2\hbar}} \hat{\eta}_{\lambda} + i \frac{1}{\sqrt{2\hbar\omega_{\lambda}}} \hat{\pi}_{\lambda} \quad (1.65)$$

$$\hat{a}_{\lambda}^{\dagger} = \sqrt{\frac{\omega_{\lambda}}{2\hbar}} \hat{\eta}_{\lambda}^{\dagger} - i \frac{1}{\sqrt{2\hbar\omega_{\lambda}}} \hat{\pi}_{\lambda}^{\dagger} \quad (1.66)$$

It is easy to calculate the commutation relations:

$$[\hat{a}_{\lambda}, \hat{a}_{\lambda'}^{\dagger}] = \delta_{\lambda, \lambda'} \quad (1.67)$$

$$[\hat{a}_{\lambda}, \hat{a}_{\lambda'}] = 0 \quad (1.68)$$

Note that, due to eqs. 1.60, 1.61 and 1.62:

$$\hat{\eta}_{\lambda} = \sqrt{\frac{\hbar}{2\omega_{\lambda}}} \left(\hat{a}_{\lambda} + \hat{a}_{-\lambda}^{\dagger} \right) \quad (1.69)$$

$$\hat{\pi}_{\lambda} = -i\sqrt{\frac{\hbar\omega_{\lambda}}{2}} \left(\hat{a}_{\lambda} - \hat{a}_{-\lambda}^{\dagger} \right) \quad (1.70)$$

Now we can show how ladder operators can be used to write the Hamiltonian.

If we multiply them together we obtain:

$$\hat{a}_\lambda^\dagger \hat{a}_\lambda = \frac{\omega_\lambda}{2\hbar} \hat{\eta}_{-\lambda} \hat{\eta}_\lambda + \frac{1}{2\hbar\omega_\lambda} \hat{\pi}_{-\lambda} \hat{\pi}_\lambda + \frac{i}{2\hbar} (\hat{\eta}_{-\lambda} \hat{\pi}_\lambda - \hat{\pi}_{-\lambda} \hat{\eta}_\lambda) \quad (1.71)$$

then, by summing this term for λ and $-\lambda$:

$$\hat{a}_\lambda^\dagger \hat{a}_\lambda + \hat{a}_{-\lambda}^\dagger \hat{a}_{-\lambda} = \frac{1}{\hbar\omega_\lambda} (\hat{\pi}_{-\lambda} \hat{\pi}_\lambda + \omega_\lambda^2 \hat{\eta}_{-\lambda} \hat{\eta}_\lambda) + \frac{i}{2\hbar} ([\hat{\eta}_{-\lambda}, \hat{\pi}_\lambda] + [\hat{\eta}_\lambda, \hat{\pi}_{-\lambda}]) \quad (1.72)$$

In the first term we can recognize $\hat{H}_{\text{pair}}^\lambda$, while for the second one we can use (1.63):

therefore,

$$\begin{aligned} \hat{a}_\lambda^\dagger \hat{a}_\lambda + \hat{a}_{-\lambda}^\dagger \hat{a}_{-\lambda} &= \frac{1}{\hbar\omega_\lambda} \hat{H}_{\text{pair}}^\lambda - 1 \\ \Leftrightarrow \hat{H}_{\text{pair}}^\lambda &= \hbar\omega_\lambda \left(\hat{a}_\lambda^\dagger \hat{a}_\lambda + \frac{1}{2} \right) + \hbar\omega_\lambda \left(\hat{a}_{-\lambda}^\dagger \hat{a}_{-\lambda} + \frac{1}{2} \right) \end{aligned} \quad (1.73)$$

where we recovered the usual Hamiltonians for uncoupled harmonic oscillators in second quantization formalism.

We can then obtain some useful commutation relations:

$$[\hat{H}, \hat{a}_\lambda^\dagger] = \hbar\omega_\lambda \hat{a}_\lambda^\dagger \quad (1.74)$$

$$[\hat{H}, \hat{a}_\lambda] = -\hbar\omega_\lambda \hat{a}_\lambda \quad (1.75)$$

where $\hat{H} = \sum_\lambda^{BZN^+} \hat{H}_{\text{pair}}^\lambda$.

The commutators eq. 1.74 and eq. 1.75 are extremely important in interpreting the action of the ladder operators. Indeed, by acting $\hat{a}_\lambda^\dagger$ on an eigenstate of $\hat{H}$, $|\psi_i\rangle$, with eigenvalue E_i , we generate a new eigenstate of $\hat{H}$, $|\tilde{\psi}\rangle$ with eigenvalue $E + \hbar\omega_\lambda$:

$$\underbrace{\hat{H} \hat{a}_\lambda^\dagger |\psi_i\rangle}_{|\tilde{\psi}\rangle} = \hat{a}_\lambda^\dagger \hat{H} |\psi_i\rangle + \hbar\omega_\lambda \hat{a}_\lambda^\dagger |\psi_i\rangle = (E_i + \hbar\omega_\lambda) \underbrace{\hat{a}_\lambda^\dagger |\psi_i\rangle}_{|\tilde{\psi}\rangle} \quad (1.76)$$

$$\hat{H} |\tilde{\psi}\rangle = (E_i + \hbar\omega_\lambda) |\tilde{\psi}\rangle \quad (1.77)$$

and similarly with $\hat{a}_\lambda$, which generates eigenstates with eigenvalue $E_i - \hbar\omega_\lambda$. We can now see the reason they are called “ladder” operators: they can be used to generate eigenstates of the vibrational Hamiltonian with increasing/decreasing energy by fixed steps of $\hbar\omega_\lambda$.

There are some boundary conditions to apply. If we take the lowest energy eigenstate of $\hat{H}$, $|0\rangle$, and we try to apply the annihilation operator $\hat{a}_\lambda$, we

expect it to return the 0-energy state, that is the null state:

$$\hat{a}_\lambda |0\rangle = 0 \quad (1.78)$$

Starting from $|0\rangle$ we can get all the other states by applying the “creation” operator $\hat{a}_\lambda^\dagger$. Note that the boundary condition of the annihilation of the physical state tells us something else very important: applying the annihilation operator on $|0\rangle$ and imposing that the result is 0, means that $|0\rangle$ must be the solution of a first-order linear ordinary differential equation, which can be proved to only admit one solution. Therefore, the lowest state is unique. This is enough for us to be sure that the ladder operators generate *all* the eigenstates of $\hat{H}$.

Instead of solving the differential equations of the harmonic oscillators, and keeping their cumbersome functional form (Hermite polynomials) we can simply fix the lowest one and say that we obtain all the other ones by applying the ladder operators.

The Hamiltonian’s eigenstates are now written as elements of the Fock space, by specifying the state of each phonon mode:

$$|n_1, \dots, n_\lambda, \dots, n_{3M \times N}\rangle \quad (1.79)$$

For the sake of simplicity, we will focus on one mode at a time and will consider all the other ones in their lowest state, if not explicitly stated:

$$|n_\lambda\rangle \equiv |0_1, \dots, n_\lambda, \dots, 0_{3M \times N}\rangle \quad (1.80)$$

Similarly, we will use single-mode Hamiltonians $\hat{H}_\lambda = \hbar\omega_\lambda (\hat{a}_\lambda^\dagger \hat{a}_\lambda + \frac{1}{2})$, because of linearity.

Let us now define the “number operator” $\hat{n}_\lambda \equiv \hat{a}_\lambda^\dagger \hat{a}_\lambda$. As it commutes with the Hamiltonian, the set of states $|n_\lambda\rangle$ is a set of eigenstates of $\hat{n}_\lambda$ as well:

$$\hat{n}_\lambda |n_\lambda\rangle = n_\lambda |n_\lambda\rangle \quad (1.81)$$

Note that we denoted from the beginning these states with the letter n because we want to refer to them based on the eigenvalue of $\hat{n}_\lambda$ associated with them.

In order to understand the physical meaning of the number operator and its

eigenvalues, let us apply the Hamiltonian to one of its eigenstates:

$$\hat{H}_\lambda |n_\lambda\rangle = E_{|n_\lambda\rangle} |n_\lambda\rangle = \hbar\omega_\lambda \left(\hat{n}_\lambda + \frac{1}{2} \right) |n_\lambda\rangle = \hbar\omega_\lambda \left(n_\lambda + \frac{1}{2} \right) |n_\lambda\rangle \quad (1.82)$$

$$\Leftrightarrow n_\lambda = \frac{E_{|n_\lambda\rangle}}{\hbar\omega_\lambda} - \frac{1}{2} \quad (1.83)$$

that is the eigenvalue of the number operator counts how many quanta of energy were added to the lowest-energy state (zero-point state or ground state).

Now we can finally fully understand the action of the ladder operators on Hamiltonian's eigenstates. First we need to calculate the following commutators:

$$[\hat{n}_\lambda, \hat{a}_{\lambda'}^\dagger] = \hat{a}_{\lambda'}^\dagger \delta_{\lambda,\lambda'} \quad (1.84)$$

$$[\hat{n}_\lambda, \hat{a}_{\lambda'}] = -\hat{a}_{\lambda'} \delta_{\lambda,\lambda'} \quad (1.85)$$

We can use eq. 1.84 to write:

$$\hat{n}_\lambda \hat{a}_\lambda |n_\lambda\rangle = \hat{a}_\lambda \hat{n}_\lambda |n_\lambda\rangle - \hat{a}_\lambda |n_\lambda\rangle = (n_\lambda - 1) \hat{a}_\lambda |n_\lambda\rangle \quad (1.86)$$

Hence $\hat{a}_\lambda |n_\lambda\rangle$ is still an eigenvector of the number operator, with eigenvalue $n_\lambda - 1$. By the definition in eq. 1.81, we know that $\hat{n}_\lambda |n_\lambda - 1\rangle = (n_\lambda - 1) |n_\lambda - 1\rangle$, therefore $\hat{a}_\lambda |n_\lambda\rangle$ must be proportional to the latter:

$$\hat{a}_\lambda |n_\lambda\rangle = c_- |n_\lambda - 1\rangle \quad (1.87)$$

In order to find the value of the normalization constant, we multiply on the left by the complex conjugate:

$$\begin{aligned} \langle n_\lambda | \hat{a}_\lambda^\dagger \hat{a}_\lambda | n_\lambda \rangle &= \langle n_\lambda | \hat{n}_\lambda | n_\lambda \rangle = n_\lambda \langle n_\lambda | n_\lambda \rangle = |c_-|^2 \langle n_\lambda - 1 | n_\lambda - 1 \rangle \\ \Leftrightarrow n_\lambda &= |c_-|^2 \\ \Leftrightarrow c_- &= \sqrt{n_\lambda} \end{aligned} \quad (1.88)$$

For $\hat{a}_\lambda^\dagger$ we can follow a similar reasoning, leading to $c_+ = \sqrt{n_\lambda + 1}$. Finally:

$$\hat{a}_\lambda |n_\lambda\rangle = \sqrt{n_\lambda} |n_\lambda - 1\rangle \quad (1.89)$$

$$\hat{a}_\lambda^\dagger |n_\lambda\rangle = \sqrt{n_\lambda + 1} |n_\lambda + 1\rangle \quad (1.90)$$

At thermal equilibrium, the expectation value of the number operator is the

Bose-Einstein occupation factor eq. 1.23.

Note that:

$$\langle m_\lambda | \hat{a}_\lambda | n_\lambda \rangle = \sqrt{n_\lambda} \langle m_\lambda | n_\lambda - 1 \rangle = \sqrt{n_\lambda} \delta_{m_\lambda, n_\lambda - 1} \quad (1.91)$$

$$\langle m_\lambda | \hat{a}_\lambda^\dagger | n_\lambda \rangle = \sqrt{n_\lambda + 1} \langle m_\lambda | n_\lambda + 1 \rangle = \sqrt{n_\lambda + 1} \delta_{m_\lambda, n_\lambda + 1} \quad (1.92)$$

This is all we need for the scope of this thesis, but for further study of the second-quantization formalism the reader is referred to textbooks like [17].

1.1.3 Phonon Green's functions

We will now address the phonon Green's functions for a crystal, which can be collapsed to the molecular case by only considering $\mathbf{q} = 0$.

For an uncoupled harmonic oscillator (in the phonon mode basis), we define its *bare Green's function*, as the solution of the equation of motion when an external impulsive force is applied:

$$\ddot{G}_{\mathbf{q}j}^0(t) + \omega_{\mathbf{q}j}^2 G_{\mathbf{q}j}^0(t) = \delta(t) \quad (1.93)$$

where $\mathbf{q}$ and j are the mode wavevector and band index, respectively.

From linear response theory, $G^0(t)$ is proportional to the autocorrelation function of displacement:

$$G_{\mathbf{q}j}^0(t) = -\frac{i}{\hbar} \theta(t) \langle [Q_{\mathbf{q}j}(t), Q_{\mathbf{q}j}(0)] \rangle_0 \quad (1.94)$$

where $\theta(t)$ is the Heaviside step function, and the expectation value is taken on the initial unperturbed state.

The bare Green's function can be easily written in frequency domain as:

$$G_{\mathbf{q}j}^0(\omega) = \frac{1}{\omega_{\mathbf{q}j}^2 - \omega^2}. \quad (1.95)$$

To describe an interacting harmonic oscillator, we include all the interactions in a linearized way with respect to the motion variable. We can do this by defining a quantity called a "self-energy", $\Sigma_{\mathbf{q}j}$.

$$\ddot{G}_{\mathbf{q}j}(t) + \omega_{\mathbf{q}j}^2 G_{\mathbf{q}j}(t) + \int_{-\infty}^t \Sigma_{\mathbf{q}j}(t - t') G_{\mathbf{q}j}(t') dt' = \delta(t) \quad (1.96)$$

The self-energy acts as a memory kernel; in other words, it expresses how the motion at each past time contributes to the present motion (from which the dependence of both time variables is derived). We also assume that $\Sigma_{\mathbf{q}j}$ does not depend on absolute times, but only on time differences. This allows us to see the extra term of the EOM as a convolution (by imposing causality and extending the integral to $+\infty$).

In frequency domain, the *full-interacting Green's function* can be expressed as:

$$G_{\mathbf{q}j}(\omega) = \frac{1}{\omega_{\mathbf{q}j}^2 - \omega^2 - \Sigma_{\mathbf{q}j}(\omega)} \quad (1.97)$$

and we can retrieve the Dyson equation:

$$G_{\mathbf{q}j}(\omega) = G_{\mathbf{q}j}^0(\omega) + G_{\mathbf{q}j}^0(\omega)\Sigma_{\mathbf{q}j}(\omega)G_{\mathbf{q}j}(\omega) \quad (1.98)$$

We define the *spectral function* of a given phonon mode from the imaginary part of the Green's function:

$$A_{\mathbf{q}j}(\omega) = -\frac{1}{\pi} \text{Im}[G_{\mathbf{q}j}(\omega)]. \quad (1.99)$$

It can be proved that

$$A_{\mathbf{q}j}(\omega) = \frac{1}{\pi} \frac{\Gamma_{\mathbf{q}j}(\omega)}{[\omega^2 - \omega_{\mathbf{q}j}^2 - \Delta_{\mathbf{q}j}(\omega)]^2 + \Gamma_{\mathbf{q}j}^2(\omega)} \quad (1.100)$$

where we defined $\Sigma_{\mathbf{q}j} = \Delta_{\mathbf{q}j} + i\Gamma_{\mathbf{q}j}$.

The self-energy components can be derived from many-body perturbation theory. For the sake of clarity, we momentarily map the wavevector and band indices to a single phonon index $(\mathbf{q}ij) \rightarrow \lambda_i$; $(-\mathbf{q}ij) \rightarrow -\lambda_i$. At the first-order of perturbation, the imaginary part of the self energy due to three-phonon scattering events, $\Gamma_{\lambda}^{(3)}(\omega)$, reads [18]:

$$\Gamma_{\lambda}^{(3)}(\omega) = \frac{\pi\hbar}{4} \sum_{ij} |\tilde{\Phi}_{\lambda, -\lambda_i, -\lambda_j}|^2 \left[\frac{1}{2} \Omega_{decay}^{(i,j)} + \Omega_{abs}^{(i,j)} \right] \quad (1.101)$$

where $\tilde{\Phi}_{\lambda\lambda_1\lambda_2}$ is the third-order IFC matrix written in the phonon mode basis:

$$\begin{aligned} \tilde{\Phi}_{\lambda\lambda_1\lambda_2} &= \sum_{L_1L_2L_3}^N \sum_{ijk}^{3M} \frac{X_{i\lambda}X_{j\lambda_1}X_{k\lambda_2}}{\sqrt{\omega_{\lambda}\omega_{\lambda_1}\omega_{\lambda_2}}} \Phi_{L_1i,L_2j,L_3k}^{(3)} e^{i(\mathbf{q}\cdot\mathbf{P}_{L_1}+\mathbf{q}_1\cdot\mathbf{P}_{L_2}+\mathbf{q}_2\cdot\mathbf{P}_{L_3})} \\ \Phi_{L_1i,L_2j,L_3k}^{(3)} &= \left. \frac{\partial^3 U}{\partial u_{L_1i} \partial u_{L_2j} \partial u_{L_3k}} \right|_0 \frac{1}{\sqrt{m_i m_j m_k}} \end{aligned} \quad (1.102)$$

Again i, j and k represent an atomic displacement along a certain direction, and the respective masses, m , are the masses of the atoms being displaced. The indices L_1, L_2 and L_3 run over the cell repetitions, and X are the phonon eigenvectors. Finally $\mathbf{q}, \mathbf{q}_1$ and $\mathbf{q}_2$ are the wave vectors of the modes λ, λ_1 and λ_2 , respectively.

The statistical factors are:

$$\begin{aligned}\Omega_{decay}^{(i,j)} &= (1 + n_i + n_j) \delta(\omega - \omega_i - \omega_j) \\ \Omega_{abs}^{(i,j)} &= (n_i - n_j) \delta(\omega + \omega_i - \omega_j)\end{aligned}\quad (1.103)$$

where n_i is the thermal occupation of phonon i , $n_i \equiv n(\omega_i, T)$. An analogous expression can be written for the fourth-order processes [18]:

$$\Gamma_{\lambda}^{(4)}(\omega) = \frac{\pi\hbar}{8} \sum_{ijk} |\tilde{\Phi}_{\lambda, -\lambda_i, -\lambda_j, -\lambda_k}|^2 \left[\Omega_{split}^{(i,j,k)} + 3\Omega_{redis}^{(i,j,k)} + 3\Omega_{coal}^{(i,j,k)} \right] \quad (1.104)$$

with the statistical factors accounting for splitting, redistribution, and coalescence of four phonons:

$$\begin{aligned}\Omega_{split}^{(i,j,k)} &= [(1 + n_i)(1 + n_j)(1 + n_k) - n_i n_j n_k] \delta(\omega - \omega_i - \omega_j - \omega_k) \\ \Omega_{redis}^{(i,j,k)} &= [n_i(1 + n_j)(1 + n_k) - (1 + n_i)n_j n_k] \delta(\omega + \omega_i - \omega_j - \omega_k) \\ \Omega_{coal}^{(i,j,k)} &= [n_i n_j(1 + n_k) - (1 + n_i)(1 + n_j)n_k] \delta(\omega + \omega_i + \omega_j - \omega_k)\end{aligned}\quad (1.105)$$

Note that we are working in the first order of perturbation (both scattering-orders happen just once), second order in the diagrammatic expansion (only two vertices, namely phonons must eventually recombine into a single one), and third and fourth order, respectively, in the potential expansion.

Isotopic mass disorder in the lattice breaks the periodicity, but it can usually be treated as a small perturbation to the perfect lattice, which acts as a source of scattering for the phonon waves. In a perturbative fashion, we use the isotopic average mass for the displacement operators, and introduce a correction term in the self-energy to account for mass disorder. We only focus on the imaginary part of the correction, as the real part is usually negligible. According to Tamura's model[19], the isotopic correction $\Gamma_{\lambda}^{(iso)}$ is given by:

$$\Gamma_{\lambda}^{(iso)}(\omega) = \frac{4}{\pi} \omega_{\lambda}^2 \sum_i \left[\delta(\omega - \omega_i) \sum_I^{3M} g_I |\boldsymbol{\eta}_{\lambda,I}^{\dagger} \cdot \boldsymbol{\eta}_{\lambda_i,I}|^2 \right] \quad (1.106)$$

where we unpacked the atomic displacement index j into atomic and Cartesian indices $j \rightarrow (J, \alpha)$, and $\eta_{ij} \rightarrow \eta_{i,J,\alpha}$ are the phonon eigenvectors defined in section 1.1.2. We also introduced the mass variance parameter g_I for atom I , which reads:

$$g_I = \sum_{\kappa} f_I^{\kappa} \left(1 - \frac{M_I^{\kappa}}{\bar{M}_I} \right)^2 \quad (1.107)$$

where f_I^{κ} , M_I^{κ} are the fractional concentration and the mass of isotope κ , respectively, and $\bar{M}_I = \sum_{\kappa} f_I^{\kappa} M_I^{\kappa}$.

The imaginary part of the self-energy is then obtained as the sum of all the desired orders and corrections:

$$\Gamma_{\lambda}(\omega) \approx \Gamma_{\lambda}^{(3)}(\omega) + \Gamma_{\lambda}^{(4)}(\omega) + \Gamma_{\lambda}^{iso}(\omega) \quad (1.108)$$

and the real part is computed via the Kramers-Kronig relations:

$$\Delta_{\lambda}(\omega) = \frac{1}{\pi} \mathcal{P} \left\{ \int_{-\infty}^{+\infty} \frac{\Gamma_{\lambda}(\omega')}{\omega' - \omega} d\omega' \right\} \quad (1.109)$$

where $\mathcal{P}\{f\}$ is the Cauchy principal value of f .

1.2 Raman spectroscopy

When the crystal is illuminated with visible light, photons undergo several processes, mostly involving electrons, because the differences between certain electronic states lie in the visible range, while vibrational states differ by significantly smaller energy quanta, in the infrared range. Nonetheless, part of the incident radiation gets scattered with a change in the momentum which is comparable to the energy of vibrational excitations. This is known as *Raman scattering*, and can lead to decrease or increase of the photon frequency, denoted by “Stokes” and “anti-Stokes” Raman scattering, respectively.

The Raman effect can be explained through the perturbation of the polarizability of the system due to the atomic vibrations. The quantitative description of the Raman cross section can be achieved by adopting a quantum approach. The spectral radiation energy emitted by a dipole oscillating at frequency ω per unit of time and solid angle is given by [Cardona2010]:

$$\frac{\partial^2 W_s(\omega)}{\partial \Omega \partial \omega} = \frac{\omega^4}{(4\pi)^2 \epsilon_0 c^3} \left\langle |\mathbf{e}_s \cdot \hat{\mathbf{P}}(\omega)|^2 \right\rangle \quad (1.110)$$

where c is the speed of light in vacuum, ε_0 is the permittivity of vacuum, ω is the angular frequency of oscillation of the dipole, and $\mathbf{e}_s$ is the observation direction.

In order to rewrite $\hat{\mathbf{P}}$ extensively, we need to introduce the polarizability operator as an expansion of the displacement operators:

$$\begin{aligned}
 \hat{\chi}_{\alpha\beta} &\approx \chi_{\alpha\beta}^0 \hat{I} + \sum_{Li} \left. \frac{\partial \chi_{\alpha\beta}}{\partial Q_{Li}} \right|_0 \hat{Q}_{Li} \\
 &= \chi_{\alpha\beta}^0 \hat{I} + \underbrace{\sum_{\lambda} \frac{1}{\sqrt{N}} \sum_{Li} \left. \frac{\partial \chi_{\alpha\beta}}{\partial Q_{Li}} \right|_0 X_{i\lambda} e^{i\mathbf{q}_{\lambda} \cdot \mathbf{P}_L}}_{\equiv \chi'_{\alpha\beta\lambda}} \hat{\eta}_{\lambda} \\
 &= \chi_{\alpha\beta}^0 \hat{I} + \sum_{\lambda} \sqrt{\frac{\hbar}{2\omega_{\lambda}}} \chi'_{\alpha\beta\lambda} (\hat{a}_{\lambda} + \hat{a}_{-\lambda}^{\dagger}) \quad (1.111)
 \end{aligned}$$

where we adopted the second-quantization formalism introduced in chapter 1. Note that we truncated the polarizability at the first order in the atomic displacements, namely we restrict the current description to the first order, and we defined the so-called “Raman tensor”, $\chi'_{\alpha\beta\lambda}$.

We consider an incoming electric field written as:

$$\mathbf{F}_i(t) = \mathbf{F}_i(0) e^{-i\omega_i t} = F_i \mathbf{e}_{i\beta}(0) e^{-i\omega_i t} \quad (1.112)$$

and that the polarization is proportional to the field via the polarizability:

$$\begin{aligned}
 \hat{P}_{\alpha}(t) &= \sum_{\beta} \hat{\chi}_{\alpha\beta}(t) F_{i\beta}(t) \\
 &= F_i \sum_{\beta} e_{i\beta} \hat{\chi}_{\alpha\beta}(t) e^{-i\omega_i t} \\
 &= F_i \sum_{\beta} \left[e_{i\beta} e^{-i\omega_i t} \left(\chi_{\alpha\beta}^0 \hat{I} + \sum_{\lambda} \sqrt{\frac{\hbar}{2\omega_{\lambda}}} \chi'_{\alpha\beta\lambda} [\hat{a}_{\lambda}(t) + \hat{a}_{-\lambda}^{\dagger}(t)] \right) \right] \\
 &= F_i \sum_{\beta} e_{i\beta} \chi_{\alpha\beta}^0 e^{-i\omega_i t} \hat{I} \\
 &\quad + F_i \sum_{\beta\lambda} e_{i\beta} \chi'_{\alpha\beta\lambda} \sqrt{\frac{\hbar}{2\omega_{\lambda}}} [\hat{a}_{\lambda} e^{-i(\omega_i + \omega_{\lambda})t} + \hat{a}_{-\lambda}^{\dagger} e^{-i(\omega_i - \omega_{\lambda})t}] \quad (1.113)
 \end{aligned}$$

where we used eqs. 1.62 and 1.111. By taking the Fourier transform of $\hat{P}_{\alpha}(t)$

we obtain its frequency decomposition:

$$\begin{aligned}\hat{P}_\alpha(\omega) &= F_i \sum_\beta e_{i\beta} \chi_{\alpha\beta}^0 \delta(\omega - \omega_i) \hat{I} \\ &+ F_i \sum_{\beta\lambda} e_{i\beta} \chi'_{\alpha\beta\lambda} \sqrt{\frac{\hbar}{2\omega_\lambda}} [\hat{a}_\lambda \delta(\omega - \omega_{i\lambda}^+) + \hat{a}_{-\lambda}^\dagger \delta(\omega - \omega_{i\lambda}^-)]\end{aligned}\quad (1.114)$$

where we introduced the shorthands $\omega_{i\lambda}^+ \equiv \omega_i + \omega_\lambda$ and $\omega_{i\lambda}^- \equiv \omega_i - \omega_\lambda$.

We can now expand eq. 1.110:

$$\begin{aligned}|\mathbf{e}_s \cdot \hat{\mathbf{P}}(\omega)|^2 &= \left| \sum_\alpha e_{s\alpha} \hat{P}_\alpha(\omega) \right|^2 \\ &= \sum_{\alpha\beta} e_{s\beta}^* e_{s\alpha} \hat{P}_\beta^\dagger(\omega) \hat{P}_\alpha(\omega)\end{aligned}\quad (1.115)$$

Then eq. 1.110 becomes:

$$\frac{\partial^2 W_s(\omega)}{\partial \Omega \partial \omega} = \frac{\omega^4}{(4\pi)^2 \varepsilon_0 c^3} \sum_{\alpha\beta} e_{s\beta}^* e_{s\alpha} \left\langle \hat{P}_\beta^\dagger(\omega) \hat{P}_\alpha(\omega) \right\rangle \quad (1.116)$$

Let us resolve the operators:

$$\begin{aligned}\hat{P}_\beta^\dagger(\omega) \hat{P}_\alpha(\omega) &= F_i^2 \left[\sum_{\gamma'} e_{i\gamma'}^* (\chi_{\beta\gamma'}^0)^* \delta(\omega - \omega_i) \hat{I} + \right. \\ &\quad \left. \sum_{\gamma\lambda} e_{i\gamma}^* \sqrt{\frac{\hbar}{2\omega_\lambda}} (\chi'_{\beta\gamma\lambda})^* [\hat{a}_\lambda^\dagger \delta(\omega - \omega_{i\lambda}^+) + \hat{a}_{-\lambda} \delta(\omega - \omega_{i\lambda}^-)] \right] \\ &\quad \times \left[\sum_{\delta} e_{i\delta} \chi_{\alpha\delta}^0 \delta(\omega - \omega_i) \hat{I} + \right. \\ &\quad \left. \sum_{\delta\lambda'} e_{i\delta} \sqrt{\frac{\hbar}{2\omega_{\lambda'}}} \chi'_{\alpha\delta\lambda'} [\hat{a}_{\lambda'} \delta(\omega - \omega_{i\lambda'}^+) + \hat{a}_{-\lambda'}^\dagger \delta(\omega - \omega_{i\lambda'}^-)] \right]\end{aligned}$$

[continues on the next page]

$$\begin{aligned}
= & F_i^2 \left[\sum_{\gamma\delta} e_{i\gamma}^* e_{i\delta} (\chi_{\beta\gamma}^0)^* \chi_{\alpha\delta}^0 \delta(\omega - \omega_i) \right. \\
& + \sum_{\gamma\delta\lambda'} e_{i\gamma}^* e_{i\delta} \sqrt{\frac{\hbar}{2\omega_{\lambda'}}} (\chi_{\beta\gamma}^0)^* \chi'_{\alpha\delta\lambda'} \delta(\omega - \omega_i) \\
& \cdot [\hat{a}_{\lambda'} \delta(\omega - \omega_{i\lambda'}^+) + \hat{a}_{-\lambda'}^\dagger \delta(\omega - \omega_{i\lambda'}^-)] \\
& + \sum_{\gamma\delta\lambda} e_{i\gamma}^* e_{i\delta} \sqrt{\frac{\hbar}{2\omega_\lambda}} (\chi'_{\beta\gamma\lambda})^* \chi_{\alpha\delta}^0 \delta(\omega - \omega_i) \\
& \cdot [\hat{a}_\lambda^\dagger \delta(\omega - \omega_i - \omega_\lambda) + \hat{a}_{-\lambda} \delta(\omega - \omega_{i\lambda}^-)] \\
& + \sum_{\gamma\delta\lambda\lambda'} e_{i\gamma}^* e_{i\delta} \frac{\hbar}{2\sqrt{\omega_\lambda\omega_{\lambda'}}} (\chi'_{\beta\gamma\lambda})^* \chi'_{\alpha\delta\lambda'} \\
& \cdot [\hat{a}_\lambda^\dagger \hat{a}_{\lambda'} \delta(\omega - \omega_{i\lambda}^+) \delta(\omega - \omega_{i\lambda'}^+) \\
& + \hat{a}_\lambda^\dagger \hat{a}_{-\lambda'}^\dagger \delta(\omega - \omega_{i\lambda}^+) \delta(\omega - \omega_{i\lambda'}^-) \\
& + \hat{a}_{-\lambda} \hat{a}_{\lambda'} \delta(\omega - \omega_{i\lambda}^-) \delta(\omega - \omega_{i\lambda'}^+) \\
& \left. + \hat{a}_{-\lambda} \hat{a}_{-\lambda'}^\dagger \delta(\omega - \omega_{i\lambda}^-) \delta(\omega - \omega_{i\lambda'}^-)] \right] \quad (1.117)
\end{aligned}$$

If we now take the expectation value of the latter expression, we notice that the first term is related to the Rayleigh scattering, the second and third terms, which contain single ladder operators, vanish because of eq. 1.91, and the last term, which is responsible for Raman processes, can only have non-vanishing terms when $\lambda = \lambda'$. Furthermore, operators that mix λ and $-\lambda$ undergo the same fate as the previous cases, producing zero overlaps. Recall the definition and action of the number operator:

$$\langle n_\lambda | \hat{a}_\lambda^\dagger \hat{a}_\lambda | n_\lambda \rangle = \langle n_\lambda | \hat{n}_\lambda | n_\lambda \rangle = n_\lambda \langle n_\lambda | n_\lambda \rangle = n_\lambda \quad (1.118)$$

$$\langle n_{-\lambda} | \hat{a}_{-\lambda} \hat{a}_{-\lambda}^\dagger | n_{-\lambda} \rangle = \langle n_{-\lambda} | \hat{n}_{-\lambda} + 1 | n_{-\lambda} \rangle = n_{-\lambda} + 1 \quad (1.119)$$

then the Raman component of the irradiated power per solid angle is:

$$\begin{aligned}
\frac{\partial^2 W_s^R(\omega)}{\partial \Omega \partial \omega} = & \frac{\hbar \omega^4 F_i^2}{32\pi^2 \varepsilon_0 c^3 \omega_\lambda} \sum_\lambda \left[|\mathbf{e}_s \cdot \boldsymbol{\chi}'_\lambda \cdot \mathbf{e}_i|^2 n_\lambda \delta(\omega - \omega_{i\lambda}^+) \right. \\
& \left. + |\mathbf{e}_s \cdot \boldsymbol{\chi}'_\lambda \cdot \mathbf{e}_i|^2 (n_{-\lambda} + 1) \delta(\omega - \omega_{i\lambda}^-) \right] \quad (1.120)
\end{aligned}$$

where we used $|\mathbf{e}_s \cdot \boldsymbol{\chi}'_\lambda \cdot \mathbf{e}_i|^2 = \sum_{\alpha\beta\gamma\delta} (e_{s\beta}^* (\chi'_{\beta\gamma\lambda})^* e_{i\gamma}^* e_{s\alpha} \chi'_{\alpha\delta\lambda} e_{i\delta})$. The first term is responsible for Anti-Stokes processes, while the second one accounts for Stokes scattering. Note that, since n_λ is the Bose-Einstein factor—which only depends on the temperature and the frequency of the mode—and given eq. 1.62, the occupation of phonons with opposite wavevector is

identical at equilibrium: $n_\lambda = n_{-\lambda}$. We divide by the power per unit of area, $W_L = \varepsilon_0 c F_i^2$ to obtain the differential scattering cross section:

$$\begin{aligned} \frac{\partial^2 \sigma^R(\omega)}{\partial \Omega \partial \omega} &= \frac{\hbar \omega^4}{32 \pi^2 \varepsilon_0^2 c^4 \omega_\lambda} \sum_\lambda \left[|\mathbf{e}_s \cdot \boldsymbol{\chi}'_\lambda \cdot \mathbf{e}_i|^2 n_\lambda \delta(\omega - \omega_{i\lambda}^+) \right. \\ &\quad \left. + |\mathbf{e}_s \cdot \boldsymbol{\chi}'_\lambda \cdot \mathbf{e}_i|^2 (n_{-\lambda} + 1) \delta(\omega - \omega_{i\lambda}^-) \right] \end{aligned} \quad (1.121)$$

Note that the differential cross-section derived above represents the total scattered power from the entire illuminated crystal. Because the phonon normal coordinates were normalized over the total number of unit cells N in Equation 1.111, this total cross-section scales linearly with the macroscopic scattering volume V . To extract an intensive material property that is independent of the specific experimental sample size, it is standard practice to define the Raman scattering efficiency, S . This is obtained by normalizing the total differential cross-section by the scattering volume. However, this is usually incorporated in the experimental prefactor. The actual Raman spectra are usually plotted against the so-called “Raman shift”, that is the difference between the frequency of the exciting light and that of the scattered one, $\omega_s = \omega_i - \omega$:

$$\begin{aligned} \frac{\partial^2 \sigma^R(\omega_s)}{\partial \Omega \partial \omega} &= \frac{\hbar (\omega_i - \omega_s)^4}{32 \pi^2 \varepsilon_0^2 c^4 \omega_\lambda} \sum_\lambda \left[|\mathbf{e}_s \cdot \boldsymbol{\chi}'_\lambda \cdot \mathbf{e}_i|^2 n_\lambda \delta(\omega_s + \omega_\lambda) \right. \\ &\quad \left. + |\mathbf{e}_s \cdot \boldsymbol{\chi}'_\lambda \cdot \mathbf{e}_i|^2 (n_{-\lambda} + 1) \delta(\omega_s - \omega_\lambda) \right] \end{aligned} \quad (1.122)$$

In most cases $\omega_i \gg \omega_s$, hence the quartic power term is often considered constant and also incorporated in the experimental constant.

We obtained the scattering cross section (proportional to the spectrum) within the harmonic approximation, where phonons do not interact, and their spectral function is a Dirac delta. However, in order to introduce the broadening resulting from interactions, we should replace the Dirac delta with some other expressions for the phonon spectral function. We can use eq. 1.100; the Stokes spectrum is:

$$\frac{\partial^2 \sigma^R(\omega_s)}{\partial \Omega \partial \omega} = \frac{\hbar (\omega_i - \omega_s)^4}{32 \pi^2 \varepsilon_0^2 c^4 \omega_\lambda} \sum_\lambda |\mathbf{e}_s \cdot \boldsymbol{\chi}'_\lambda \cdot \mathbf{e}_i|^2 (n_{-\lambda} + 1) A_\lambda(\omega_s) \quad (1.123)$$

and analogously for the anti-Stokes one.

Finally, we must consider that momentum must also be conserved:

$$\begin{aligned} \text{Stokes : } \quad \mathbf{k}_i &= \mathbf{k}_s + \mathbf{q}_\lambda \\ \text{Anti Stokes : } \quad \mathbf{k}_i &= \mathbf{k}_s - \mathbf{q}_\lambda \end{aligned} \quad (1.124)$$

where $\mathbf{k}_i$ is the wavevector of the exciting light, and $\mathbf{k}_s$ is the wavevector of the scattered light, and $\mathbf{q}_\lambda$ is the wavevector of the phonon. The photons used to excite the crystal in the typical Raman experiment carry a momentum of $\sim 1.24 \cdot 10^{-27} \text{ kg} \cdot \text{m/s}$ (corresponding to $\sim 500 \text{ nm}$), while the width of the Brillouin Zone of a crystal can be estimated to be around $\pi \cdot 10^{-9} \text{ m}^{-1}$, with a corresponding momentum of $3.31 \cdot 10^{-25} \text{ kg} \cdot \text{m/s}$. For this reason, $\mathbf{q}_\lambda$ in eq. 1.124 is forced to be very small, hence Raman processes probe phonons very close to the center of the Brillouin Zone, which we can safely assume to be exactly Γ . Practically, this means we restrict the summation in eq. 1.123 to the $3M + 3$ phonons with zero wavevector.

1.2.1 Unpolarized experiments on isotropic samples

In the case of unpolarized source and detector, and isotropic sample (like liquids, glasses, and powders), the laser excitation acts as an incoherent mixture of two orthogonal polarization states spanning the transverse plane defined by $\mathbf{k}_i$. Similarly, the detector captures all scattered intensity by summing over two orthogonal basis vectors spanning the transverse plane of the scattered wavevector $\mathbf{k}_s$. The unpolarized intensity is therefore the double sum over these transverse geometric bases:

$$I_{\text{unpol}} \propto \sum_{\mathbf{e}_i \perp \mathbf{k}_i} \sum_{\mathbf{e}_s \perp \mathbf{k}_s} \langle |\mathbf{e}_s \cdot \boldsymbol{\alpha} \cdot \mathbf{e}_i|^2 \rangle \quad (1.125)$$

Note that, if the sum were replaced by a continuous angular integral over the transverse planes, cross-terms would naturally vanish due to trigonometric integration over the full 2π period, so only two orthogonal basis vectors are required for both sums.

For a macroscopic powder or liquid sample, the system is statistically isotropic. The molecules/crystallites are randomly oriented, meaning the probability of scattering along any laboratory axis is identical. This spatial averaging forces the diagonal and off-diagonal tensor components to collapse into just two distinct values, which we define as the parallel ($I_{\parallel}$) and

orthogonal ($I_{\perp}$) intensities:

$$\langle |\alpha_{ii}|^2 \rangle = \langle |\alpha_{jj}|^2 \rangle = I_{\parallel} \quad (1.126)$$

$$\langle |\alpha_{ij}|^2 \rangle = I_{\perp} \quad (1.127)$$

with $\alpha_{ij} = \mathbf{e}_i \cdot \boldsymbol{\alpha} \cdot \mathbf{e}_j$, for all i, j .

Note that we implicitly assumed the direction of propagation of the incoming light does not matter. However, because the wavevectors of the probed phonons are not exactly null, the eigenvectors that describe the triggered motion do depend on the propagation direction of the light (parallel to its wavevector). The difference between eigenvectors (and eigenvalues) for small $\mathbf{q}$ -vectors lies in the LO-TO splitting, as described in chapter 1. To account for it, Γ -phonons with the appropriate correction must be used to calculate and average the spectra for infinitely many values of $\mathbf{k}_i$ (and resulting $\mathbf{q}_{\lambda}$) on the sphere of radius $|\mathbf{k}_i|$. In the material presented in the next chapters, the LO-TO splitting is negligible, and the averaging is not performed.

Backscattering geometry (180°):

We choose the incident light to travel along the z -axis ($\mathbf{k}_i \parallel \mathbf{e}_z$) and the scattered light to be collected back along the z -axis ($\mathbf{k}_s \parallel -\mathbf{e}_z$). Both transverse planes lie strictly in x and y . The double sum spans $\mathbf{e}_i, \mathbf{e}_s \in \{\mathbf{e}_x, \mathbf{e}_y\}$:

$$\begin{aligned} I_{\text{unpol}}^{180^\circ} &= \langle |\alpha_{xx}|^2 \rangle + \langle |\alpha_{xy}|^2 \rangle + \langle |\alpha_{yx}|^2 \rangle + \langle |\alpha_{yy}|^2 \rangle \\ &\propto I_{\parallel} + I_{\perp} + I_{\perp} + I_{\parallel} = 2I_{\parallel} + 2I_{\perp} \end{aligned} \quad (1.128)$$

Orthogonal geometry (90°):

The incident light travels along the x -axis ($\mathbf{k}_i \parallel \mathbf{e}_x$), yielding a transverse plane of y and z . The scattered light is collected along the y -axis ($\mathbf{k}_s \parallel \mathbf{e}_y$), yielding a transverse plane of x and z . The double sum spans $\mathbf{e}_i \in \{\mathbf{e}_y, \mathbf{e}_z\}$ and $\mathbf{e}_s \in \{\mathbf{e}_x, \mathbf{e}_z\}$:

$$I_{\text{unpol}}^{90^\circ} = \langle |\alpha_{yx}|^2 \rangle + \langle |\alpha_{yz}|^2 \rangle + \langle |\alpha_{zx}|^2 \rangle + \langle |\alpha_{zz}|^2 \rangle \quad (1.129)$$

$$I_{\text{unpol}}^{90^\circ} = I_{\perp} + I_{\perp} + I_{\perp} + I_{\parallel} = I_{\parallel} + 3I_{\perp} \quad (1.130)$$

Rotational Invariants

Finally, we can introduce the fundamental rotational invariants of the intrinsic molecular tensor: the isotropic mean polarizability (a) and the anisotropy (γ^2):

$$a = \frac{1}{3} \text{Tr}\{\alpha\} \quad (1.131)$$

$$\gamma^2 = \frac{1}{2} \left[|\alpha_{xx} - \alpha_{yy}|^2 + |\alpha_{yy} - \alpha_{zz}|^2 + |\alpha_{xx} - \alpha_{zz}|^2 + 3(|\alpha_{xy}|^2 + |\alpha_{yz}|^2 + |\alpha_{zx}|^2) \right] \quad (1.132)$$

We can now rewrite the intensities in terms of the invariants:

$$I_{\parallel} = a^2 + \frac{4}{45} \gamma^2 \quad \text{and} \quad I_{\perp} = \frac{1}{15} \gamma^2 \quad (1.133)$$

Evaluating the final geometries yields:

$$I_{\text{unpol}}^{180^\circ} = 2 \left[a^2 + \frac{4}{45} \gamma^2 \right] + 2 \left[\frac{1}{15} \gamma^2 \right] = 2a^2 + \frac{14}{45} \gamma^2 \quad (1.134)$$

$$I_{\text{unpol}}^{90^\circ} = \left[a^2 + \frac{4}{45} \gamma^2 \right] + 3 \left[\frac{1}{15} \gamma^2 \right] = a^2 + \frac{13}{45} \gamma^2 \quad (1.135)$$

For a further description of the possible experimental geometries, we refer to [20].

1.3 Density Functional Theory

Finding the analytical solutions of Schrödinger's equation for anything more complex than a hydrogen-like atom is virtually impossible, so numerical methods are used instead to determine solutions within certain approximations.

In the Hartree-Fock (HF) method, the system is described as a symmetrized linear combination (Slater determinant) of some single-particle orbitals, one for each electron, and the electronic interaction is described as average coulombic and exchange integrals, in the mean-field fashion. The single-particle wavefunctions, expressed as a linear combination of some basis set, are found by minimizing the energy under orthonormality constraint. This approach relies on the *variational theorem*, which states that “the energy of any trial wavefunction, under a given Hamiltonian, is greater or equal

to the lowest-energy eigenstate of the Hamiltonian". One of the greatest limitations of the HF method is the lack of correlation between electrons. Several methods were developed to add back the correlation, the so-called "post-HF methods", such as the Møller-Plesset perturbation theory (MP2) or Coupled Cluster (CC), which, however, come at a much higher computational cost: while HF scales at $\mathcal{O}(N^4)$, MP2 scales at $\mathcal{O}(N^5)$ and CC at $\mathcal{O}(N^7)$, with N being the number of electrons.

We owe Kohn and Sham the most successful computational technique for the ground-state calculation: Density Functional Theory (DFT). Before facing it, let us rewrite the electronic Hamiltonian in eq. 1.4:

$$\hat{H}_e(\mathbf{r}, \mathbf{R}) = \hat{T}_e(\mathbf{r}) + \hat{V}_{ee}(\mathbf{r}) + \hat{V}_{ext}(\mathbf{r}, \mathbf{R}) \quad (1.136)$$

where $\hat{V}_{ext}(\mathbf{r}, \mathbf{R}) = \hat{V}_{nn}(\mathbf{R}) + \hat{V}_{en}(\mathbf{r}, \mathbf{R})$ is the external potential. The respective eigenvalue equation reads:

$$\hat{H}_e\psi_i = E_i\psi_i \quad (1.137)$$

The first thing we note is that when the wavefunction ψ_i changes, the respective energy changes as well; in other words the energy is a *functional* of the wavefunction:

$$E_i = E[\psi_i] \quad (1.138)$$

The foundation of DFT is the Hohenberg-Kohn theorems, which states that:

- The external potential determines the ground state (GS) electron density **and vice versa**.
- In any quantum state, the external potential uniquely determines the wavefunction.

As a consequence of these two points and eq. 1.138, we conclude that the energy of the GS is a functional of the sole GS electron density, n_0 :

$$E_0 = E[n_0] \quad (1.139)$$

The power of Hohenberg-Kohn's theorem lies in the fact that, in order to compute the GS energy, we do not need the cumbersome wavefunction, which depends on $3N$ variables, but just the GS electron density, which is a function of three variables.

