

Modélisation conjointe du silicium, de l'oxygène et de la silice

par

Karim ZONGO

THÈSE PRÉSENTÉE À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION
DU DOCTORAT EN GÉNIE
Ph.D.

MONTRÉAL, LE 28 AVRIL 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC



Karim Zongo, 2025



Cette licence Creative Commons signifie qu'il est permis de diffuser, d'imprimer ou de sauvegarder sur un autre support une partie ou la totalité de cette oeuvre à condition de mentionner l'auteur, que ces utilisations soient faites à des fins non commerciales et que le contenu de l'oeuvre n'ait pas été modifié.

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE:

Mme. Claudiane Ouellet-Plamondon, directrice de thèse
Département de génie de la construction, École de technologie supérieure, Montréal, QC

M. Laurent Karim Béland, codirecteur
Department of Mechanical and