Though we introduced a great simplification of the electronic problem, we still do not know what the functional looks like. We can start by writing the electron density as:

$$n(\mathbf{r}) = \sum_i |\psi_i(\mathbf{r})|^2 = \sum_i \langle \psi_i | \mathbf{r} \rangle \langle \mathbf{r} | \psi_i \rangle \quad (1.140)$$

Now the GS energy is:

$$E[n_0] = \int n_0(\mathbf{r}) V_{ext}(\mathbf{r}) d\mathbf{r} + \langle \psi[n_0] | \hat{T}_e(\mathbf{r}) + \hat{V}_{ee}(\mathbf{r}) | \psi[n_0] \rangle \quad (1.141)$$

where we used the fact that the wavefunction itself is a functional of the electron density: $\psi_0 = \psi[n]$. Unlike the first term, the kinetic and electron-electron interaction terms do not depend explicitly on the electron density in a known, exact form, so we must find a way to treat them. Kohn and Sham's solution is to exactly map the interacting system onto a fictitious auxiliary system of non-interacting fermions that yields the exact same ground-state density. By computing the exact kinetic energy of this auxiliary system, all remaining complex many-body effects—including the non-classical electron interactions and the kinetic energy difference between the interacting and non-interacting systems—are rigorously gathered into a single term called the "exchange-correlation functional", E_{xc} :

$$\begin{aligned} E[n_0] = & \int n_0(\mathbf{r}) V_{ext}(\mathbf{r}) d\mathbf{r} - \sum_i \int \phi_i^*(\mathbf{r}) \frac{\nabla^2}{2} \phi_i(\mathbf{r}) d\mathbf{r} \\ & + \frac{1}{2} \iint \frac{n_0(\mathbf{r}) n_0(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' + E_{xc}[n_0] \end{aligned} \quad (1.142)$$

where the orbitals $\{\phi_i\}$ are orthonormal and build the exact electron density: $\sum_i |\phi_i(\mathbf{r})|^2 = n_0(\mathbf{r})$. The exchange and correlation functional contains all that was left from the kinetic and Hartree terms, namely a bit of kinetic energy and the non-classical exchange and correlation interaction between electrons.

Even assuming we knew the exact form of E_{xc} , we still do not know how to determine the GS electron density. The *Hohenberg-Kohn variational principle* states that:

$$\left. \frac{\delta E[n]}{\delta n} \right|_{n_0} = 0 \quad (1.143)$$

where the left-hand term is a functional derivative. Recalling the quantum variational theorem, we can easily conclude that the GS electron density is

the point of global minimum for the functional E .

By minimizing E under the constraint of orthonormality of the auxiliary orbitals, we obtain the Kohn-Sham equation:

$$\left[-\frac{1}{2}\nabla^2 + V_{ext}(\mathbf{r}) + V_H(\mathbf{r}) + V_{xc}(\mathbf{r}) \right] \varphi_i(\mathbf{r}) = \varepsilon_i \varphi_i(\mathbf{r}) \quad (1.144)$$

where the set of orbitals $\{\varphi_i\}$ is obtained by linearly combining the auxiliary orbitals $\{\phi_i\}$ to retrieve an eigenvalue equation; $V_H(\mathbf{r}) = \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}'$ is the Hartree potential, and the exchange and correlation potential V_{xc} is given by:

$$V_{xc}(\mathbf{r}) = \left. \frac{\delta E_{xc}[n]}{\delta n} \right|_{n(\mathbf{r})} \quad (1.145)$$

Therefore, we can obtain the GS electron density by diagonalizing the Kohn-Sham Hamiltonian, built for a system of non-interacting electrons, via $n_0(\mathbf{r}) = \sum_i |\varphi_i(\mathbf{r})|^2$, and the GS energy by plugging the orbitals and n_0 into the Kohn-Sham functional in 1.142.

Unfortunately, the Kohn-Sham eigenvalue equations are self-consistent, as the solutions $\{\varphi_i\}$ and n_0 appear in the operators that need to be diagonalized. As a result, each orbital depends on all the others. As for many other computational techniques, we address the problem by starting from a reasonable set of orbitals to write the Kohn-Sham operators, then we diagonalize the Kohn-Sham Hamiltonian, obtaining a new set of orbitals, and repeat these steps until the electron density (or related quantities) no longer changes, achieving self-consistency.

For more details on the numeric implementation, including the formulation and several generations of exchange and correlation functionals, we refer to [21].

The computational cost of DFT scales with the cube of the number of electrons, which is orders of magnitude less than HF and post-HF methods.

1.3.1 Density Functional Perturbation Theory (DFPT)

Many interesting phenomena and properties of matter can be described as the response of the system to some external perturbation. For example, the structure of an infrared spectrum is due to the presence of vibrational states, which arise from the perturbation of the GS induced by the lattice vibrations. When the perturbation is small, we can rely on Perturbation Theory to model

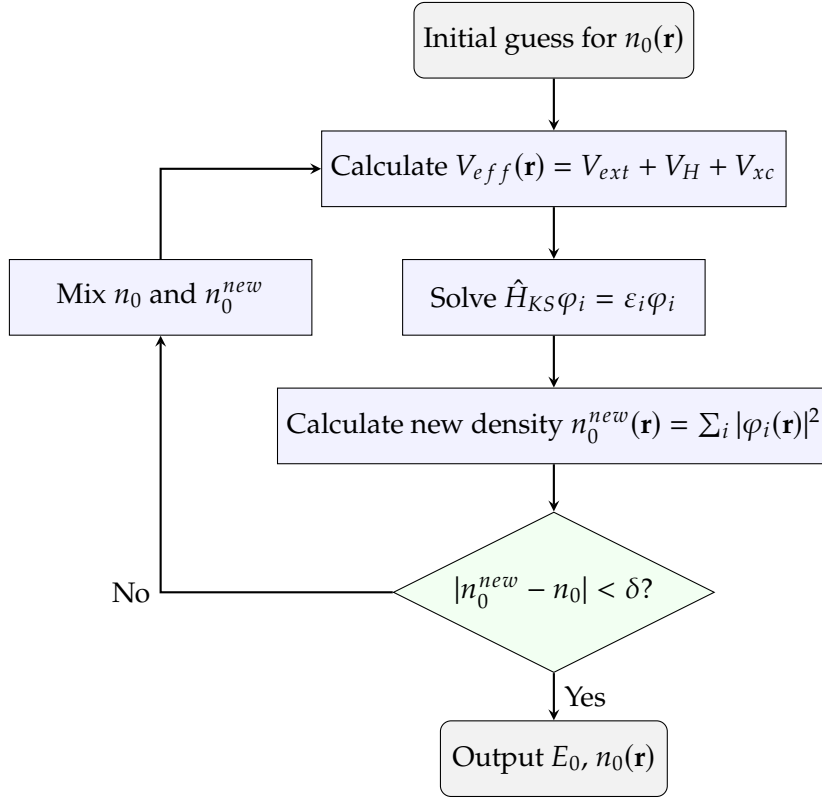


Figure 1.1: Schematic representation of the self-consistent field (SCF) cycle used to solve the Kohn-Sham equations.

the response and consequent properties, and in the context of DFT we use the so-called “Density Functional Perturbation Theory” (DFPT). The core idea of perturbation theory is that a correction is added to the *unperturbed* Hamiltonian to account for the perturbation:

$$\hat{H} = \hat{H}^{(0)} + \hat{H}^{(1)} = \hat{H}^{(0)} + \lambda \hat{V} \quad (1.146)$$

where $\hat{V}$ represents the perturbing field, and λ is a small parameter.

$$\hat{H}^{(0)}\Psi_i^{(0)} = E_i^{(0)}\Psi_i^{(0)} \quad (1.147)$$

$$\hat{H}\Psi_i = E_i\Psi_i \quad (1.148)$$

For sufficiently weak perturbations, we can expand the energy and the wavefunctions as power series in λ :

$$E_i = E_i^{(0)} + \lambda E_i^{(1)} + \lambda^2 E_i^{(2)} + \dots \quad (1.149)$$

$$\Psi_i = \Psi_i^{(0)} + \lambda \Psi_i^{(1)} + \lambda^2 \Psi_i^{(2)} + \dots \quad (1.150)$$

with

$$E_i^{(n)} = \frac{1}{n!} \left. \frac{d^n E_i}{d\lambda^n} \right|_{\lambda=0} \quad (1.151)$$

$$\Psi_i^{(n)} = \frac{1}{n!} \left. \frac{d^n \Psi_i}{d\lambda^n} \right|_{\lambda=0} \quad (1.152)$$

We will not derive the equations of DFPT, but simply highlight two points of practical importance for this thesis:

- **The Hellmann-Feynmann theorem:** first order derivatives of the energy are obtainable from the GS wavefunction by differentiating the Hamiltonian:

$$\frac{\partial E_i}{\partial \lambda} = \sum_i \left\langle \Psi_i^{(0)} \left| \frac{\partial \hat{H}}{\partial \lambda} \right| \Psi_i^{(0)} \right\rangle \quad (1.153)$$

where $\{\Psi_i^{(0)}\}$ are the solutions of the *unperturbed* Hamiltonian. A common application of this theorem is that of forces and stresses, which are the first derivatives with respect to atomic displacement and cell strain, respectively. The functionals are differentiated once, and the matrix elements in eq. 1.153 are evaluated by the software as needed, once the unperturbed (GS) wavefunctions are calculated. This is equivalent—but faster, more accurate and more robust—than performing a finite differentiation in the zero-perturbation limit.

- **The $2n + 1$ rule:** it can be proved [22] that once the perturbed wavefunctions at order n are calculated, the energy derivatives are then accessible up to $(2n + 1)$ -th order. For instance, second and third derivatives of the energy (such as IFCs, dielectric tensor and Raman tensor) are obtainable by knowing the wavefunctions perturbed at the first order. For example, the (atomic) Raman tensor, is the change in the dielectric susceptibility due to a change in the atomic displacement, namely a third derivative of the energy:

$$\chi'_{\alpha\beta Q_{Li}} = \frac{\partial^3 U}{\partial F_\alpha \partial F_\beta \partial Q_{Li}} \quad (1.154)$$

with F being an electric field and Q the atomic displacement (see section 1.2 for notation clarification). Instead of an expensive calculation of the third-order variation of the wavefunction, one can simply compute the first derivative of it with respect to both perturbations (electric field and atomic displacement).

1.4 Molecular Dynamics

The macroscopic behavior of a material is fundamentally dictated by the time-dependent dynamical trajectories of its atoms, encompassing both equilibrium states and non-equilibrium processes. For instance, by knowing how the microscopic stress and the velocity of each atom evolve in time at equilibrium, macroscopic transport coefficients such as the thermal conductivity can be rigorously determined. Similarly, under non-equilibrium conditions, natural thermal fluctuations can induce profound conformational changes in complex molecular systems like enzymes, generating transient active sites for reactions that would otherwise be hindered by high activation barriers. Simulating how atoms move and interact in time is therefore a powerful tool for scientists.

Leveraging the Born-Oppenheimer approximation, the natural birthplace of any time-dependent atomistic model lies in classical dynamics. The temporal evolution of the system is fundamentally governed by Lagrangian mechanics. In its most direct application, Molecular Dynamics (MD), this framework reduces to the numerical integration of Newton's second law of motion. In other words, the trajectory of a body (e.g., an atom) is determined by its initial velocity and the forces acting on it:

$$\mathbf{r}(t) = \mathbf{r}(0) + \mathbf{v}(0)t + \int_0^t \int_0^{t'} \frac{\mathbf{F}(t'')}{m} dt'' dt' \quad (1.155)$$

In atomistic simulations, the atomic force is a complicated function of time, determined by the positions of all the interacting atoms, which themselves are the result of the dynamics under the atomic forces at previous times. Although deterministic, classical dynamics is chaotic, hence we cannot hope to calculate the trajectory analytically, as pointed out by Poincaré in his study of the three-body problem[23]. Nevertheless, the problem can be practically attacked by subdividing time into small finite steps Δt . The position of a particle at the future time $t + \Delta t$ and at the past time $t - \Delta t$ can be approximated via a Taylor series expansion:

$$\mathbf{r}(t + \Delta t) \approx \mathbf{r}(t) + \mathbf{v}(t) \Delta t + \frac{1}{2} \mathbf{a}(t) \Delta t^2 + \frac{1}{6} \mathbf{b}(t) \Delta t^3 \quad (1.156)$$

$$\mathbf{r}(t - \Delta t) \approx \mathbf{r}(t) - \mathbf{v}(t) \Delta t + \frac{1}{2} \mathbf{a}(t) \Delta t^2 - \frac{1}{6} \mathbf{b}(t) \Delta t^3 \quad (1.157)$$

where $\mathbf{a}(t)$ is the acceleration, and $\mathbf{b}(t)$ is the so-called *jerk*. By summing those two equations and rearranging the terms we retrieve the classic *Verlet*

algorithm for the propagation of positions forward in time:

$$\mathbf{r}(t + \Delta t) \approx 2\mathbf{r}(t) - \mathbf{r}(t - \Delta t) + \mathbf{a}(t) \Delta t^2 \quad (1.158)$$

This equation allows us to construct a complete trajectory $\mathbf{r}(t)$ step-wise, using only the current and previous positions, and the current atomic forces (proportional to the acceleration), with an $\mathcal{O}(\Delta t^4)$ error which can be reduced by shrinking the time step. Since the dynamics is governed by Hamiltonian mechanics, the volume of phase space is conserved, and the error oscillates within some boundaries and does not accumulate. This property is known as “symplecticity”.

Therefore, in order to carry out an MD simulation, we must be able to determine the forces acting on each atom, at every time step. Keeping our description restricted to conservative fields, the atomic forces at a given time are obtained as the negative gradient of the Potential Energy Surface (PES):

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} U(\mathbf{r}_1, \dots, \mathbf{r}_N) \quad (1.159)$$

Non-conservative fields, like viscosity, which do not only depend on the atomic positions, are introduced as additional force terms contributing to the acceleration.

Many so-called “classical force fields” (FFs) have been developed to describe the PES with variable accuracy and transferability. Classical FFs are usually relatively simple functions which depend on computational or even experimental parameters. Sometimes, FFs are built in a modular fashion by adding together classical FFs devoted to describe different interactions (e.g., van der Waals forces, interactions with the solvent).

In the simplest cases —i.e., ground state of crystals or isolated molecules, or heterostructures— we can accurately compute the PES locally (at each time step) with *ab initio* methods, like DFT; this approach is termed “*ab initio* MD” (AIMD). Considering that MD simulations can last millions of time steps, the cost of AIMD is very high, limiting the viability of workflows involving MD simulations. Fortunately, nowadays a new class of FFs is available, which is able to achieve accuracies close to *ab initio*, but at a fraction of the cost: Machine Learning Force Fields, and specifically the Machine Learning Interatomic Potentials (MLIPs), which reproduce the PES without non-conservative terms.

In solid-state physics simulations, periodic boundary conditions are often applied on the cell and the interactions are computed adopting the *minimum image convention*, that is considering the closest replica of an atom, when calculating the distances for many-body interaction terms. We already anticipated that MD is widely used to compute equilibrium and even non-equilibrium properties. Thermodynamic quantities, like pressure and temperature, must often be carefully defined and controlled, as they are in real experiments. However, simple integration of Newton's second law does not account for thermodynamic constraints, and only ensures the conservation of particles, energy, and volume, namely the system naturally evolves in the *microcanonical ensemble*, NVE. In order to fix other quantities, their conservation must be explicitly enforced.

1.4.1 Thermostats

A *thermostat* is a bath of kinetic energy which slows down or accelerates the atoms to achieve a desired temperature. A thermostat can be implemented in several ways, depending on the properties of the simulation needed.

For example, the **Langevin thermostat** [24] works by adding a viscosity term, which constantly subtracts kinetic energy, and a stochastic momentum injector. The spread of the random noise is fixed to be proportional to the viscosity coefficient, to ensure that the fluctuation-dissipation theorem is verified and the expected temperature is attained. The presence of random kicks prevents the system from getting stuck into shallow local minima, ensuring ergodicity (see chapter 1.4.4); however, randomness also destroys time-reversibility and the simulation is no longer deterministic. Also, because the thermostat adds non-conservative fields, the Langevin integrator is not symplectic; however, the system is prevented from drifting through algorithms which avoid numerical integration for the non-conservative part (like the BAOAB integrator [25]).

The **Nosé-Hoover thermostat** [26, 27], on the other hand, explicitly introduces a fictitious extra degree of freedom in the system, representing a thermal bath, which actively exchanges energy with the atoms to keep the temperature fixed. This ensures that the evolution is deterministic and time-reversible. However, the bath degree of freedom needs parametrization via an effective mass, which is not trivial to choose. Indeed, a wrong choice of the effective mass can lead to strong instability or a slow dynamic which does not thermalize usefully.

1.4.2 Barostats

To enforce constant pressure, we need a *barostat*. The basic concept behind a barostat is to control the volume of the simulation box, to reduce or build up the pressure until a target value is reached. Again, there are several implementations of barostats, with their respective advantages and disadvantages.

The **Andersen barostat** [28] works in a similar manner as the Nosé-Hoover thermostat: a fictitious degree of freedom is added to the Hamiltonian, with an effective mass. It is defined so as to increase the volume when the pressure is greater than the desired one, and shrink the volume in the opposite case; the atomic positions are then rescaled according to the new volume, proportionally to $\sqrt[3]{V^{new}/V^{old}}$. The effective mass determines how strongly the bath is coupled to the atomic system. While the Andersen barostat applies isotropic stress by construction, other implementations were developed to enforce an anisotropic stress tensor, such as the **Parrinello-Rahman barostat** [29].

1.4.3 Thermodynamic ensembles

An extended and thorough study of the various thermodynamic ensembles can be found in statistical mechanics textbooks, such as [30]. We will just review the most common ones, and when they are used.

Microcanonical ensemble, NVE

The microcanonical ensemble is the region of phase space where the total energy, the volume and the number of particles are fixed quantities. A system that evolves in the NVE ensemble cannot exchange energy with any external system, hence this ensemble is appropriate to describe isolated systems.

Canonical ensemble, NVT

In the canonical ensemble the total energy can fluctuate, while the temperature is kept constant, as well as the number of particles and the volume. Naturally, as T and V can be chosen independently, the pressure P cannot be fixed and must change. In such conditions, the Helmholtz free energy,

F , is well defined, and is the thermodynamic state function which governs the probability of the system to be found in a specific point in phase space. It can be used to simulate systems which are isolated mechanically, but exchange heat with a reservoir. In MD, NVT sampling is attained by using a thermostat.

Isobaric-isothermal ensemble, NPT

Most experiments are carried out at constant pressure and with free exchange of heat. Under such conditions, and because of Gibbs phase rule, the volume of the system cannot be fixed. For instance, a (thermodynamicly irreversible) phase transition at constant pressure and temperature will result in a change of volume of the material. The statistics of the NPT ensemble is determined by the Gibbs free energy. A barostat and a thermostat must be combined to sample the NPT (isobaric-isothermal) ensemble.

Grand Canonical ensemble, μVT

The μVT ensemble is used to simulate chemical reactions or processes where the number of particles changes, such as desorption, or reactions in solution where a precipitate is formed, at constant temperature and without exchange of mechanical energy. In all such cases, the total chemical potential μ is conserved. The probability distribution of this ensemble is governed by the Grand Potential. The μVT ensemble is not easy to implement in MD (while it is widely sampled in *Monte Carlo* simulations [31, 32]). In fact, continuous numerical integration of interatomic forces is inherently incompatible with discontinuous particle exchanges. Instantaneously inserting a new particle often causes atomic overlaps that generate massive repulsive force spikes, which instantly destabilize and crash the simulation's time integrator.

1.4.4 The Ergodic Hypothesis

One of the main results of statistical mechanics is that the value of an observable A , if measured at equilibrium, is the thermal average of the observable, $\langle A \rangle$, computed for all possible states in the phase space Ω :

$$\langle A \rangle = \int_{\Omega} P(\mathbf{x}, \mathbf{p}) A(\mathbf{x}, \mathbf{p}) d\mathbf{x} d\mathbf{p} \quad (1.160)$$

where $\mathbf{x}$ and $\mathbf{p}$ are the ensemble of all variables in the phase space, and P is the probability distribution function which governs the statistics, based on the conservation laws enforced, that is the chosen thermodynamic ensemble. Unfortunately, the probability distribution is rarely easy to compute—and integrate—analytically, so it is very hard to compute thermal averages accurately this way.

However, there is an assumption, which is often true, under which the equilibrium value of an observable can be calculated: the *Ergodic Hypothesis*. This hypothesis postulates that a system left free to evolve in time will eventually explore all the possible states in its phase space, and spend in each state $(\mathbf{x}, \mathbf{p})$ an amount of time proportional to the respective $P(\mathbf{x}, \mathbf{p})$. Therefore, for a system in equilibrium, the time average of a physical observable calculated over a sufficiently long trajectory is equivalent to the ensemble average calculated over all possible states of the system:

$$\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_0^\tau A(t) dt = \langle A \rangle \quad (1.161)$$

In practice, this allows us to extract thermodynamic properties from a single, evolving, simulation box. However, the assumption of ergodicity can be compromised if the simulation time is shorter than the characteristic time required for the system to explore its relevant phase space. This issue frequently arises in systems with high activation barriers—such as glassy materials, structural phase transitions, or complex biomolecules—where the system remains kinetically trapped in a local minimum of the PES for the duration of the simulation. In such cases, the time average fails to converge to the ensemble average, and enhanced sampling techniques are required to restore ergodicity and obtain meaningful statistics.

1.4.5 Equilibration of MD simulations

To initiate an MD simulation, we require an initial arrangement of atoms in space. While this is sufficient to compute forces and integrate the equations of motion, several technical issues must be addressed to ensure physical validity. For instance, if one simulates a crystal at constant energy (NVE) and initializes the system at its perfect equilibrium geometry, the Potential Energy Surface (PES) is at a minimum, the atomic forces are zero, and the system remains static. To avoid this, it is standard practice to perform a *random displacement* (or “rattling”) of the atomic positions.

Furthermore, if initial velocities are not specified, they are typically assigned according to a Maxwell-Boltzmann distribution at the target temperature. In isothermal or isobaric simulations, the system may initially be far from the target ensemble. Consequently, the velocities and pressure during the first time steps are not representative of the target state; it takes a finite time for the thermostat and barostat to drive the system to the desired conditions.

Therefore, the simulation is divided into two distinct phases: *equilibration* and *production*. During the equilibration phase, the system evolves under the influence of the chosen thermostat and barostat to shed the artifacts of the initial configuration. We monitor this phase by tracking the time evolution of observables such as potential energy, temperature, and pressure. The system is considered equilibrated when these quantities show no systematic drift and oscillate stably around their target averages. Once this stationary state is reached, the equilibration trajectory is discarded, and we proceed to the production phase, where the trajectory is considered consistent with the target thermodynamic ensemble and is used for subsequent analysis.

1.5 Machine Learning

With the advent of modern computing, an alternative paradigm for establishing relationships between variables of interest has taken hold. Traditional physical models rely on minimal analytical equations where each parameter corresponds to a concrete physical quantity. In contrast, *Machine Learning* (ML) links observables through complex networks of intermediate, non-physical variables interacting in highly non-linear ways. While this approach allows for highly accurate mapping between arbitrary input and output variables, the underlying physics ends up implicitly encoded within the internal parameters of the model, making physical intuition difficult to extract. On the other hand, ML allows for efficient calculation of observables from input variables, in cases where an analytical mapping is not available. The ensemble of intermediate relations and their parameterization constitute a “Machine Learning model”.

In order to capture the underlying relation between input and output variables, a set of relevant known examples must be given. The mapping function is then constructed by optimizing the internal parameters of the model to fit this data, in a process known as “training”.

1.5.1 Linear fit

The simplest form of ML model is linear interpolation. Suppose we want to model the final grade, y , of high school students, based on the grades of the previous years, $\{x_1, x_2, x_3, x_4\}$. Note that we suppose that *there is* a relation between the input variables (previous grades) and the output (y). The simplest way to combine the input variables is the linear combination:

$$y = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 \quad (1.162)$$

where $\{w_n\}$ are the linear coefficients to be found. We take a survey and collect our “training set”:

Table 1.1: Dataset of previous grades (x_1 to x_4) and the resulting final grade (y) for five students.

Student	x_1	x_2	x_3	x_4	y
1	7.0	7.0	7.5	9.0	8.0
2	6.0	6.5	6.0	8.0	7.0
3	9.0	8.5	8.0	9.5	8.9
4	5.0	6.0	5.5	7.0	6.2
5	8.0	7.5	9.0	6.5	7.4

It is not possible to find a set of coefficients that exactly reproduce the datapoints in the training set, because the number of parameters is less than the number of datapoints. This is because the relation is not actually perfectly linear. Whichever set $\{w_n\}$ is chosen, the predicted final grade of the datapoint i , $\tilde{y}^i$, will always be affected by an error:

$$y^i - \tilde{y}^i = \epsilon_i \quad (1.163)$$

Under the assumption that the chosen functional form (linear combination) is valid, the best set of parameters is found by minimizing a “loss function”, L :

$$\{w_n\} = \operatorname{argmin} \{L\} \quad (1.164)$$

An example of loss function is the *root mean square error* (RMSE):

$$L = \sqrt{\frac{1}{N} \sum_{i=1}^N \epsilon_i^2} \quad (1.165)$$

Once the set of coefficients is determined, the model is defined and can be used to compute the final grade of a new student given the previous years’ grades.

By applying the ordinary least squares method to the training set in Table 1.1,

we find the optimized parameter set:

$$w_1 = 0.137, \quad w_2 = 0.195, \quad w_3 = 0.206, \quad w_4 = 0.458 \quad (1.166)$$

This yields a minimal residual error of $\text{RMSE} \approx 0.0095$.

Because this model is linear, these parameters provide immediate interpretability. The magnitude of the coefficients directly shows the impact of each feature: a student's performance in their most recent years is the strongest predictor of their final grade.

While this model is transparent, it cannot capture higher-order relationships or sudden changes in a student's trajectory. This necessitates the transition towards more complex—and hence flexible—models.

1.5.2 Neural Networks (NN)

Real progress in the ML field was made by generalizing the linear fit to an arbitrarily long sequence of nested relations, namely “Neural Networks” (NN). However, a simple chain of nested linear combinations remains mathematically equivalent to a single linear transformation. To break this limitation and capture highly non-linear phenomena, the output of each linear combination must be passed through a non-linear transformation, known as an “activation function” (σ).

We can visualize this by expanding our previous student grade example. Instead of mapping the previous grades directly to the final grade g_f , the inputs are first passed to an intermediate layer of hidden variables—or *neurons*— $\{h_m\}$. Each neuron executes a localized linear fit of the inputs, modified by an intrinsic parameter called a *bias* (b_m), and filters the result through the activation function:

$$h_m = \sigma \left(\sum_{n=1}^4 w_{mn} g_n + b_m \right) \quad (1.167)$$

where w_{mn} represents the weight matrix linking the input features to the hidden layer.

The final grade is then calculated by combining the outputs of this hidden layer:

$$\tilde{y} = \sum_m w_m h_m + b_f \quad (1.168)$$

Input Layer Hidden Layer Output Layer

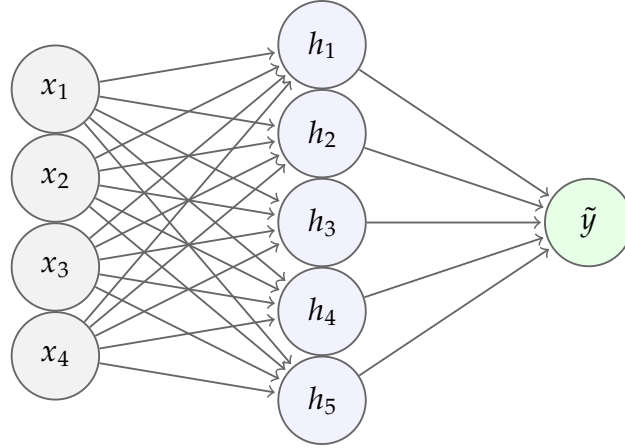


Figure 1.2: Schematic of a fully-connected feed-forward neural network. The input layer accepts the features (x_n), maps them to a hidden layer (H_m), which ultimately recombines to yield the final prediction ($\tilde{y}_f$)

By stacking multiple hidden layers between the input and the output, the network forms a *Deep Neural Network* (DNN). As the architecture grows, the total number of internal parameters (weights and biases) scales from a handful to millions. While this vast parameter space allows the network to approximate virtually any continuous multi-dimensional function, the explicit physical meaning of any single weight is lost within the collective network, transforming the model into a predictive *black box*. For a deep NN, the m -th neuron of the l -th layer is calculated as:

$$h_m^{(l)} = \sigma \left(\sum_{n=1}^{N^{(l-1)}} w_{mn}^{(l)} h_n^{(l-1)} + b_m^{(l)} \right) \quad (1.169)$$

where N_l is the number of neurons in the l -th layer. It can also be written in matrix form:

$$\mathbf{h}^{(l)} = \sigma \left(\mathbf{w}^{(l)} \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)} \right) \quad (1.170)$$

Input Layer Hidden Layer 1 Hidden Layer 2 ... Hidden Layer K Output Layer

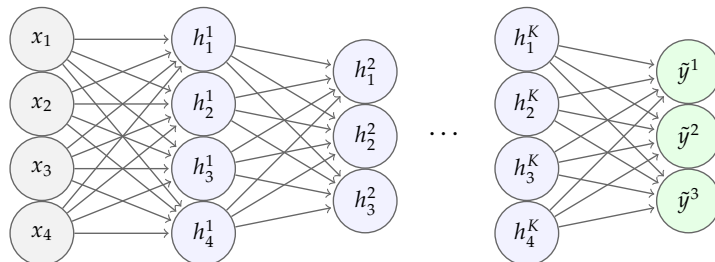


Figure 1.3: Deep neural network mapping input parameters through an arbitrary number of K hidden layers to yield multi-dimensional predictions.

1.5.3 Backpropagation and Model Optimization

The training of the network is achieved by optimizing its vast parameter space—consisting of all weight matrices $\mathbf{W}$ and bias vectors $\mathbf{b}$ —to minimize the loss function. Because of the multi-layered, nested structure of deep networks, calculating how a change in an early hidden layer's weight affects the final output error is non-trivial. This is done through the *backpropagation* algorithm.

Backpropagation utilizes the chain rule to systematically compute the partial derivatives of the loss function L with respect to every individual parameter in the network. The error is first evaluated at the output layer and then propagated backward through the hidden layers. For a given weight $w_{mn}^{(l)}$ in layer l , the gradient is calculated as:

$$\frac{\partial L}{\partial w_{mn}^{(l)}} = \frac{\partial L}{\partial h_m^{(l)}} \frac{\partial h_m^{(l)}}{\partial w_{mn}^{(l)}} \quad (1.171)$$

Once these gradients are computed across the training set, an optimization algorithm—most commonly *Gradient Descent*—updates the parameters in the direction that decreases the loss:

$$w_{mn}^{(l)} \leftarrow w_{mn}^{(l)} - \eta \frac{\partial L}{\partial w_{mn}^{(l)}} \quad (1.172)$$

where η is the learning rate, a *hyperparameter* controlling the step size of the update. Through thousands of iterative cycles of forward propagation, error evaluation, and backpropagation, the model eventually converges to a state where it can generalize predictions to unseen data.

1.5.4 Activation functions

Activation functions play a crucial role in NNs: their derivatives and limits determine the boundaries and the local differential structure of the model which has to approximate the target output variables as functions of the input ones.

There are several activation functions; a few historical examples are the following:

- Identity function: $f(x) = x$;

- Heaviside step function: $f(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases};$
- Rectified Linear Unit (ReLU): $f(x) = \max(0, x);$
- Sigmoid (or Logistic): $f(x) = (1 + e^{-x})^{-1}.$

Each activation function has its advantages and flaws, and the choice for an NN architecture should take into account the nature of the expected prediction (e.g. if it is binary classification or continuous inference). The graphs of each activation function are plotted in Fig. 1.4.

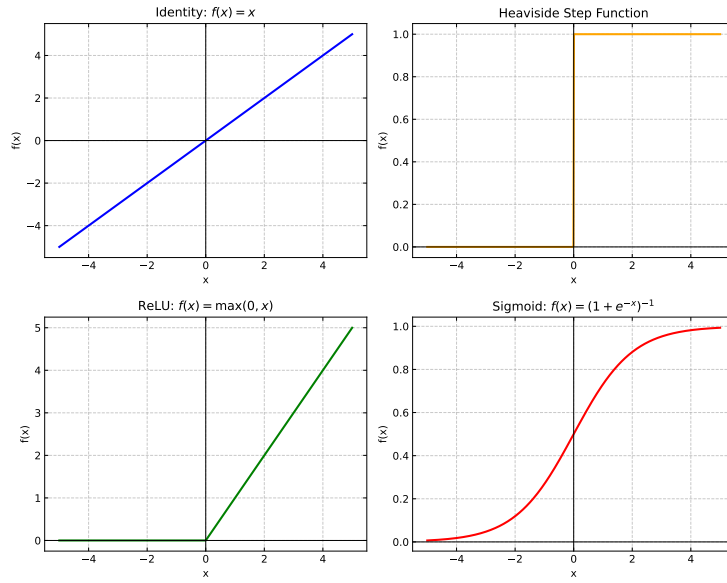


Figure 1.4: Graphs of some of the most important activation functions.

1.5.5 Training

The process of training a Neural Network consists of optimizing its parameters to minimize the chosen loss function L across the training set. In classical linear regression, such as the example in Table 1.1, the problem is typically *overdetermined* because the number of independent data points M exceeds the number of adjustable parameters P ($M > P$). This constraint prevents the model from perfectly fitting the data, forcing it to capture the global trend.

In contrast, deep NNs are often *underdetermined* or *overparameterized*, meaning the number of parameters vastly exceeds the available training data points ($P \gg M$). Because a network with millions of parameters possesses enough mathematical flexibility to fit virtually any set of points exactly, it is prone to minimizing the loss function by memorizing the specific data

points—including random noise or numerical inaccuracies—rather than learning the true underlying relationship.

To prevent the parameters from adopting arbitrarily complex or large values that lead to unphysical predictions, *regularization* techniques are introduced into the training process. The most common form is L_2 regularization (or weight decay), which modifies the loss function by adding a penalty proportional to the square of the weights:

$$L_{\text{reg}} = L + \lambda \sum_l \|\mathbf{w}^{(l)}\|^2 \quad (1.173)$$

where λ is a regularization hyperparameter and $\|\mathbf{A}\|^2$ is the sum of the squared elements of the matrix $\mathbf{A}$. By penalizing the square of the parameters, this technique severely punishes large individual weights, forcing the optimization algorithm to distribute the values more evenly across the network. This prevents the model from relying excessively on any single feature, effectively constraining its overall flexibility and ensuring a smoother, more physically reliable output landscape.

Finally, in addition to the learnable parameters, ML models usually need a higher-level parametrization, in other words, the so-called “hyperparameters” must be chosen. Hyperparameters are not learned during the training, but are fixed from the beginning; they can influence the training efficiency and path (like the learning rate or the number of neurons in each layer of a NN), as well as the numerical stability of the optimization. It is important to choose appropriate values of the hyperparameters for the chosen model, and there are methods to systematically find good values, such as the Bayesian optimization [33].

1.5.6 Validation

To quantitatively assess whether a model has truly learned the underlying relationship or has simply memorized the training data, the available dataset must be partitioned into distinct subsets: the training set, the “validation set”, and optionally the “test set”.

During the optimization process, the network updates its weights and biases based exclusively on the errors evaluated on the training set. Simultaneously, the model’s performance is evaluated on the validation set, which consists of data points the network has never seen during backpropagation. This split allows for the detection of *overfitting*. Overfitting is characterized by a

divergence in performance: while the training loss continuously decreases toward zero, the validation loss reaches a minimum and subsequently begins to increase, indicating that the model is losing its capacity for generalization.

To exploit this behavior, the *early stopping* protocol is employed: throughout the training phase, the validation loss is monitored at every epoch. Optimization is artificially stopped at the exact point where the validation error reaches its absolute minimum, discarding subsequent iterations where overfitting would occur.

The remaining unseen data in the test set is then used as a final, unbiased benchmark to evaluate the model's true accuracy before production.

***k*-fold cross-validation**

The arbitrary splitting of the dataset can accidentally lead to biased assessments of the model's performance. Randomly shuffling the dataset could result in a lucky or unlucky alignment of data points, especially for small-sized datasets. To overcome this issue, cross-validation protocols have been developed; *k-fold* cross-validation is a frequent one. According to this protocol, the dataset is randomly partitioned into k equal-sized subsets (folds), and the model is trained k times; in the i -th iteration, the training is done on all data points but the i -th fold, which is instead used to validate the model, according to some error metrics. Finally, the k metrics values are collected and analyzed, either by averaging them, or taking the worst, for a more conservative approach. The model is then retrained on the whole dataset, with the previously obtained error metrics serving as an unbiased estimate of how well the model generalizes.

1.5.7 Machine Learning Interatomic Potentials

Machine Learning Interatomic Potentials (MLIPs) are particular applications of ML models for the prediction of the energy landscape of nuclei (PES), as a function of the nuclear positions. Machine Learning Interatomic Potentials revolutionized *ab initio* calculations, allowing the prediction of the ground state properties in a fraction of the time needed by first-principles algorithms, yet with very high accuracy.

Automatic differentiation allows MLIPs to predict not only the energy, but also its derivatives with respect to atomic positions, like forces and stress.

Representation

The primary challenge in applying machine learning to atomistic systems is representing the spatial arrangement of atoms in a format suitable for a model’s internal processing architecture. To achieve this, the raw Cartesian positions and chemical identities of all atoms within the simulation box are mapped onto specialized functions known as *descriptors*.

A common approach is to expand the atomic environment using a set of localized mathematical basis functions centered on each atom. The Atomic Cluster Expansion (ACE) framework [34] relies on this exact concept: the local neighbor density around a central atom is projected onto a basis composed of spherical harmonics multiplied by radial functions. These single-particle projections are subsequently combined into tensor products to explicitly capture high-order many-body correlations. Crucially, these cluster functions are mathematically coupled to ensure strict invariance under translation, rotation, and the permutation of identical chemical species. Finally, these symmetric many-body terms are combined via learnable coefficients to evaluate individual atomic contributions to the global target properties.

We will now briefly present two MLIPs that we used in our workflow.

Moment Tensor Potential (MTP)

Moment Tensor Potentials (MTPs), introduced by Shapeev [35], belong to the class of descriptor-based machine learning interatomic potentials. MTPs construct a local representation of the atomic neighborhood and systematically contract these representations to form invariant basis functions.

The description of the local environment of an atom i begins with the definition of moment tensors. For every neighboring atom j within a cutoff radius R_c , the relative position vector $\mathbf{r}_{ij}$ and scalar distance $r_{ij} = |\mathbf{r}_{ij}|$ are used to compute the moment tensor of rank ν :

$$M_{\mu,\nu} = \sum_j f_{\mu}(r_{ij}) \underbrace{\mathbf{r}_{ij} \otimes \cdots \otimes \mathbf{r}_{ij}}_{\nu \text{ times}} \quad (1.174)$$

where $f_{\mu}(r_{ij})$ are radial basis functions, typically defined as a combination of Chebyshev polynomials and a cutoff envelope that ensures a smooth decay to zero at R_c . The index μ denotes the radial function channel, and ν dictates the tensorial rank, which directly captures the angular distribution

of the neighbors.

Because the potential energy of a physical system must be invariant under global rotations and translations, these covariant tensors cannot be used directly to predict energy. Instead, invariant basis functions, B_α , are systematically generated by taking tensor dot products (contractions) of the moment tensors. For example, a rank-1 tensor can be contracted with itself to yield a scalar invariant (equivalent to a two-body or three-body term), while higher-order contractions of multiple tensors naturally encode complex many-body atomic interactions.

The total potential energy of the system is then obtained as a linear combination of these invariant basis functions:

$$E = \sum_{i=1}^N V(\mathcal{N}_i) = \sum_{i=1}^N \sum_{\alpha} \xi_{\alpha} B_{\alpha}(\mathcal{N}_i) \quad (1.175)$$

where $V(\mathcal{N}_i)$ is the local atomic energy contribution of atom i given its neighborhood $\mathcal{N}_i$, and ξ_{α} are the learnable linear regression coefficients.

The completeness and accuracy of the MTP are controlled by the "level" of the potential, which restricts the maximum complexity of the contracted basis functions. Due to its polynomial nature and linear fitting parameters, MTP provides an exceptional balance between computational efficiency and *ab initio* accuracy.

1.5.8 Message Passing Neural Network with Atomic Cluster Expansion (MACE)

While descriptor-based MLIPs like MTP are localized to a fixed cutoff radius, message-passing neural networks can extend the effective interaction range beyond a single cutoff by iteratively propagating information across the graph. MACE, developed by Batatia et al. [36], is a state-of-the-art local/medium-range MLIP that fuses the systematic many-body expansion of the ACE formalism with the expressive, iterated message-passing architecture of Message Passing Neural Networks (MPNNs). Standard MACE remains a local model: each message-passing layer aggregates information from neighbors within a fixed cutoff, and the effective receptive field grows with the number of layers. It therefore captures short- and medium-range many-body interactions very accurately, while true long-range effects such as electrostatics or dispersion generally require explicit long-range extensions or range-corrected variants.

In a standard MPNN, atoms are treated as nodes in a graph, and information (messages) is iteratively exchanged between neighboring nodes, allowing the network to learn increasingly complex many-body representations over multiple layers. MACE builds on this framework by using ACE-style equivariant messages, so that each layer updates atomic features using information from the local atomic environment.

During each message-passing layer t , the features of atom i , denoted as $\mathbf{h}_i^{(t)}$, are updated by aggregating messages from all neighbors j :

$$\mathbf{m}_{ij}^{(t)} = \phi \left(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t)}, \mathbf{r}_{ij} \right) \quad (1.176)$$

where the message function ϕ preserves the geometric equivariance of the features.

The critical innovation of MACE lies in how it processes these aggregated messages. Rather than relying solely on multi-layer perceptrons (MLPs) to mix features, MACE constructs a high-order atomic cluster expansion at every node within each layer. By taking tensor products of the equivariant messages and utilizing Clebsch-Gordan tensor product reductions, MACE systematically computes multi-body correlations (e.g., 3-body, 4-body interactions) explicitly in its architecture.

Owing to the mathematically rigorous inclusion of high-order many-body interactions within the network layers, MACE exhibits remarkable extrapolative capabilities. Consequently, MACE can model highly disordered systems, defect formations, and complex phase spaces—such as amorphous structures—with unprecedented fidelity, while maintaining robust generalizability beyond its training set.

Dataset generation

The most crucial element to keep in mind when working with Machine Learning methods is the correlation between the quality of the training data and the inference performance of the model. It is important to underline that the ML inference can have at most the same accuracy as that of the training set. Nowadays, DFT is considered the gold-standard for MLIPs solid-state training data.

Two abstract regions of the configurational space can be defined: the interpolation region, within which the properties of interest can be obtained by interpolating the training data point with the trained model, and the

extrapolation region, where the ML properties might significantly differ from the real ones. These regions can often be easily mapped back to a more geometrical classification. Depending on the curvature of the PES, data points close to each other in the configurational space have similar properties. Descriptors are usually designed to be flexible enough to adapt to local changes of gradient and curvature within the near-equilibrium region, but will fail if pushing the exploration toward steeper points, or very different local minima. Most of the time, MLIPs are required to accurately describe a narrow but significant region of the configurational space, rather than being able to span a wide domain, drastically losing interpolation accuracy. For this reason, it is common practice to finetune a pretrained model, already able to roughly describe a vast region, to specialize its predictions on a reduced zone of interest.

In any case, anyone who wants to train an MLIP must face the problem of data procurement. Thanks to the effort of many scientists, more and more large DFT datasets are now available and free to use, spanning a wide range of materials, like the Alexandria [37] and the MPtrj [38] datasets. However, high accuracy requires specific data for the specific class of materials under study. There are algorithms to generate relevant and *representative* data. We will see two examples, relevant for this thesis.

Machine Learning Assisted Canonical Sampling

The Machine Learning Assisted Canonical Sampling (MLACS) algorithm aims at generating a dataset representative of the canonical ensemble of a system. This is crucial for the extraction of finite-temperature thermodynamic and vibrational properties like free energy and anharmonic phonons. In principle, it is possible to use *ab initio* molecular dynamics to generate such data, but its massive computational cost over long propagation times makes it prohibitively expensive. To circumvent this bottleneck, MLACS relies on cheap Machine Learning Molecular Dynamics (MLMD) simulations coupled with a limited number of DFT calculations, to self-consistently generate configurations on the fly. For the derivation and formal justification of the variational principle sustaining the self-consistency, we refer to [39, 40].

The algorithm consists of the following iterative scheme:

1. Train an MLIP on an initial low-quality dataset
2. Run a short MLMD simulation

3. Select a few configurations from the trajectory and compute their DFT GS properties
4. Add the new configurations to the dataset and retrain the MLIP
5. Repeat from point 2 until convergence

Random distortions

A common way of generating new data when a small dataset is available, is to randomly displace and strain the preexisting cells. Note that the canonical ensemble is not exactly represented; this negatively affects the data efficiency of the model and its performance for NVT uses.

Active learning: committee uncertainty and embedding-diversity

One of the main sources of data inefficiency when training an MLIP is the presence of structures in the training set which do not bring significant information. Such data points slow down and bias the training towards a preferentially better fit of the regions they belong to in the configurational space, because these regions are overrepresented. Furthermore, including the “useless” structures in the dataset would still require calculating their DFT properties, which is then an additional computational cost. There are several selection strategies to avoid this inefficiency; we will see *committee uncertainty* and *embedding-diversity*.

Committee uncertainty has the goal of ranking the data points based on how much the model is able to infer their properties: the larger the error, the more information is added by including a structure, and the more relevant it is. However, to know the error of the MLIP prediction on a configuration, one has to calculate the reference (DFT) properties, which is undesired for non-relevant structures. To avoid running useless DFT simulations, and yet being able to rank the data points, a *committee* of models is created, where each member differs from the others by a non-physical parameter, like the seed for the random generator. This approach produces a spectrum of predictions for every data point, where the spread of the values reflects the disagreement among the committee members. While the variability in the prediction clearly depends and arises from shuffling the chosen parameter (seed) within the committee, uncertainty can still be ranked and their order relation can be ascribed to the intrinsic ability of the MLIP to infer the structures.

In order to validate this assumption, it is possible to perform a *calibration* of the committee on a smaller labeled dataset. The average error and the uncertainty of the committee are calculated for each data point, along with their Spearman rank correlation coefficients, which measure the correlation between the variations in the error and the variations in the uncertainty. If the Spearman coefficient is large enough, the committee can be safely used to rank structures based on uncertainty. Note that the labeled dataset must span a space similar to the unlabeled one.

Once the ranking is done, embedding-diversity can be used to further rank the structures based on their mutual distance in the descriptor space. Assuming each configuration can be represented by an N -dimensional vector in some embedding space, the vector $\mathbf{x}_i$ is calculated for each data point, and it is standardized:

$$\tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i - \bar{\mathbf{x}}}{\sigma} \quad (1.177)$$

where $\bar{\mathbf{x}}_i$ is the average vector over all structures in the dataset, and σ is an N -dimensional vector whose i -th component is the standard deviations of the i -th component of the vectors of all structures.

After (randomly) choosing a starting structure, $\tilde{\mathbf{x}}_0$, the most different structure is chosen as the one having the largest distance:

$$\tilde{\mathbf{x}}_1 = \operatorname{argmax} (|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_0|) \quad (1.178)$$

Then, for every remaining structure, the minimum distance from the “selected points” (that is the structures already put aside, in this case $\tilde{\mathbf{x}}_0$ and $\tilde{\mathbf{x}}_1$) is computed:

$$d_i = \min (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j) ; \tilde{\mathbf{x}}_j \in S \quad (1.179)$$

(where S is the set of selected points). The structure with the largest distance d is then added to the selection and the distances of the remaining structures are recomputed as in eq. 1.179 according to the newly augmented selection set. The order of addition of the structures in the selection set corresponds to their diversity ranking.

1.6 Temperature Dependent Effective Potential

In chapter 1.1, we first introduced lattice dynamics as a collection of vibrational excitations within the harmonic approximation, later extending this

description to finite-lifetime quasiparticles by incorporating the higher-order curvature of the Potential Energy Surface (PES). However, this perturbative approach lacks a fundamental component: temperature. At finite temperatures, larger vibrational amplitudes drive atoms far from the 0 K equilibrium geometry, causing the standard Taylor expansion around the reference minimum to break down. Because the system continuously explores regions of the PES where anharmonicity is strong—a mechanism that directly underpins macroscopic phenomena like thermal expansion—the atoms experience interactions that differ significantly from those at the bottom of the potential well. The Temperature Dependent Effective Potential (TDEP) method addresses this limitation by explicitly modeling the effective interatomic forces directly at finite temperatures.

The renormalization of interactions is achieved by defining an effective PES at a given temperature T :

$$U(\mathbf{u}, T) = U(0_T) + \frac{1}{2} \sum_{ij} \Phi_{ij}^{(2)}(T) u_i u_j + \frac{1}{6} \sum_{ijk} \Phi_{ijk}^{(3)}(T) u_i u_j u_k + \dots \quad (1.180)$$

where 0_T is the geometry of minimum-energy² in the unit cell of thermodynamic equilibrium, and $\Phi_{ij}^{(2)}(T)$ is the effective second-order IFC (IFC2) at temperature T (similarly for higher orders). Note that this expression is perfectly analogous to eq. 1.180.

2: Here we refer to the internal energy of the thermalized unit cell, including internal degrees of freedom.

The effective IFCs are optimized to fit the extent of the PES sampled at temperature T . To do so, a representative set of (*position, force*) pairs must be sampled, and the optimized IFCs are found by minimizing the sum of the squared errors between the reference forces, derived from the sampling, and those predicted by the model.

$$\{\Phi^{(n)}\} = \operatorname{argmin} \left[\sum_{I,i,\alpha} \left| \Delta F_{Ii\alpha}^{(n)} \right|^2 \right] \quad (1.181)$$

where I, i and α are the sampling, atom and Cartesian indices, respectively. The error on the forces, $\Delta F_{Ii\alpha}^{(n)}$ is computed according to a perturbative scheme: first the IFC2s are fit minimizing the error between the reference forces and those computed truncating the model at the second order. Subsequently, the IFC3s are fit by minimizing the error between the force residuals from the previous order and the 3rd-order force terms, $F_{Ii\alpha}^{(3)} = \sum_{j,k} \Phi_{ijk}^{(3)}(T) u_j u_k$. The procedure is analogous for higher orders. In this way, anharmonic effects are incorporated at the lowest order possible.

It is possible to include LO-TO corrections to IFCs by providing the Born effective charges and the dielectric tensor (see chapter 1), but in our case these are not renormalized with T .

The optimized sets of IFCs, $\{\Phi^{(2)}(T)\}$, $\{\Phi^{(3)}(T)\}$, ..., are directly related to spectral and thermal quantities at the given temperature T . As we saw in chapter 1.1.3, the phonon spectral function can be determined from the IFCs of order beyond the second, and can be used for any purpose, including the calculation of the Raman spectral peak shape (see A_λ in eq. 1.123).

Note on first-order optimization

TDEP works under the assumption that the system starts from a thermal equilibrium, and the thermalized unit vectors are known and fixed. Once the temperature-dependent cell vectors are available, the atom positions must be thermalized, namely their motion in time must determine a minimum average energy over time. Sometimes this is not ensured by the chosen sampling method; in such cases, a non-vanishing first-order term appears in eq. 1.180, and the framework breaks. To overcome this issue, we can fit the $\{\Phi_i^{(1)}\}$, which are simply the forces acting on the atoms at the given geometry, and move the atoms accordingly. By reiterating this until convergence of the positions, we finally obtain the actually relaxed positions. Note that in these steps the unit cell vectors are fixed, and only the reference atomic positions are updated. The new unit cell is then used to run the TDEP fitting procedure, with no first-order terms.

Note on convergence

The calculation of atomic forces in the sums of eq. 1.180 is practically implemented by applying a cutoff radius around every atom for each perturbation order ($rc2$, $rc3$, ...). The higher the cutoff radii, the more interactions are included in the fitting procedure. There are also other parameters to be fixed before fitting, like the size of the sampling.

The size and proportions of the supercell should also be chosen carefully, as they determine the maximum cutoff radius (the minimum image convention is adopted). For example, a unit cell which is very elongated along z , with layers in the xy plane, might need a large number of repetitions along the plane, so as to extend the cutoff to fully include an acceptable number

of interlayer interactions, and adequately describe out-of-plane phonon modes.

It is crucial to converge the final quantities (phonon dispersion bands, DOS, spectral function, thermal conductivity, etc.) with respect to each parameter.

1.6.1 Sampling for TDEP: Molecular Dynamics

One way of sampling representative (*position-force*) points at the desired temperature is through a Molecular Dynamics in the NVT ensemble (fixed unit cell shape and volume). Classical MD lacks the quantum phonon occupation, hence it fails to capture the residual anharmonicity at low temperature. Furthermore, beyond the occupation issue, nuclear quantum effects (NQE) are also neglected, which can be significant for light atoms, like Hydrogen. However, provided that NQEs are not dominant, and for high temperature (classical limit), the canonical space can be sampled with MD. Note that, in order to provide relevant information to the effective potential model, snapshots should be decorrelated, namely only a subset should be used as sampling (e.g. every τ fs, with τ being larger than the decorrelation time τ_d).

While *ab initio*-MD (AIMD) is the most accurate way of evolving positions in time (classically), Machine Learning-driven MD (MLMD) is cheaper and potentially not much worse in terms of accuracy. It is possible to approximate with ML at different levels; for instance, one can run the MD with ML force fields, and only compute the DFT properties of the decorrelated subset of snapshots, or the inferred ML ground state forces could be directly fed to TDEP.

The target quantities should be converged with respect to the length of the simulation, and the sampling stride of the trajectory or the decorrelation time.

1.6.2 Sampling for TDEP: Self Consistent Stochastic TDEP, sTDEP

Instead of classical MD, a relevant configurational sampling can be obtained by the random generation of structures from an effective canonical ensemble at a given temperature, followed by the computation of their exact ground-state forces. Such stochastic generation is achieved by determining

the Probability Density Function (PDF) that governs the thermodynamics of the system. This statistical approach to phase-space sampling is fundamentally connected to a broader family of self-consistent techniques designed to capture strong anharmonicity. Notable among these are the stochastic Self-Consistent Harmonic Approximation (SSCHA) [41, 42] and earlier, foundational variants of Self-Consistent Phonon (SCPH) [43] theory, which rely on similar probabilistic principles by variationally optimizing an effective harmonic density matrix to minimize the system's free energy.

Based on statistical mechanics, the PDF of the canonical ensemble is given by:

$$f(\mathbf{u}) = \frac{1}{Z} \exp(-\beta U(\mathbf{u})) \quad (1.182)$$

where $\mathbf{u}$ are the displacements, $\beta = k_B T$, U is the energy at every point in the phase space, and $Z = \int \exp(-\beta U(\mathbf{u}))$ is the partition function. By replacing U with the Taylor expansion we performed earlier, truncated at the second order, we obtain an analytical expression of the PDF which depends on the IFCs at all orders. The current TDEP implementation, which computes the PDF and generates structures from it, truncates the PES at the second order. In this case the PDF becomes:

$$f(\mathbf{u}) \propto \sum_{ij} \exp\left(-\beta \Phi_{ij}^{(2)} \mathbf{u}_i \mathbf{u}_j\right) \quad (1.183)$$

which can be rewritten in the phonon mode space (pure quadratic terms), yielding a multivariate gaussian distribution. The Bose-Einstein phonon occupation can also be enforced, to correctly represent the zero-point oscillation amplitudes.

This approach presents a major challenge: to accurately describe the PDF, it requires the temperature-dependent IFC2s, which are the very first quantities TDEP aims at determining. To solve this self-consistent approach, an iterative procedure is used.

1. From an initial guess of the IFC2s, compute the initial PDF $f^{(0)}$.
2. Generate a certain number N_0 of configurations, $\{S\}_0$ by randomly sampling the position space according to $f^{(0)}$.
3. Compute the DFT ground state energies and forces of $\{S\}_0$ and use it to fit TDEP IFC2s, $\{\Phi^{(2)}\}_0$.
4. Compute a new PDF, $f^{(1)}$ with $\{\Phi^{(2)}\}_0$.
5. Generate N_1 structures according to $f^{(1)}$: $\{S\}_1$.
6. Compute the DFT energies and forces of $\{S\}_1$, and fit TDEP: $\{\Phi^{(2)}\}_1$.

7. Repeat until convergence of the IFCs or target quantities.

The number of configurations to generate at each iteration, $N_0, N_1, \dots$, should be small at the beginning, when the PDF is roughly approximated, and can grow with the number of iterations. Note that the first-order optimization is not mandatory, but might be necessary at every sTDEP iteration, when computing the new IFC2s. A minor concern in optimizing the structure at each iteration should be considered: if LO-TO correction is needed for the IFC2s to use in the calculation of the PDF, both the dielectric tensor and the Born effective charges (BECs) change upon rearrangement of atoms, and should be recomputed. However, it is usually not necessary, because the first-order optimization is supposed to be just a refinement, and the atomic positions should not change dramatically. On average and to first approximation the dielectric properties remain quite constant with T (in most systems, exceptions include ferroelectrics and high- k dielectrics).

For more details and formal justification, we refer to [44].

1.7 Characterization of amorphous materials

Because amorphous networks lack periodicity, characterization cannot rely on reciprocal-space band structures or single-crystal symmetry operations. Instead, we deploy a topological and chemical analysis framework:

1.7.1 Radial and Angular Distributions

We compute partial Radial Distribution Functions (RDF, $g_{\alpha\beta}(r)$) and bond Angle Distribution Functions (ADF, $P_{\alpha\beta\gamma}(\theta)$) to determine characteristic interatomic bond lengths, cutoffs, and angular distribution profiles [45]:

$$g_{\alpha\beta}(r) = \frac{V}{4\pi r^2 N_\alpha N_\beta} \sum_{i=1}^{N_\alpha} \sum_{j=1}^{N_\beta} \delta(r - |\mathbf{r}_i - \mathbf{r}_j|), \quad (1.184)$$

where V is the supercell volume, and N_α, N_β are the counts of species α and β .

The first minima of the partial RDFs define the coordination sphere cutoffs ($r_c^{\alpha\beta}$) for the subsequent angular distribution characterization. The triplet

angular distribution $P_{\alpha\beta\gamma}(\theta)$, centered on a central atom of species β , is defined as:

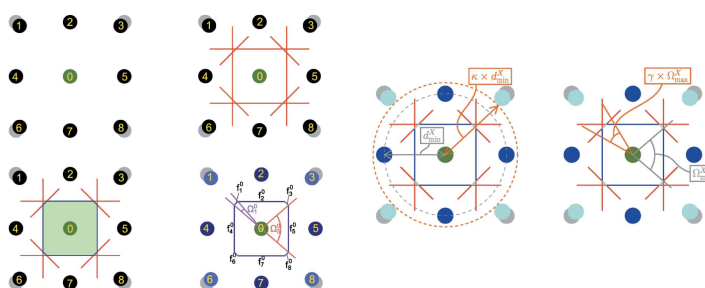
$$P_{\alpha\beta\gamma}(\theta) = \sum_{j \in \Omega_\beta} \sum_{\substack{i \in \Omega_\alpha \\ i \neq j \\ r_{ij} < r_c^{\alpha\beta}}} \sum_{\substack{k \in \Omega_\gamma \\ k \neq i \\ k \neq j \\ r_{jk} < r_c^{\beta\gamma}}} \delta(\theta - \theta_{ijk}) \quad (1.185)$$

where θ_{ijk} is the angle formed by the vectors $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and $\mathbf{r}_{jk} = \mathbf{r}_k - \mathbf{r}_j$, and Ω_x is the set of atoms belonging to species x .

1.7.2 Local Coordination Environments: ChemEnv Analysis

We used the ChemEnv algorithms and package [46] for the coordination analysis. The environment of each atom, determined through a Voronoi analysis, is scored with a continuous measure of similarity (CSM) to a list of coordination polyhedra. For a given central atom, the polyhedron associated with the lowest CSM (that is, largest similarity) is considered to represent the best description of the coordination. This is the simplest way, or *strategy*, to classify local environments. ChemEnv offers several more sophisticated strategies, based on the combination of different sets of weights for each polyhedron, which return more complete results. For further details on ChemEnv and the underlying algorithms, we refer to the corresponding publication [46].

Figure 1.5: Voronoi analysis and cutoff criteria employed by ChemEnv. Adapted from "ChemEnv: a fast and robust coordination environment identification tool" by Waroquiers *et al.*, 2020, Acta Cryst. - Structural Science Crystal Engineering Materials, doi.org/10.1107/S2052520620007994. CC BY 2.0.



1.7.3 Intermediate-Range Order: Ring Statistics

While the RDF and ADF successfully describe the short-range order, they are largely insensitive to the network's topology at longer length scales. To properly characterize the medium-range order in amorphous networks, we rely on the connectivity of atomic rings. Ring statistics have been used

extensively to decode the topological structure of amorphous silicon, silica, and complex chalcogenide glasses [47–49].

In theoretical terms, an amorphous supercell can be mapped onto a mathematical graph where vertices represent atoms and edges represent chemical bonds (defined by the cutoff distances $r_c^{\alpha\beta}$ extracted from the first minimum of the RDFs). A “ring” is simply a closed loop of bonded atoms. However, enumerating rings in a highly connected, periodic three-dimensional network is topologically ambiguous. Several distinct definitions have been established in the literature to avoid overcounting overlapping loops:

- *King’s shortest-path rings*: The oldest and computationally simplest definition. A ring is identified as the shortest closed path originating from a given node and returning to it [47]. While straightforward, this method can artificially obscure larger, physically meaningful rings if a smaller ring shares the same central node.
- *Guttman rings*: A more restrictive definition where a ring is considered valid only if the shortest path between *any* two of its constituent nodes lies exclusively along the perimeter of the ring itself [50]. This prevents the inclusion of rings that possess internal “shortcuts”.
- *Primitive (or irreducible) rings*: The most rigorous and widely adopted definition for covalent networks. A primitive ring is a closed path that cannot be decomposed into a sum of two or more smaller closed paths [48, 51]. This definition ensures that only the fundamental rings are counted.

For our structural analysis, we focus on the distribution of primitive rings, as they provide the most accurate representation of the network’s fundamental connectivity.

The resulting ring size distribution, $R_c(n)$, counts the number of valid rings containing n nodes.

1.7.4 Vibrational Properties: Vibrational Density of States (vibDOS)

By computing the mass-weighted velocity autocorrelation function (VACF) from microcanonical (NVE) equilibrium trajectories:

$$C_v(t) = \frac{\sum_i m_i \langle \mathbf{v}_i(t) \cdot \mathbf{v}_i(0) \rangle}{\sum_i m_i \langle \mathbf{v}_i(0) \cdot \mathbf{v}_i(0) \rangle} \quad (1.186)$$

and evaluating its Fourier transform, we extract the total and partial vibrational density of states ($D(\omega)$). Because the MD supercell captures finite-wavevector vibrations, this approach inherently yields the total phonon density of states integrated over the entire Brillouin zone. In the crystalline phase, the vibrational density of states $D(\omega)$ features a well-defined structured profile punctuated by distinct phononic band gaps and localized maxima. In the amorphous phase, the diversity in local coordination, bond lengths, and bond angles causes these features to smear out. Consequently, tracking the evolution of $D(\omega)$ allows us to observe how the separated vibrational bands of crystalline MoS₂ broaden, merge, and redistribute into a continuous spectrum as topological disorder increases.

Temperature-dependent Raman spectra of 2H-MoS₂

2

2.1 Introduction and motivation

MoS₂ is the best known molybdenum sulfide; it consists of layers of Mo atoms surrounded by S atoms in a trigonal prism coordination. Adjacent prisms have the same orientation and share an S-S edge. While neighboring Mo and S atoms form covalent bonds, interlayer van der Waals interactions allow for the formation of soft stacks. Owing to the labile interlayer bonds, layers can shear with relatively low effort; this is the reason why MoS₂ is suited for lubricant formulations and can be obtained with different numbers of layers by exfoliation.

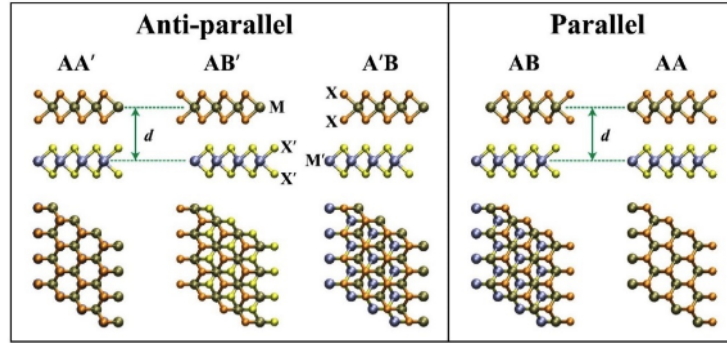
Raman spectroscopy stands out as one of the most powerful, non-destructive characterization techniques for transition metal dichalcogenides (TMDs) like MoS₂. Because the optical phonon frequencies are highly sensitive to the restoring forces between atoms, Raman scattering routinely provides precise structural fingerprints, probing layer thickness, applied strain, defect density, and stacking configurations. Beyond static structural characterization, it also serves as a direct probe of the material's dynamical and thermal properties. In modern nanoscale applications, MoS₂-based devices inevitably experience finite-temperature effects and local heating, making it critical to understand how the lattice dynamics evolve away from the absolute zero limit.

Understanding this thermal response requires moving beyond the standard harmonic approximation. At finite temperatures, the harmonic picture—where phonons propagate independently with infinite lifetimes—fundamentally breaks down due to lattice anharmonicity. As atoms undergo larger amplitude displacements, higher-order terms in the potential energy surface become significant, allowing phonons to interact, decay, and exchange energy.

Anharmonicity manifests in the Raman spectrum through two primary observables. First, the resonant frequencies shift due to both the thermal expansion of the lattice (volume effects) and the real part of the anharmonic phonon self-energy (dynamic effects). Second, the Raman peaks experience intrinsic broadening. This linewidth expansion is a direct reflection of the finite phonon lifetimes dictated by the imaginary part of the self-energy,

as phonons scatter via three-phonon and higher-order processes. Accurately predicting these temperature-dependent frequency shifts and spectral broadenings requires an explicit, rigorous treatment of the anharmonic potential. Capturing these many-body interactions across a massive configurational phase space is computationally prohibitive with standard *ab initio* methods, motivating the integration of Machine Learning Interatomic Potentials (MLIPs) with advanced finite-temperature sampling techniques utilized in this work.

Figure 2.1: Common stacking geometries of transition metal chalcogenides. Reproduced from "Anisotropic Interlayer Force Field for Group-VI Transition Metal Dichalcogenides" by Jang *et al.*, 2023, arXiv:2307.11331v2. CC BY 4.0.



MoS₂ can be synthesized and found naturally in a variety of stacking conformations. Each stacking lends the material unique electronic and spectroscopic properties, also because of the change in the symmetry of the system. For example, while the “H” stacking (see AA’ in Fig. 2.1) leads to a centrosymmetric unit cell, the “AA” stacking does not, inducing some selection rules in the vibrational modes that can be excited with IR or Raman spectroscopy. Furthermore, depending on the shift between planes, different stackings exhibit different interlayer distances; this has a considerable effect on the vibrational frequencies, and thus carries a clear spectral fingerprint.

Among the molybdenum disulfide phases, this chapter will focus on the most stable one: the 2H stacking as a bulk.

2.2 2H-MoS₂

2H-MoS₂ is a semiconductor with an indirect gap of ~1.23 eV [52], and its optical properties, including the band gap, strongly depend on the number of layers [53, 54]. 2H-MoS₂ belongs to the centrosymmetric $P6_3/mmc$ space group, with a unit cell containing six atoms distributed across two layers. These layers are arranged in an AA’ stacking sequence, where the prime indicates a 180° in-plane rotation of the second layer relative to the first.

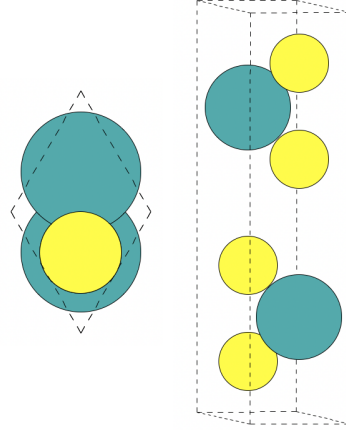


Figure 2.2: MoS₂ unit cell. The lattice vectors and positions are from Materials Project (mp-2815).

Our unit vectors are defined as:

$$\mathbf{a} = a \left(\frac{\sqrt{3}}{2} \mathbf{e}_x - \frac{1}{2} \mathbf{e}_y \right) \quad (2.1)$$

$$\mathbf{b} = a \left(\frac{\sqrt{3}}{2} \mathbf{e}_x + \frac{1}{2} \mathbf{e}_y \right) \quad (2.2)$$

$$\mathbf{c} = c \mathbf{e}_z \quad (2.3)$$

By symmetry requirement, only three degrees of freedom are present: the two cell parameters, a and c , and the vertical distance between each S atom and the Mo atom bonded to it, δ . The latter, then, is the only parameter needed to define the relative position of every atom in the primitive cell:

$$\begin{aligned} \text{S}_1 &: \left(\frac{1}{2}, \frac{1}{\sqrt{12}}, \frac{1}{4} - \delta \right) \\ \text{Mo}_1 &: \left(\frac{1}{2}, -\frac{1}{\sqrt{12}}, \frac{1}{4} \right) \\ \text{S}_2 &: \left(\frac{1}{2}, \frac{1}{\sqrt{12}}, \frac{1}{4} + \delta \right) \\ \text{S}_3 &: \left(\frac{1}{2}, -\frac{1}{\sqrt{12}}, \frac{3}{4} - \delta \right) \\ \text{Mo}_2 &: \left(\frac{1}{2}, \frac{1}{\sqrt{12}}, \frac{3}{4} \right) \\ \text{S}_4 &: \left(\frac{1}{2}, -\frac{1}{\sqrt{12}}, \frac{3}{4} + \delta \right) \end{aligned}$$

Note that this unit cell belongs to the D_{6h} point group.

Our goal is to compute the Raman spectrum of 2H-MoS₂ at finite temperature, explicitly accounting for anharmonicity. While previous theoretical studies have successfully evaluated the vibrational properties of MoS₂, they

have predominantly relied on the standard harmonic approximation via Density Functional Perturbation Theory (DFPT) or restricted their scope to zero-temperature perturbative corrections [55]. Accessing the true finite-temperature regime requires extensive phase-space sampling, but traditional approaches like *ab initio* Molecular Dynamics (AIMD) are computationally prohibitive for the large supercells and long simulation times required to converge Raman lineshapes. We overcome this limitation by deploying Machine Learning-driven statistical sampling. By replacing the costly *ab initio* force evaluations with a highly accurate Machine Learning Interatomic Potential (MLIP), we achieve rigorous, large-scale exploration of the finite-temperature configurational space. Coupling this extensive sampling with the Temperature Dependent Effective Potential (TDEP) method allows us to capture the exact anharmonic frequency shifts and broadenings that static models miss. As discussed in Chapter 1.2 (eq. 1.123), the simulated Raman spectrum is constructed from two essential building blocks: the fully anharmonic phonon spectral function and the Raman activity. The robust evaluation of these quantities using our ML-driven framework is the primary focus of this chapter.

2.3 Phonons at 0 K

We computed the phonon frequencies and eigenvectors at 0 K with DFPT (using Abinit). In Fig. 2.3 we show the phonon dispersion bands along the conventional high-symmetry path from [56]¹.

1: Note that our definition of the unit cell is not aligned to the cell from [56], thus a transformation must be applied.

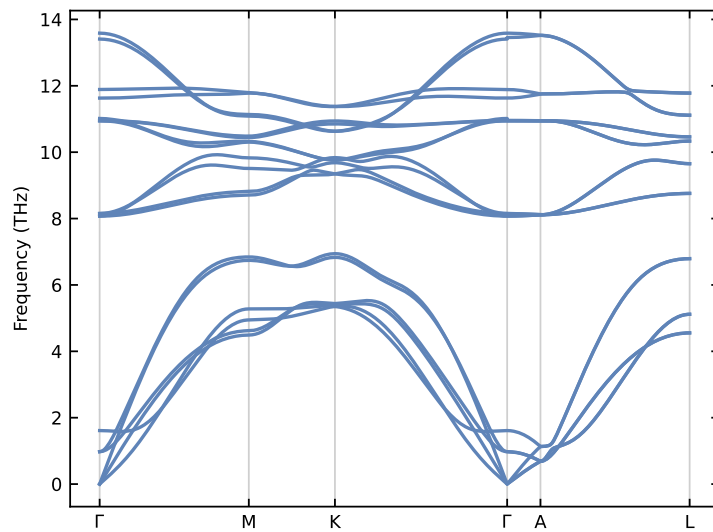


Figure 2.3: Phonon dispersion band of 2H-MoS₂ at 0 K computed with DFPT.

Note that the LO-TO correction for 2H-MoS₂ is very small; in fact, the Born

effective charges are $\sim 1 e$ and $\sim 0.5 e$ for Mo and S, respectively.

The vibrational mode representation, Γ_{vib} can be decomposed into a direct sum of the D_{6h} irreducible representations [55]:

$$\Gamma_{\text{vib}} = A_{1g} \oplus 2A_{2u} \oplus 2B_{2g} \oplus B_{1u} \oplus E_{1g} \oplus 2E_{1u} \oplus 2E_{2g} \oplus E_{2u} \quad (2.4)$$

As in all centrosymmetric systems, by symmetry, only modes that are symmetric with respect to the inversion center, denoted by “g”, can be Raman-active, and only the antisymmetric ones, denoted by “u”, can be probed with IR spectroscopy. Furthermore, only modes that transform like quadratic forms of the Cartesian axes are Raman active. According to the D_{6h} character table [57], only the irreducible representations A_{1g} , E_{1g} and E_{2g} do so; hence only the modes with the same symmetry are optically active in a Raman experiment.

Atomic motions for the Raman-active modes of 2H-MoS₂ are depicted in Fig. 2.4. According to the A_{1g} mode, S atoms oscillate in the c -direction (out of plane), while the Mo atoms stay still. In contrast, the E_{1g} and E_{2g} modes have polarization in the plane for all atoms.

Symmetry	Polarization	Exp. ω		DFPT ω	
		THz	cm ⁻¹	THz	cm ⁻¹
E_{1g}	In-plane	8.57	286	8.07	269.3
E_{2g}^1	In-plane	11.48	383	10.95	365.4
A_{1g}	Out-of-plane	12.23	408	11.89	396.6

Experimental frequencies for bulk 2H-MoS₂ are taken from literature [58]. The calculated values were obtained via DFPT at 0 K.

Table 2.1: Experimental and calculated frequencies for the Raman-active optical phonon modes of bulk 2H-MoS₂.

Because the E_{1g} mode has very low Raman activity, Raman experiments usually focus on the other two; following the literature, we excluded this mode from the study.

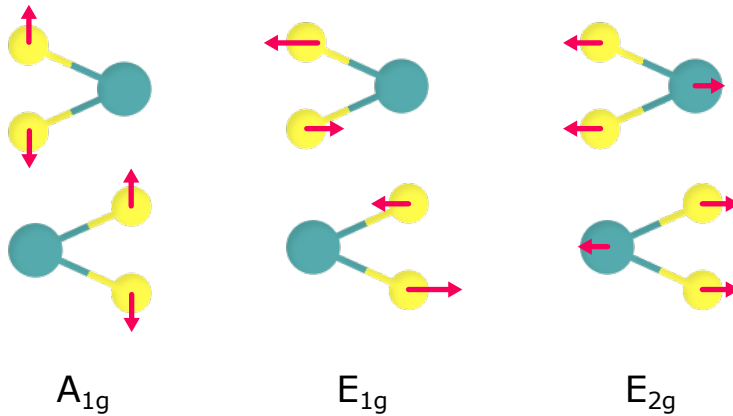


Figure 2.4: Schematic representation of the polarization of the Raman active modes in 2H-MoS₂.

2.4 Generation of a dataset and training of the MLIP (M1)

In order to train an MLIP for 2H-MoS₂, a good-quality and representative dataset is required; we generated one using the MLACS algorithm (see chapter 1.5.8) as implemented in [40]. In order to train an MLIP able to approximate the PES in a wide range of temperature, we randomly selected a value of T between 100 and 700 K at every MLACS iteration and used that value to run an NPT MD simulation at 0 pressure. For the first 10 iterations the MD length was set to only 100 fs, from the 11th iteration on, MD was run for 1000 fs. Of every MD trajectory, we always picked the last snapshot (more advanced selection criteria were not yet implemented in the package).

We used a supercell obtained through:

$$\mathbf{S} = \begin{pmatrix} 4 & 4 & 0 \\ -2 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{U} \quad (2.5)$$

where $\mathbf{U}_{ij}$ is the j -th component of the i -th unit cell vector (and similarly for the supercell $\mathbf{S}$). This produces an orthogonal supercell with 96 atoms.

We computed the reference ground state properties of the new configurations via DFT with the software ABINIT [59–62], at the Generalized Gradient Approximation (GGA) level of theory in the Perdew-Burke-Ernzerhof (PBE) formulation [63], with Grimme’s D3 van der Waals description [64] (cutoff set to 18.897 Bohr), and using the optimized norm-conserving Vanderbilt (ONCV) [65] fully relativistic pseudopotentials from the PseudoDojo database [66] (version 0.4, standard accuracy tier).

As MLIP, we used the Moment Tensor Potential [35] as implemented in [67], with the following hyperparameters: MTP level: 16, cutoff distance: 8 Å, 8 Chebyshev radial basis functions, maximum number of iterations: 1000, weights for energy, forces and stress: {1,1,1}.

After 339 configurations were generated, we performed a 10-fold cross-validation (see chapter 1.5.6) with increasing subsets of the dataset to assess whether the MLACS procedure had converged or more iterations were needed. In this process, we trained MTP with the same hyperparameters as during MLACS, except the MTP level, which was set to 20. The average RMSE over the 10 folds is shown in Fig. 2.5 (and the coefficient of determination

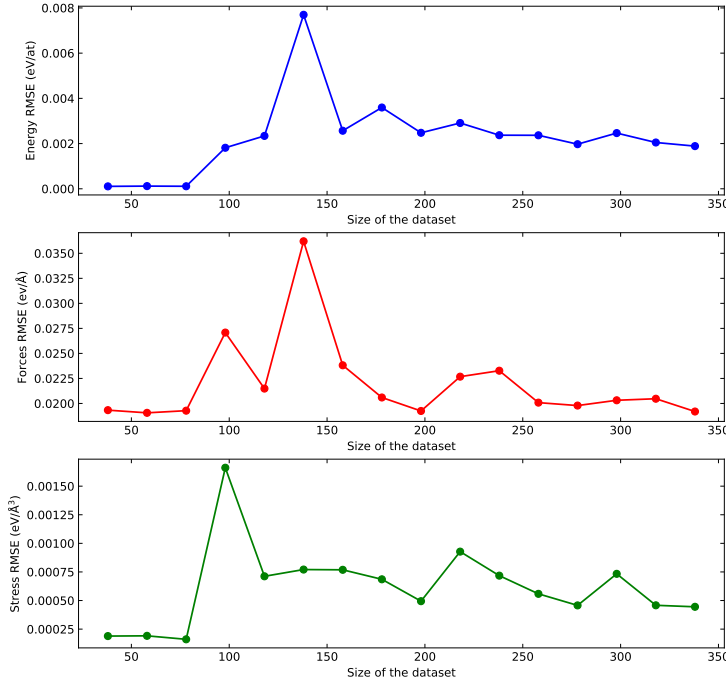


Figure 2.5: Convergence study of the RMSE of energy, forces and stresses across the MLACS procedure, with respect to the number of iterations, that is the size of the dataset (1 new structure at every iteration). The average of the RMSE is taken over the 10 folds used to cross-validate the MLIP.

in Fig. 2.6) for energy, forces, and stress as a function of the size of the dataset.

For the whole dataset, we obtained average RMSEs of 0.003 eV/atom, 0.022 eV/Å, and 0.0005 eV/Å³ for energy, forces, and stress, respectively. These error metrics indicate a highly converged and reliable potential. In the literature for machine learning interatomic potentials, energy errors below 5 meV/atom and force errors below 30 meV/Å are widely established as state-of-the-art benchmarks required to accurately reproduce the Born-Oppenheimer potential energy surface of solids [68]. More importantly, the force error of 22 meV/Å is sufficiently small to ensure that artificial noise does not wash out the physical thermal fluctuations driving the lattice dynamics. Because the faithful extraction of higher-order interatomic force constants in the TDEP formalism is highly sensitive to the accuracy of the forces, this level of precision is strictly necessary to guarantee the correct prediction of phonon lifetimes and the corresponding finite-temperature Raman spectra.

A final training was then performed on the whole dataset (instead of leaving a fold out), obtaining our production MLIP, which will be denoted by “**M1**”. The assessment of its inference capabilities on the training set is reported in the parity plots in Fig. 2.7. We used **M1** to replace ground state DFT in every following step, including MD simulations and sTDEP.

To corroborate the trained model, we performed a cell optimization with

Figure 2.6: Convergence study of the R^2 coefficient for energy, forces and stresses across the MLACS procedure, with respect to the number of iterations, that is the size of the dataset (1 new structure at every iteration). The average of the R^2 is taken over the 10 folds used to cross-validate the MLIP.

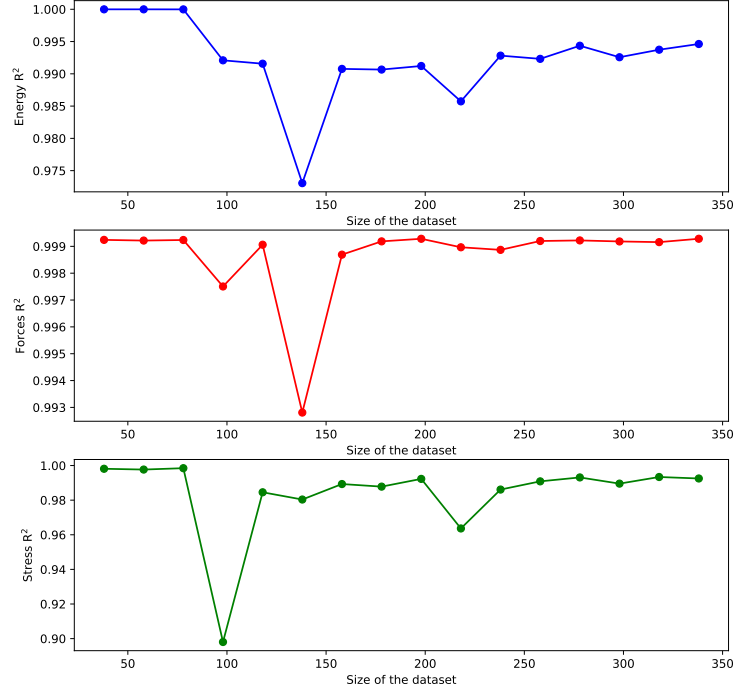
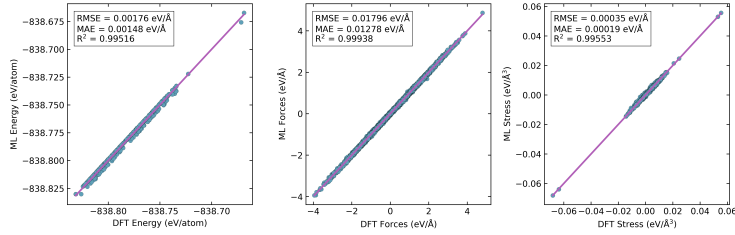


Figure 2.7: Parity plots of the MLIP M1 for energy, forces and stresses. The inference was done on the training set.



DFT and MLMD; the cell parameters obtained with MLMD differ from those obtained from DFT by less than $0.00012 \text{ \AA}$ and $0.027 \text{ \AA}$ for a and c , respectively.

All MD simulations for 2H-MoS₂, for the dataset generation and TDEP sampling, were performed with the LAMMPS software [69].

2.5 Cell thermalization

We determined the lattice parameters of the unit cell at 100, 200, 300, 400, 500, 600 and 700 K as the thermal average over NPT evolutions at 0 pressure. For each temperature, we ran 40 independent NPT simulations sharing the same starting point, discarding the first 5 ps of equilibration and averaging over 95 ps. Then, we averaged the cell parameters over the 40 NPT instances. We monitored the convergence of the cell parameters with respect to the number of NPT simulations by computing differences of a and c obtained with 1, 2, 3 and 4 instances less, and averaging them. Let i be the number of

NPT instances ($\in [4, 40]$); we evaluate the averages $\tilde{\Delta}_i^a$ and $\tilde{\Delta}_i^c$:

$$\tilde{\Delta}_i^a = \frac{1}{4} \sum_{j=i}^4 \left| \frac{1}{i} \sum_{k=1}^i \tilde{a}_k - \frac{1}{i-j} \sum_{k=1}^{i-j} \tilde{a}_k \right| \quad (2.6)$$

$$\tilde{\Delta}_i^c = \frac{1}{4} \sum_{j=i}^4 \left| \frac{1}{i} \sum_{k=1}^i \tilde{c}_k - \frac{1}{i-j} \sum_{k=1}^{i-j} \tilde{c}_k \right| \quad (2.7)$$

where $\tilde{a}_k$ is the thermal average of a for the k -th instance and similarly for c).

Since the trend of $\tilde{\Delta}_i^a$ and $\tilde{\Delta}_i^c$ vs the number of instances could be influenced by the order in which the instances are added, we repeated the averaging procedure for 10 shufflings of the instances and averaged the trends.

We considered 40 instances to be a converged value for every temperature, with an average absolute difference below 10^{-5} Å (see Fig. 2.8).

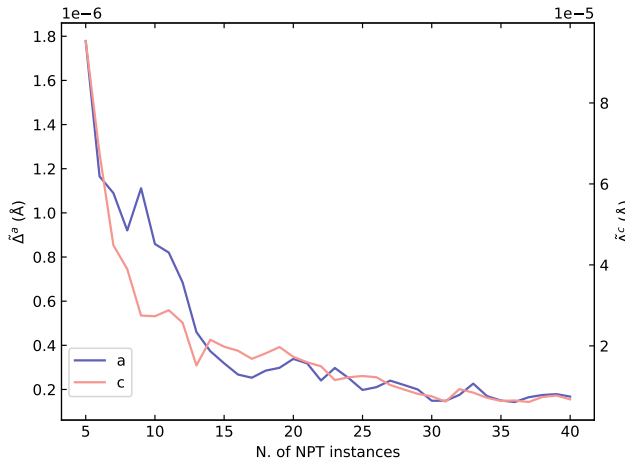
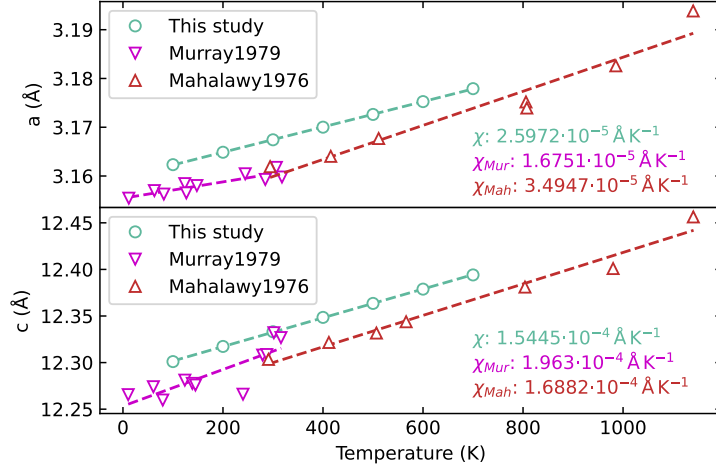


Figure 2.8: Convergence of the residuals $\tilde{\Delta}^a$ and $\tilde{\Delta}^c$ vs. the number of NPT iterations.

The lattice parameters that we obtained at every temperature are shown in Fig. 2.9, as compared to experimental measurements from [70, 71]. As expected, both parameters grow upon heating, with a linear coefficient of $2.5972 \cdot 10^{-5}$ Å/K and $1.5445 \cdot 10^{-4}$ Å/K for a and c respectively. Our calculated parameters align very well with experiments, with a respective overestimation of only $\sim 0.16\%$ and $\sim 0.2\%$; this is well within the expected error of PBE on cell parameters [72, 73].

It is very important to attain accurate estimation of the equilibrium lattice constants, as they represent the central geometry of the Taylor expansion of the PES, and therefore all derived quantities depend on them.

Figure 2.9: Lattice parameters of 2H-MoS₂ vs temperature from our study and experimental work by El-Mahalawy *et al.* [70] and Murray *et al.* [71]. The data was fit to a linear function in T (dashed lines), and χ are the angular coefficients.



2.6 Phonons at finite temperature

In order to fit effective Interatomic Force Constants (IFCs) with TDEP and compute phonons and their spectral function at finite temperature, we sampled the canonical ensemble with classical MD (see chapter 1.6.1) and with sTDEP (see chapter 1.6.2), to compare the results. For both methods, we built our supercell according to eq. 2.5, with the multiplicity matrix:

$$\begin{pmatrix} 8 & 8 & 0 \\ -8 & 8 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

which yielded an orthogonal box with 2304 atoms. This choice was the result of a careful convergence study. Due to computational constraints, we prioritized in-plane convergence over the out-of-plane dimension. This choice is physically justified by the mathematical limitations of the TDEP formalism, where the maximum permissible interaction cutoff radius is dictated by the largest sphere that can be inscribed within the supercell boundaries. Because 2H-MoS₂ possesses a naturally large c/a ratio, this interaction cutoff is strictly bottlenecked by the in-plane dimensions. Consequently, expanding the supercell along the basal plane to drive the simulation box toward a cubic geometry is the optimal strategy for maximizing the spatial range of the extracted force constants. While elongating the supercell along the c -axis would yield additional symmetry-equivalent atomic environments for statistical averaging, this benefit is ultimately redundant in our workflow; the robust, large-scale configurational sampling enabled by the MLIP independently guarantees stringent statistical convergence without requiring unnecessary out-of-plane repetitions.

We carried out a convergence study to ensure that our choice of supercell size was sufficient to yield reliable results. Let us adopt the following notation for the multiplicity matrix:

$$mmn \equiv \begin{pmatrix} m & m & 0 \\ -m & m & 0 \\ 0 & 0 & n \end{pmatrix} \quad (2.8)$$

We extracted the IFC2s and IFC3s at 400 K for the supercells obtained with the matrices [333, 444, 555, 666, 772, 882, 883]. We adopted the maximum cutoff for the second-order cutoff (rc2) for every supercell, while keeping the third-order one (rc3) fixed, rc3=7 Å. In this regard, we emphasize that every supercell size allows for a different maximum cutoff. The extraction of the IFCs was done with an MD sampling (see chapter 2.6.1), including the LO-TO correction. The IFCs were then used to compute the spectral function (with third order scattering) at Γ , which served as our convergence criterion. As shown in Fig. 2.10, we evaluated the convergence

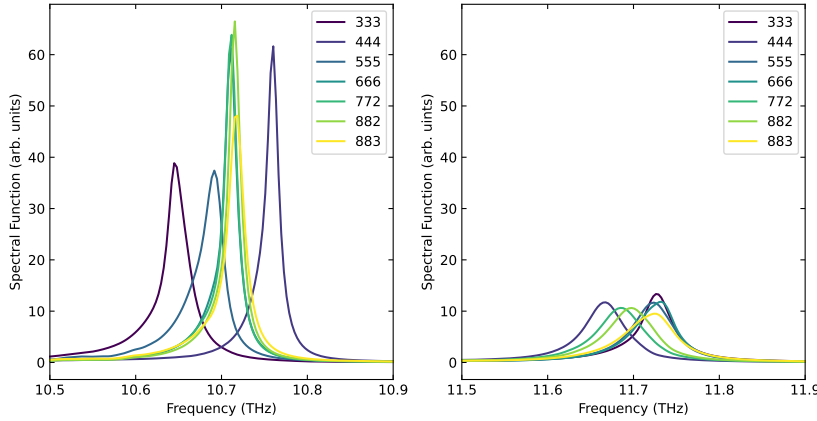


Figure 2.10: Spectral functions of the A_{1g} (right) and E_{2g} (left) modes for different supercell sizes.

of the spectral functions for the Raman-active E_{2g}^1 and A_{1g} modes across the different supercell dimensions. For the in-plane E_{2g}^1 mode, the peak position is already fully converged in the $6 \times 6 \times 6$ supercell, while the linewidth exhibits an expected marginal variation upon increasing the out-of-plane repetitions ($8 \times 8 \times 2 \rightarrow 8 \times 8 \times 3$). Because reciprocal-space integration is evaluated on a fixed dense grid and the third-order cutoff radius is held constant at 7 Å, this minor broadening arises from the residual renormalization of the second-order force constants by extended long-range contributions, which marginally adjusts the baseline for the residual extraction of the third-order force constants. The A_{1g} mode, which involves out-of-plane atomic displacements and is governed by interlayer van der Waals interactions, exhibits a naturally higher sensitivity to the out-of-plane

supercell geometry. However, comparing the largest configurations ($8 \times 8 \times 2$ to $8 \times 8 \times 3$), the residual discrepancy in the calculated peak position is approximately 0.02 THz, and the variation in the spectral linewidth is strictly bounded below 0.02 THz. These residual numerical errors fall within the standard experimental resolution of Raman spectrometers (0.03 THz [74]) and are significantly smaller than the intrinsic thermal broadening at finite temperatures. Furthermore, because van der Waals forces decay rapidly with interlayer separation, extending the supercell along c without a proportional in-plane expansion cannot accommodate larger real-space interaction shells under the minimum-image convention. Consequently, any remaining finite-size effects are physically negligible, rendering the $8 \times 8 \times 3$ representation sufficiently converged for accurately predicting finite-temperature Raman lineshapes.

Note that the self-energy we used to calculate the spectral functions only accounts for phonon-phonon interactions of the third order, namely two phonons coalescing into a new one and a phonon decomposing into two new phonons. Inclusion of higher orders is discussed later, while interactions between phonons and other (quasi-) particles, like electrons, are not the subject of our work. Furthermore, since the Raman experiment only probes phonons at the center of the Brillouin zone, we computed the spectral functions at $\mathbf{q} = \Gamma$.

2.6.1 MD sampling

We used Machine-Learning Molecular Dynamics (MLMD) to collect a representative sampling for the extraction of IFCs between 100 and 700 K. To compute the forces at each time step we used our trained MTP model, **M1**. For each temperature, we started from the thermalized 883 cell and we ran an NVT MLMD simulation with Langevin thermostat and Grønbech-Jensen/Farago integrator [75]. We let the system equilibrate for 50 ps, and then we simulated 850 ps of evolution, with a time step of 1 fs. Snapshots were saved every 50 steps.

To assess convergence with respect to the sampling size, we calculated the spectral functions of the E_{2g} and A_{1g} modes using a fixed $64 \times 64 \times 64$ $\mathbf{q}$ -point integration grid, with the second- and third-order cutoff radii set to 12.65 Å and 5 Å respectively. For this convergence analysis, we subsampled the full 850-ps NVT trajectory by extracting one frame every 200 fs (i.e., selecting

every fourth saved step). This stride yielded a total of 4250 structures, deliberately striking a balance between minimizing statistical auto-correlation between adjacent frames and reducing computational overhead. However, for the final production calculations presented in the subsequent sections, we removed this subsampling and utilized the full resolution of the saved trajectory. While processing the complete dataset is computationally more expensive, it maximizes the configurational phase space provided to the TDEP algorithm for the final extraction of the interatomic force constants.

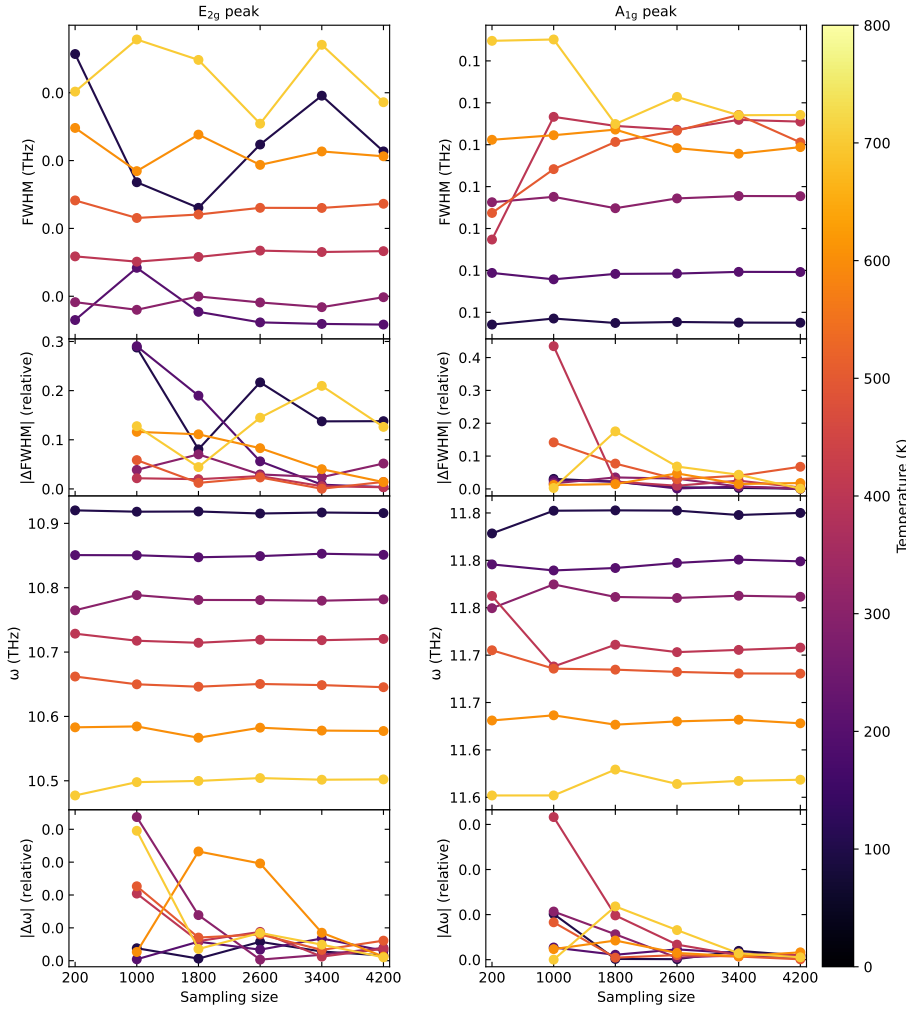


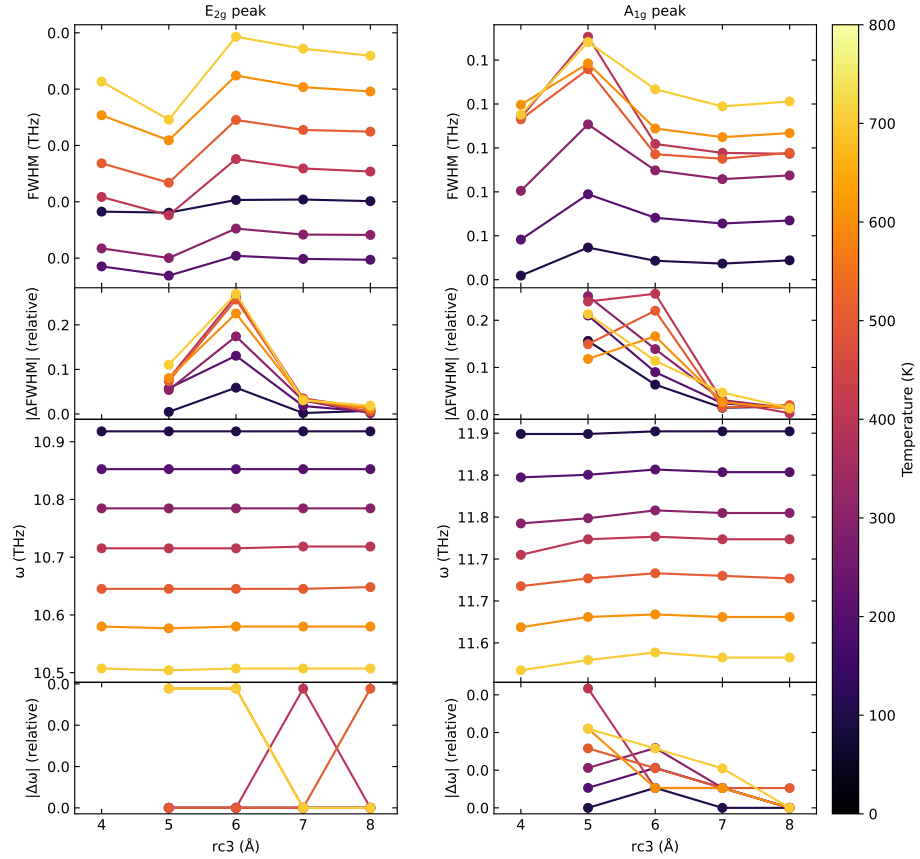
Figure 2.11: convergence study of E_{2g} and A_{1g} peaks in the spectral function of 2H-MoS₂ obtained with MD-TDEP vs. the sampling size.

In Fig. 2.11, we present the convergence of the FWHM and peak center of the spectral functions as a function of the configurational sample size. We observe that the linewidth of the E_{2g} mode converges more slowly at 100 K and 700 K. This behavior is primarily attributed to these temperatures lying at the extreme boundaries of our MLIP training set; at these limits, slightly larger uncertainties in the predicted forces naturally translate into broader statistical fluctuations during the extraction of the interatomic force constants. Conversely, the FWHM of the A_{1g} mode and the peak frequencies

for both modes demonstrate robust convergence across all investigated sample sizes.

We also monitored the spectral functions with varying third-order cutoff radii, as depicted in Fig. 2.12. For both peaks and at all temperatures, $rc3=8$ Å led to well converged peaks, in terms of both width and position.

Figure 2.12: Convergence study of E_{2g} and A_{1g} peaks in the spectral function of 2H-MoS₂ obtained with MD-TDEP vs. the third-order cutoff radius ($rc3$).



Once the convergence study was done, we proceeded with the fit of IFCs on the whole saved trajectory (17000 structures) with $rc2=12.65$ Å (which was the maximum cutoff radius our super cell allowed for) and $rc3=8$ Å. We used TDEP to interpolate the IFC2s on a $64 \times 64 \times 64$ q-grid and computed the phonons and the DOS at all temperatures, as shown in Fig. 2.13. We note that all branches shift towards lower frequencies upon increasing the temperature. This is due to the atoms oscillating with larger amplitudes. Because the restoring force is smaller at larger interatomic distances (i.e. the PES is anharmonic), atoms experience lower restoring forces, on average, taking more time to complete a full period of oscillation, i.e. vibrate with lower frequency. At the same time, the average distance becomes larger, and this causes the crystal to expand. The volume and the frequencies are related by the Grüneisen parameter, $\gamma = V d\omega/dV$.

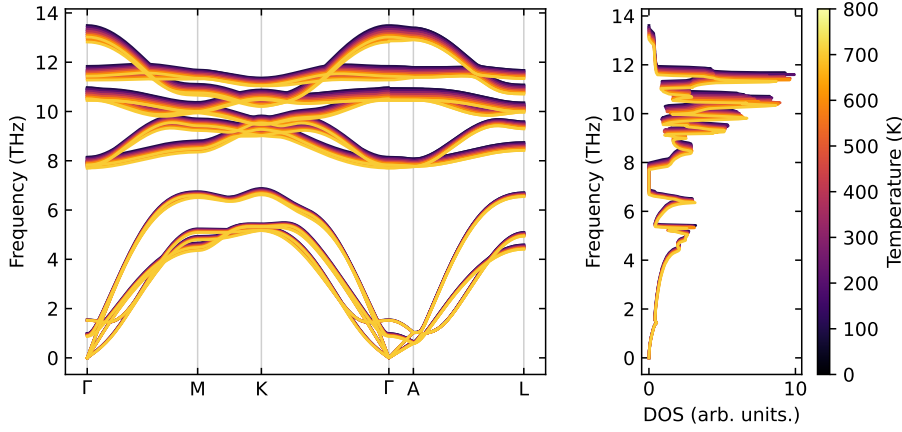


Figure 2.13: Phonon dispersion bands and density of states (DOS) of 2H-MoS₂ between 100 and 700 K obtained with MD-TDEP.

We also computed the spectral functions of the E_{2g} and A_{1g} modes, which are displayed in Fig. 2.14.

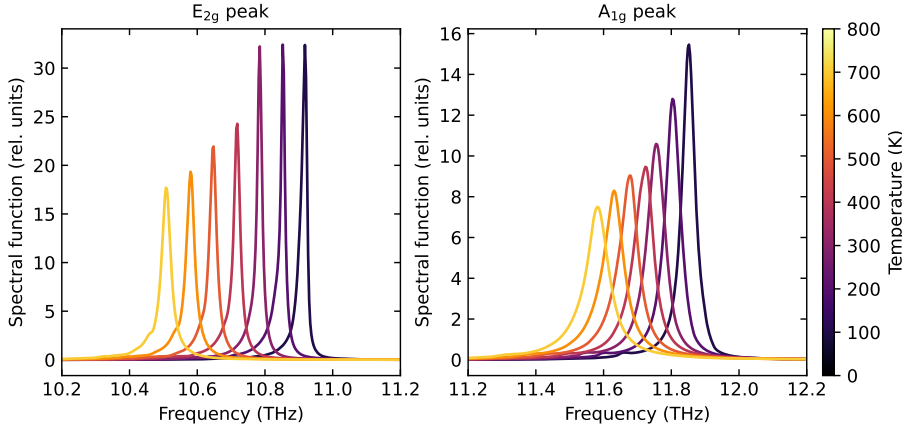
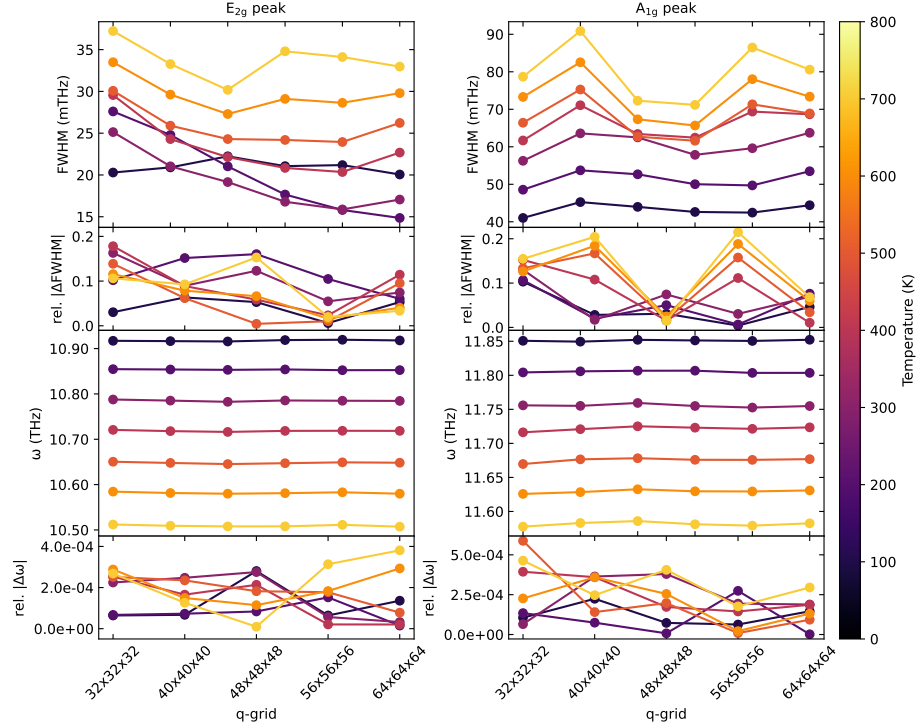


Figure 2.14: Spectral functions of 2H-MoS₂ obtained with MD-TDEP vs. temperature. The left panel is related to the E_{2g} mode, while the right panels is related to the A_{1g} mode.

We performed the summation in eq. 1.101 on a $64 \times 64 \times 64$ q-point grid, after assessing the convergence of FWHMs and frequencies with respect to the grid density (the convergence study is reported in Fig. 2.15).

The relevant peaks exhibit a physical red-shift with temperature, consistent with what was previously found for the phonon dispersion bands. Furthermore, for both modes the peaks broaden with temperature, as a result of the increasing probability of scattering upon heating. However, for the E_{2g} mode at low temperature the broadening does not exhibit a continuous growth upon heating. Normally, since the phonon population increases with temperature, the probability of scattering events increases as well, and the phonon lifetimes—roughly proportional to the inverse of the FWHM of the spectral function peaks—diminishes accordingly. However, there is another factor that plays an important role in the extent of space of allowed events: the vibrational frequencies also change with temperature. In order to understand whether this is relevant in our case, we computed

Figure 2.15: Convergence study of the FWHMs of E_{2g} and A_{1g} peaks in the spectral function obtained with MD-TDEP.



the three-phonon density of states (3P-DOS), that is a measure of how many allowed scattering events can happen with the production or annihilation of a phonon at a certain frequency. For the production/annihilation of phonons at Γ , the three-phonon density of states is given by:

$$P_3(\omega) = \frac{1}{N} \sum_{\mathbf{q}} \sum_{ij}^{3N_{at}} \left[(n_i(\mathbf{q}) + n_j(\mathbf{q}) + 1) \delta(\omega - \omega_i(\mathbf{q}) - \omega_j(\mathbf{q})) + 2(n_i(\mathbf{q}) - n_j(\mathbf{q})) \delta(\omega - \omega_i(\mathbf{q}) + \omega_j(\mathbf{q})) \right] \quad (2.9)$$

where $n_i(\mathbf{q})$ is the Bose-Einstein occupation factor for the phonon with wavevector $\mathbf{q}$ and that belongs to the branch i .

The 3P-DOS at all temperatures are shown in Fig. 2.16. We note that, while in the vicinity of the A_{1g} frequency the 3P-DOS separate well at different temperatures, in the region between 10.5 and 11 THz, namely the E_{2g} domain, the 3P-DOS oscillates, making the distinction in temperature much harder. This can explain at least partially the uneven trend observed in the peak widths.

2.6.2 Stochastic sampling

We sampled the canonical space of 2H-MoS₂ at every temperature of interest according to the sTDEP algorithm described in chapter 1.6.2. As initial guess

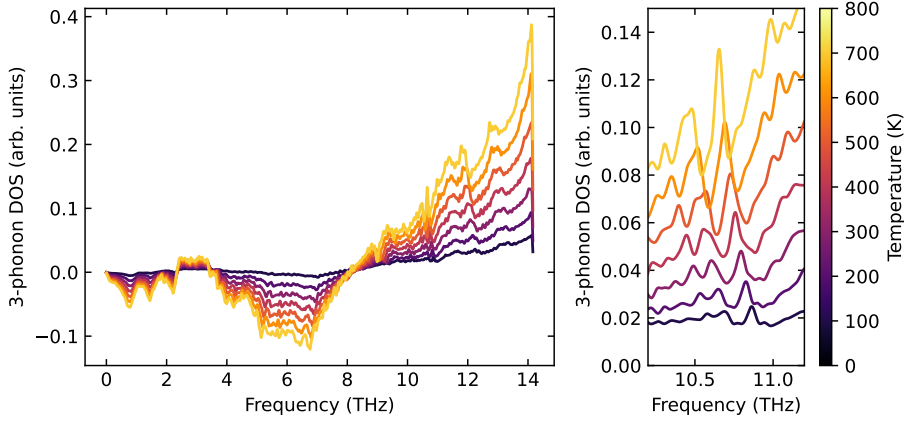


Figure 2.16: 3-phonon density of states of 2H-MoS₂ obtained with MD-TDEP vs. temperature.

for the IFC2s we use a tailored unconditionally stable pair potential: for any pair of atoms i, j separated by a distance vector $\mathbf{r}_{ij}$, a radial potential is defined by its first and second derivatives evaluated at the equilibrium geometry:

$$\left. \frac{\partial U}{\partial r_{ij}} \right|_{eq} = 0 \quad (2.10)$$

$$\left. \frac{\partial^2 U}{\partial r_{ij}^2} \right|_{eq} = \frac{\gamma}{r_{ij}^4} \quad (2.11)$$

where $r = |\mathbf{r}_{ij}|$ and γ is a parameter to optimize. It is easy to show that the interatomic force constant related to the force acting on atom i along α due to a displacement of atom j along β , $\Phi_{i\alpha j\beta}$, is given by:

$$\frac{\partial^2 U}{\partial u_{i\alpha} \partial u_{j\beta}} = -\frac{\gamma}{r_{ij}^6} (\mathbf{r}_{ij})_{\alpha} (\mathbf{r}_{ij})_{\beta} \quad (2.12)$$

With these tentative IFC2s (and the geometry of the cell), the maximum vibrational frequency can be calculated as a function of γ , and *vice versa*. Hence, γ is chosen to make the maximum frequency match a realistic value. We set the maximum frequency to 15 THz. Note that this does not need to be too accurate, as it is just a starting point; from the next iterations real IFC2s are calculated (with TDEP, see chapter 1.6.2) and used to calculate the PDF. Once some new configurations were generated, the atomic forces were computed with the **M1** model, for them to be used in the TDEP fitting in order to determine the updated IFC2s. We limited the second-order interactions to the maximum cutoff, that is 12.65 Å, for the 883 supercell. The number of configurations generated randomly from the PDF for every sTDEP iteration is reported in Fig. 2.17. As anticipated in chapter 1.6.2, since at the beginning the PDF is not accurate, only a few configurations are

generated; as the PDF becomes more representative of the model system, the number of new configurations is increased.

To assess the global convergence of the sTDEP procedure with respect to the number of iterations, we extracted the IFC3s (with $rc3 = 5 \text{ \AA}$) and monitored both the peak centers (ω) and the full width at half maximum (FWHM) of the spectral functions for the E_{2g} and A_{1g} modes. The results are reported in Fig. 2.18.

Overall, the peak centers exhibit excellent numerical stability, converging with a relative error strictly below 0.1% and 0.2% for the E_{2g} and A_{1g} modes, respectively, across all temperatures. For the spectral linewidths, we observe a disparity in the statistical sampling requirements between the two modes. The FWHM of the out-of-plane A_{1g} mode rapidly stabilizes, with the relative error between successive iterations dropping below 0.4%. In contrast, the degenerate in-plane E_{2g} modes exhibit persistent, low-amplitude statistical fluctuations across the iterative process.

Rather than requiring strict asymptotic flatness—which is practically unattainable due to the intrinsic statistical noise inherent to sampling the large in-plane configurational phase space—we define convergence through a strict physical criterion. Evaluating the absolute magnitude of these residual fluctuations reveals that, by the 27th iteration, the variation in the E_{2g} FWHM is robustly bounded within 0.025 THz. Because this residual uncertainty is firmly below the 0.03 THz resolution limit of standard Raman spectrometers and significantly smaller than the intrinsic thermal broadening of the mode, additional iterations offer no physically meaningful refinement to the simulated spectrum. Consequently, the procedure was truncated at 27 iterations, as it fully satisfies the physical accuracy bounds required to reliably predict the finite-temperature Raman lineshapes.

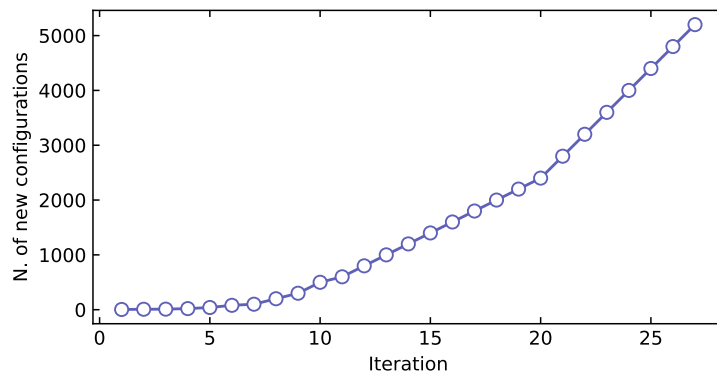


Figure 2.17: Number of configurations generated at each sTDEP iteration.

As shown in Fig. 2.17, at the 27th sTDEP iteration, 5200 new structures were

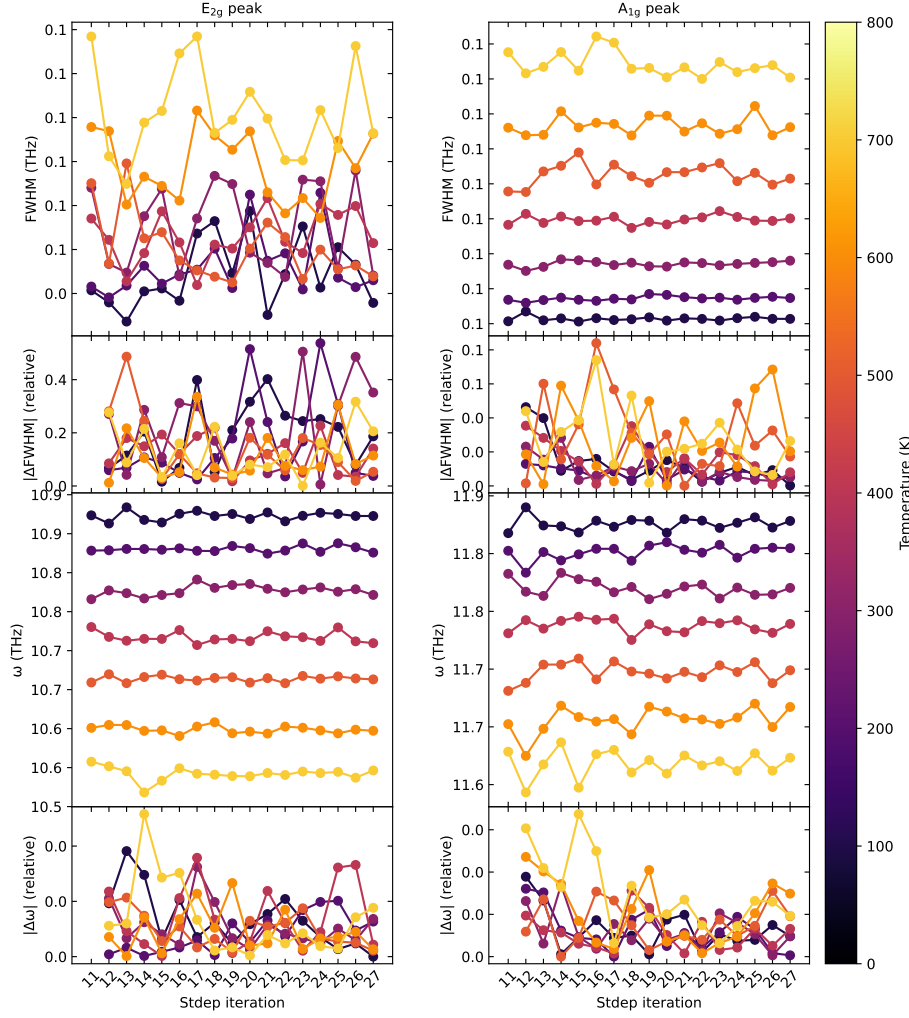


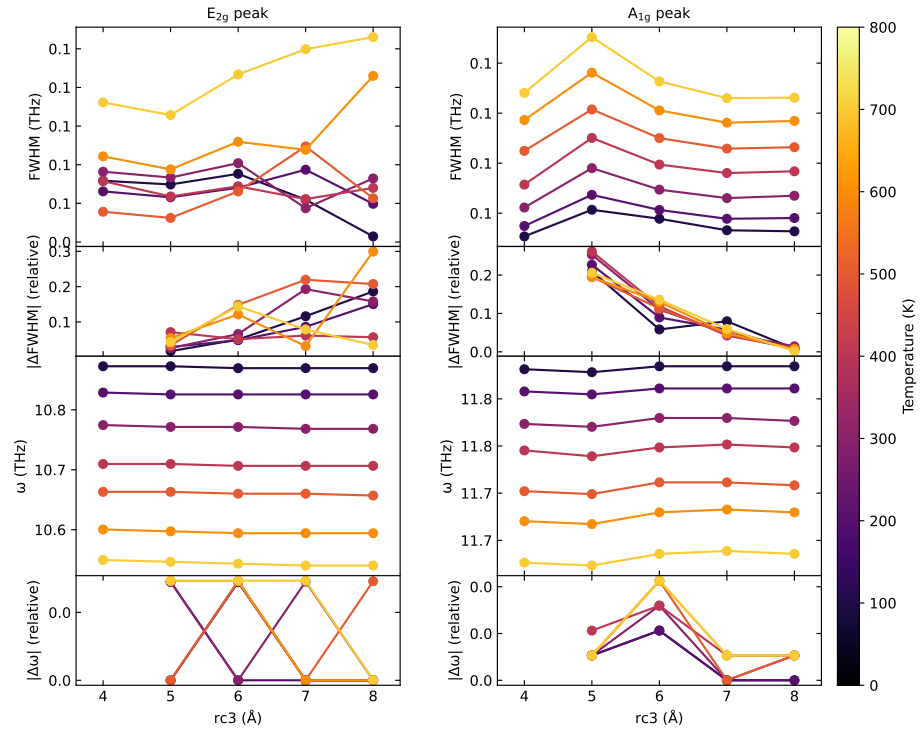
Figure 2.18: Convergence of FWHMs and peak centers of the peaks in the spectral functions of the E_{2g} and A_{1g} modes vs the sTDEP iterations, between 100 and 700 K. The relative absolute error at the i -th iteration is defined as the absolute difference between the FWHM at iteration i and that at iteration $i - 1$, divided by the latter (analogously for the center peak ω).

generated, serving as the configurational dataset for the final interatomic force constant (IFC) extraction. We determined the second- and third-order force constants using cutoff radii of $rc2 = 12.65 \text{ \AA}$ and $rc3 = 8 \text{ \AA}$. The second-order cutoff inherently corresponds to the maximum interaction range permitted by the supercell dimensions under the minimum-image convention. To validate the sufficiency of the third-order cutoff, we monitored the behavior of the spectral function peak centers and linewidths across a range of $rc3$ values, as reported in Fig. 2.19.

The peak centers (ω) demonstrate immediate and robust convergence, with relative errors bounded below 0.003% and 0.0005% for the E_{2g} and A_{1g} modes, respectively. For the spectral linewidths, the FWHM of the out-of-plane A_{1g} mode exhibits a strictly converging trend, achieving a relative absolute error below 0.11% at all temperatures. While the sensitive in-plane E_{2g} mode exhibits a monotonic drift at larger $rc3$ values, an analysis of the absolute variation reveals that this effect is physically marginal. Specifically,

expanding the third-order cutoff up to 8 Å induces a maximum absolute variation in the E_{2g} linewidth of approximately 0.0187 THz. Because this variation remains strictly below the typical resolution limit of experimental Raman setups, any higher-order long-range interactions beyond 8 Å are physically inconsequential to the final observable lineshapes. Consequently, setting $rc3 = 8$ Å represents an optimal and physically rigorous truncation, capturing all relevant anharmonic scattering channels without attempting to fit purely residual noise from weakly interacting distant neighbors.

Figure 2.19: convergence study of E_{2g} and A_{1g} peaks in the spectral function of 2H-MoS₂ obtained with sTDEP vs. the third-order cutoff radius ($rc3$).



We extracted the IFCs from the dataset from the 27th sTDEP iteration, with $rc2=12.65$ Å and $rc3=8$ Å. In Fig. 2.20 we report the phonon band dispersions and the DOS at each temperature. Again, we observe the typical red-shift of frequency caused by thermal expansion and the renormalization of phonons due to anharmonicity.

In Fig. 2.21 we show the normalized spectral functions. We note that the spectral functions are not symmetric. In finite-lifetime phonon treatments, the spectral function is often assumed to be a Lorentzian with a finite broadening, which is symmetric far from the peak. This is a good approximation only close to the renormalized harmonic frequency or with a constant self-energy. When the frequency-dependent self-energy is used, the spectral function becomes asymmetric, generally speaking, regardless of the position of its peak. As expected, the peaks of both modes generally broaden with temperature; however, surprisingly, the E_{2g} mode exhibits a slightly

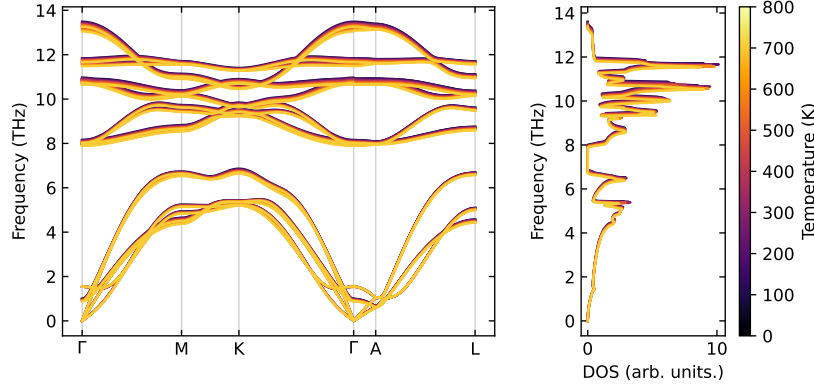


Figure 2.20: Phonon dispersion bands and density of states (DOS) of 2H-MoS₂ between 100 and 700 K obtained with sTDEP.

non-monotonic change in width for intermediate temperatures (300, 400 and 500 K).

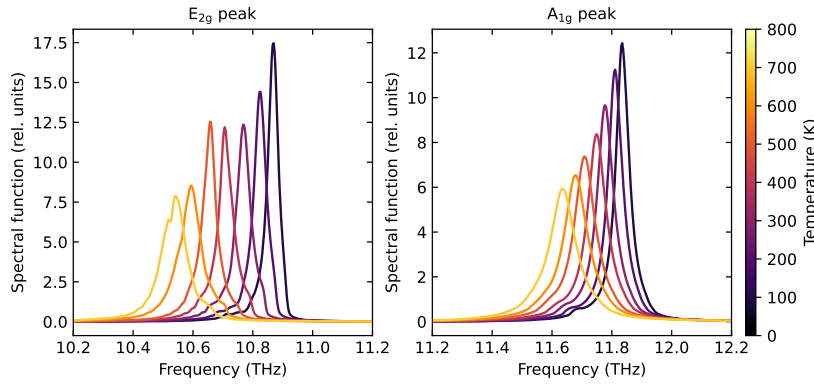


Figure 2.21: Spectral functions of 2H-MoS₂ obtained with sTDEP vs. temperature. The left panel is related to the E_{2g} mode, while the right panels is related to the A_{1g} mode.

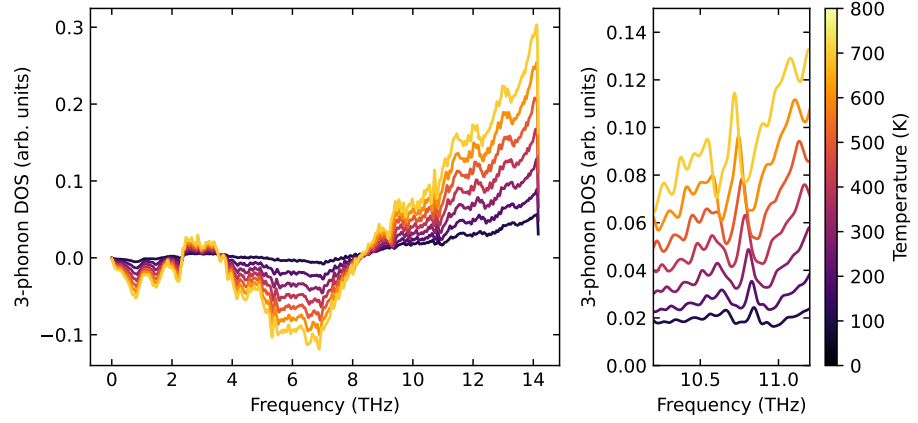
As we did for the previous mode, we computed the 3P-DOS to better understand the role of renormalization of the phase space. In Fig. 2.22 we report P_3 for all temperatures. Again, the 3P-DOS oscillates between 10.5 and 11 THz, disrupting the expected monotonic evolution of the broadening in temperature.

2.6.3 Comparison and experimental reference

In this section we will compare the results obtained, with the MD sampling and the stochastic one, to experimental references.

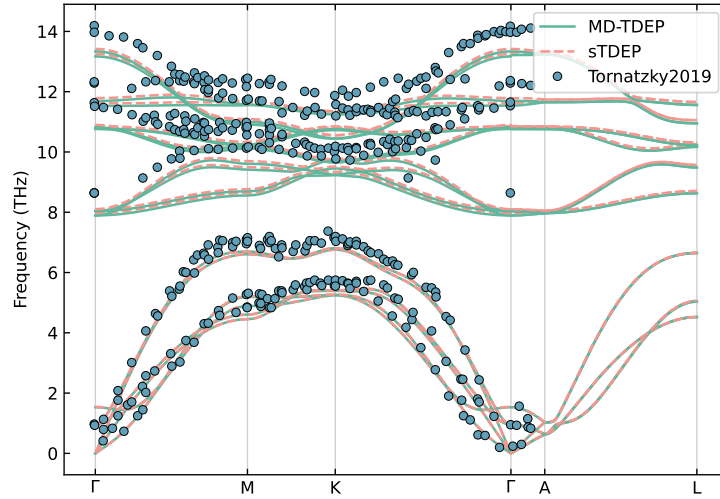
In Fig. 2.23 the phonon dispersion at 300 K is presented, as obtained with MD-TDEP, sTDEP and from the experimental work of Tornatzky *et al.* [76]. The two sampling methods gave similar bands, and both underestimated the frequencies as compared to the experimental ones, by a factor which increases with the frequency itself. This does not come as a surprise, as GGA functionals are known to systematically produce softer modes. This

Figure 2.22: 3-phonon density of states of 2H-MoS₂ obtained with sTDEP vs. temperature.



is also consistent with our slight overestimation of the lattice parameters shown in Fig. 2.9.

Figure 2.23: Phonon dispersion bands of 2H-MoS₂ as obtained at 300 K with MD-TDEP and sTDEP, and from the experimental work by Tornatzky *et al.* [76]



We compared the FWHM of the spectral functions calculated with the two sampling methods and experiments. We extracted the FWHMs directly from our calculated spectral functions, while experimental ones were obtained by Cao *et al.* [77] as the FWHM of a Lorentzian distribution function they used to fit their experimental Raman spectra. Similarly, our frequencies represent the position of the maximum of each peak in the spectral function, while the experimental frequencies correspond to the centers of the Lorentzian fitting curves. Both quantities are shown in Fig. 2.24.

We performed a linear fit of the frequencies vs temperature for both modes; the fitting coefficients are reported in Table 2.2. While our frequencies decrease with temperature for both sampling methods, sTDEP led to smaller thermal softening than MD sampling. Since the same unit cell was used for both MD-TDEP and sTDEP², this discrepancy must be ascribed to the quantum phonon occupation in the stochastic method, which has a finite

2: The first-order optimization performed for sTDEP had negligible effects on the frequencies.

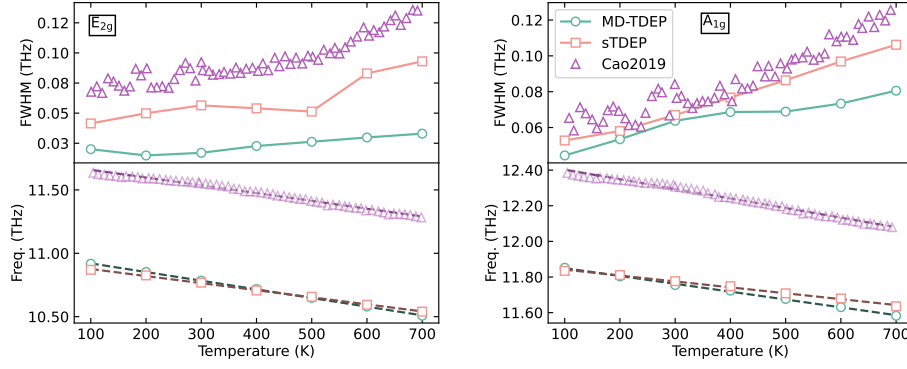


Figure 2.24: Frequencies and FWHMs of the E_{2g} and A_{1g} modes vs. temperature, as obtained with MD-TDEP and sTDEP, and from the experimental work of Cao and Chen [77].

offset at 0 K, and increases more slowly than it does in the classical case (MD sampling). The rate of change in the occupation is related to that of the mean square displacement, which determines how much more anharmonicity is sampled due to an increase in temperature. The phonon linewidth of the E_{2g} mode appears as a sensitive probe of quantum occupation effects.

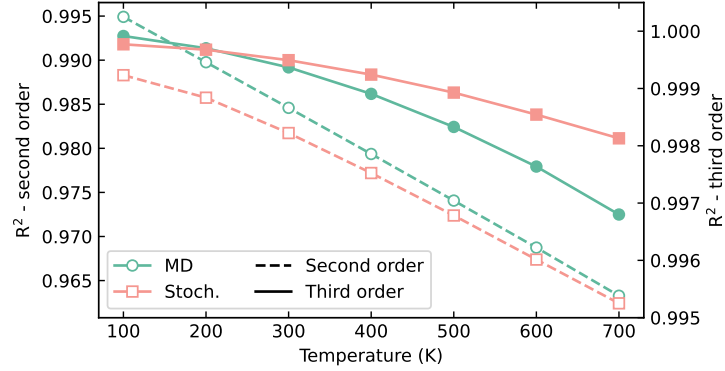
	E_{2g}		A_{1g}	
	χ_0	χ_1	χ_0	χ_1
MD-TDEP	1.099	-6.835	1.189	-4.401
sTDEP	1.093	-5.581	1.187	-3.313
Cao2019[77]	1.172	-6.144	1.246	-5.389

Table 2.2: Linear parametrization of the temperature dependent Raman shifts from this study and experimental data in the literature. The χ_0 and χ_1 coefficients are given in $\times 10$ THz and 10^{-4} THz/K, respectively.

As for the FWHMs, both sampling methods produced smaller widths than those measured experimentally. This discrepancy could be ascribed to interactions beyond the third order, which are not accounted for in our calculations, and to additional experimental broadening sources, such as inhomogeneity of the sample and noise in measurements. The experimental factors are very hard to rule out, and make a strict quantitative comparison a hard task. Nevertheless, our general trend vs. temperature is similar to the experimental one. The FWHMs obtained with the MD and stochastic sampling follow a similar, generically increasing trend in temperature. However, sTDEP leads to systematically broader peaks than MD sampling. We decided to investigate the difference between the two sampling methods by comparing the IFCs fit error metric. In Fig. 2.25 we show the R^2 of the fitting at the second and third order. Because the two methods attain different displacement amplitudes, they also sample different anharmonicity extents. In general, higher anharmonicity pushes down the R^2 . At low temperature, the quantum occupation ($1/2 + n(T)$) induces a larger displacement, hence more anharmonic forces, relevant for the sTDEP case, leading to a lower

R^2 for this sampling method. This effect persists until roughly the Debye temperature. At high temperature, the MD effectively explores asymmetric displacements (while sTDEP is still bound to Gaussian—hence symmetrical—displacements), leading to a faster decreasing R^2 than sTDEP, as it can be observed at the third order.

Figure 2.25: Determination coefficient, R^2 of the TDEP fitting procedure vs temperature, at the second and third order, from MD and stochastic sampling.



To further understand the role of anharmonicity, it would be useful to include fourth order IFCs in the picture. Currently, the TDEP framework only allows for the extraction of the fourth-order imaginary self-energy evaluated at the harmonic phonon frequency, $\Gamma_{\lambda}^{(4)}(\omega_{\lambda})$. Obtaining the complete frequency-dependent fourth-order self-energy is computationally demanding and not yet implemented; similarly, its real counterpart $\Delta_{\lambda}^{(4)}(\omega)$ remains inaccessible at any ω . As a first approximation, we decided to incorporate the available fourth-order contribution as a frequency-independent shift (a rigid translation) applied to the imaginary part:

$$\Sigma_{\lambda} \approx \Delta_{\lambda}^{(3)}(\omega) + i \left(\Gamma_{\lambda}^{(3)}(\omega) + \Gamma_{\lambda}^{(4)}(\omega_{\lambda}) \right) \quad (2.13)$$

In principle, one could attempt to isolate the pure fourth-order contribution to the phonon self-energy by combining outputs from two distinct tools in the TDEP suite: subtracting the pure third-order linewidth (from the lineshape program in TDEP suite) from the combined third- and fourth-order linewidth evaluated at the harmonic frequency (from the thermal_conductivity program).

However, as demonstrated in Fig. 2.26, this approach proves unviable. Because the two programs rely on fundamentally different numerical integration algorithms, their direct subtraction introduces severe inconsistencies. Specifically, this algorithmic mismatch artificially yields combined linewidths that are narrower than the pure third-order contribution alone—a physically counterintuitive result. Since combining these mismatched im-

plementations compromises both absolute and relative interpretations of the physical entities, this ad-hoc extraction was discarded. A rigorous and consistent treatment of this residual anharmonicity must await future implementations of a unified, frequency-dependent fourth-order self-energy solver.

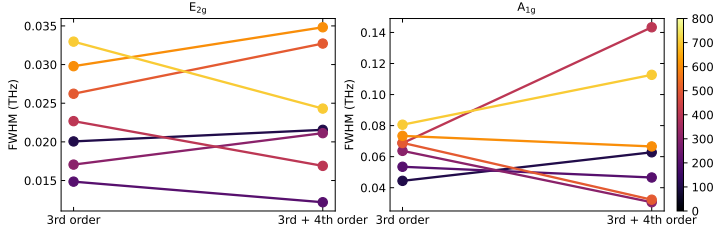


Figure 2.26: FWHM of the E_{2g} and A_{1g} modes with the third- and fourth-order contributions (see text), across all temperatures.

As for the relative widths of the two peaks, our calculations yield smaller E_{2g}/A_{1g} linewidth ratios compared to experiment. This discrepancy is consistent with the presence of inhomogeneous broadening in experimental spectra: an identical inhomogeneous contribution exerts a larger relative broadening on the narrower E_{2g} peak, thereby reducing the overall ratio. Furthermore, our theoretical description of anharmonicity accounts solely for thermal expansion and phonon–phonon scattering, omitting other channels such as electron–phonon interactions that may contribute to spectral broadening [78–80]. Moreover, as established previously, these linewidths exhibit significant sensitivity to numerical and computational parameters.

Finally, we computed the Raman susceptibility tensor χ' for the E_{2g} and A_{1g} modes according to eq. 1.111 and averaged the Cartesian components following eq. 1.134. To compare our computed intensities directly with experimental data under the assumption that N_{peak} degenerate modes contribute equally with unpolarized intensity I_{unpol}^λ , the integrated spectral intensity is expressed as:

$$\int_{\text{peak}} S(\omega), d\omega = k N_{\text{peak}} \frac{I_{\text{unpol}}^\lambda [n_\lambda(T) + 1]}{\omega_\lambda} \quad (2.14)$$

where $S(\omega) = \frac{\partial^2 \sigma^R(\omega_s)}{\partial \Omega \partial \omega}$ represents the Stokes component of the Raman spectrum (eq. 1.123), and k absorbs all frequency-independent experimental and geometric prefactors. Note that we considered $(\omega_i - \omega_s)$ in eq. 1.123 as constant, since the excitation energy is much larger than the shifts probed, hence it is included in k as well. We extracted the experimental Raman

intensities by isolating I_{unpol}^λ :

$$I_{\text{unpol}}^\lambda = \frac{\omega_\lambda \int_{\text{peak}} S(\omega) d\omega}{kN_{\text{peak}} [n_\lambda(T) + 1]} \quad (2.15)$$

taking ω_λ as the experimental peak centroid. While the experimental peak position naturally incorporates many-body self-energy renormalizations across all interaction orders, our theoretical evaluation relies solely on the temperature-dependent second-order effective frequencies derived from TDEP. In carrying out this comparison, we implicitly assume that the higher-order dynamic corrections to the real part of the phonon self-energy (Δ_λ) are small compared to the baseline quasiparticle frequency ($\Delta_\lambda \ll \omega_\lambda$), a condition well verified by our theoretical results.

To benchmark these intensities, we analyzed the temperature-dependent experimental Raman spectra of few-layer MoS₂ from Sahoo *et al.* [81]. Following baseline subtraction, each optical mode was fit with a standard Lorentzian profile:

$$f(\omega) = \frac{A}{\pi} \frac{\Gamma}{(\omega - \omega_\lambda)^2 + \Gamma^2} \quad (2.16)$$

where A is the integrated peak area, 2Γ is the Full Width at Half Maximum (FWHM), and ω_λ is the peak center.

The ratio between the calculated and experimental intensities as a function of temperature is presented in Fig. 2.27. The intensity ratios obtained via stochastic sampling (sTDEP) lie slightly below those extracted from molecular dynamics trajectories (MD-TDEP), but asymptotically converge toward the same values at elevated temperatures. This behavior aligns with our previous observation that the in-plane E_{2g} mode exhibits greater sensitivity to the sampling scheme than the out-of-plane A_{1g} mode. In contrast to the experimental trend, the theoretical E_{2g} intensity grows faster with temperature than the A_{1g} intensity. This evolution mirrors the frequency shifts: over a 600 K window, the E_{2g} frequency undergoes a more pronounced softening than A_{1g} in both TDEP implementations (Figs. 2.14 and 2.21).

While our computed intensity ratios share the same order of magnitude as experiment, the measured values exhibit an opposite, decreasing temperature trend. This discrepancy stems primarily from the absence of electronic transitions in our static Raman formalism. Relative Raman cross-sections depend strongly on the excitation laser energy; any experimental excitation above the optical band gap (1.23 eV for bulk 2H-MoS₂ [52]) triggers

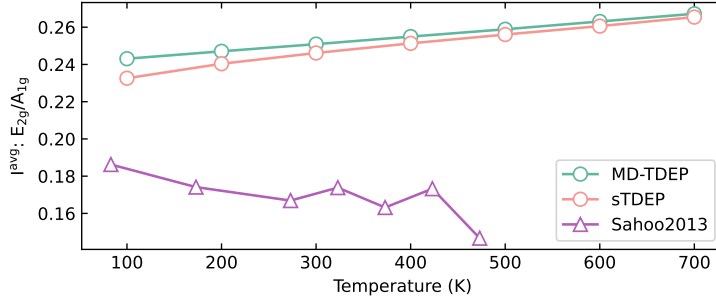


Figure 2.27: Ratio between the isotropically averaged Raman intensity of the E_{2g} and the A_{1g} modes, as obtained with the MD and stochastic sampling methods, vs temperature, and from the experimental work by Sahoo *et al.* [81].

resonant Raman scattering, selectively enhancing specific vibrational channels. Because our theoretical intensities are evaluated strictly within the non-resonant, static limit ($\omega_{\text{laser}} \rightarrow 0$), these resonant enhancements are not captured. Beyond electronic resonance, external factors such as substrate choice, local dielectric environment, optical interference, and collection geometry heavily influence measured intensity ratios, limiting the ability to perform a strict quantitative comparison with standard literature values.

The complete simulated Raman spectra, generated using both MD-TDEP and sTDEP frameworks, are presented in Fig. 2.28 and Fig. 2.29. These final observable lineshapes are constructed by convoluting the temperature-dependent static Raman intensities with the anharmonic phonon spectral functions. Consequently, the variations in relative peak heights between the two methods are the direct macroscopic manifestation of the distinct dynamical broadening mechanisms (FWHMs) analyzed in the preceding sections.

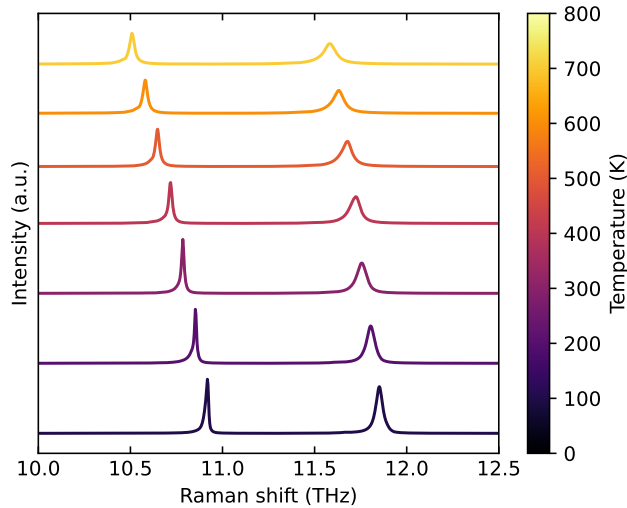


Figure 2.28: Raman spectra of 2H-MoS₂ calculated from MD-TDEP results.

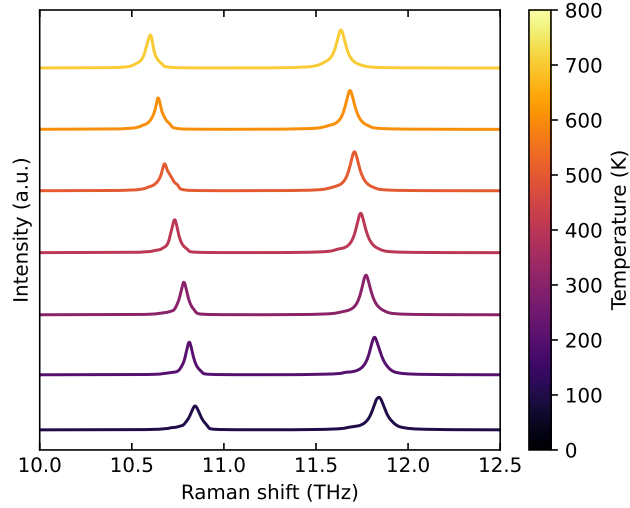


Figure 2.29: Raman spectra of 2H-MoS₂ calculated from sTDEP results.

2.7 Conclusions and perspectives

The integration of Machine-Learning Interatomic Potentials (MLIPs) with the Temperature Dependent Effective Potential (TDEP) method provides a robust, highly scalable framework for predicting the finite-temperature vibrational properties of 2H-MoS₂. By replacing computationally prohibitive *ab initio* molecular dynamics with a rigorously trained Moment Tensor Potential (**M1**), we achieved the extensive phase-space sampling necessary to extract higher-order interatomic force constants and explicitly construct the anharmonic phonon self-energy.

Our comparative analysis between classical molecular dynamics (MD-TDEP) and stochastic sampling (sTDEP) highlights the critical role of quantum statistical mechanics in modeling lattice dynamics. While both methods successfully captured the physical thermal red-shift and linewidth broadening of the Raman-active E_{2g}^1 and A_{1g} modes, sTDEP demonstrated a distinct theoretical advantage at low temperatures. By inherently incorporating the quantum zero-point energy occupation, sTDEP sampled a broader extent of the anharmonic potential energy surface near the zero-temperature limit, yielding more realistic low-temperature broadenings and a reduced thermal softening rate compared to the purely classical MD trajectories.

Despite capturing the primary third-order phonon-phonon scattering channels, our simulated spectral linewidths systematically underestimated experimental measurements. This discrepancy elucidates the physical boundaries of the current theoretical paradigm. The attempt to integrate a static fourth-order correction revealed numerical incompatibilities within the available TDEP suite, emphasizing that a rigorous, frequency-dependent evaluation of

fourth-order phonon scattering remains a necessary mathematical development to capture the full intrinsic anharmonic broadening. Furthermore, the divergence between our calculated Raman intensity ratios and experimental trends exposes the limitations of the static Placzek approximation. Because experimental probes often utilize excitation energies that trigger electronic transitions, capturing the true temperature-dependent spectral intensities requires moving beyond the static polarizability tensor to include fully resonant electron-phonon coupling effects.

Alternatively, an increasingly established methodology bypasses perturbative phonon formulations entirely by utilizing MLIPs to directly predict tensorial properties alongside atomic forces [82, 83]. By training a machine-learning model to learn the macroscopic polarizability and evaluating it on-the-fly during large-scale molecular dynamics, the complete Raman spectrum is extracted directly from the Fourier transform of the polarizability time-autocorrelation function. This time-domain approach inherently captures all orders of structural anharmonicity dynamically, without relying on truncated Taylor expansions of the interatomic potential.

Building on this foundational methodology, several promising avenues for future research emerge. Developing computational frameworks that natively integrate fully dynamic Raman tensors, electron-phonon scattering, or direct autocorrelation approaches will be crucial for closing the remaining gap with experimental lineshapes. Furthermore, because the MLIP infrastructure efficiently scales to thousands of atoms, this methodology is perfectly positioned to investigate symmetry-breaking phenomena in low-dimensional configurations.

3.1 Introduction and Motivation

Transition metal dichalcogenides (TMDs) have emerged as primary candidates in the search for cost-effective, earth-abundant electrocatalysts to replace scarce and polluting platinum-group metals in the hydrogen evolution reaction (HER) [3, 15]. In pristine, stoichiometric crystalline 2H-MoS₂, the catalytic activity is heavily limited: the basal plane is chemically inert, and active proton-reduction sites are almost exclusively restricted to undercoordinated sulfur edge facets [3]. Consequently, improving catalytic performance requires defect engineering, phase transformations (such as inducing the metallic 1T phase), or reducing domain boundaries.

In this context, amorphous molybdenum sulfides (*a*-MoS_{*x*}, with $2 \leq x \leq 6$) have emerged as high-performing catalysts capable of delivering HER current densities comparable to noble metals at modest overpotentials in acidic conditions [13, 14]. They are mostly synthesized via low-temperature wet-chemical precipitation, electrochemical deposition [14], or physical vapor deposition [84]. The absence of long-range periodic translational symmetry allows for a high density of inherently accessible, unsaturated sites throughout the bulk and surface of the material [13, 85].

Despite substantial catalytic interest, the microscopic atomic structure of *a*-MoS_{*x*} remains a subject of ongoing debate, with proposals ranging from [Mo₃S₁₃]²⁻-based coordination polymers to other Mo-S network motifs [86–89]. Because conventional X-ray diffraction (XRD) on amorphous samples yields featureless patterns, experimental structural characterization relies on local-probe techniques such as Extended X-ray Absorption Fine Structure (EXAFS), X-ray Absorption Near Edge Structure (XANES), and Pair Distribution Function (PDF) analysis from total scattering [13, 86]. From these measurements, the short-range order appears to be dominated by Mo-S bonds (~ 2.40 – 2.45 Å) and characteristic Mo-Mo distances (~ 2.75 – 2.82 Å), alongside sulfur-sulfur bridging motifs [86].

To bridge the gap between these averaged experimental spectra and the true atomistic complexity of *a*-MoS_{*x*}, computational modeling must extract a wider array of topological and chemical observables. Through large-scale molecular dynamics, one can move beyond one-dimensional radial

distribution functions to evaluate the local coordination number of every atom, explicitly mapping the density of undercoordinated, catalytically active sites. Furthermore, the network topology can be classified through ring statistics and cluster connectivity, differentiating between isolated Mo dimers, triangular Mo₃ cores, and extended polymeric chains. The chemical nature of the sulfur network—distinguishing between isolated sulfides (S²⁻), terminal disulfides (S₂²⁻), and bridging ligands—is systematically captured by tracking S-S pairing motifs.

Structural models proposed in the literature span a spectrum between two main topologies:

- *Molecular cluster models*: predominant in sulfur-rich compositions ($x \approx 3$), these models depict a -MoS _{x} as an interconnected coordination network consisting of triangular [Mo₃S₁₃]²⁻ or [Mo₃S₇]⁴⁻ building blocks bridged by disulfide ligands (S₂²⁻) and apical sulfido ligands (S²⁻) [86–89].
- *Disordered continuous networks*: Approaching the stoichiometric disulfide limit ($x \approx 2$), the structure is described as a distorted, cross-linked amorphous network of MoS _{n} polyhedra, sharing edges and corners, and containing short-range Mo–Mo metal bonds and sulfur homopolar bonds, deviating significantly from the classical trigonal prismatic layer coordination [13, 85, 87, 88].

Developing predictive structure–property relationships requires atomistic models of the disordered bulk and surface. Such atomistic simulations must capture the topological and chemical diversity across multiple length scales without relying on crystalline symmetry constraints, making this task challenging for *ab initio* methods like DFT.

To address these challenges, this chapter investigates the structural and vibrational properties of amorphous molybdenum sulfides across three key stoichiometries: a -MoS₂, a -MoS₄, and a -MoS₆. As a starting point, we adopt the structural models recently generated by Si and Ma [90] via *ab initio* molecular dynamics (AIMD). Building upon their foundational work, we employ these configurations to initialize the training of a Machine-Learning Interatomic Potential (MLIP). As discussed in the following section, transitioning to an ML-driven approach allows us to extend the accessible spatial and temporal scales of the simulations. This enables the application of slower melt–and–quench cooling rates to further explore the structural phase space and thermal relaxation of these complex amorphous networks.

3.2 State of the Art: Computational generation of amorphous structures

Computational generation of amorphous materials typically proceeds via three main paradigms:

- *Reverse Monte Carlo (RMC)*: A stochastic fitting method where atomic coordinates inside a supercell are iteratively adjusted to match experimental total-scattering structure factors $S(q)$ or pair distribution functions $g(r)$ [91, 92]. While effective at reproducing measured experimental profiles, RMC is prone to non-uniqueness: it explores configurations driven by a target metric rather than a physical Hamiltonian, which can yield unphysical bond angles, coordination geometries, or high-energy steric clashes unless heavily constrained by empirical penalty terms.
- *Random Packing and Stochastic Relaxation*: Atoms or molecular fragments are placed randomly within a periodic box subject to minimum-distance hard-sphere criteria, followed by continuous relaxation at target densities [93–95]. While computationally inexpensive, this method often fails to capture the true topological connectivity and intermediate-range ring structures that form during natural kinetic nucleation.
- *Melt-and-quench molecular dynamics*: Regarded as the standard for generating realistic glassy and amorphous networks [96]. The system is initialized at high temperatures above the melting point to erase crystal memory, thoroughly thermalized in the liquid phase, and quenched down to target temperatures at controlled (usually constant) cooling rates (dT/dt).

Historically, the application of the melt-and-quench method to complex chalcogenides has been limited by the tradeoff between accuracy and computational cost [96]. Indeed, while *ab initio* Molecular Dynamics (AIMD) based on Density Functional Theory (DFT) accurately describes the covalent and metallic bonding, it is computationally restricted to small supercells ($\sim 10^2$ atoms) and short trajectories ($\sim 10^1$ – 10^2 ps). This imposes unphysically high computational quenching rates ($\sim 10^{15}$ K/s [90]) —many orders of magnitude faster than both rapid experimental liquid-quenching methods (10^4 – 10^6 K/s [97]) and controlled laboratory annealing ($\sim 10^{-2}$ K/s [98]). For this reason, the system can be trapped in high-energy local minima with unphysical defect densities [96, 99]. On the other hand, classical empirical or

reactive force fields (such as classical Stillinger–Weber [100], Tersoff [101], or ReaxFF [102]) allow large systems and slow cooling rates, but can struggle to capture the complex bonding chemistry of amorphous molybdenum sulfides.

Machine-Learning Interatomic Potentials (MLIPs) bridge this gap by offering near-DFT accuracy at a fraction of the computational cost, scaling linearly with the number of atoms $\mathcal{O}(N)$ [35] (while DFT scales with the cube of the number of bands), as anticipated in section 1.5.7. By training the MLIP on representative configurations of liquid, distorted, and defective phases, ML-driven MD (MLMD) enables the simulation of large supercells, as well as slower, more realistic cooling rates.

Training an MLIP for amorphous systems introduces two formidable challenges: the absence of representative crystalline reference states, and the wide extent of the configurational and chemical space of disordered structural networks.

Indeed, the recent literature increasingly acknowledges that the amorphous state remains a central, unresolved frontier even for modern machine-learned interatomic potentials [103]. While current “foundational” MLIPs exhibit remarkable accuracy and transferability for ordered crystalline systems, their reliability frequently degrades when applied to non-crystalline solids [103]. This discrepancy arises because standard training datasets are typically biased toward periodic structures, failing to adequately capture the vast, rugged potential energy surface of the glassy state—which is populated by an almost infinite combination of undercoordinated defects, distorted polyhedra, and transient bonds. Consequently, when baseline models are deployed in high-temperature melt-and-quench molecular dynamics, they inevitably encounter out-of-domain topologies. Forced to extrapolate beyond their training distribution, the models predict unphysical forces [99, 104], leading to severe simulation instability during the critical liquid-phase thermalization.

This still remains a hard challenge, and new models are formulated and tested on a daily basis to improve the predictive performance of MLIPs on amorphous systems [99, 105, 106].

3.3 MLIP, and training data

As anticipated in the introduction of this chapter, we took the structures obtained by Si and Ma [90] for a -MoS₂, aMoS₄ and aMoS₆ as our starting point¹. The authors provided us with a 200-structure-long DFT-MD trajectory for each stoichiometry, representing the quenching phase at 300 K of the melt-and-quench protocol they carried out.

To rigorously benchmark the selected MLIP formalisms, we established a controlled training and evaluation framework designed to separate baseline interpolation accuracy from high-temperature extrapolation capability. Note that we only considered a -MoS₂ for the choice of the MLIP model, as much more experimental and simulation data is available.

The benchmarked models were trained exclusively on a foundational dataset. This set was constructed starting from 200 reference a -MoS₂ configurations, which were then heavily augmented by applying systematic geometric distortions and volumetric strains. A minimum-interatomic-distance filter² was subsequently applied to remove unphysical steric clashes, yielding a robust training set of roughly 1750 structures containing a vast array of local glassy coordination geometries.

To evaluate the models, we required a strictly out-of-distribution test set that represented the complex, high-temperature dynamics of the liquid phase. To achieve this, an intermediate exploratory MLIP was temporarily trained on a -MoS₂ and used to drive an independent melt-and-quench molecular dynamics simulation³. From the resulting trajectory, we extracted two distinct evaluation sets. The first was constructed via uniform temporal sampling to capture the broad thermal distribution of the trajectory. The second set was designed as a targeted test set: frames were extracted based on high committee disagreement⁴ and passed through a SOAP-based diversity filter (as detailed in section 1.5.8).

Crucially, none of these MD-derived structures were included in the benchmarking training set. This strict separation ensured that the final benchmark evaluated each model's true capacity to generalize to the unmapped, high-temperature potential energy surface of the liquid and amorphous phases.

We benchmarked four MLIP models: Moment Tensor Potentials (MTP) [35, 67], MatterSim [107] (finetuning the foundation model v1.0.0-5M), NequIP [108], and MACE [36] (finetuning the foundation model matpes-pbe-omat-ft).

1: We recomputed the energy, forces and stresses with DFT to obtain a uniform dataset

2: Structures with any interatomic distance below 1.8 Å were discarded.

3: The MD was performed with ASE (in the NVT ensemble, for stability, with a Langevin thermostat), as interfaced to our in-house package for the training and deployment of MLIP models, MlipKit, covered in section 4.1. As the intermediate MLIP did not allow for very high temperatures to be reached, the system was heated up to 1300 K.

4: The committee was composed of 10 instances of the same MLIP used to run the MD that generated the trajectory; every instance differing from the others by the seed of the random generator.

In Figures 3.1 and 3.2 the force parity plots are reported for each of the four models, inferring the uniform and active-learning selection, respectively. The predictive accuracy of the four potentials was first evaluated on the uniformly sampled trajectory. This phase is populated by high-speed atomic collisions that generate transient, extreme forces reaching up to ± 6 eV/Å. On this set, MACE yielded the highest accuracy, achieving the lowest force RMSE of 0.330 eV/Å and the highest correlation ($R^2 = 0.858$). MTP provided the second-best performance with an RMSE of 0.361 eV/Å ($R^2 = 0.831$), while NequIP and MatterSim exhibited higher force RMSEs of 0.367 eV/Å ($R^2 = 0.819$) and 0.375 eV/Å ($R^2 = 0.816$), respectively. The models were subsequently evaluated on the active-learning test set (Fig. 3.2). Interestingly, all potentials exhibited nominally lower absolute force errors on this set compared to the uniformly sampled trajectory. While seemingly counterintuitive for a dataset selected specifically for its "difficulty," this reduction in absolute error stems directly from the differing force distributions of the two sets. Because the steep repulsive wall governing short-range interactions is a simple, well-sampled physical feature, the baseline MLIPs reliably capture it. Consequently, the independent committee members exhibit strong agreement on these high-magnitude force predictions, resulting in low overall uncertainty. The active-learning algorithm therefore bypasses these thermally extreme, but physically straightforward, collision frames. Instead, the committee disagreement isolates out-of-domain topological complexity. The selected local environments generate moderate force magnitudes—strictly bounded within ± 4 eV/Å—but cause high inter-model disagreement. Even on this strictly topologically complex subset, the relative ranking of the models remained invariant. MACE maintained its slight predictive superiority, achieving a force RMSE of 0.273 eV/Å ($R^2 = 0.877$), outperforming NequIP (0.306 eV/Å), MTP (0.317 eV/Å), and MatterSim (0.320 eV/Å). Based on this consistent accuracy and generalization capability across both the thermal and topological evaluation regimes, the MACE model was selected to drive the production melt-and-quench simulations.

The hypothesis that the active-learning committee targets local structural complexity rather than high-temperature thermal noise is empirically confirmed by analyzing the temperature distribution of the extracted frames, reported in Fig. 3.3. The uniform extraction mirrors the full molecular dynamics schedule, capturing configurations across all stages, including the high-temperature liquid melt extending beyond 1000 K. As previously noted, this liquid phase generates extreme but physically trivial repulsive

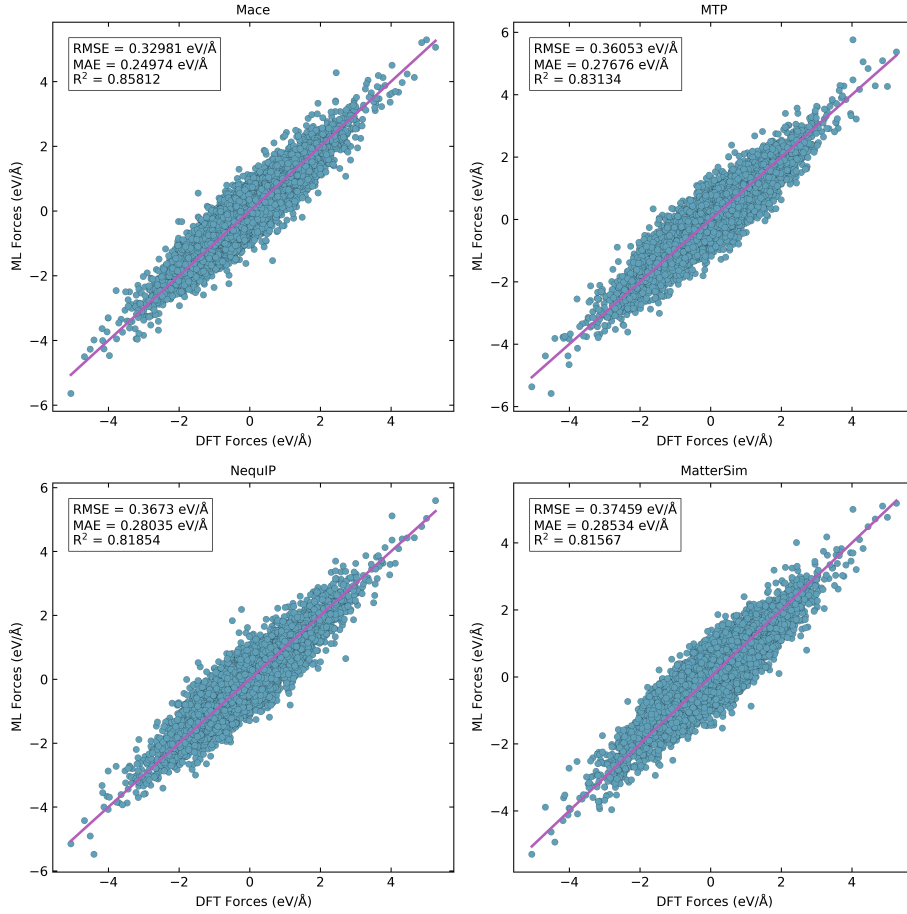


Figure 3.1: Parity plots comparing DFT-computed atomic forces against MLIP predictions on the uniformly sampled held-out trajectory. The extended force distribution (± 6 eV/Å) reflects the inclusion of extreme, high-temperature thermal collisions.

forces.

Remarkably, the active-learning algorithm completely truncated its sampling above ~ 900 K. Instead, the committee’s uncertainty peaks sharply in two specific regimes: at ~ 300 K, corresponding to the frozen structural defects of the amorphous state, and at ~ 750 – 850 K. By restricting these regimes of maximum topological variance, the active-learning set provided a targeted benchmark.

For our production runs, in order to increase the amount of information, even in spite of efficiency, we generated more distorted configurations from the reference ones, and computed their DFT properties. Our datasets then contained 2914, 2949 and 2295 configurations for a -MoS₂, a -MoS₄ and a -MoS₆, respectively. For each of them, 20% of the frames were used as validation set, and the rest served as training set. We finetuned the foundation model `matpes-pbe-omat-ft` on the three augmented datasets, obtaining our three production MLIPs: **M2** for a -MoS₂, **M3** for a -MoS₄, and **M4** for a -MoS₆.

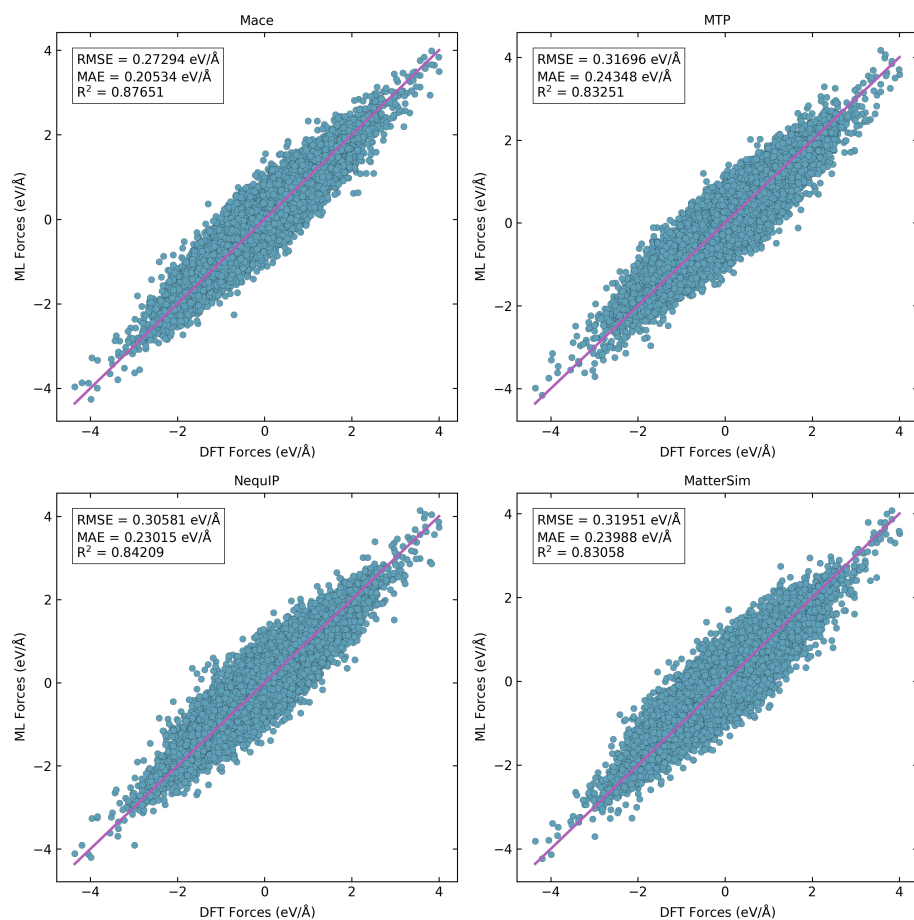


Figure 3.2: Parity plots for the active-learning test set, selected via committee disagreement and embedding diversity filtering. The narrowed force distribution (± 4 eV/Å) indicates the filtering of trivial thermal collisions, isolating topologically complex amorphous environments.

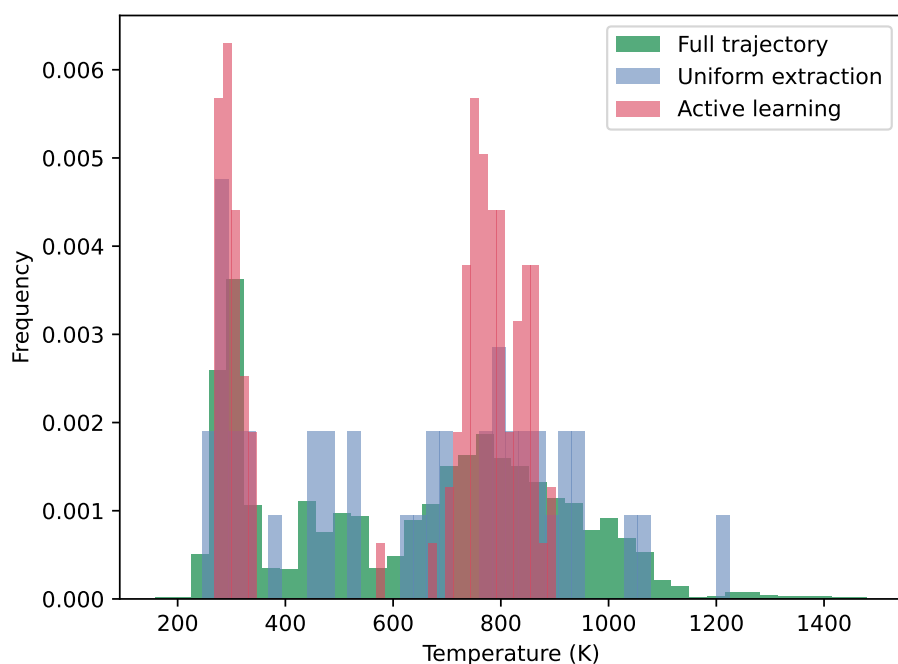


Figure 3.3: Temperature distributions of the configurations extracted from the MD trajectory. The uniform extraction (blue) naturally tracks the temporal distribution of the full trajectory (green), including the high-temperature melt (> 1000 K). In contrast, the active learning extraction (salmon) systematically rejects the high-temperature liquid phase, preferentially sampling the topologically complex amorphous (300 K) and glass-transition (~ 800 K) regimes.

3.4 Amorphous structure generation

To generate realistic $a\text{-MoS}_x$ structures, we decided to use the melt-and-quench approach. The network topology can be heavily influenced by the thermal history, especially in kinetically driven processes, like many MD simulations, where the cooling rates are many orders of magnitude higher than the experimental ones. Aware of such limitations, we set up 9 distinct MLMD melt-and-quench protocols, defining three melt temperatures, $T_{\text{melt}} = 1500, 2000$, and 2500 K, and three cooling rates: $\gamma_1 = 1.0 \cdot 10^{13}$, $\gamma_2 = 0.5 \cdot 10^{13}$ and $\gamma_3 \approx 0.33 \cdot 10^{13}$ K/s. For each combination of T_{melt} and γ , three independent stochastic realizations were simulated to ensure robust ensemble averaging, yielding a total of 27 production trajectories per stoichiometry. Each MD simulation, performed in the NPT ensemble, was constituted by a 10 ps equilibration phase at 800 K, followed by the heating (applied at a rate of $1.4 \cdot 10^{13}$ K/s, namely 100 ps), then the system was kept at the melting temperature for 40 ps, and successively quenched down to 300 K; finally, the system was equilibrated for 40 ps. The protocols are depicted in 3.4. We simulated supercells of 1200, 1500 and 1512 atoms for $a\text{-MoS}_2$, $a\text{-MoS}_4$ and $a\text{-MoS}_6$, respectively.

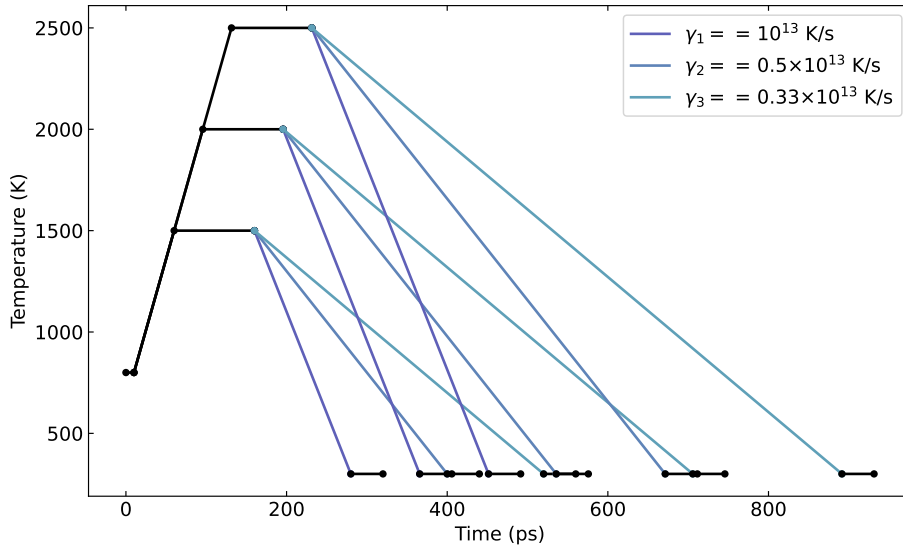


Figure 3.4: Melt-and-quench protocols: three melting temperatures and three quenching rates.

We will first address $a\text{-MoS}_2$, and later $a\text{-MoS}_4$ and $a\text{-MoS}_6$, as they came with different complications.

3.4.1 $a\text{-MoS}_2$ structural exploration

As anticipated by the fundamental challenges of modeling amorphous systems, our MLIP M2 struggled with the extreme configurational diversity

of the high-temperature liquid state. While **M2** successfully stabilized the solid-state configurations, simulations initialized at $T_{\text{melt}} = 2500$ K exhibited catastrophic instability. At this temperature, the high kinetic energy pushed the liquid into highly distorted local configurations. Forced to extrapolate, the MLIP predicted non-physical atomic forces, leading to the rapid structural explosion of all nine 2500 K samples.

To rectify this, we employed an active learning strategy, based on the committee uncertainty and embedding diversity criteria discussed in section 1.5.8. We ran an independent molecular dynamics trajectory on a smaller supercell (150 atoms) at a lower, stable melt temperature of 1500 K. Then we trained a 10-member committee with the same hyperparameters used to train **M2** (only the seed of the random generator changed) and calibrated it on a trajectory obtained with $T_{\text{melt}} = 1500$ K and γ_3 (sample 1), obtaining a Spearman coefficient of 0.88. We extracted the 500 structures on which the committee exhibited the largest uncertainty, and then we selected the 300 most diverse ones⁵. Finally, we computed their corresponding Density Functional Theory (DFT) energies, forces, and stresses, and added them to the training dataset.

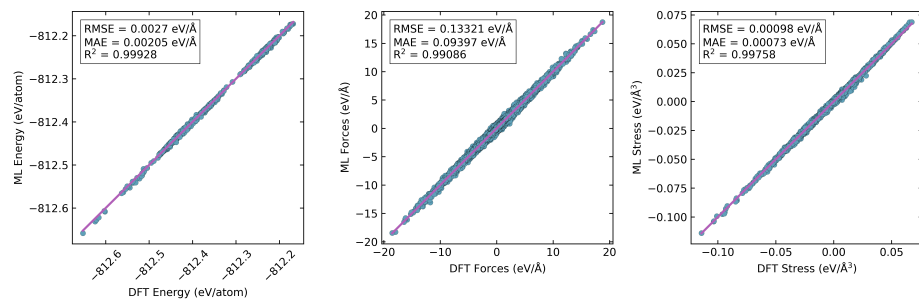
5: We used the SOAP descriptors with $r_{\text{cut}}=6$, $n_{\text{max}}=8$, and $l_{\text{max}}=6$, see section 1.5.7.

Table 3.1: Diffusion coefficient in 10^{-5} cm²/s at each T_{melt} (using the ASE tool) from the molten, averaged over 18 samples (3 γ and 3 repetitions).

	Temperature (K)		
	1500	2000	2500
Mo	0.390336	2.37885	4.94268
S	0.505190	3.12390	6.92002

Retraining the MLIP on this augmented dataset proved highly effective in resolving the instability. In the subsequent production iteration, the model successfully navigated the rugged potential energy surface of the extreme high-temperature liquid. Melts at 2500 K were stabilized, and all 9 independent trajectories successfully reached the 300 K target temperature without unphysical expansions. This refined MLIP, **M2.1**, was subsequently used to generate all the structural and vibrational data discussed in the following. The parity plots of **M2.1** are shown in Fig. 3.5.

Figure 3.5: Parity plots for energy, forces and stresses of the model **M2.1**, trained on α -MoS₂. The inference was done on the validation set (20% of the total data).



Because of the evidence of the sensitivity of such MLMD simulations to thermal stress, we decided to not push T_{melt} further up (to recover the higher temperatures achieved in the study by Si and Ma [90]). However, in order to assess proper melting even at 1500 K, we calculated the diffusion coefficient in the melted phase. The values we obtained, shown in Table 3.1, are of the order of 10^{-5} cm²/s. This lies in the typical range of experimental measurements for liquid metals close to their melting point, $10^{-5} - 10^{-6}$ cm²/s [109].

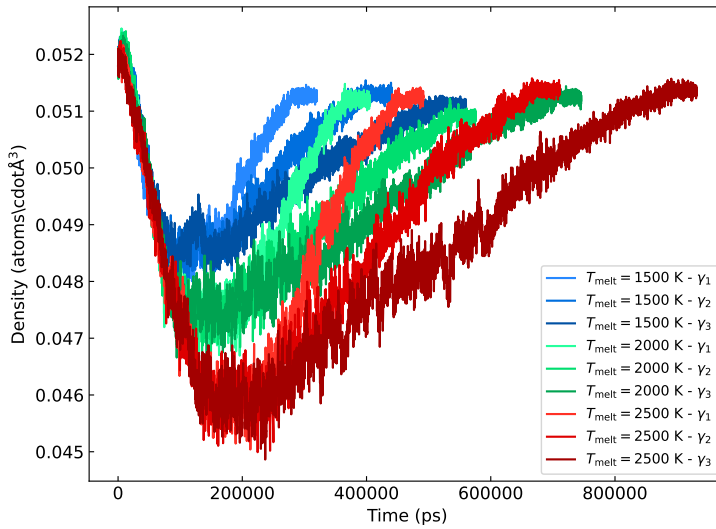


Figure 3.6: Evolution of the atomic density across the nine thermal protocols, averaged over the three replicas. The initial decrease corresponds to thermal expansion during heating and melting, while the subsequent recovery tracks the densification during the quench. All trajectories converge to a consistent final density regardless of the specific thermal history.

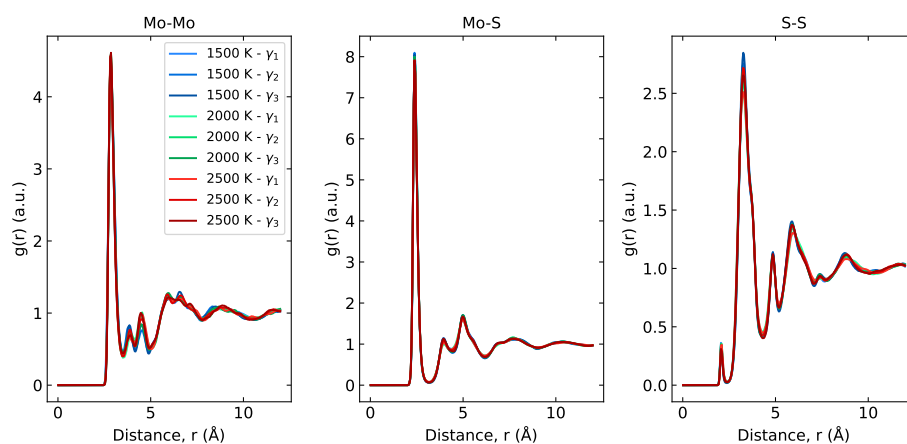
The macroscopic stability of the simulated amorphous phases was validated by tracking the atomic density throughout the melt-and-quench simulations, as shown in Fig. 3.6.

The density evolution accurately reflects the thermodynamic phases of the protocol. The network undergoes significant thermal expansion during the heating phase, reaching a lower-density liquid state at the target T_{melt} . During the subsequent quench, the system densifies at rates proportional to the applied γ . Despite the divergent temporal pathways and different liquid-state densities, all nine protocols converge to a tightly clustered final atomic density at 300 K. This macroscopic convergence proves that the final quenched state is volumetrically robust and is not kinetically trapped in highly expanded, under-coordinated configurations by the faster cooling rates.

To assess the sensitivity of the melt-and-quench process to the different melting temperatures and cooling rates, we computed the radial distribution functions using the LAMMPS software (RDFs, see section 1.7.1) on the last phase of the melt-and-quench protocol, namely the final equilibration.

For every pair (T_{melt}, γ), we averaged the RDFs over the three replicas. The RDFs are depicted in Fig. 3.7. The distributions of the radial atomic distances exhibit invariance to the melting temperature or the quench rate. The primary peaks overlap perfectly across all nine protocols, and for every atomic pair, confirming that the fundamental covalent bond lengths are strictly dictated by local chemistry rather than thermal history. Crucially, the RDFs exhibit deep first minima, indicating a strict, unambiguous separation between the primary covalent coordination shell and the broader non-bonded network. Because this short-range order is completely locked in, these invariant first minima provide physically rigorous, parameter-free cutoffs for evaluating the three-dimensional geometry of the glass. Applying

Figure 3.7: Sample-averaged radial distribution functions (RDFs) for α -MoS₂ across all nine thermal histories. The insets highlight the strict invariance of the primary coordination peaks.



these cutoffs, the angular distribution functions (ADFs) were computed, again with the LAMMPS software, to assess the three-dimensional geometry of the glass, and averaged over samples (Fig. 3.8). Consistent with the radial distributions, the fundamental angular order remains robust. The peak positions across S-Mo-S (intra-polyhedra), Mo-S-Mo (inter-polyhedra), and Mo-Mo-Mo (extended network) triplets are strictly invariant. On the other hand, the robustness of the S-S-Mo ADF (also an intra-polyhedron descriptor) is affected by the scarcity of covalently bonded S atoms (see the height of the relative first peak in the S-S RDF). However, the position of the peaks remains fixed across the protocols, and still provides a good characterization metric.

In order to further investigate the local geometry, we used the ChemEnv package (see 1.7.2) to classify the coordination of each atom. We first tried with multi-weight strategies. These are designed to score all the tabulated polyhedra for each atom according to multiple criteria, and combine the scores with specific weights. However, we found that such strategies were too complex and failed to reproduce the average coordination number for sulfur

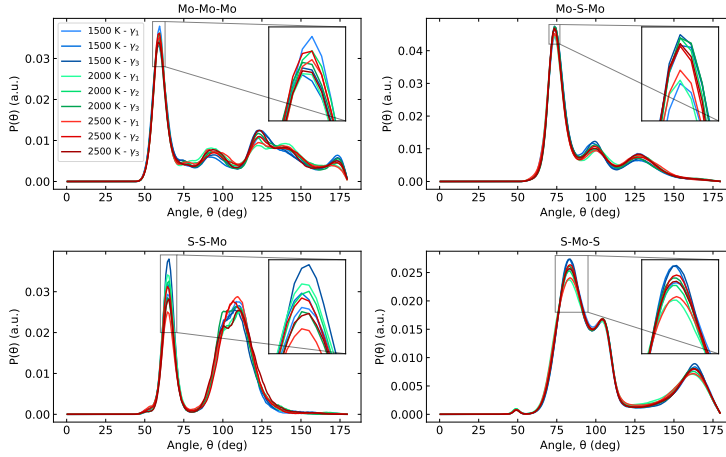


Figure 3.8: Sample-averaged angular distribution functions (ADFs) for a - MoS_2 evaluated across the nine thermal histories. The insets magnify the primary peaks to illustrate the strict positional invariance and minimal intensity variance across all quench pathways.

in our systems, and therefore we decided to keep the simplest strategy, which assigns a single polyhedra to each central atom, namely the polyhedron with the highest similarity to the actual coordination (lowest CSM). In 3.9, we report the fraction of the most common coordination polyhedra for every protocol, averaging across the three replicas. The polyhedral analysis confirms the short-range order invariance. The relative fractions of both Mo and S environments remain essentially constant across all protocols, confirming that the local structural motifs are inherent in our amorphous phase, and do not depend on the melting temperature or quench velocity, as already suggested by RDFs and ADFs. A slower quench rate does favor octahedral coordination slightly.

To extend the analysis to the medium-range order, we also computed the number of irreducible rings⁶ (see section 1.7.3). The sample-averaged results, only considering the alternating Mo-S rings, are shown in Fig. 3.10. The populations of small rings (4-, 6-, and 8-membered) are not significantly affected by the different protocol settings. However, large 10- and 12-membered rings exhibit considerable variance depending on the specific thermal history; notably, the ($T_{\text{melt}} = 1500 \text{ K}$, γ_1) trajectory diverges from the others at larger ring sizes. This indicates that while local polyhedra assemble identically, their large-scale packing into extended chains (and voids) remains highly susceptible to kinetic trapping during the quenching phase. This may be useful for applications if correlations between meso-structure and catalytic activity can be demonstrated: the melt temperature and quenching speed have a clear impact. Although the fastest speeds used here may not be accessible in practice, the trend would be interesting to explore experimentally.

6: We used the Julia implementation and Python API from Morrow's repository [110].

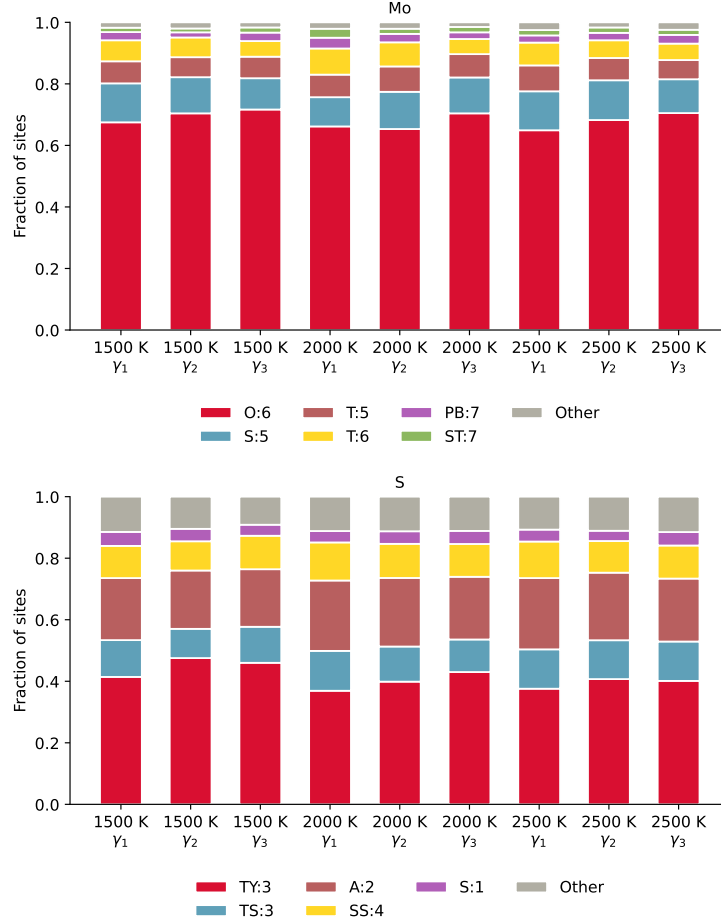


Figure 3.9: Fraction of local coordination environments for Mo (top) and S (bottom) across the nine thermal histories. The relative populations of the polyhedral geometries remain quite stable, confirming the short-range order is independent of the quench protocol. Note that only heteropolar bonds are included in the coordination sphere. Polyhedra that individually do not account for at least 3% of the total are grouped under “Others”.

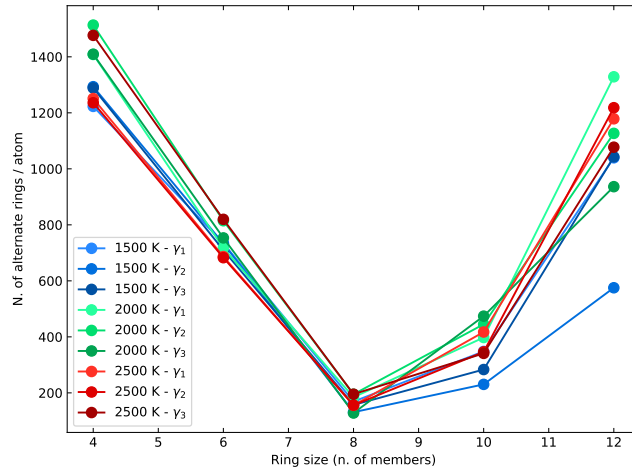


Figure 3.10: Distribution of alternate rings in the amorphous network. While small-to-medium rings exhibit consistent populations, large mesoscopic rings (10- and 12-membered) display high sensitivity to the thermal history, especially for low melt temperatures.

7: The vibDOS was computed from an NVE MLMD run at 300 K, using the last configuration of the NPT melt-and-quench trajectory as starting point.

Finally, we complement our structural analysis with the vibrational density of states (vibDOS, see section 1.7.4)⁷. The results are reported in Fig. 3.11. The high-frequency optical bands (> 2 THz), which are dominated by stiff, localized intra-polyhedral Mo–S stretching, overlap perfectly across all quench rates and melting temperatures. Conversely, the so-called “boson peak” [111] at low frequencies (< 1 THz), arising from the disorder-induced

localization of long-wavelength modes, exhibits subtle shifts depending on the cooling rate for structures quenched from higher melting temperatures. Because the boson peak stems from large-scale structural inhomogeneities that disrupt the propagation of acoustic vibrations, its thermal variance is compatible with the topological fluctuations observed in the mesoscopic ring statistics.

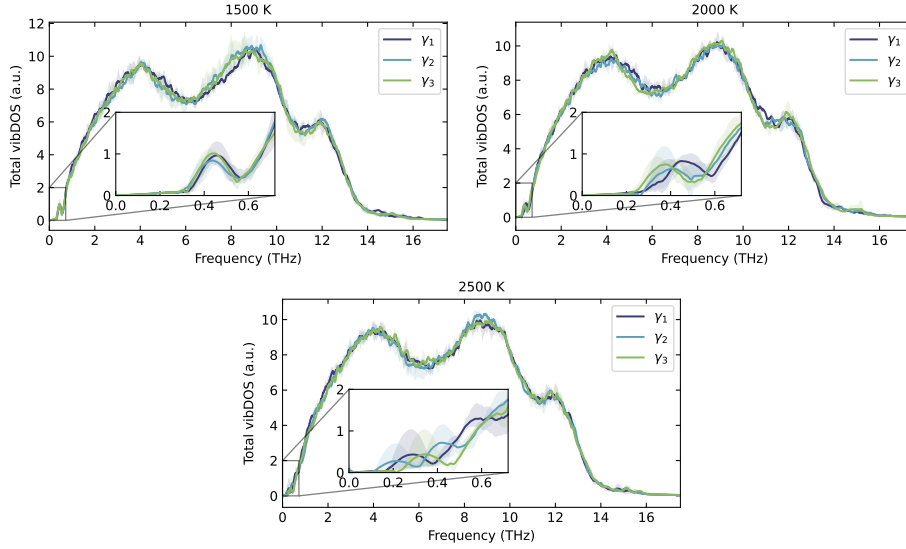


Figure 3.11: Total vibrational density of states (vibDOS) for the amorphous networks. Solid filling = extent from the minimum to the maximum over all samples. All the curves are smoothed with a Savitzky-Golay filter (from SciPy). The high-frequency modes overlap perfectly for all melting temperatures, while the insets highlight thermal variance in the low-frequency acoustic mode “Boson peak” (< 1 THz).

Because the fundamental local chemistry is perfectly conserved, but the mesoscopic packing retains slight kinetic memory, the structural and dynamic descriptors will hereafter be investigated for an intermediate protocol ($T_{\text{melt}} = 2000$ K, γ_2). This configuration avoids extreme kinetic outliers while preserving the universal short-range order, serving as an ideal master baseline for detailed benchmarking against crystalline and amorphous references from the literature.

We compared our amorphous structures with the 2H and 1T⁸ crystalline phases of MoS₂ and the amorphous reference from [90], in order to assess the disruption of the short-range order due to the loss of periodicity.

As a primary metric of structural validation, Table 3.2 compares the atomic densities of the final quenched configurations from our trajectory alongside those from the study by Si and Ma [90] and the crystalline references. As expected, the amorphous phases are consistently less dense than the crystalline 2H phase. This density deficit is a physical hallmark of amorphization in layered transition metal dichalcogenides: destroying the efficiently packed, 2D van der Waals stacked layers to form a continuous, disordered 3D random network inherently introduces topological voids and increases the

8: The 2H structure we used corresponds to the thermalized unitcell at 300 K, from our previous study (see section 2.5), while the 1T one is from Materials Project, mp-1238797.

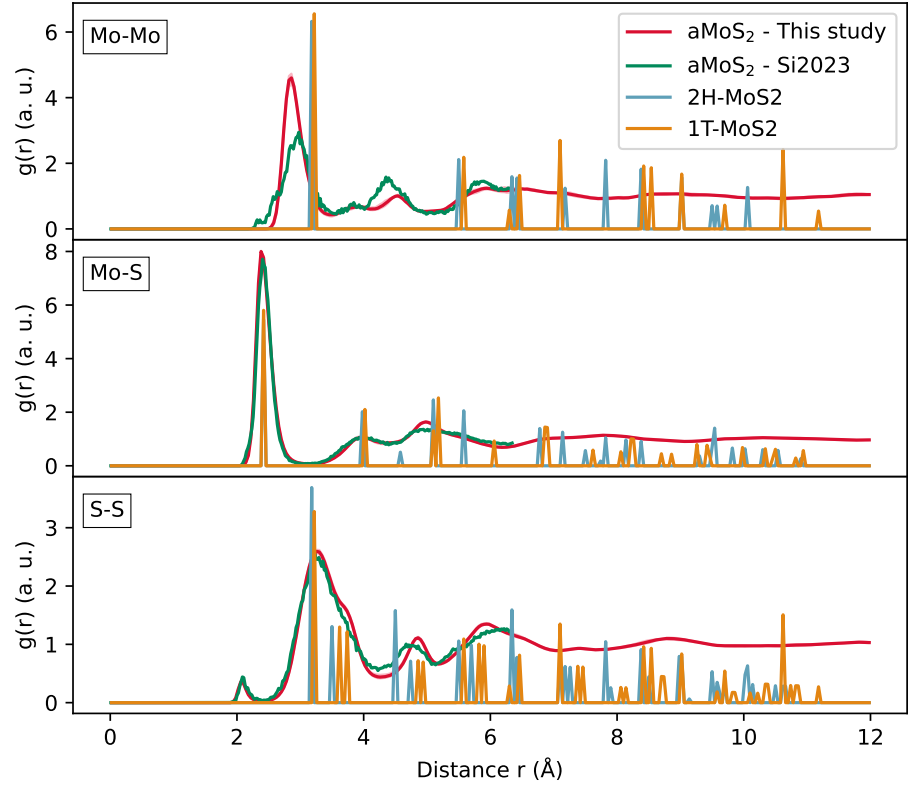
free volume of the system.

Table 3.2: Atomic densities of the amorphous and crystalline phases. Note that the densities of our a -MoS₄ and a -MoS₆ are the same as those from Si and Ma, because we ran the melt-and-quench simulations in the NVT ensemble (see sections 3.4.2,3.4.3).

	ρ_{at} (atoms·rÅ ⁻³)
a -MoS ₂ - This study	0.0510
a -MoS ₂ - Si2023	0.0518
a -MoS ₄ - Si2023	0.0470
a -MoS ₆ - Si2023	0.0455
2H-MoS ₂	0.0560
1T-MoS ₂	0.0527

To start the analysis of the short-range order, we show in Fig. 3.12 the radial distribution functions (RDFs). To facilitate visual comparison alongside the unnormalized amorphous distributions, the crystalline RDFs were appropriately rescaled.

Figure 3.12: Radial distribution functions comparing our a -MoS₂ configurations to the Si2023 reference structure and the 2H and 1T crystalline phases. The amorphous distributions are obtained from the equilibration phase at 300 K (for both our and Si's structures). Minimum-to-maximum filling is reported for our a -MoS₂ RDF. The crystalline distributions, calculated with the ASE package, are rescaled for visibility.



The radial distributions explicitly capture the transition from crystalline periodicity to amorphous disorder. In the Mo-S pair distribution, the primary peak position and absolute intensity align perfectly between our a -MoS₂ structures and those from Si and Ma. This demonstrates that the fundamental metal-chalcogen covalent bond length is strictly preserved in the amorphous state, matching the position of the crystalline baselines exactly.

On the other hand, the S–S and Mo–Mo pair distributions reveal extreme chemical and topological deviations from the crystal lattices. Most notably, both amorphous networks exhibit a distinct S–S pre-peak at approximately 2.1 Å, which is entirely absent in the 2H and 1T phases. This peak is a clear indication of the formation of homopolar covalent disulfide (S_2^{2-}) bonds during the quench. Furthermore, the first Mo–Mo coordination shell in the amorphous state shifts to a significantly shorter distance relative to the crystalline structures. This is a clear sign of the densification of adjacent polyhedra, contrasting with the overall lower macroscopic density of the amorphous phases (see Table 3.2). This suggests that the melt glassifies into a more porous global network.

Crucially, the generated *a*-MoS₂ model exhibits exceptional agreement across all pairs with the reference structure from Si and Ma. The precise overlap of both the coordination shell distances and their (unnormalized) intensities confirms that the MLIP **M2.1** successfully captures the universal short-range chemistry of the amorphous network. Finally, beyond 6 Å, the distinct coordination peaks of the crystalline lattices persist, whereas the amorphous distributions rapidly converge to unity, as expected in the complete absence of long-range order.

By comparing the ADFs reported in Fig. 3.13, we note that the narrow peaks typical of the crystalline phases underwent two distinct modifications. While some smeared into broad distributions as expected from structural disordering (such as the Mo–Mo–Mo peak at 60°), others completely disappeared, like the characteristic S–Mo–S peak at 135°, indicating the loss or severe distortion of specific coordination geometries. Concurrently, new peaks emerged across all evaluated triplets, such as the features at ~ 92° and 140° for Mo–Mo–Mo, and the prominent peaks at 64° and 112° (appearing as a broad shoulder) for S–S–Mo. This again suggests either the heavy distortion of preexisting polyhedra or the population of new coordinations.

Interestingly, the primary ~ 82° intra-polyhedral S–Mo–S peak, present in both crystalline phases, is retained in the amorphous networks. This confirms that the fundamental triangular faces of the constituent polyhedra (whether prismatic or octahedral) are largely preserved, albeit heavily distorted. However, the characteristic 2H peak at 135° is marginally retained in Si’s and Ma’s structures but completely obliterated in our *a*-MoS₂ model. This peak is a strict geometric consequence of perfect trigonal prismatic symmetry, which is disrupted during the melt and angularly redistributed into a broad population approaching the 160° region.

9: In the 1T phase, the octahedra are compressed along the vertical axis owing to the layered structure, splitting the ideal 90° adjacent angle into ~ 84° and ~ 97°.

Furthermore, the near-right angle⁹ characteristic of the 1T octahedral geometry (~ 97°) appears as a prominent shoulder in the Si2023 reference and resolves into a clear peak in our *a*-MoS₂ model.

Finally, the acute S–Mo–S peak near 50° is a direct structural signature of the newly formed homopolar S–S bridges, which bind to single Mo centers in a side-on (η^2) coordination geometry. Given an S–S distance of ~ 2.0 Å (derived from the small pre-peak in the RDF) and this 50° angle, simple trigonometry deduces an average distance of ~ 2.36 Å between each bonded S atom and the central Mo, in excellent agreement with the primary Mo–S coordination shell. This exact disulfide coordination was also proposed in the experimental study by Tran *et al.* [86] on amorphous MoS₃.

The Mo–S–Mo angles reflect the reciprocal orientation of adjacent polyhedra. While the ordered structures result in single peaks at ~ 81.5° and ~ 83.5° for 2H and 1T, respectively, the amorphous phases exhibit a more diverse population of this angular feature. The main peak is shifted to ~ 73°, indicating a compression of the inter-polyhedral linkage, again in contrast to the dense packing of the crystalline phases (see Table 3.2), which corroborates the hypothesis of the formation of larger void regions in the glasses.

The angular distribution of S–S–Mo angles starkly differs between the amorphous and crystalline structures. The main crystalline peaks at ~ 50° are mostly lost upon amorphization, in favor of the emergence of two new large populations. The first one, at ~ 65°, perfectly complements the η^2 capping (assuming isosceles S–Mo–S triangles). The second broad and highly prominent peak at ~ 106° exhibits the typical shape of two partially overlapping peaks, one at ~ 103° and the other at ~ 111°. These discrete populations capture the inter-polyhedral orientations of the disulfide pairs bridging adjacent Mo centers, indicating two preferred stereochemical bridging modes that cross-link the amorphous network, completely overriding the discrete angular constraints of the crystalline lattices.

The Mo–Mo–Mo ADF, yielding information on the medium-range order, describes the symmetrical mutual arrangement of the metallic sub-network in the crystals and the curvature of the polyhedral chains in the glass. Notably, the relative intensity of the peak at 60° is dominant in the amorphous phases. This massive population is directly linked to the tight folding and clustering of the polyhedral chains, which collapse out of the flat crystalline plane into dense, localized triangular motifs.

The high agreement between the ADFs obtained in our study and those

from Si and Ma further corroborates the solidity of the MLIP **M2.1**. Once again, it proves capable of reproducing average geometrical features related to both short- and medium-range interactions.

Conversely, our much lower quench rates provide a significant expansion of the kinetic phase space compared to Si and Ma. The agreement of the basic structural data is therefore also a confirmation that their short range structures are not biased by the extreme quench rates commonly used in AIMD.

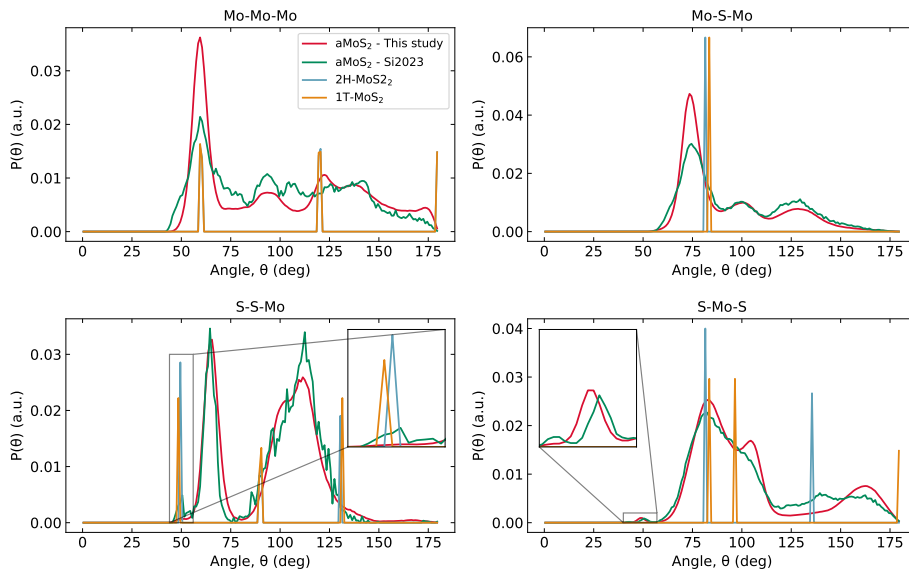


Figure 3.13: Angular distribution functions comparing our *a*-MoS₂ configurations to the Si2023 reference structure and the 2H and 1T crystalline phases. The amorphous distributions are obtained from the equilibration phase at 300 K (for both our and Si’s structures). The crystalline distributions, calculated with the ASE package, are rescaled for visibility. The central atom in the triplet is the second listed in the labels.

We proceeded to classify the coordination of every atom with ChemEnv, in order to establish a more comprehensive quantitative geometrical characterization of our structures. The results are shown in Fig. 3.14. Focusing first on the Mo centers, the complete suppression of the 135° S–Mo–S peak mathematically supports the disruption of the crystalline T:6 (trigonal prismatic) environments. In the amorphous state, the network is instead heavily dominated by O:6 (octahedral). Other coordinations, like S:5 (square pyramidal), T:6 and T:5 (trigonal bipyramidal) are also represented, but to much lower extent, accounting for less than 40%. Note that, owing to the severe structural frustration of the glass, the theoretical angles expected in such polyhedra are substantially distorted. This widespread polyhedral warping drives the emergence of the broad continuous populations in ADFs. This can be appreciated from the bottom panel of Fig. 3.14, where the CSM are plotted, along with their distribution. We note that the two mostly represented coordinations in the study by Si and Ma, octahedral and square pyramidal, are more distorted than those obtained with our melt-and-quench simulations. The variability of the distortion, represented

by the error bars in the violin plots, confirms that the systems do not resolve into precise distortions of each ideal polyhedra, but rather into a smeared range of distortions.

Importantly, restricting the analysis to Mo–S bonds explains the extreme CSM values observed for the Mo centers. While the algorithm does not explicitly evaluate the homopolar S–S bridges, it detects the geometric consequence: two Mo–S vectors pinched into a highly constrained 50° S–Mo–S angle. Because standard IUPAC reference polyhedra cannot accommodate such an acute internal angle, the presence of the side-on disulfide coordinates inherently limits the symmetry of the underlying geometry, severely inflating the CSM.

A similar reasoning can be used for the sulfur environments, which can be mapped to the Mo–S–Mo inter-polyhedral angles distribution. Because homopolar bonds are excluded, the sulfur speciation strictly quantifies how many Molybdenum centers a given sulfur atom bridges. The 2H and 1T crystalline structures rely exclusively on TY:3 (triangular non-coplanar) sulfur geometries bridging three Mo atoms. In the *a*-MoS₂ networks, this homogeneous speciation shatters into a highly diverse polyhedral mix that includes S:1 (terminal), A:2 (angular), SS:4 (see-saw), T:4 (tetrahedral), and TS:3 (T-shaped) environments. The emergence of A:2 coordinations (sulfurs bridging two Mo atoms) alongside heavily distorted TY:3 centers drives the massive smearing of the Mo–S–Mo distribution across the 60° to 110° range. Simultaneously, the emergence of SS:4 and TS:3 environments—which inherently feature wider equatorial angles to accommodate four or three Mo neighbors, respectively—mechanistically explains the new corner-sharing population observed at 125° to 130°. This confirms that the diverse Mo-bridging speciation of the S centers directly facilitates the three-dimensional cross-linking of the glassy network. We note that the CSMs of S atoms are much higher than those of Mo atoms. This extreme capacity for structural frustration is largely accommodated by the smaller atomic radius of sulfur, which permits more highly distorted local configurations. Finally, the emergence of terminal S atoms (S:1 coordination) is highly relevant from an application perspective, as such undercoordinated sites act as critical active centers for the catalytic properties of materials containing similar network motifs.

While the overall polyhedral speciation is similar between the two amorphous models, the ChemEnv analysis reveals a critical distinction regarding

the recovery of crystalline coordinations. The T:6 (trigonal prismatic) coordination is heavily suppressed in both amorphous models.

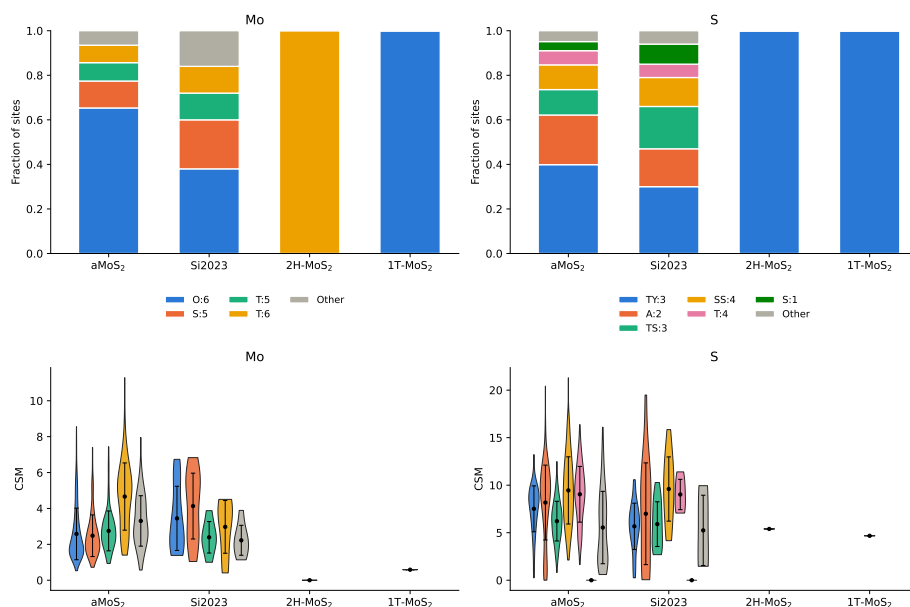


Figure 3.14: ChemEnv analysis of a - MoS_2 as obtained in our study and from Si and Ma, and crystalline references. The distributions in the violin plots are normalized so as to extend to a fixed maximum width. The error bars of a specific coordination represent the standard deviation over all atoms which are determined to belong to coordination. For our a - MoS_2 , polyhedra that individually do not account for at least 3% of the total are grouped under “Others”. Given the limited size of the cell, for the structure from Si and Ma we applied a threshold of 5%, in order to preserve readability.

Finally, the medium-range topology of the network is quantified through the distribution of irreducible alternating rings (Fig. 3.15). The ordered crystalline lattices are strictly constrained to small, densely packed ring structures. Specifically, the 2H phase is heavily dominated by 6-membered rings with a smaller fraction of 4-membered rings, while the 1T phase consists exclusively of 4-membered rings under this alternating definition.

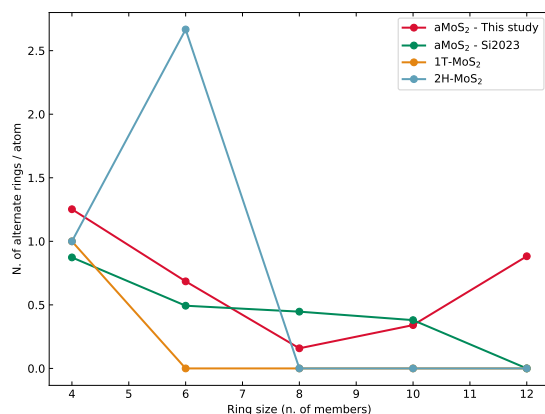


Figure 3.15: Irreducible ring statistics of a - MoS_2 as obtained in our study and from Si and Ma, and crystalline references. Only alternated rings (Mo-S-Mo) are considered.

The amorphization process fundamentally alters this local connectivity. While both amorphous networks retain a foundation of 4- and 6-membered rings necessary for basic polyhedral assembly, they also develop significant

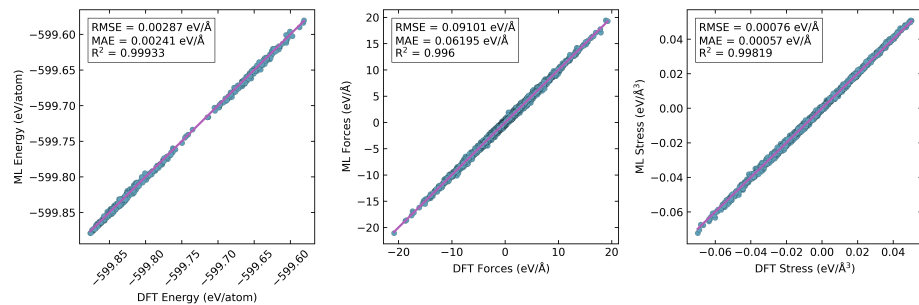
populations of larger rings. These 8-, 10-, and 12-membered rings represent extended topological chains that enclose substantial structural voids within the isotropic three-dimensional network. The emergence of these expanded ring structures provides the direct topological mechanism for the density anomaly discussed in the RDF and ADF analysis: although adjacent Mo polyhedra locally densify, the macroscopic density of the amorphous phase decreases due to the widespread formation of these porous mesoscopic voids.

Comparing the two amorphous models highlights the structural impact of the thermodynamic history. Both models successfully capture the transition toward mesoscopic porosity, developing comparable populations of 8- and 10-membered rings. However, the *a*-MoS₂ model generated in this study uniquely exhibits a prominent population of 12-membered rings, which are entirely absent in the Si and Ma reference. This structural divergence corroborates the advantages of our intermediate thermal protocol. The significantly lower quench rates applied in this study allow the network to explore a more extensive kinetic phase space, providing the necessary relaxation time for the formation of larger rings and extended void regions that extreme AIMD quench rates artificially suppress.

3.4.2 *a*-MoS₄ structural exploration

Most of the NPT simulations across all protocols and replicas were unstable, and led the system to uncontrolled expansion. For this reason, we decided to run NVT melt-and-quench simulations (still with the **M3** model), which proved to be more stable and ran to completion. The parity plots of **M3** are shown in Fig. 3.16.

Figure 3.16: Parity plots for energy, forces and stresses of the model **M3**, trained on *a*-MoS₄. The inference was done on the validation set (20% of the total data).



To assess the sensitivity of the melt-and-quench process to the different melting temperatures and cooling rates for the sulfur-rich stoichiometry, we computed the radial distribution functions (RDFs) on the final equilibrated configurations of the *a*-MoS₄ trajectories. For every pair (T_{melt} , γ), the RDFs

were averaged over the three replicas and are depicted in Fig. 3.17. The distributions of the radial atomic distances exhibit strict invariance to both the melting temperature and the quench rate. The primary peaks overlap perfectly across all nine protocols for every atomic pair, confirming that the fundamental bond lengths are dictated entirely by the local chemistry rather than the thermal history.

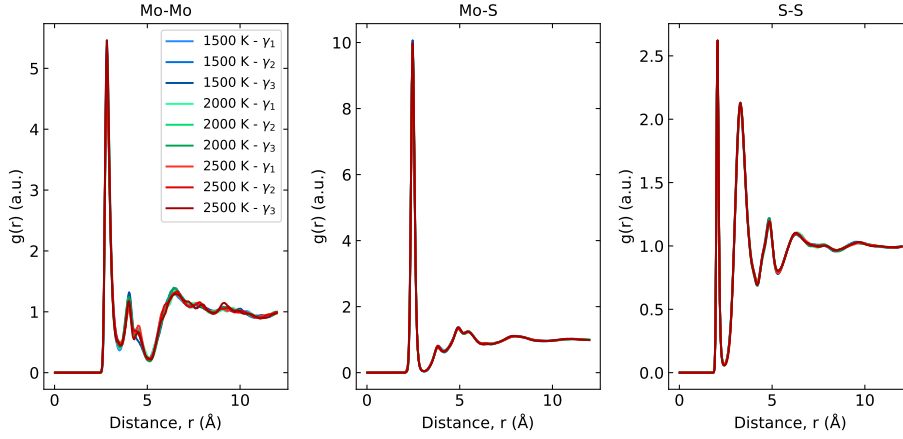


Figure 3.17: Sample-averaged radial distribution functions (RDFs) for a - MoS_4 evaluated across the nine thermal histories. The primary peaks illustrate strict positional invariance across all quench pathways.

Applying the cutoffs derived from the first minima of the RDFs, the angular distribution functions (ADFs) were computed to assess the three-dimensional geometry of the network (Fig. 3.18). Consistent with the radial distributions, the fundamental angular order remains robust. The peak positions across all triplet combinations (S-Mo-S, Mo-S-Mo, Mo-Mo-Mo, and S-S-Mo) remain rigidly fixed across the protocols (e.g., maintaining invariant peak locations down to the exact $\sim 50^\circ$ and $\sim 80^\circ$ maxima). This perfect overlap confirms that the local angular topology is completely stable against variations in the cooling protocol, within the range we covered.

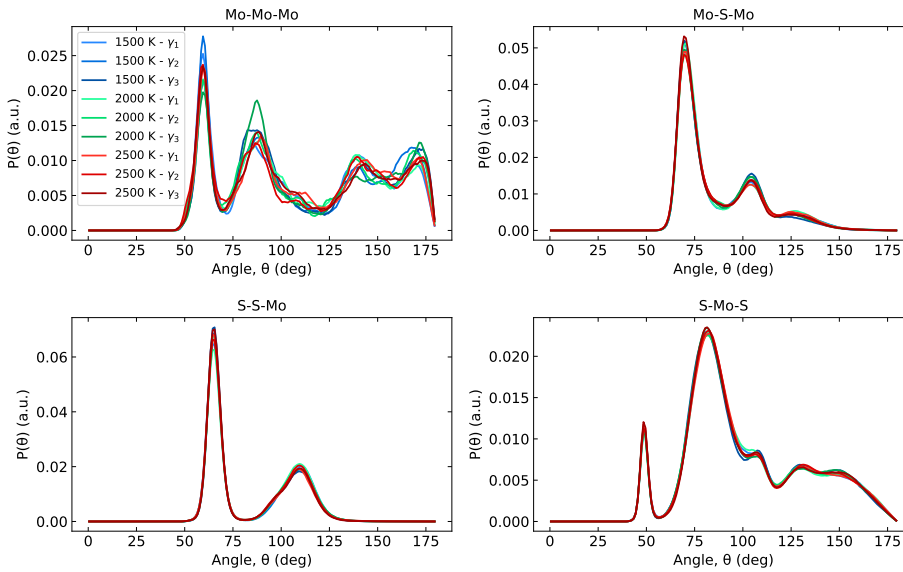


Figure 3.18: Sample-averaged angular distribution functions (ADFs) for a - MoS_4 evaluated across the nine thermal histories, highlighting minimal intensity variance and perfect overlap across all cooling rates.

In Fig. 3.19 we report the ChemEnv speciation for all protocols. The polyhedral analysis corroborates the short-range order invariance observed in the RDFs and ADFs. The relative fractions of both Mo and S coordination environments remain essentially constant across all nine protocols, confirming that the local structural motifs are inherent to the α -MoS₄ phase and do not depend on the specific melting temperature or quench velocity.

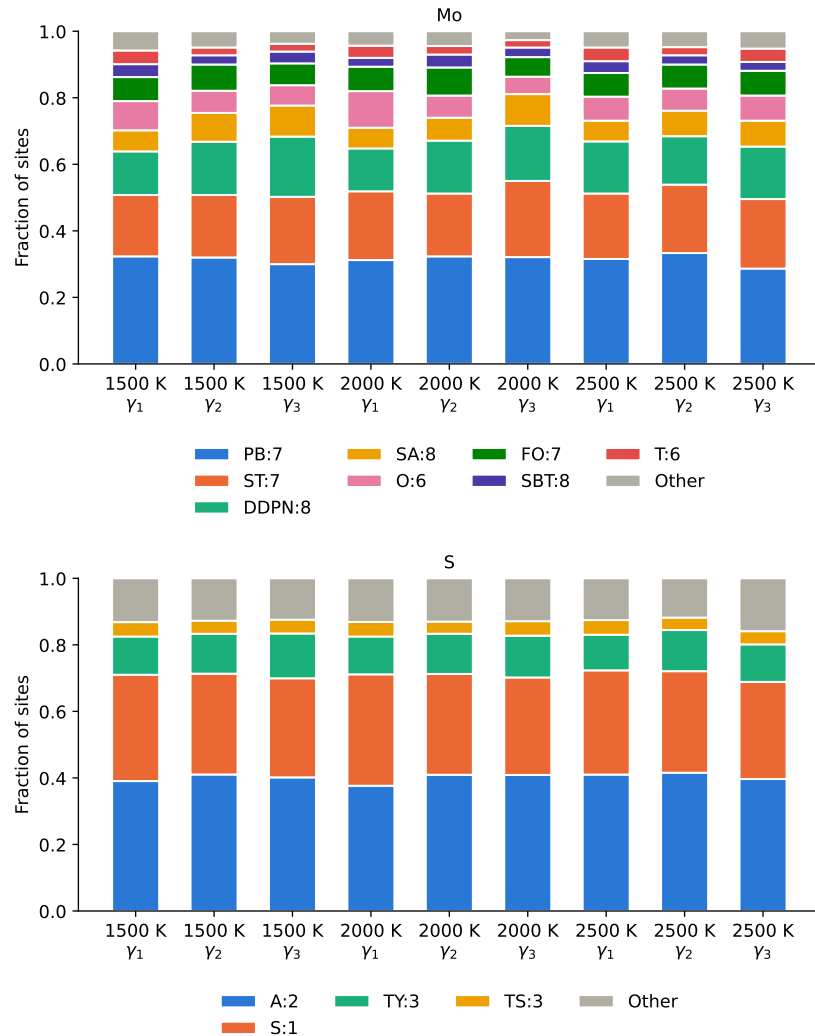


Figure 3.19: Fraction of local coordination environments for Mo (top) and S (bottom) in α -MoS₄ across the nine thermal histories. The relative populations of the polyhedral geometries remain completely stable. Polyhedra that individually do not account for at least 3% of the total are grouped under “Others”.

From the irreducible rings distribution, shown in Fig. 3.20), we notice that the populations of small rings (such as 4- and 6-membered rings) are not significantly affected by the different protocol settings. However, similar to previous observations, the larger mesoscopic rings (such as 10- and 12-membered rings) exhibit considerable variance depending on the specific thermal history. This indicates that while the local polyhedra assemble identically, their large-scale mesoscopic packing remains susceptible to kinetic trapping during the quenching phase.

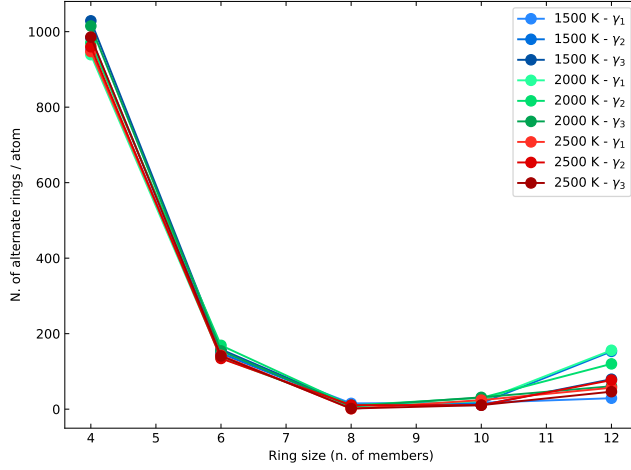


Figure 3.20: Distribution of alternate Mo-S rings in the a -MoS₄ network. While small rings exhibit consistent populations, large mesoscopic rings (10- and 12-membered) display variance based on the thermal history.

Finally, we report the vibDOS in Fig. 3.21. The high-frequency optical bands overlap perfectly across all quench rates and melting temperatures, confirming the stability of the stiff, localized bonds. Conversely, the low-frequency modes (< 1 THz) exhibit subtle changes depending on the cooling rate. However, because these low-frequency acoustic modes stem from large-scale structural inhomogeneities, their thermal variance is entirely compatible with the topological fluctuations observed in the extended mesoscopic ring statistics.

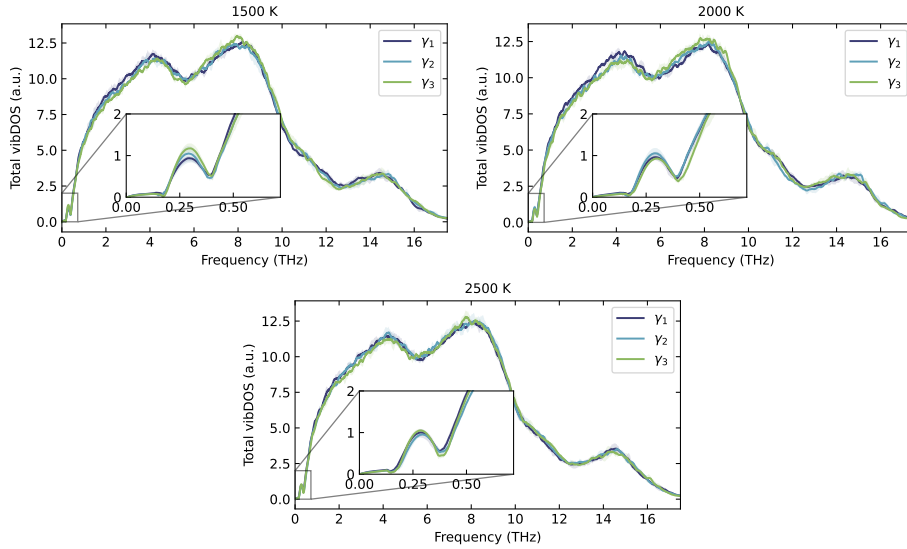
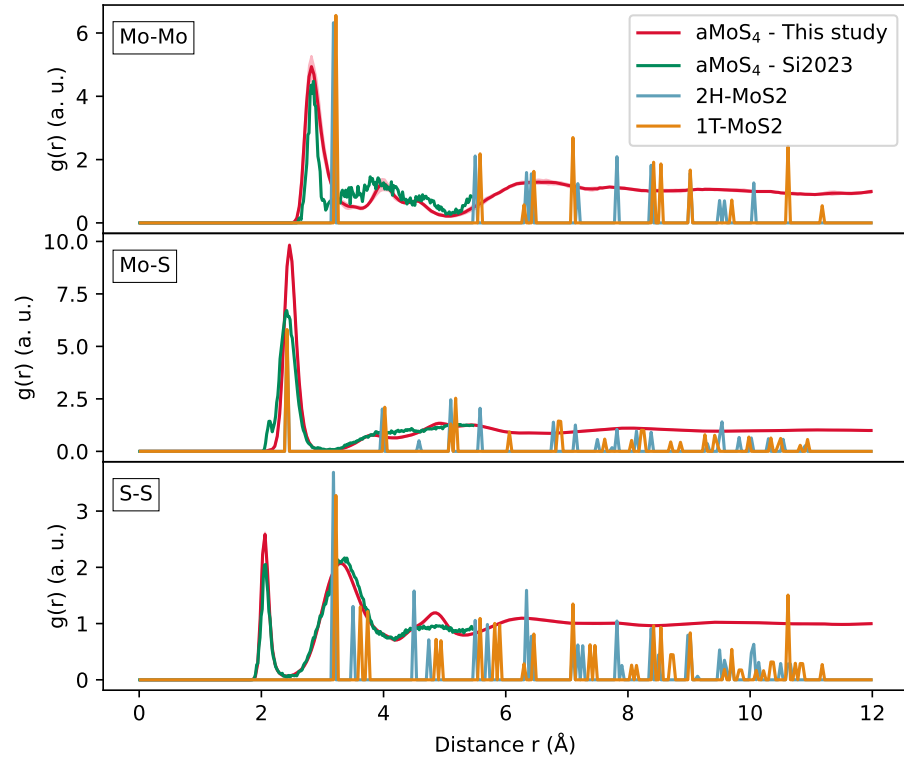


Figure 3.21: Total vibrational density of states (vibDOS) for the a -MoS₄ networks.

Having established the robustness of the short- and medium-range order against variations in the quenching protocol, we now benchmark the structural topology of the a -MoS₄ phase against the standard crystalline references (2H and 1T) and the computational glass models reported by Si and Ma [90]. For this detailed structural analysis, we utilize the trajectory generated by our intermediate protocol ($T_{\text{melt}} = 2000$ K, γ_2).

Figure 3.22: Radial distribution functions comparing our *a*-MoS₄ configurations to the Si2023 reference structure and the 2H and 1T crystalline phases. The amorphous distributions are obtained from the equilibration phase at 300 K (for both our and Si's structures). Minimum-to-maximum filling is reported for our *a*-MoS₄ RDF. The crystalline distributions, calculated with the ASE package, are rescaled for visibility.



10: Again, the crystal phases of MoS₂ are rescaled for visibility.

The RDFs¹⁰, in Fig. 3.22, immediately reveal how the excess sulfur forces the amorphous network to deviate entirely from standard crystalline constraints. The *a*-MoS₄ RDF is dominated by a massive, sharp S-S peak at ~ 2.05 Å. This is in stark contrast with the lowest S-S distance peak at ~ 3.1 Å found in 2H-MoS₂, which arise from van der Waals non-bonding interactions, and which is still present, as the second peak, in *a*-MoS₄ RDF. This proves the widespread formation of covalent disulfides (S₂²⁻) and polysulfide chains, a chemical feature that the conventional crystalline phases cannot accommodate. Similarly, the Mo-Mo RDF displays a prominent clustering peak at ~ 2.7 – 2.8 Å, which is significantly shorter than the standard Mo-Mo distance in the crystalline phase of MoS₂ (~ 3.16 Å). Interestingly, a small peak present in the Mo-S RDF of Si and Ma, at slightly lower distances than the main peak, is redistributed in the main peak of our structures. Apart from that and a small peak arising at ~ 4.8 Å in the S-S RDF of our *a*-MoS₄, the RDFs that we obtained align very well to those from Si and Ma.

The structural divergence from the crystalline phases of MoS₂ is also displayed by the angular distribution functions (ADFs), as reported in Fig. 3.23. The 2H crystalline phase dictates a rigid trigonal prismatic geometry characterized by sharp S-Mo-S angles at roughly 82° and 136°, while the 1T phase forces an octahedral coordination with characteristic angles of 90° (split into two distorted peaks) and 180°. The *a*-MoS₄ network abandons

both of these highly symmetric motifs. While retaining the main peak of 2H-MoS₂, as a broad population, it also exhibits a sharp maximum at $\sim 50^\circ$. Such peak was already observed in *a*-MoS₂, and is compatible with the η^2 -capping coordination. This extreme angular variation indicates that the Mo polyhedra are severely distorted by the steric demands of the sulfur-rich environment.

The Mo–S–Mo ADF, related to the inter-polyhedron angles, strongly deviates from the crystalline cases too. In agreement with Si and Ma, our amorphous phases exhibit two peaks, at smaller and larger angles than the single one found in the crystals.

As for the S–S–Mo angles, again those of *a*-MoS₄ neatly depart from the crystalline ones, resolving in a prominent main peak at $\sim 65^\circ$, compatible with the η^2 -coordination of disulfide, and a broader one at $\sim 109^\circ$. Our peak position and broadening are similar to those from Si and Ma, but with a slight difference in the height ratio.

Finally, the Mo–Mo–Mo angle distribution, which reflects the medium-range structure along polyhedra chains, exhibits a main peak at $\sim 60^\circ$ shared among all materials. While both our structures and those from Si and Ma deviate from the narrow and discrete peaks of the crystals, ours gives a much broader distribution over the angle axis, smearing into four broad bumps at 60° , 87° , 137° and 160° .

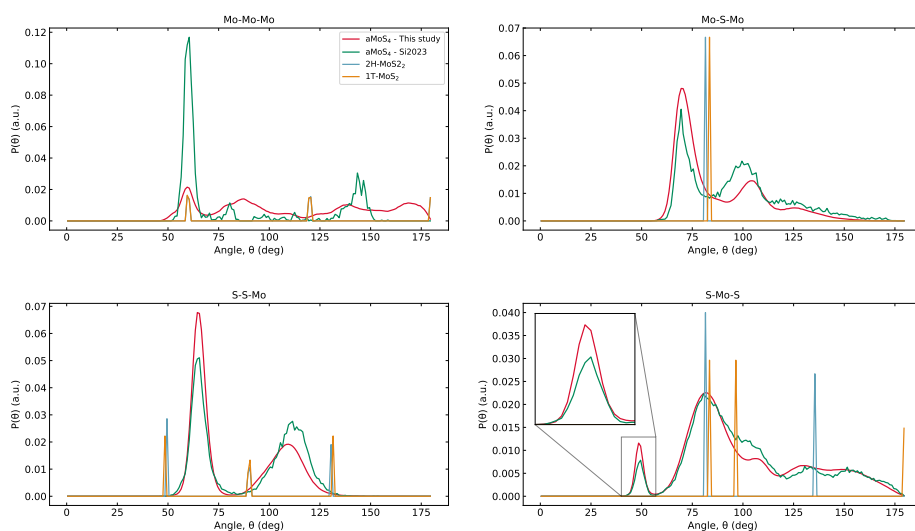
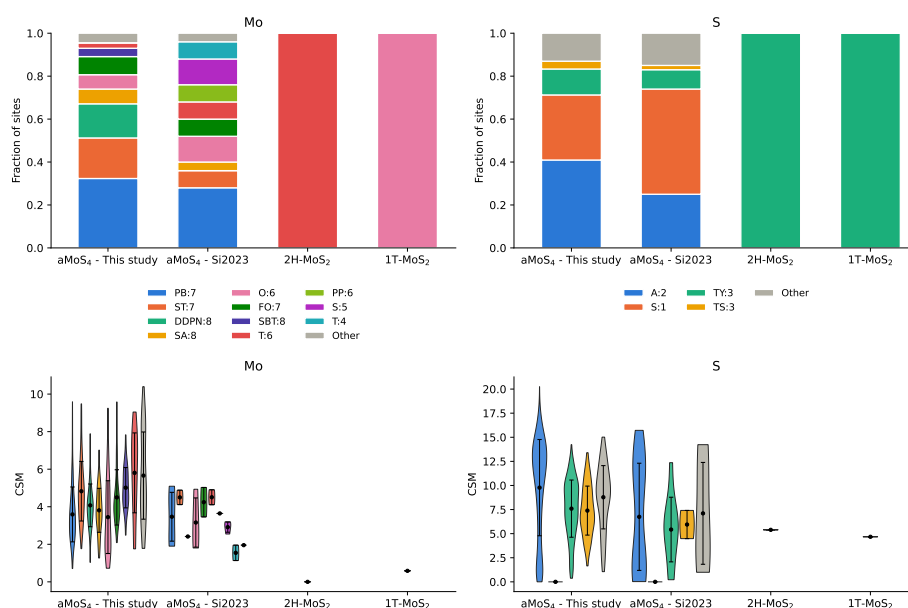


Figure 3.23: Angular distribution functions comparing our *a*-MoS₄ configurations to the Si2023 reference structure and the 2H and 1T crystalline phases. The amorphous distributions are obtained from the equilibration phase at 300 K (for both our and Si’s structures). The crystalline distributions, calculated with the ASE package, are rescaled for visibility. The central atom in the triplet is the second listed in the labels.

From the ChemEnv analysis reported in Fig. 3.24, we observe that our amorphous phase, similarly to the structures from Si and Ma, shows a total disruption of the octahedral (O:6) and trigonal prismatic (T:6) Mo coordinations typical of the crystalline phases. Instead, the pentagonal

bipyramidal (PB:7) geometry is preferred, accounting for $\sim 35\%$ of the total Mo environments. We notice that, while the glass from Si and Ma retain lower coordinations of Mo, like T:4 and S:5, to a considerable extent ($\sim 20\%$), our a -MoS₄ mainly settle into higher-coordination geometries, departing from the crystalline cases more strongly. Finally, the abundant presence of angular (A:2) and terminal (S:1) sulfur geometries suggests the loss of cross-linking of polyhedra chains, in favor of linear chains, which can better accommodate the increased amount of sulfur. From the bottom panels of Fig. 3.24, we also note that the coordinations, both low and high, are severely distorted, as a result of the extra sulfur atoms.

Figure 3.24: ChemEnv analysis of a -MoS₄ as obtained in our study and from Si and Ma, and crystalline references. The distributions in the violin plots are normalized so as to extend to a fixed maximum width. The error bars of a specific coordination represent the standard deviation over all atoms which are determined to belong to a given coordination. For our a -MoS₄, polyhedra that individually do not account for at least 3% of the total are grouped under “Others”. Given the limited size of the cell, for the structure from Si and Ma we applied a threshold of 5%, in order to preserve readability.



As depicted in Fig. 3.25, the alternate Mo-S topological network of our a -MoS₄ mainly resolves into small rings with 4 and 6 members, similarly to what is observed in the simulation by Si and Ma. Notably, 2H-MoS₂ exhibits a strikingly larger population of 6-membered rings, while the 1T phase and the amorphous cases do not. Finally, we observe a considerable population of 12-membered rings in our a -MoS₄, which are entirely absent in the Si and Ma amorphous reference. This resurgence of large mesoscopic loops in our model can be directly attributed to the differences in the extended metallic framework, as previously indicated by the angular distribution functions. The broader angular flexibility and diverse high-coordination polyhedra in our a -MoS₄ model allow the Mo and S atoms to maintain a more continuous, extended polymeric network capable of folding and

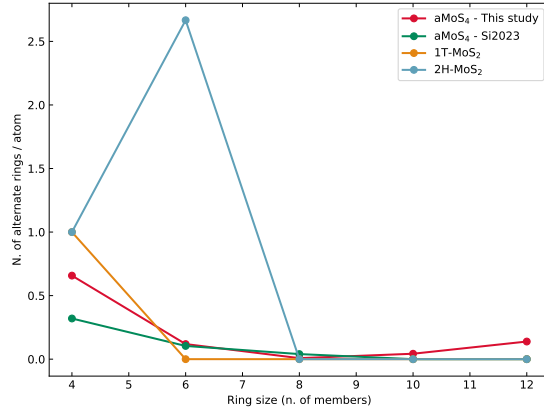


Figure 3.25: Irreducible ring statistics of a -MoS₄ as obtained in our study and from Si and Ma, and crystalline references. Only alternated rings (Mo-S-Mo) are considered.

closing large 12-membered loops across the mesostructure. Furthermore, as demonstrated in our earlier analysis of sensitivity to the MD protocol, the population of these large, 10- and 12-membered mesoscopic rings is highly susceptible to kinetic trapping during the quenching phase.

3.4.3 a -MoS₆ structural exploration

To evaluate the structural topology at the extreme end of the sulfur-rich stoichiometry, we subjected the a -MoS₆ phase to the same systematic analysis. Also in this case the NPT simulations were unstable, so we ran the melt-and-quench in the NVT ensemble, with the **M4** model. The parity plots of **M4** are shown in Fig. 3.26. We first assessed the sensitivity of the

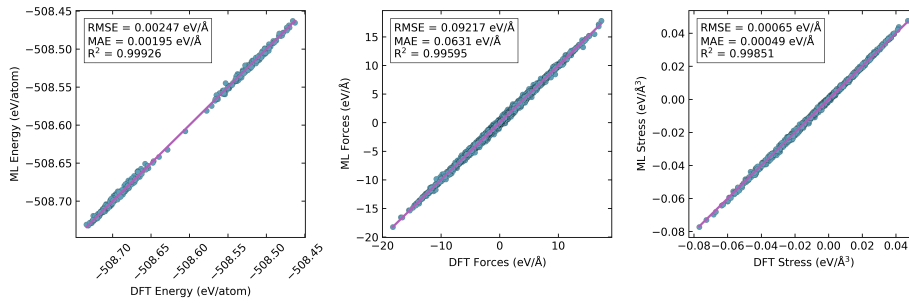
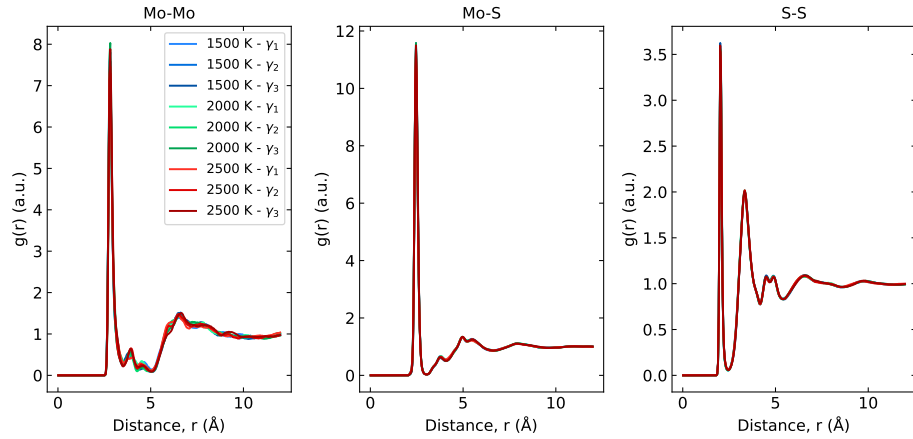


Figure 3.26: Parity plots for energy, forces and stresses of the model **M4**, trained on a -MoS₆. The inference was done on the validation set (20% of the total data).

a -MoS₆ melt-and-quench process to the different melting temperatures and cooling rates. The sample-averaged radial distribution functions (RDFs) evaluated across the nine thermal histories are depicted in Fig. 3.27. As for the previous stoichiometries, the distributions are invariant upon melting at higher temperature, or applying different cooling rates.

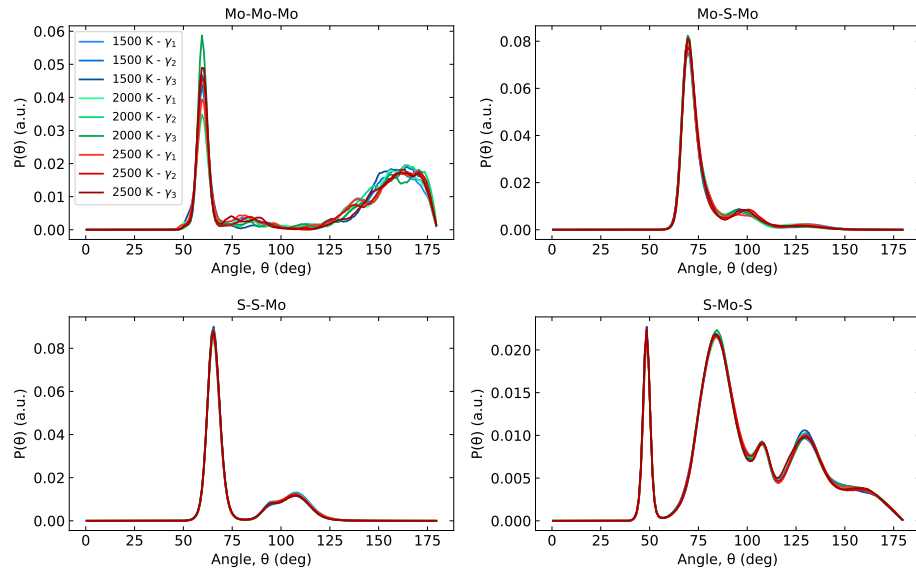
Adopting the cutoffs derived from the RDF minima, the angular distribution functions (ADFs) (Fig. 3.28) confirm that the fundamental angular order remains robust. The peak positions across all triplet combinations remain rigidly fixed across the nine protocols. The perfect overlap highlights minimal

Figure 3.27: Sample-averaged radial distribution functions (RDFs) for α -MoS₆ evaluated across the nine thermal histories.



intensity variance and proves the local angular topology is completely stable against variations in the cooling rate.

Figure 3.28: Sample-averaged angular distribution functions (ADFs) for α -MoS₆ evaluated across the nine thermal histories.



The ChemEnv classification reported in Fig. 3.29 corroborates this short-range order invariance. The relative fractions of both Mo and S coordination polyhedra remain stable across all nine protocols. However, it is possible to appreciate an increase in the population of dodecahedra with triangular faces (DDPN:8) with slower quenching. This effect is evident when melting at 1500 K, but becomes more subtle when $T_{\text{melt}} = 2000$ K, and even inverts with $T_{\text{melt}} = 2000$ K.

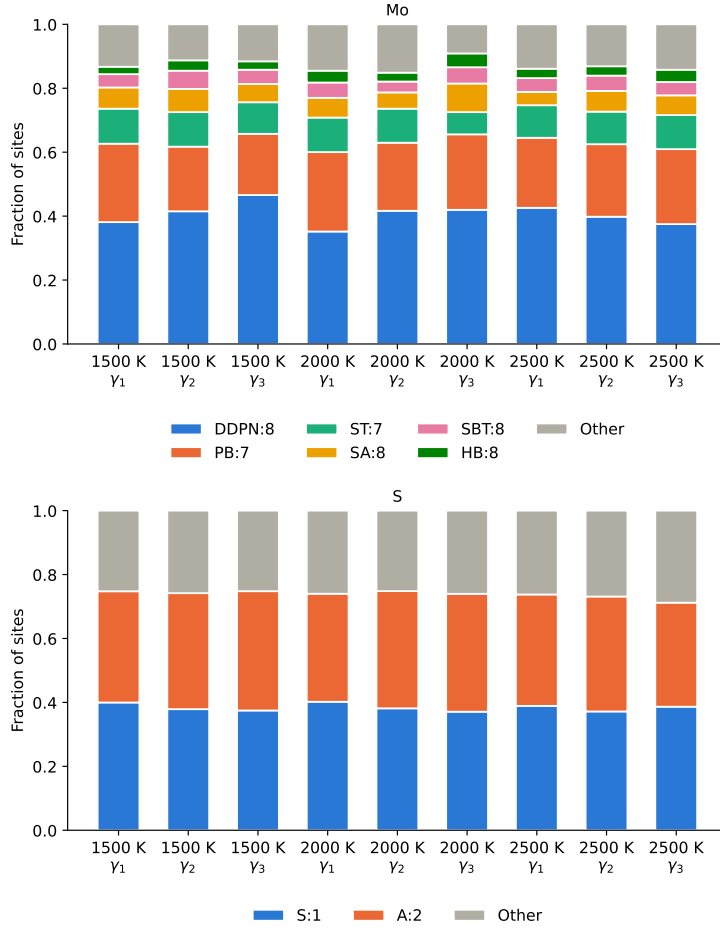


Figure 3.29: Fraction of local coordination environments for Mo (top) and S (bottom) in a -MoS₆ across the nine thermal histories. Polyhedra that individually do not account for at least 3% of the total are grouped under “Others”.

Interestingly, the irreducible ring distribution (Fig. 3.30) reveals a departure from the kinetic sensitivity observed in the lower stoichiometries. The large mesoscopic rings (8-, 10-, and 12-membered) are entirely suppressed across all thermal histories, dropping to zero. The network is strictly dominated by small 4-membered rings, whose populations exhibit negligible variance regardless of the protocol. This indicates that the extreme sulfur excess so heavily interrupts the alternating Mo–S connectivity that large-scale mesoscopic packing is no longer possible.

Finally, the total vibrational density of states (vibDOS) (Fig. 3.31) confirms the dynamic stability of the network. The high-frequency optical bands overlap perfectly across all quench rates, while the low-frequency acoustic modes (< 1 THz) exhibit only subtle thermal variance.

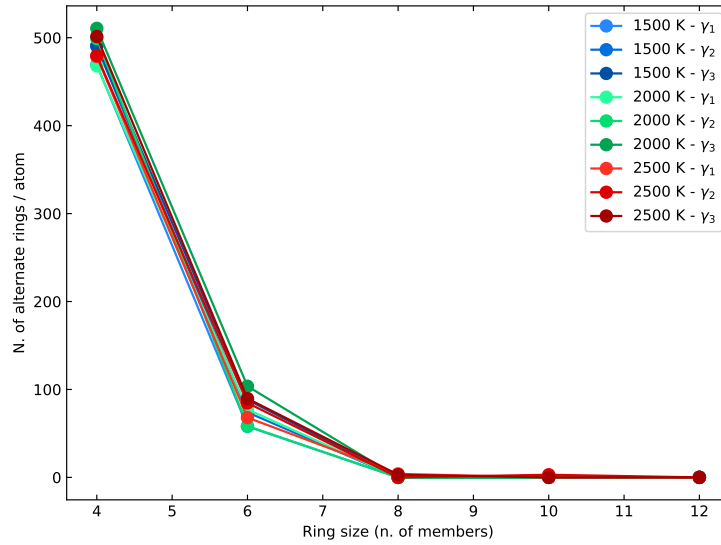


Figure 3.30: Distribution of alternate Mo-S rings in the a -MoS₆ network.

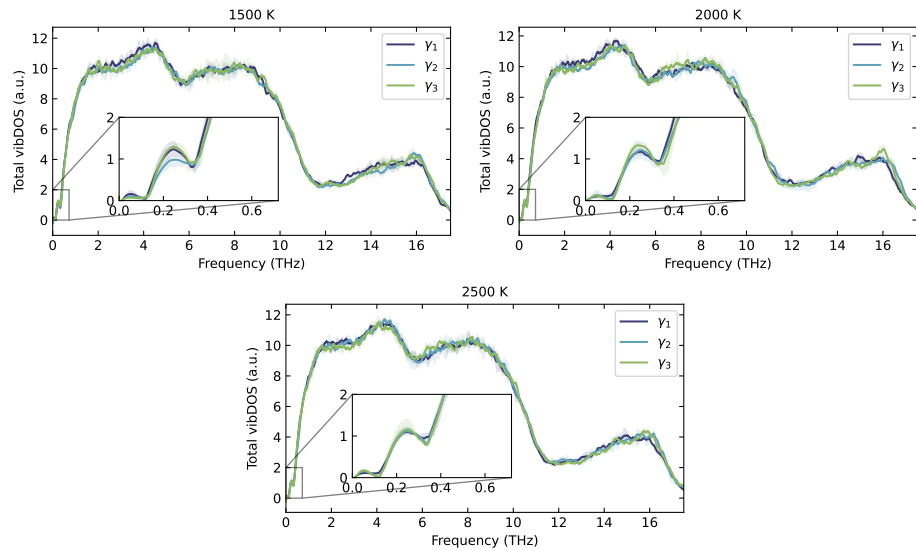


Figure 3.31: Total vibrational density of states (vibDOS) for the a -MoS₆ networks.

We then benchmarked the structural topology of a -MoS₆ against the 2H and 1T crystalline references and the amorphous models reported by Si and Ma. Again, we utilized the trajectory generated by our intermediate protocol ($T_{\text{melt}} = 2000$ K, γ_2).

The RDFs (Fig. 3.32) show that the massive S-S covalent bonding peak at ~ 2.05 Å becomes even more dominant, proving the extreme proliferation of polysulfide chains which are not observed in the low stoichiometry ratio of the crystalline phases. The Mo-Mo RDF displays the characteristic clustering peak at ~ 2.7 Å, aligning closely with the primary peak of the Si and Ma structure. The Mo-S RDF of our a -MoS₆ shows a higher main peak, but follows a similar broadened profile which quickly converges to 1.

The ADFs (Fig. 3.33) further emphasize the structural divergence. The S-

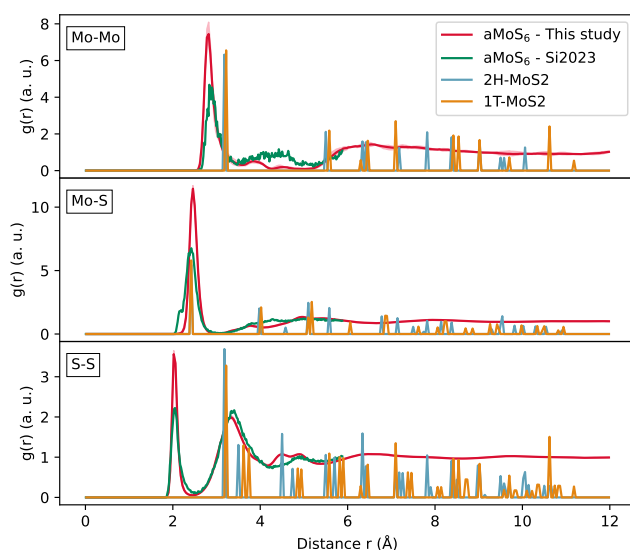


Figure 3.32: Radial distribution functions comparing our *a*-MoS₆ configurations to the Si2023 reference structure and the 2H and 1T crystalline phases.

Mo–S distribution exhibits the sharp $\sim 50^\circ$ maximum typical of η^2 -capping coordinations, completely abandoning the highly symmetric discrete angles of the 2H and 1T crystals.

The Mo–S–Mo ADF of our *a*-MoS₆ exhibits a very prominent peak at $\sim 70^\circ$, and a smaller cluster around 96° . This is in slight contrast with what we observed in the structure from Si and Ma, that is a main peak at $\sim 74^\circ$ and a smaller but consistent uniform spread which decays to zero after 150° .

The S–S–Mo distribution, again has a large peak at $\sim 65^\circ$, compatible with the haptic (η^2) coordination, and broader and more modest population at 108° . This is in good agreement with the reference amorphous structures.

Interestingly, the Mo–Mo–Mo angular distribution shows important differences between our *a*-MoS₆ and that from Si and Ma. While in the first one there is a main peak at $\sim 60^\circ$, the reference shows two peaks, at $\sim 52^\circ$ and $\sim 64^\circ$. Furthermore, of the three broad bumps Si and Ma found at 83° , 115° and 140° , the first two are mostly lost, and the last one vanishes completely, and the population massively shifts towards higher angles ($\sim 170^\circ$).

The ChemEnv analysis is reported in Fig. 3.34. In our *a*-MoS₆, the Mo atoms completely abandon the 6-fold coordinated geometries in favor of highly packed environments, with the 8-fold coordinated dodecahedron with triangular faces (DDPN:8) becoming the dominant polyhedron ($\sim 40\%$ of the total). On the contrary, the model from Si and Ma exhibits a high population of lower coordinations, with T:4, S:5, and T:5 accounting for $\sim 30\%$.

This discrepancy can be attributed to the vastly different kinetic histories of the two models. The reference structures generated via *ab initio* molecular

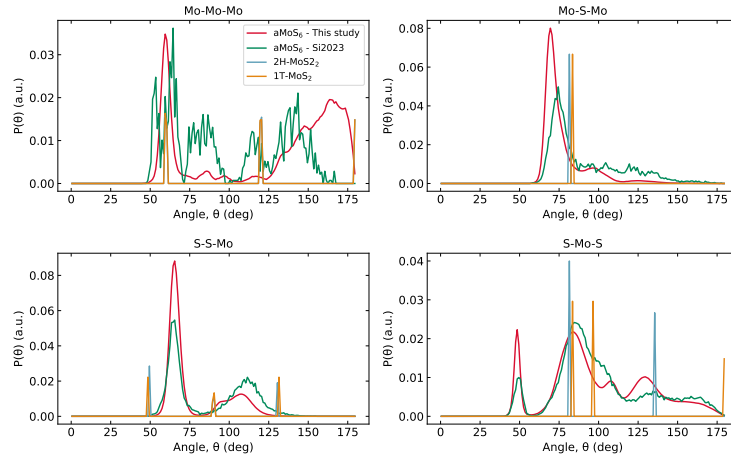


Figure 3.33: Sample-averaged angular distribution functions (ADFs) for a -MoS₆, comparing our active-learning trajectory with the Si2023 reference and the crystalline baselines.

dynamics rely on ultrafast quenching rates ($\sim 10^{15}$ K/s). At these extreme speeds, the network lacks sufficient time for local structural relaxation, kinetically trapping the molybdenum centers in undercoordinated, high-energy configurations. In contrast, the significantly slower cooling rates enabled by our MLMD approach allow the Mo atoms to fully saturate their coordination spheres. From a purely chemical perspective, this high coordination is rigorously consistent with the extreme $x = 6$ stoichiometry: accommodating a massive excess of sulfur—which our model identifies almost entirely as terminal (S:1) and angular (A:2) species—sterically and electronically demands highly packed Mo centers to maintain cohesive cross-linking.

The Continuous Symmetry Measure (CSM) distributions confirm that these high-coordination polyhedra are severely distorted for both amorphous models.

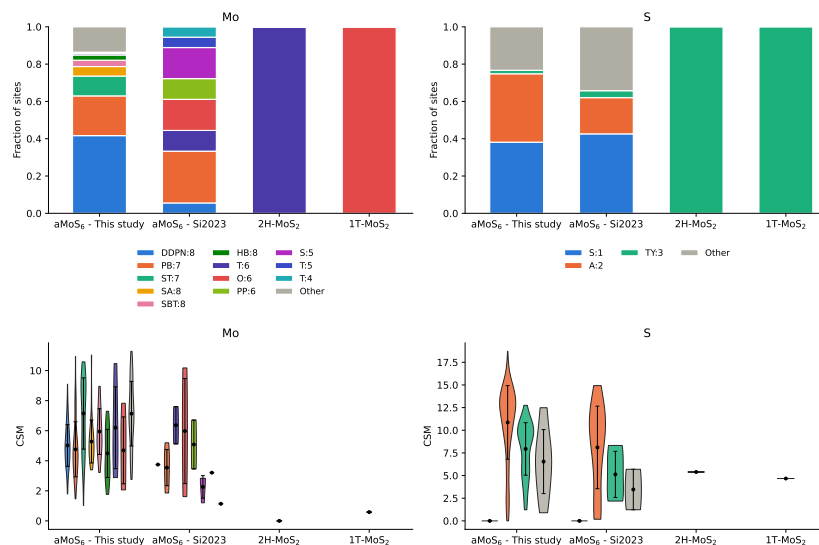


Figure 3.34: ChemEnv analysis of a -MoS₆ as obtained in our study, compared to Si and Ma and crystalline references. For our a -MoS₆, polyhedra that individually do not account for at least 3% of the total are grouped under “Others”. Given the limited size of the cell, for the structure from Si and Ma we applied a threshold of 5%, in order to preserve readability.

Finally, the irreducible ring statistics (Fig. 3.35) demonstrate the complete macroscopic collapse of the alternating metal-sulfur network. Both our a -MoS₆ model and the Si2023 structure resolve almost exclusively into isolated 4-membered rings. Notably, the 12-membered rings that were present in our a -MoS₄ network are entirely absent here. This strongly suggests that at the $x = 6$ stoichiometry, the massive excess of flexible polysulfide chains heavily fragment the alternating Mo–S–Mo framework, preventing the network from folding into large continuous mesoscopic loops. This may also be observable experimentally, in the mechanical stability of the amorphous material.

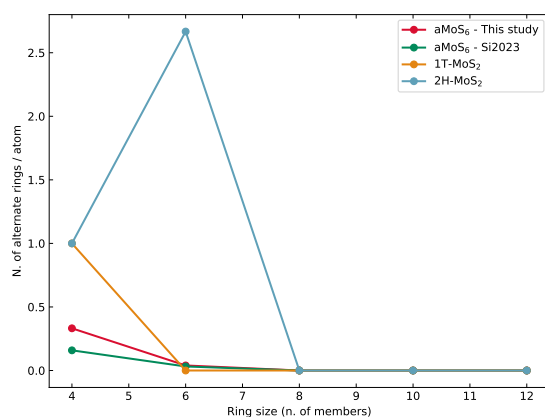


Figure 3.35: Irreducible ring statistics of a -MoS₆ as obtained in our study, from Si and Ma, and from crystalline references.

3.4.4 Stoichiometry comparison

Having analyzed the a -MoS₂, a -MoS₄, and a -MoS₆ phases individually, we can now synthesize these observations to extract the fundamental topological evolution of the amorphous network as a function of stoichiometry.

The most immediate chemical consequence of increasing the sulfur content is the breakdown of the classical Mo–S–Mo alternating lattice in favor of homopolar S–S bonding. As the stoichiometry increases from $x = 2$ to $x = 4$ and $x = 6$, the radial distribution functions (Fig. 3.36) reveal the emergence and subsequent explosion of the covalent S–S peak at ~ 2.05 Å. This signals the creation of an expanding matrix of polysulfide chains or bridging units.

This macroscopic fragmentation of the metal-sulfur network is perfectly tracked by the local ChemEnv speciation (Fig. 3.37). From the perspective of the sulfur atoms, the addition of excess sulfur drives a strict, monotonic reduction in coordination. The 3-coordinated nodes typical of a -MoS₂ are replaced entirely by angular (A:2) and terminal (S:1) species in a -MoS₄

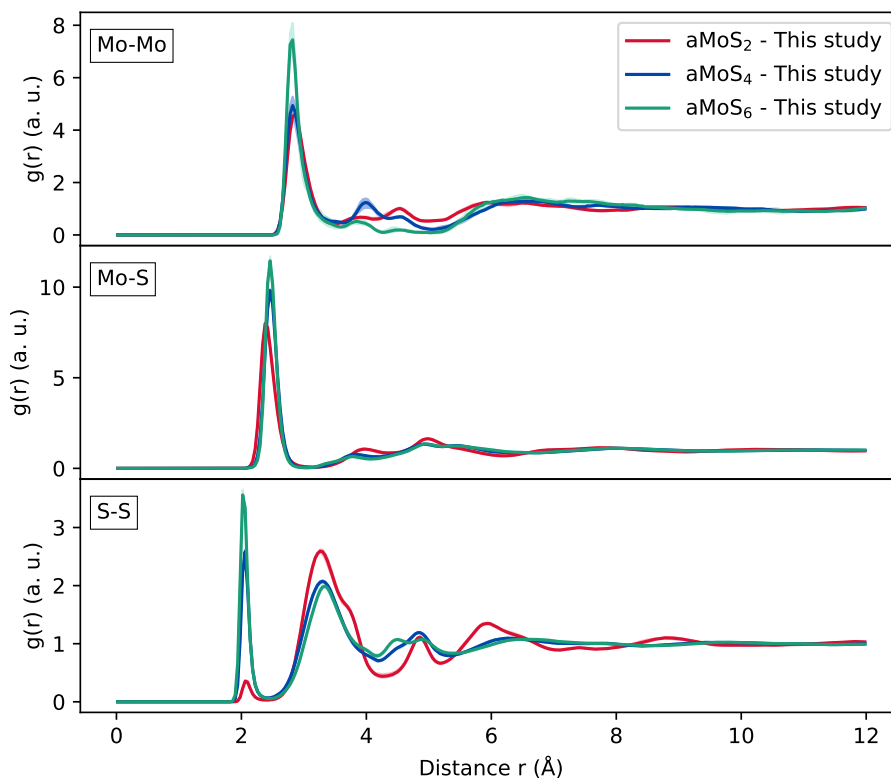


Figure 3.36: Comparison of the sample-averaged radial distribution functions (RDFs) across the three stoichiometries ($x = 2, 4, 6$). Note the emergence and rapid growth of the S-S covalent bonding peak at ~ 2.05 Å.

and $a\text{-MoS}_6$. Conversely, the molybdenum atoms are forced into increasingly crowded, high-coordination polyhedra to accommodate the bidentate binding of these polysulfide chains. The primary Mo geometry shifts from 6-coordinated environments in $a\text{-MoS}_2$, to 7-coordinated pentagonal bipyramids (PB:7) in $a\text{-MoS}_4$, and finally culminates in the massive dominance of 8-coordinated dodecahedra with triangular faces (DDPN:8) in $a\text{-MoS}_6$. Furthermore, the increase of the sulfur content leads to progressively more distorted coordinations, as shown by the increase of the CSM in the bottom panels of Fig. 3.37. The geometric constraints of packing these bulky, high-coordination polyhedra physically explain the progressive broadening and severe distortion of the angular distribution functions (ADFs) (Fig. 3.38). Most notably, the S-Mo-S and S-S-Mo distributions reveal the emergence of sharp acute angles at $\sim 50^\circ$ and $\sim 65^\circ$, respectively. These features, which scale proportionally with the sulfur content, are the characteristic signature of η^2 -capping disulfides bonded tightly to a central Mo atom.

This transition from a rigidly cross-linked metal-chalcogenide lattice to a flexible, chain-dominated polymer dictates the medium-range mesostructure (Fig. 3.39). In $a\text{-MoS}_2$, the irreducible ring statistics display a broad distribution of extended, alternating Mo-S sequences. In $a\text{-MoS}_4$, the proliferation of S-S chains effectively "short-circuits" these extended metal-sulfur

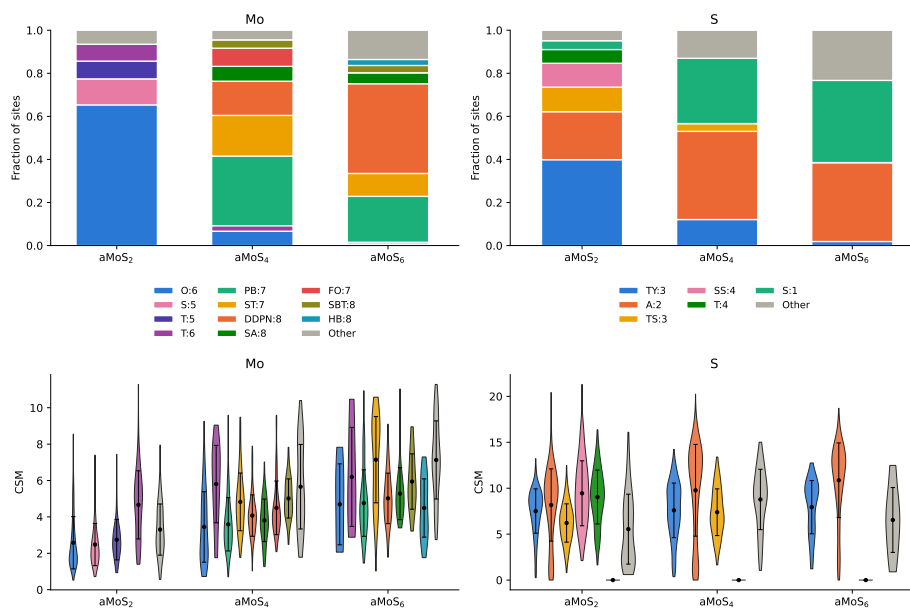


Figure 3.37: ChemEnv classification of local coordination environments across the three stoichiometries. The Mo environments shift toward higher coordination (from 6 to 8), while S environments transition entirely to 1- and 2-coordinated species.

connections, causing the structural voids to collapse into tight 4-membered rings, though the network retains enough connectivity to fold into large 12-membered mesoscopic loops. By $x = 6$, the sheer volume of polysulfide chains completely suppresses any extended alternating connectivity; the 12-membered loops vanish, leaving a fragmented Mo-S topology restricted almost entirely to isolated 4-membered rings.

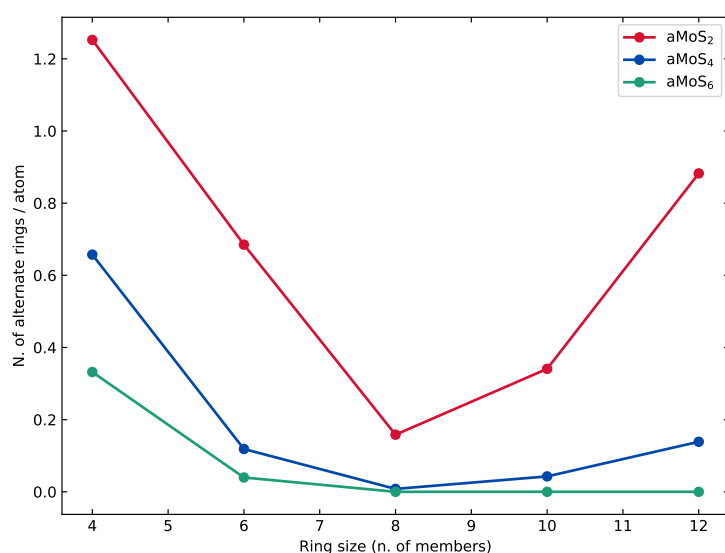


Figure 3.39: Irreducible ring statistics across the three stoichiometries. Note the collapse of extended alternating connectivity as sulfur content increases.

Finally, the dynamic signature of this stoichiometric evolution is captured by the vibrational density of states (Fig. 3.40). As the sulfur concentration increases from $x = 2$ to $x = 6$, new high-frequency optical modes between 14 and 17 THz emerge and intensify. These modes, completely absent in the stoichiometric baseline, serve as the definitive dynamic fingerprint of

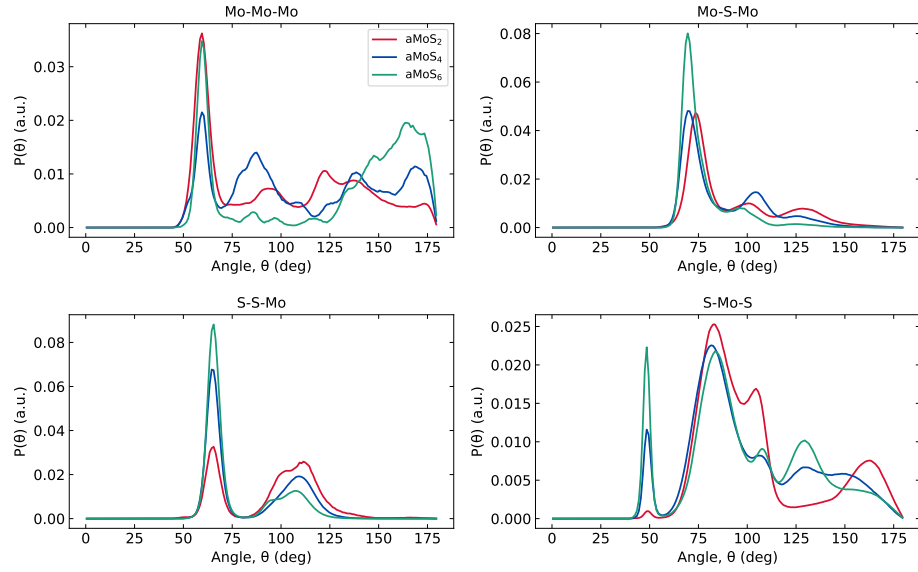


Figure 3.38: Angular distribution functions (ADFs) across the three stoichiometries, highlighting the shift toward acute, η^2 -capping geometries and the severe broadening of the Mo polyhedra.

the rigid S–S covalent stretching within the polysulfide chains, scaling directly with the macroscopic sulfur excess. Indeed, as shown by the partial density of states in Fig. 3.41, the region beyond 14 THz exhibits absolutely no vibrational contribution from molybdenum. As a confirmation, we can observe two neat peaks in the experimental Raman spectra of amorphous MoS_3 by in refs. [86] and [112], which were assigned to terminal, bridging and shared S–S stretching. Concurrently, the low-frequency acoustic modes (< 1 THz) shift significantly to lower frequencies as stoichiometry increases, which is compatible with the transition towards a less rigid, highly polymeric network.

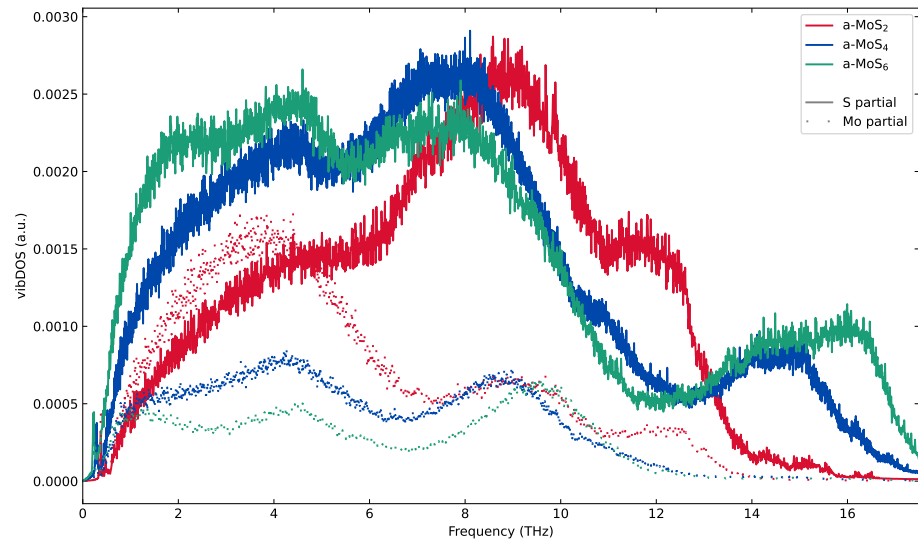


Figure 3.41: Partial vibrational density of states (vibDOS) across the three stoichiometries.

Ultimately, increasing the sulfur stoichiometry from $x = 2$ to $x = 6$ does not simply introduce localized defects into an existing lattice; it drives a complete

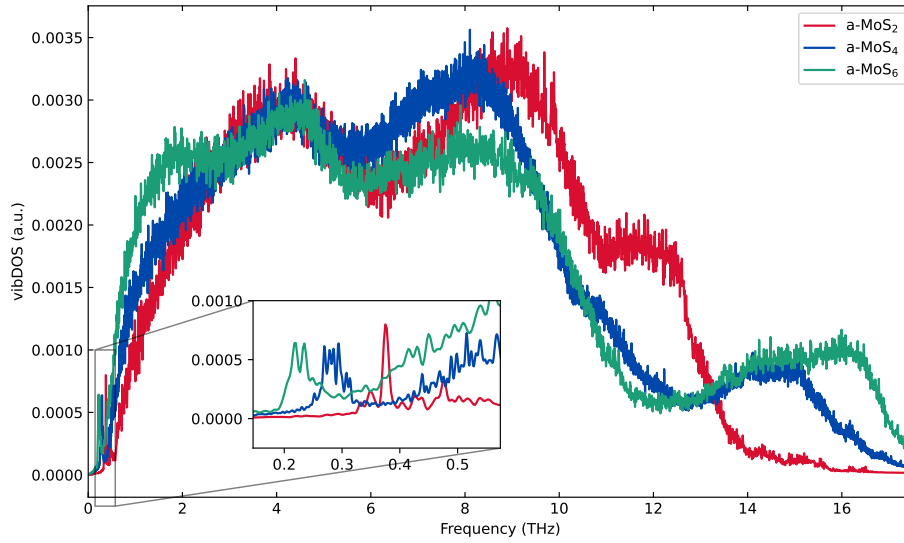


Figure 3.40: Total vibrational density of states (vibDOS) across the three stoichiometries.

topological restructuring. The material evolves from a continuously cross-linked, three-dimensional covalent glass into a severely depolymerized mesostructure, where isolated, highly coordinated molybdenum centers are suspended within a flexible matrix of polysulfide chains.

3.5 Conclusions and perspectives

The application of Machine-Learning Interatomic Potentials, augmented by a targeted active-learning framework, has proven highly effective in overcoming the traditional computational bottlenecks associated with modeling amorphous transition metal dichalcogenides. By explicitly training the potential to navigate the complex, out-of-domain potential energy surface of high-temperature liquid melts, we successfully extended both the spatial and temporal scales beyond the practical limits of conventional *ab initio* molecular dynamics. This methodological advancement enabled a systematic structural and vibrational investigation of amorphous molybdenum sulfides ($a\text{-MoS}_x$) across varying stoichiometries and kinetic histories.

Our structural analysis reveals a fundamental dichotomy in the thermodynamic and kinetic behavior of these disordered networks. Across all stoichiometries ($x = 2, 4, 6$), the primary short-range order—defined by fundamental bond lengths and acute η^2 -capping geometries—is mostly thermodynamically driven, exhibiting invariance to the kinetic parameters of melting temperature and quenching rate. However, as the structural complexity increases with sulfur content, subtle kinetic sensitivities emerge even at the local level; notably, the specific polyhedral speciation in $a\text{-MoS}_6$

exhibits structural memory of the liquid melt, where extreme temperatures and slower quench rates drive the network toward highly diversified, frustrated polyhedral assemblies rather than a single dominant coordination. Furthermore, the medium-range mesostructure, characterized by the assembly of extended alternate rings and the emergence of low-frequency acoustic vibrational modes, demonstrates distinct kinetic sensitivity across all phases. Slower quenching protocols facilitate the formation of larger, 10- and 12-membered mesoscopic loops and extended void regions, which are artificially suppressed by the faster quench rates typical of AIMD.

Furthermore, tracking the stoichiometric evolution confirmed a severe topological transition: as the sulfur content increases, the rigid, three-dimensionally cross-linked metal-chalcogenide glass ($x = 2$) depolymerizes into a highly frustrated network dominated by high-coordination molybdenum centers bridged by an extensive matrix of flexible polysulfide chains ($x = 6$).

Building upon these foundational bulk models, several critical avenues for future research emerge. First, a rigorous structural analysis of the pore network and interstices would provide crucial insights into active site accessibility. These structural metrics could ultimately be integrated with DFT calculations of active site reactivity to H/H₂ into macroscopic kinetic models to predict overall catalytic performance.

Furthermore, to definitively bridge the gap between these atomistic predictions and experimental characterization, future computational efforts should focus on simulating spectroscopic observables. Calculating core-level binding energies and simulating X-ray Absorption Near Edge Structure (XANES) or Extended X-ray Absorption Fine Structure (EXAFS) spectra directly from these MLMD trajectories would allow for a direct, quantitative assignment of local oxidation states and coordination environments against available experimental data. Additionally, computing the macroscopic polarizability to simulate Raman spectra would provide a crucial bridge to experimental vibrational characterization. This would explicitly validate our assignment of the high-frequency optical bands to polysulfide S-S stretching modes, directly correlating our computational vibDOS fingerprints with the characteristic Raman shifts routinely observed in laboratory samples.

While this study demonstrates the topological advantages of utilizing slower, more realistic cooling rates, a systematic mapping of the kinetic trapping threshold remains to be explored. By using our trained MLIP to replicate the

extremely fast quenching protocols ($> 10^{14}$ K/s) employed in earlier AIMD studies, future work could isolate the precise cooling regime where large mesoscopic rings collapse, providing a complete kinetic phase diagram of the amorphization process. Additionally, beyond its traditional application in structural exploration, the Minima Hopping algorithm [113, 114] could be utilized as a diagnostic tool to investigate this kinetic trapping. By leveraging its dynamic escape temperature to probe local barrier heights, future studies could systematically map the potential energy surface to determine the energetic barriers that trap the large-scale mesoscopic packing and ring structures observed in the amorphous network, as well as serve as an additional phase space exploration tool.

Finally, integrating these structural insights with the material's primary application—electrocatalysis—represents the most pressing perspective. While this chapter focuses on bulk topologies, the high density of undercoordinated terminal sulfur (S:1) and highly strained angular (A:2) geometries identified in our models are precisely the structural motifs hypothesized to drive the exceptional Hydrogen Evolution Reaction (HER) activity of $a\text{-MoS}_x$. Expanding this methodology to generate explicit amorphous surfaces and nanoparticles, and subsequently evaluating the hydrogen adsorption free energies (ΔG_{H}) across these diverse active sites, will ultimately enable the formulation of predictive, atomistic structure–activity relationships for amorphous chalcogenide catalysts.

4.1 Description of the MlipKit software package

Many different types of Machine-Learned Interatomic Potentials have been tested and compared in this thesis. MLIPs are called in many different steps of the workflows, such as MD, MLACS, TDEP, on top of the MLIP training itself, inference, and cross-validation. Using the potentials through their own specific CLI or Python APIs makes building the workflow a challenge of its own. For example, Moment Tensor Potentials (MTP) utilize compiled C++ binaries (mlip-2) and custom .cfg text formats, whereas MatterSim provides a Python API with its own methods. Without a unified abstraction layer, data must be manually passed between external MLIP engines and analysis codes (like phonopy), resulting in massive I/O overhead and formatting mismatches. To overcome this limitation, we have developed a Python interface, *MlipKit*, which provides a unified and extensible API for the training and deployment of MLIPs.

The structural core of MlipKit is the `MlipModel` abstract base class. To integrate a new MLIP backend, a developer only needs to inherit from this class and implement three mandatory methods:

- ▶ `_train()`,
- ▶ `_compute_properties_single()`,
- ▶ `_compute_properties_list()`.

The explicit separation of single-structure and list-based property evaluations is a deliberate design choice to maximize computational efficiency. Evaluating a single structure is optimized for on-the-fly execution, such as step-by-step molecular dynamics where the state continuously updates. Conversely, the list-based evaluation is designed for batch processing. When evaluating thousands of displaced structures, invoking a Python wrapper or spawning a subprocess for every individual structure would result in I/O overhead that eclipses the actual MLIP inference time. By processing lists natively, MlipKit batches the I/O and executes inference over the entire dataset in a single call.

To ensure workflows remain reproducible across different computing clusters, `MlipModel` implements robust serialization. A model's state, including

its structural metadata, hyperparameters, and directory pointers, can be saved to and loaded from a JSON file using `_to_dict()` and `_from_json()`. This allows a potential trained on a high-performance GPU node to be seamlessly reloaded on a local workstation for post-processing without rebuilding the Python object from scratch.

MlipKit natively bridges custom models to the broader computational chemistry ecosystem via the Atomic Simulation Environment (ASE) by automatising the generation of standard ASE calculators. When a user calls `make_ase_calculator()`, the software dynamically wraps the specific MLIP backend inside an `MlipkitCalculator` class (inheriting from `ase.calculators.calculator.Calculator`). This lightweight proxy delegates the physics calculations to the underlying MLIP model, while saving and storing the results in the ASE format. This architectural choice instantly unlocks native ASE molecular dynamics, structural geometry optimizations, and vibration analyses. Any external code capable of interacting with an ASE Atoms object can immediately utilize MTP or MatterSim through MlipKit, requiring zero custom wrapper scripts.

In complex computational workflows, a common source of silent errors occurs when a model's hyperparameters or training datasets are modified, but the potential files are not retrained to reflect those changes. To guarantee data integrity, MlipKit employs a rigorous, automated hashing system. Whenever a model is trained, MlipKit computes a SHA-256 cryptographic hash of both the training dataset and the resulting potential weights. The dataset hashing serializes the atomic numbers, cell parameters, and positions of every configuration. Similarly, the raw bytes of the generated potential files are parsed and hashed. These hashes are stored both as attributes of the `MlipModel` instance and persistently in a metadata file. The same is done for the training hyperparameters. If a user subsequently alters a hyperparameter dictionary or updates the training dataset, MlipKit immediately detects the hash mismatch and the model is automatically flagged as "untrained", preventing the user from accidentally calculating physical properties using an outdated MLIP.

4.2 Usage example

Here follows a simple example of instantiation, training and deployment of an MTP model with ASE Atoms objects:

```

1 from ase.io import read
2 from mlipkit.MTP.mtp_model import MTPModel
3 from mlipkit.mlip_models import MlipModel
4
5 # 1. Load the dataset (ASE format)
6 dataset = read('Training_set.traj', index=':')
7
8 # 2. Define MTP training hyperparameters
9 hyperparameters = dict(
10     mlip_bin='path/to/mlp',
11     untrained_pot_file_dir='path/to/untrained_mtps',
12     mtp_level=12,
13     min_dist='find', # Automatically calculates MIC min distance
14     max_dist=8,
15     radial_basis_type='RBChebyshev',
16     radial_basis_size=8,
17     bin_pref='mpirun')
18
19 # 3. Initialize the model
20 mlip = MTPModel(
21     root_dir='model_dir',
22     name='mtp_model',
23     train_set=dataset,
24     hyperparameters=hyperparameters)
25
26 # 4. Train the model (automatically saves to JSON and hashes data)
27 mlip.train_model()
28
29 # 5. Reloading the model for downstream workflows
30 reloaded_model = MlipModel.load_model('model_dir/mtp_model.json')
31
32 # 6. Generate the ASE Calculator bridging the MLIP to the ecosystem
33 compute_params = {'lammps_bin': 'path/to/lmp-mpi', 'bin_pref': '
    mpirun'}
34 calculator = reloaded_model.make_ase_calculator(parameters=
    compute_params)
35
36 # The calculator can now be attached to any ASE Atoms object
37 # for molecular dynamics or phonon evaluations.
38 atoms = dataset[0]
39 atoms.calc = calculator
40 energy = atoms.get_potential_energy()

```


Building a machine-learning interatomic potential (MLIP) that is accurate over the whole region of configuration space relevant to a finite-temperature simulation is, to a large extent, a problem of *data generation*: the potential is only as good as the set of *ab initio* configurations it has been trained on. PioMLACS (Python Interoperable MLACS) is a small Python package written during this thesis to automate that data generation. It implements an iterative active-learning loop in the spirit of the Machine Learning Assisted Canonical Sampling (MLACS) method [castellano2022, castellano2025] (see section 1.5.8), but is designed so that *any* MLIP can be plugged into it without touching the code of the loop itself, thanks to its native support for MlipKit models (see chapter 4.1).

5.1 Decoupling the loop from the potential

The MLIPs that `mlacs` can train are a fixed set of models (SNAP, MTP and ACE [castellano2025]), each implemented as a dedicated class in the package’s own source code. Using a different potential is therefore not a matter of configuration: it requires writing a new interface inside `mlacs` itself.

This became a practical limitation in the present work. The family of potentials of interest has grown rapidly to include graph-neural network models and pre-trained “foundation” potentials such as MatterSim, which are most efficiently used by *fine-tuning* on a small system-specific dataset rather than by training from scratch. None of these could be used inside `mlacs` without modifying it. Piomlacs addresses this by separating two concerns that are intertwined in `mlacs`:

1. the active-learning loop (train $\rightarrow$ MD $\rightarrow$ select $\rightarrow$ DFT $\rightarrow$ extend the dataset), which is generic; and
2. the MLIP, which is specific and is treated as a black box.

The loop only requires the MLIP to be an object of a class (`MlipModel`) from MlipKit (see chapter 4). Supporting a new potential therefore means writing a new MlipKit model, and PioMLACS does not change at all.

The scope of PioMLACS is deliberately narrower than that of `mlacs`: it implements the data-generation loop only. Its products are a DFT-labeled training set and a potential trained on it; the reweighting and thermodynamic-integration tools of `mlacs` are not reimplemented.

5.2 Design

The core of the package is the function `run_mlacs`. Given a starting structure and/or a pre-existing dataset $\mathcal{D}$, it performs N_{iter} iterations of the following cycle:

1. **Training:** The MLIP is (re)trained on the current dataset $\mathcal{D}$.
2. **MD:** An MD simulation driven by the freshly trained MLIP is run in the NVT or NPT ensemble through ASE. At every iteration the temperature is drawn uniformly from $[T_{\text{start}}, T_{\text{stop}}]$ (and, for NPT, the pressure from $[p_{\text{start}}, p_{\text{stop}}]$), so that successive iterations probe different thermodynamic conditions.
3. **Selection:** A user-defined set of frames is extracted from the saved MD trajectory, according to an *extraction policy* (see below).
4. **Labeling:** Energy, forces and stress of each selected frame are computed with DFT (implemented as an ASE calculator).
5. **Update:** The labeled frames are appended to $\mathcal{D}$ and written to disk.

If no dataset is provided, the loop is bootstrapped by computing with DFT a number of randomly rattled copies of the starting structure.

The extraction policy is a list with one entry per iteration; each entry is a list of frame indices in the MD trajectory, with Python's negative-indexing convention. For example, `[-1]` selects only the last frame, while `[-100, -50, -1]` selects three frames spaced along the end of the run. This makes it possible to add few configurations in early iterations, when the potential is still rough and the MD is short-lived, and more configurations later, once the dynamics is stable.

Two objects are passed to `run_mlacs`: (i) the `MLipModel` object (from `MlipKit`), and (ii) the DFT calculator, namely any ASE calculator.

All files are written under a single root directory. The dataset is appended to `Dataset.traj` after every DFT calculation, and the latest MD trajectory is stored in `MD/mlmd.traj`. A timestamped summary of the run is written to `log_piomlacs`, while the output of the three expensive steps is redirected

to `log_mlip`, `log_md` and `log_dft`. If `Dataset.traj` is found in the root directory when `run_mlacs` is called, its content is loaded and the run continues from it, so an interrupted job can simply be resubmitted.

5.3 Usage example

In the following, we show a complete PioMLACS run in which a pre-trained MatterSim potential is fine-tuned on a system-specific dataset generated on the fly. Ten iterations of NVT MD are performed at temperatures between 100 K and 700 K; one configuration per iteration is sent to DFT in the first four iterations, and three in the remaining ones.

```

1 from ase.io import read
2 from mlipkit.mattersim.mattersim_model import MatterSimModel
3 from piomlacs.piomlacs import run_mlacs
4
5 # --- MLIP: a MatterSim foundation model, to be fine-tuned
6 # -----
7 mlip = MatterSimModel(
8     root_dir='mlip',
9     name='mattersim_finetuned',
10    hyperparameters=dict(
11        load_model_path='mattersim-v1.0.0-5M.pth', # pre-trained
12        weights
13        device='cuda',
14        epochs=1000,
15        batch_size=16,
16        learning_rate=2e-4,
17        include_forces=True,
18        include_stresses=True,
19        early_stop_patience=10),
20    parameters_compute_properties=dict(device='cuda'))
21
22 # --- DFT: any ASE calculator
23 # -----
24 dft_calculator = ... # e.g. an Abinit calculator
25
26 # --- Active-learning loop
27 # -----
28 n_iterations = 10
29 run_mlacs(
30     root_dir='run_01',

```

```

27     mlip=mlip,
28     dft_calculator=dft_calculator,
29     starting_conf=read('init_conf.traj'),
30     n_starting_conf_rattlings=10,
31     ensemble='NVT',
32     md_parameters=dict(timestep_fs=1, langevin_friction=0.02),
33     n_iterations=n_iterations,
34     n_steps_per_iteration=4000,
35     saving_interval=50,
36     n_throw=20,
37     T_start=100, T_stop=700,
38     extraction_policy=[[-1]] * 4 + [[-100, -50, -1]] * 6)

```

As anticipated, because PioMLACS only interacts with the MLIP through the MlipKit interface, switching to a different potential amounts to replacing the object `mlip`. For instance, the same run can be performed with a Moment Tensor Potential trained from scratch by replacing the import on line 2 and the definition on lines 6–18 with

```

from mlipkit.MTP.mtp_model import MTPModel

mlip = MTPModel(
    root_dir='mlip',
    name='mtp',
    hyperparameters=dict(mlip_bin='path/to/mlp',
                          untrained_pot_file_dir='path/to/
untrained_mtps',
                          mtp_level=12, min_dist=2, max_dist=8,
                          radial_basis_type='RBChebyshev',
                          radial_basis_size=8),
    parameters_compute_properties=dict(lammps_bin='path/to/lmp'))

```

while the call to `run_mlacs` is left unchanged. At the end of the run, `run_01/Dataset.traj` contains all the DFT-labelled configurations, and the MLIP object holds the potential trained at the last iteration, ready to be used for production simulations.

A central grand challenge in modern computational materials science is bridging the gap between the predictive power of first-principles quantum mechanics and the macroscopic reality of functional materials. Real-world applications are governed by finite-temperature fluctuations, intrinsic disorder, and extended mesoscopic topologies— regimes where traditional *ab initio* methods encounter impassable computational limits. Throughout this thesis, we have demonstrated that Machine-Learning Interatomic Potentials (MLIPs) provide a rigorous and highly scalable solution to this fundamental bottleneck. By decoupling the evaluation of the potential energy surface from the explicit, on-the-fly calculation of the electronic ground state, we extended the reach of atomistic simulations to systematically investigate the anharmonic dynamics of crystalline 2H-MoS₂ and the vast configurational landscape of amorphous molybdenum sulfides (*a*-MoS_x). Through the creation of *MlipKit* and *Piomlacs*, we bridged the gap between raw MLIP backends and the broader materials simulation ecosystem. By implementing a unified abstraction layer and an automated active-learning generation loop, we enabled the seamless, self-consistent generation of highly representative training datasets. This infrastructure proved crucial for pushing MLIPs beyond their standard crystalline training domains, successfully stabilizing high-temperature liquid melts and complex glassy topologies.

Equipped with these reliable potentials, we mapped the finite-temperature lattice dynamics of crystalline 2H-MoS₂. By driving the Temperature Dependent Effective Potential (TDEP) method with MLIPs, we bypassed the static harmonic approximation to explicitly compute the anharmonic phonon self-energies. A direct comparison between classical molecular dynamics and stochastic sampling (sTDEP) highlighted the crucial role of quantum zero-point energy in capturing the true extent of the phase space at low temperatures. While our models successfully reproduced the physical thermal red-shifts and lineshape broadenings, the residual discrepancies with experimental Raman spectra exposed the physical limitations of the static Placzek approximation and the truncation of the self-energy at the third order.

In the second part, we addressed the topological challenge presented by amorphous MoS_x. Here, the MLIPs were forced to navigate the extreme configurational diversity of high-temperature liquid melts and their subsequent

kinetically driven vitrification. By accessing cooling rates inaccessible to standard *ab initio* molecular dynamics, we mapped a profound stoichiometric transition. We showed that moving from $x = 2$ to $x = 6$ does not merely substitute atomic sites, but triggers a depolymerization of the cross-linked metal-chalcogenide glass into a matrix of Mo polyhedra suspended in polysulfide chains. Crucially, we found that while the local η^2 -capping chemistry is thermodynamically locked, the mesoscopic architecture—specifically the formation of extended, porous ring structures—is deeply dependent on the kinetic history of the quench.

By establishing this reliable atomistic baseline, this work lays the foundation for several ambitious research directions. First, bridging the remaining gap between simulated and experimental spectroscopic observables requires moving beyond static perturbative approximations. For crystalline dynamics, the inherent limitations of truncating the phonon self-energy and relying on static polarizability tensors suggest a transition toward fully time-dependent methodologies and the explicit inclusion of electron-phonon terms. Training MLIPs to simultaneously predict tensorial properties—such as macroscopic polarizability or dipole moments—would allow entire Raman or infrared spectra to be extracted directly from time-autocorrelation functions, naturally encompassing all orders of anharmonicity and unlocking the inclusion of resonant electron-phonon effects.

Second, the kinetic sensitivity observed in the mesoscopic ring structures of amorphous MoS_x demands a systematic exploration of the kinetic phase diagram. Future studies could leverage Minima Hopping algorithms not merely to find global minima, but as a dynamic diagnostic tool to quantify the specific energy barriers that trap these large-scale topological voids. Benchmarking these barriers against trajectories generated at ultra-fast quench rates will precisely define the kinetic thresholds separating highly defective physical glasses from artificially quenched computational models. Finally, the ultimate realization of these structural insights lies in predictive catalysis. The topological voids and highly strained, undercoordinated sulfur geometries (S:1 and A:2) natively generated in our bulk models are the precursors to catalytically active surface sites. Expanding our MLMD framework to simulate explicit solid-liquid interfaces and compute hydrogen adsorption free energies across these diverse structural motifs will directly link atomistic network topology to macroscopic hydrogen evolution reaction performance, paving the way for the rational, computationally-guided design of next-generation amorphous electrocatalysts.

Bibliography

Here are the references in citation order.

- [1] Weihao Zheng et al. 'Light Emission Properties of 2D Transition Metal Dichalcogenides: Fundamentals and Applications'. In: *Advanced Optical Materials* 6.21 (2018), p. 1800420. (Visited on 03/05/2026) (cited on page 1).
- [2] Thomas Mueller and Ermin Malic. 'Exciton physics and device application of two-dimensional transition metal dichalcogenide semiconductors'. In: *npj 2D Mater Appl* 2.1 (Sept. 10, 2018), p. 29. (Visited on 03/05/2026) (cited on page 1).
- [3] T. F. Jaramillo et al. 'Identification of Active Edge Sites for Electrochemical H₂ Evolution from MoS₂ Nanocatalysts'. In: *Science* 317.5834 (2007), pp. 100–102. DOI: [10.1126/science.1141483](https://doi.org/10.1126/science.1141483) (cited on pages 1, 95).
- [4] Berit Hinnemann et al. 'Biomimetic Hydrogen Evolution: MoS₂ Nanoparticles as Catalyst for Hydrogen Evolution'. In: *J. Am. Chem. Soc.* 127.15 (Apr. 1, 2005), pp. 5308–5309. (Visited on 03/05/2026) (cited on page 1).
- [5] Dalilah Mohamad Khaidar et al. 'Transition metal dichalcogenides-based catalysts for CO₂ conversion: An updated review'. In: *International Journal of Hydrogen Energy* 68 (May 28, 2024), pp. 35–50. (Visited on 03/05/2026) (cited on page 1).
- [6] Qiang Fu et al. '2D Transition Metal Dichalcogenides: Design, Modulation, and Challenges in Electrocatalysis'. In: *Advanced Materials* 33.6 (2021), p. 1907818. (Visited on 03/05/2026) (cited on page 1).
- [7] Mohammad R. Vazirisereshk et al. 'Solid Lubrication with MoS₂: A Review'. In: *Lubricants* 7.7 (July 2019), p. 57. (Visited on 03/12/2026) (cited on page 1).
- [8] A. Shankara et al. 'Study of solid lubrication with MoS₂ coating in the presence of additives using reciprocating ball-on-flat scratch tester'. In: *Sadhana* 33.3 (June 1, 2008), pp. 207–220. (Visited on 03/12/2026) (cited on page 1).

- [9] Dimitrios G Papageorgiou et al. 'Mechanical properties of transition metal dichalcogenides: towards high-performance polymer nanocomposites'. In: *2D Mater.* 12.1 (Dec. 2024), p. 012002. (Visited on 03/12/2026) (cited on page 1).
- [10] D. Yang, P. Westreich, and R. F. Frindt. 'Transition metal dichalcogenide/polymer nanocomposites'. In: *Nanostructured Materials* 12.1 (Jan. 1, 1999), pp. 467–470. (Visited on 03/12/2026) (cited on page 1).
- [11] Mojtaba Ahmadi et al. '2D transition metal dichalcogenide nanomaterials: advances, opportunities, and challenges in multi-functional polymer nanocomposites'. In: *J. Mater. Chem. A* 8.3 (Jan. 22, 2020), pp. 845–883. (Visited on 03/12/2026) (cited on page 1).
- [12] Qing Peng and Suvranu De. 'Outstanding mechanical properties of monolayer MoS₂ and its application in elastic energy storage'. In: *Phys. Chem. Chem. Phys.* 15.44 (Oct. 23, 2013), pp. 19427–19437. (Visited on 03/12/2026) (cited on page 1).
- [13] J. D. Benck et al. 'Amorphous Molybdenum Sulfide Catalysts for Electrochemical Hydrogen Production: Insights into the Origin of their Catalytic Activity'. In: *ACS Catal.* 2.9 (2012), pp. 1916–1923. doi: [10.1021/cs300451q](https://doi.org/10.1021/cs300451q) (cited on pages 1, 95, 96).
- [14] D. Merki et al. 'Amorphous Molybdenum Sulfide Films as Catalysts for Electrochemical Hydrogen Production in Water'. In: *Chem. Sci.* 2 (2011), pp. 1262–1267. doi: [10.1039/C1SC00117E](https://doi.org/10.1039/C1SC00117E) (cited on pages 1, 95).
- [15] J. K. Nørskov et al. 'Trends in the Exchange Current for Hydrogen Evolution'. In: *J. Electrochem. Soc.* 152.3 (2005), J23–J26. doi: [10.1149/1.1856988](https://doi.org/10.1149/1.1856988) (cited on pages 1, 95).
- [16] Max Born and Kun Huang. *Dynamical theory of crystal lattices*. Clarendon press, Oxford, 1954 (cited on page 15).
- [17] Gerald D. Mahan. *Many-Particle Physics*. 3rd. New York: Springer, 2000 (cited on page 21).
- [18] Tianli Feng and Xiulin Ruan. 'Quantum mechanical prediction of four-phonon scattering rates and reduced thermal conductivity of solids'. In: *Phys. Rev. B* 93 (4 Jan. 2016), p. 045202. doi: [10.1103/PhysRevB.93.045202](https://doi.org/10.1103/PhysRevB.93.045202) (cited on pages 22, 23).
- [19] Shin-ichiro Tamura. 'Isotope scattering of dispersive phonons in Ge'. In: *Physical Review B* 27 (1983), pp. 858–866 (cited on page 23).

- [20] Derek A. Long. *The Raman Effect: A Unified Treatment of the Theory of Raman Scattering by Molecules*. Chichester, UK: John Wiley & Sons, Ltd, 2002 (cited on page 31).
- [21] Richard M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. 2nd. Cambridge: Cambridge University Press, 2020 (cited on page 34).
- [22] Xavier Gonze. ‘Adiabatic density-functional perturbation theory’. In: *Phys. Rev. A* 52 (2 Aug. 1995), pp. 1096–1114. doi: [10.1103/PhysRevA.52.1096](https://doi.org/10.1103/PhysRevA.52.1096) (cited on page 36).
- [23] Henri Poincaré. ‘Sur le problème des trois corps et les équations de la dynamique’. In: *Acta mathematica* 13.1 (1890), pp. 1–270 (cited on page 37).
- [24] Axel Brünger, Charles L. Brooks, and Martin Karplus. ‘Stochastic boundary conditions for molecular dynamics simulations of ST2 water’. In: *Chemical Physics Letters* 105.5 (1984), pp. 495–500. doi: [https://doi.org/10.1016/0009-2614\(84\)80098-6](https://doi.org/10.1016/0009-2614(84)80098-6) (cited on page 39).
- [25] Benedict Leimkuhler and Charles Matthews. ‘Rational Construction of Stochastic Numerical Methods for Molecular Sampling’. In: *Applied Mathematics Research eXpress* 2013.1 (Jan. 2013), pp. 34–56. doi: [10.1093/amrx/abs010](https://doi.org/10.1093/amrx/abs010) (cited on page 39).
- [26] S. Nosé. ‘A unified formulation of the constant temperature molecular dynamics methods’. In: *The Journal of Chemical Physics* 81.1 (1984), pp. 511–519. doi: [10.1063/1.447334](https://doi.org/10.1063/1.447334) (cited on page 39).
- [27] W. G. Hoover. ‘Canonical dynamics: Equilibrium phase-space distributions’. In: *Physical Review A* 31.3 (1985), pp. 1695–1697. doi: [10.1103/PhysRevA.31.1695](https://doi.org/10.1103/PhysRevA.31.1695) (cited on page 39).
- [28] Hans C. Andersen. ‘Molecular dynamics simulations at constant pressure and/or temperature’. In: *The Journal of Chemical Physics* 72.4 (Feb. 1980), pp. 2384–2393. doi: [10.1063/1.439486](https://doi.org/10.1063/1.439486) (cited on page 40).
- [29] M. Parrinello and A. Rahman. ‘Crystal Structure and Pair Potentials: A Molecular-Dynamics Study’. In: *Phys. Rev. Lett.* 45 (14 Oct. 1980), pp. 1196–1199. doi: [10.1103/PhysRevLett.45.1196](https://doi.org/10.1103/PhysRevLett.45.1196) (cited on page 40).
- [30] Mark E. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, Aug. 2023 (cited on page 40).

- [31] Mahnaz Firouzi et al. 'Klinkenberg effect on predicting and measuring helium permeability in gas shales'. In: *International Journal of Coal Geology* 123 (2014). Special issue: Adsorption and fluid transport phenomena in gas shales and their effects on production and storage, pp. 62–68. doi: <https://doi.org/10.1016/j.coal.2013.09.006> (cited on page 41).
- [32] Ali Poursaeidesfahani et al. 'Efficient Application of Continuous Fractional Component Monte Carlo in the Reaction Ensemble'. In: *Journal of Chemical Theory and Computation* 13.9 (Sept. 2017), pp. 4452–4466. doi: [10.1021/acs.jctc.7b00092](https://doi.org/10.1021/acs.jctc.7b00092) (cited on page 41).
- [33] Bobak Shahriari et al. 'Taking the Human Out of the Loop: A Review of Bayesian Optimization'. In: *Proceedings of the IEEE* 104.1 (2016), pp. 148–175. doi: [10.1109/JPR0C.2015.2494218](https://doi.org/10.1109/JPR0C.2015.2494218) (cited on page 49).
- [34] Ralf Drautz. 'Atomic cluster expansion for accurate and transferable interatomic potentials'. In: *Phys. Rev. B* 99 (1 Jan. 2019), p. 014104. doi: [10.1103/PhysRevB.99.014104](https://doi.org/10.1103/PhysRevB.99.014104) (cited on page 51).
- [35] A. V. Shapeev. 'Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials'. In: *Multiscale Model. Simul.* 14.3 (2016), pp. 1153–1173. doi: [10.1137/15M1054183](https://doi.org/10.1137/15M1054183) (cited on pages 51, 70, 98, 99).
- [36] Ilyes Batatia et al. 'MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields'. In: (2023) (cited on pages 52, 99).
- [37] Théo Cavignac et al. 'AI-Driven expansion and application of the Alexandria database'. In: *Journal of Physics: Materials* 9.2 (2026), p. 025014. doi: [10.1088/2515-7639/ae6620](https://doi.org/10.1088/2515-7639/ae6620) (cited on page 54).
- [38] Bowen Deng et al. 'CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling'. In: *Nature Machine Intelligence* 5.9 (2023), pp. 1031–1041. doi: [10.1038/s42256-023-00716-3](https://doi.org/10.1038/s42256-023-00716-3) (cited on page 54).
- [39] Alois Castellano et al. 'Ab initio canonical sampling based on variational inference'. In: *Phys. Rev. B* 106 (16 Oct. 2022), p. L161110. doi: [10.1103/PhysRevB.106.L161110](https://doi.org/10.1103/PhysRevB.106.L161110) (cited on page 54).
- [40] Alois Castellano et al. 'Machine learning assisted canonical sampling (Mlacs)'. In: *Computer Physics Communications* 316 (2025), p. 109730. doi: <https://doi.org/10.1016/j.cpc.2025.109730> (cited on pages 54, 70).

- [41] Ion Errea, Matteo Calandra, and Francesco Mauri. ‘Anharmonic free energies and phonon dispersions from the stochastic self-consistent harmonic approximation: Application to platinum and palladium hydrides’. In: *Phys. Rev. B* 89 (6 Feb. 2014), p. 064302. doi: [10.1103/PhysRevB.89.064302](https://doi.org/10.1103/PhysRevB.89.064302) (cited on page 60).
- [42] Raffaello Bianco et al. ‘Second-order structural phase transitions, free energy curvature, and temperature-dependent anharmonic phonons in the self-consistent harmonic approximation: Theory and stochastic implementation’. In: *Phys. Rev. B* 96 (1 July 2017), p. 014111. doi: [10.1103/PhysRevB.96.014111](https://doi.org/10.1103/PhysRevB.96.014111) (cited on page 60).
- [43] N. R. Werthamer. ‘Self-Consistent Phonon Formulation of Anharmonic Lattice Dynamics’. In: *Phys. Rev. B* 1 (2 Jan. 1970), pp. 572–581. doi: [10.1103/PhysRevB.1.572](https://doi.org/10.1103/PhysRevB.1.572) (cited on page 60).
- [44] Nina Shulumba, Olle Hellman, and Austin J. Minnich. ‘Intrinsic localized mode and low thermal conductivity of PbSe’. In: *Phys. Rev. B* 95 (1 Jan. 2017), p. 014302. doi: [10.1103/PhysRevB.95.014302](https://doi.org/10.1103/PhysRevB.95.014302) (cited on page 61).
- [45] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford: Clarendon Press, 1987 (cited on page 61).
- [46] David Waroquiers et al. ‘ChemEnv: a fast and robust coordination environment identification tool’. In: *Acta Crystallographica Section B* 76.4 (Aug. 2020), pp. 683–695. doi: [10.1107/S2052520620007994](https://doi.org/10.1107/S2052520620007994) (cited on page 62).
- [47] S. V. King. ‘Ring Configurations in a Random Network Model of Vitreous Silica’. In: *Nature* 213.5081 (1967), pp. 1112–1113. doi: [10.1038/2131112a0](https://doi.org/10.1038/2131112a0) (cited on page 63).
- [48] D. S. Franzblau. ‘Computation of Ring Statistics for Network Models of Solids’. In: *Phys. Rev. B* 44.10 (1991), pp. 4925–4930. doi: [10.1103/PhysRevB.44.4925](https://doi.org/10.1103/PhysRevB.44.4925) (cited on page 63).
- [49] S. Le Roux and P. Jund. ‘Ring Statistics Analysis of Topological Networks: New Approach and Application to Amorphous GeS₂ and SiO₂ Systems’. In: *Comput. Mater. Sci.* 49.1 (2010), pp. 70–83. doi: [10.1016/j.commatsci.2010.04.023](https://doi.org/10.1016/j.commatsci.2010.04.023) (cited on page 63).
- [50] L. Guttman. ‘Ring structure of the crystalline and amorphous forms of silicon dioxide’. In: *J. Non-Cryst. Solids* 116.2-3 (1990), pp. 145–147. doi: [10.1016/0022-3093\(90\)90686-G](https://doi.org/10.1016/0022-3093(90)90686-G) (cited on page 63).

- [51] K. Goetzke and H. J. Klein. 'Properties and efficient algorithmic determination of the irreducible ring circuit distribution in glasses'. In: *J. Non-Cryst. Solids* 127.2 (1991), pp. 215–220. doi: [10.1016/0022-3093\(91\)90145-V](https://doi.org/10.1016/0022-3093(91)90145-V) (cited on page 63).
- [52] K. K. Kam and B. A. Parkinson. 'Detailed photocurrent spectroscopy of the semiconducting group VIB transition metal dichalcogenides'. In: *The Journal of Physical Chemistry* 86.4 (Feb. 1982), pp. 463–467. doi: [10.1021/j100393a010](https://doi.org/10.1021/j100393a010) (cited on pages 66, 90).
- [53] Andrea Splendiani et al. 'Emerging Photoluminescence in Monolayer MoS₂'. In: *Nano Letters* 10.4 (Apr. 2010), pp. 1271–1275. doi: [10.1021/nl903868w](https://doi.org/10.1021/nl903868w) (cited on page 66).
- [54] Kin Fai Mak et al. 'Atomically Thin MoS₂: A New Direct-Gap Semiconductor'. In: *Phys. Rev. Lett.* 105 (13 Sept. 2010), p. 136805. doi: [10.1103/PhysRevLett.105.136805](https://doi.org/10.1103/PhysRevLett.105.136805) (cited on page 66).
- [55] A. Molina-Sánchez and L. Wirtz. 'Phonons in single-layer and few-layer MoS₂ and WS₂'. In: *Phys. Rev. B* 84 (15 Oct. 2011), p. 155413. doi: [10.1103/PhysRevB.84.155413](https://doi.org/10.1103/PhysRevB.84.155413) (cited on pages 68, 69).
- [56] Wahyu Setyawan and Stefano Curtarolo. 'High-throughput electronic band structure calculations: Challenges and tools'. In: *Computational Materials Science* 49.2 (2010), pp. 299–312. doi: <https://doi.org/10.1016/j.commatsci.2010.05.010> (cited on page 68).
- [57] Mildred S. Dresselhaus, Gene Dresselhaus, and Ado Jorio. *Group Theory: Application to the Physics of Condensed Matter*. Springer, 2008 (cited on page 69).
- [58] Tsachi Livneh and Jonathan E. Spanier. 'A Comprehensive Multiphonon Spectral Analysis in MoS₂'. In: *2D Materials* 2 (2015) 035003 (2014). doi: [10.1088/2053-1583/2/3/035003](https://doi.org/10.1088/2053-1583/2/3/035003) (cited on page 69).
- [59] Matthieu J. Verstraete et al. 'Abinit 2025: New capabilities for the predictive modeling of solids and nanomaterials'. In: *The Journal of Chemical Physics* 163.16 (Oct. 2025), p. 164126. doi: [10.1063/5.0288278](https://doi.org/10.1063/5.0288278) (cited on page 70).
- [60] Aldo H. Romero et al. 'ABINIT: Overview and focus on selected capabilities'. In: *The Journal of Chemical Physics* 152.12 (Mar. 2020), p. 124102. doi: [10.1063/1.5144261](https://doi.org/10.1063/1.5144261) (cited on page 70).

- [61] Xavier Gonze et al. ‘The Abinit project: Impact, environment and recent developments’. In: *Computer Physics Communications* 248 (2020), p. 107042. doi: <https://doi.org/10.1016/j.cpc.2019.107042> (cited on page 70).
- [62] Benoit Van Troeye, Marc Torrent, and Xavier Gonze. ‘Interatomic force constants including the DFT-D dispersion contribution’. In: *Phys. Rev. B* 93 (14 Apr. 2016), p. 144304. doi: [10.1103/PhysRevB.93.144304](https://doi.org/10.1103/PhysRevB.93.144304) (cited on page 70).
- [63] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. ‘Generalized Gradient Approximation Made Simple’. In: *Phys. Rev. Lett.* 77 (18 Oct. 1996), pp. 3865–3868. doi: [10.1103/PhysRevLett.77.3865](https://doi.org/10.1103/PhysRevLett.77.3865) (cited on page 70).
- [64] Stefan Grimme et al. ‘A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu’. In: *The Journal of Chemical Physics* 132.15 (Apr. 2010), p. 154104. doi: [10.1063/1.3382344](https://doi.org/10.1063/1.3382344) (cited on page 70).
- [65] D. R. Hamann. ‘Optimized norm-conserving Vanderbilt pseudopotentials’. In: *Phys. Rev. B* 88 (8 Aug. 2013), p. 085117 (cited on page 70).
- [66] M.J. van Setten et al. ‘The PseudoDojo: Training and grading a 85 element optimized norm-conserving pseudopotential table’. In: *Computer Physics Communications* 226 (2018), pp. 39–54. doi: <https://doi.org/10.1016/j.cpc.2018.01.012> (cited on page 70).
- [67] Ivan S. Novikov et al. *The MLIP package: Moment Tensor Potentials with MPI and Active Learning*. 2020. URL: <https://arxiv.org/abs/2007.08555> (cited on pages 70, 99).
- [68] Yunxing Zuo et al. ‘Performance and Cost Assessment of Machine Learning Interatomic Potentials’. In: *The Journal of Physical Chemistry A* 124.4 (Jan. 2020), pp. 731–745. doi: [10.1021/acs.jpca.9b08723](https://doi.org/10.1021/acs.jpca.9b08723) (cited on page 71).
- [69] Aidan P. Thompson et al. ‘LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales’. In: *Computer Physics Communications* 271 (2022), p. 108171. doi: <https://doi.org/10.1016/j.cpc.2021.108171> (cited on page 72).

- [70] S. H. El-Mahalawy and B. L. Evans. ‘The thermal expansion of 2H-MoS₂, 2H-MoSe₂ and 2H-WSe₂ between 20 and 800°C’. In: *Journal of Applied Crystallography* 9.5 (1976), pp. 403–406. doi: <https://doi.org/10.1107/S0021889876011709> (cited on pages 73, 74).
- [71] R. Murray and B. Evans. ‘The thermal expansion of 2H-MoS₂ and 2H-WSe₂ between 10 and 320 K’. In: *Journal of Applied Crystallography* 12.3 (1979), pp. 312–315. doi: <https://doi.org/10.1107/S0021889879012528> (cited on pages 73, 74).
- [72] Gábor I. Csonka et al. ‘Assessing the performance of recent density functionals for bulk solids’. In: *Phys. Rev. B* 79 (15 Apr. 2009), p. 155107. doi: [10.1103/PhysRevB.79.155107](https://doi.org/10.1103/PhysRevB.79.155107) (cited on page 73).
- [73] Simuck F. Yuk et al. ‘Putting error bars on density functional theory’. In: *Scientific Reports* 14.1 (Aug. 2024), p. 20219. doi: [10.1038/s41598-024-69194-w](https://doi.org/10.1038/s41598-024-69194-w) (cited on page 73).
- [74] Changgu Lee et al. ‘Anomalous Lattice Vibrations of Single- and Few-Layer MoS₂’. In: *ACS Nano* 4.5 (May 2010), pp. 2695–2700 (cited on page 76).
- [75] Niels Grønbech-Jensen and Oded Farago. ‘A simple and effective Verlet-type algorithm for simulating Langevin dynamics’. In: *Molecular Physics* 111.8 (2013), pp. 983–991. doi: [10.1080/00268976.2012.760055](https://doi.org/10.1080/00268976.2012.760055) (cited on page 76).
- [76] Hans Tornatzky et al. ‘Phonon dispersion in MoS₂’. In: *Phys. Rev. B* 99 (14 Apr. 2019), p. 144309. doi: [10.1103/PhysRevB.99.144309](https://doi.org/10.1103/PhysRevB.99.144309) (cited on pages 85, 86).
- [77] Zi-Yu Cao and Xiao-Jia Chen. ‘Phonon scattering processes in molybdenum disulfide’. In: *Applied Physics Letters* 114.5 (Feb. 2019), p. 052102. doi: [10.1063/1.5082932](https://doi.org/10.1063/1.5082932) (cited on pages 86, 87).
- [78] Nina Giroto, Fabio Caruso, and Dino Novko. ‘Ultrafast Nonadiabatic Phonon Renormalization in Photoexcited Single-Layer MoS₂’. In: *The Journal of Physical Chemistry C* 127.33 (Aug. 2023), pp. 16515–16524. doi: [10.1021/acs.jpcc.3c03664](https://doi.org/10.1021/acs.jpcc.3c03664) (cited on page 89).
- [79] Nina Giroto Erhardt et al. ‘Electron-mediated anharmonicity and its role in the Raman spectrum of graphene’. In: *npj Computational Materials* 11.1 (Apr. 2025), p. 114. doi: [10.1038/s41524-025-01610-9](https://doi.org/10.1038/s41524-025-01610-9) (cited on page 89).

- [80] Zherui Han et al. 'Raman Linewidth Contributions from Four-Phonon and Electron-Phonon Interactions in Graphene'. In: *Phys. Rev. Lett.* 128 (4 Jan. 2022), p. 045901. doi: [10.1103/PhysRevLett.128.045901](https://doi.org/10.1103/PhysRevLett.128.045901) (cited on page 89).
- [81] Satyaprakash Sahoo et al. 'Temperature-Dependent Raman Studies and Thermal Conductivity of Few-Layer MoS₂'. In: *The Journal of Physical Chemistry C* 117.17 (May 2013), pp. 9042–9047. doi: [10.1021/jp402509w](https://doi.org/10.1021/jp402509w) (cited on pages 90, 91).
- [82] Stefano Falletta et al. 'Unified differentiable learning of electric response'. In: *Nature Communications* 16.1 (2025), p. 4031. doi: [10.1038/s41467-025-59304-1](https://doi.org/10.1038/s41467-025-59304-1) (cited on page 93).
- [83] Bradley A. A. Martin et al. 'General Learning of the Electric Response of Inorganic Materials'. In: *PRX Intelligence* 1 (1 July 2026), p. 013006. doi: [10.1103/b116-xy8k](https://doi.org/10.1103/b116-xy8k) (cited on page 93).
- [84] M. Kibsgaard et al. 'Structural Transformation Identification of Sputtered Amorphous MoS_x as an Efficient Hydrogen-Evolving Catalyst during Electrochemical Activation'. In: *ACS Catal.* 9 (2019), pp. 2720–2728. doi: [10.1021/acscatal.8b04884](https://doi.org/10.1021/acscatal.8b04884) (cited on page 95).
- [85] C. G. Morales-Guio and X. Hu. 'Amorphous Molybdenum Sulfides as Hydrogen Evolution Catalysts'. In: *Acc. Chem. Res.* 47 (2014), pp. 2603–2611. doi: [10.1021/ar500213c](https://doi.org/10.1021/ar500213c) (cited on pages 95, 96).
- [86] P. D. Tran et al. 'Coordination Polymer Structure and Revisited Hydrogen Evolution Catalytic Mechanism for Amorphous Molybdenum Sulfide'. In: *Nat. Mater.* 15 (2016), pp. 728–734. doi: [10.1038/nmat4623](https://doi.org/10.1038/nmat4623) (cited on pages 95, 96, 112, 132).
- [87] C. Chang et al. 'Amorphous Molybdenum Sulfide and Its Mo–S Motifs: Structural Characteristics, Synthetic Strategies, and Comprehensive Applications'. In: *Nano Res.* 15 (2022), pp. 8639–8658. doi: [10.1007/s12274-022-4507-z](https://doi.org/10.1007/s12274-022-4507-z) (cited on pages 95, 96).
- [88] M. L. Grutza et al. 'Hydrogen Evolution Catalysis by Molybdenum Sulfides: Insights and Challenges'. In: *Sustainable Energy Fuels* 2 (2018), pp. 527–540. doi: [10.1039/C8SE00155C](https://doi.org/10.1039/C8SE00155C) (cited on pages 95, 96).
- [89] A. Baloglou et al. '[Mo₃S₁₃]²⁻ as a Model System for Hydrogen Evolution Catalysis by MoS_x: Probing Protonation Sites in the Gas Phase by IRMPD Spectroscopy'. In: *ChemPhysChem* 19 (2018), pp. 2993–3001. doi: [10.1002/cphc.201800471](https://doi.org/10.1002/cphc.201800471) (cited on pages 95, 96).

- [90] Jinkai Si and Ying Ma. 'Ab Initio Molecular Dynamics Simulations of Amorphous Metal Sulfides as Cathode Materials for Lithium–Sulfur Batteries'. In: *The Journal of Physical Chemistry C* 127.49 (Nov. 2023), pp. 23541–23550. doi: [10.1021/acs.jpcc.3c04835](https://doi.org/10.1021/acs.jpcc.3c04835) (cited on pages 96, 97, 99, 105, 109, 119).
- [91] R. L. McGreevy and L. Pusztai. 'Reverse Monte Carlo Simulation: A New Technique for the Determination of Disordered Structures'. In: *Mol. Simul.* 1 (1988), pp. 359–367. doi: [10.1080/08927028808080958](https://doi.org/10.1080/08927028808080958) (cited on page 97).
- [92] C. L. Farrow et al. 'PDFfit2 and PDFgui: Computer programs for studying nanostructure in crystals'. In: *J. Phys. Condens. Matter* 19.33 (2007), p. 335219. doi: [10.1088/0953-8984/19/33/335219](https://doi.org/10.1088/0953-8984/19/33/335219) (cited on page 97).
- [93] J. D. Bernal. 'Geometry of the Structure of Monatomic Liquids'. In: *Nature* 185 (1960), pp. 68–70. doi: [10.1038/185068a0](https://doi.org/10.1038/185068a0) (cited on page 97).
- [94] J. L. Finney. 'Random Packings and the Structure of Simple Liquids. I. The Geometry of Random Close Packing'. In: *Proc. R. Soc. London A* 319 (1970), pp. 479–493. doi: [10.1098/rspa.1970.0189](https://doi.org/10.1098/rspa.1970.0189) (cited on page 97).
- [95] G. S. III Cargill. 'Dense Random Packing of Hard Spheres as a Structural Model for Noncrystalline Metallic Solids'. In: *J. Appl. Phys.* 41.5 (1970), pp. 2248–2250. doi: [10.1063/1.1659231](https://doi.org/10.1063/1.1659231) (cited on page 97).
- [96] David A Drabold. 'Topics in the theory of amorphous materials'. In: *The European Physical Journal B* 68.1 (2009), pp. 1–21 (cited on page 97).
- [97] Akihisa Inoue. 'Stabilization of metallic supercooled liquid and bulk amorphous alloys'. In: *Acta Materialia* 48.1 (2000), pp. 279–306. doi: [https://doi.org/10.1016/S1359-6454\(99\)00300-6](https://doi.org/10.1016/S1359-6454(99)00300-6) (cited on page 97).
- [98] K. Vollmayr, W. Kob, and K. Binder. 'Cooling-Rate Effects in Amorphous Silica: A Computer-Simulation Study'. In: *Phys. Rev. B* 54 (1996), pp. 15808–15827. doi: [10.1103/PhysRevB.54.15808](https://doi.org/10.1103/PhysRevB.54.15808) (cited on page 97).
- [99] Volker L Deringer and Gábor Csányi. 'Machine learning based interatomic potential for amorphous carbon'. In: *Physical Review B* 95.9 (2017), p. 094203 (cited on pages 97, 98).

- [100] F. H. Stillinger and T. A. Weber. ‘Computer Simulation of Local Order in Condensed Phases of Silicon’. In: *Phys. Rev. B* 31 (1985), pp. 5262–5271. doi: [10.1103/PhysRevB.31.5262](https://doi.org/10.1103/PhysRevB.31.5262) (cited on page 98).
- [101] J. Tersoff. ‘New Empirical Approach for the Structure and Energy of Covalent Systems’. In: *Phys. Rev. B* 37 (1988), pp. 6991–7000. doi: [10.1103/PhysRevB.37.6991](https://doi.org/10.1103/PhysRevB.37.6991) (cited on page 98).
- [102] A. C. T. van Duin et al. ‘ReaxFF: A Reactive Force Field for Hydrocarbons’. In: *J. Phys. Chem. A* 105 (2001), pp. 9396–9409. doi: [10.1021/jp004368u](https://doi.org/10.1021/jp004368u) (cited on page 98).
- [103] Natascia L. Fragapane and Volker L. Deringer. *Amorphous materials as a frontier challenge for universal interatomic potentials*. 2026. URL: <https://arxiv.org/abs/2607.11384> (cited on page 98).
- [104] Shuwei Li et al. ‘Melt-Quench Failures and Practical Solutions for Universal Machine-Learning Interatomic Potentials in Amorphous Structure Generation’. In: (2026) (cited on page 98).
- [105] Alfonso Pedone, Marco Bertani, and Matilde Benassi. ‘Revisiting Machine Learning Potentials for Silicate Glasses: The Missing Role of Dispersion Interactions’. In: *Journal of Chemical Theory and Computation* 21.9 (Apr. 2025), pp. 4769–4778. doi: [10.1021/acs.jctc.5c00218](https://doi.org/10.1021/acs.jctc.5c00218) (cited on page 98).
- [106] Dongheon Lee et al. ‘Crystallization of amorphous GeTe simulated by neural network potential addressing medium-range order’. In: *Computational Materials Science* 181 (2020), p. 109725. doi: <https://doi.org/10.1016/j.commatsci.2020.109725> (cited on page 98).
- [107] Han Yang et al. ‘MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures’. In: *arXiv preprint arXiv:2405.04967* (2024) (cited on page 99).
- [108] Simon Batzner et al. ‘E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials’. In: *Nature Communications* 13.1 (2022), p. 2453. doi: [10.1038/s41467-022-29939-5](https://doi.org/10.1038/s41467-022-29939-5) (cited on page 99).
- [109] Takamichi Iida and Roderick I. L. Guthrie. *The Physical Properties of Liquid Metals*. Oxford: Clarendon Press, 1988 (cited on page 105).
- [110] Joe Morrow. *RingsStatisticsMatter.jl: Julia implementation of algorithm for counting primitive rings in an atomistic structure*. <https://github.com/MorrowChem/RingsStatisticsMatter.jl>. 2023 (cited on page 107).

- [111] Yuan-Chao Hu and Hajime Tanaka. 'Origin of the boson peak in amorphous solids'. In: *Nature Physics* 18.6 (2022), pp. 669–677. doi: [10.1038/s41567-022-01628-6](https://doi.org/10.1038/s41567-022-01628-6) (cited on page 108).
- [112] K.S. Liang et al. 'Amorphous MoS₃ and WS₃'. In: *Journal of Non-Crystalline Solids* 42.1 (1980), pp. 345–356. doi: [https://doi.org/10.1016/0022-3093\(80\)90035-6](https://doi.org/10.1016/0022-3093(80)90035-6) (cited on page 132).
- [113] Stefan Goedecker. 'Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems'. In: *The Journal of Chemical Physics* 120.21 (June 2004), pp. 9911–9917. doi: [10.1063/1.1724816](https://doi.org/10.1063/1.1724816) (cited on page 135).
- [114] Maximilian Amsler and Stefan Goedecker. 'Crystal structure prediction using the minima hopping method'. In: *The Journal of Chemical Physics* 133.22 (Dec. 2010), p. 224104. doi: [10.1063/1.3512900](https://doi.org/10.1063/1.3512900) (cited on page 135).

Acknowledgments

This PhD project was part of the Horizon Europe MSCA Doctoral network (grant n.101073486), EUSpecLab, funded by the European Union.

This research was made possible through the computational resources provided by High-Performance Computing (HPC) facilities from Consortium d'Equipements de Calcul Intensif (FRS-FNRS Belgium Grant No. 2.5020.11), EuroHPC-JU award EHPC-EXT-2023E02-050 on MareNostrum 5 at Barcelona Supercomputing Center (BSC), Spain, and by the Lucia Tier-1 of the Federation Wallonie-Bruxelles (Walloon Region (grant agreement No. 1117545)).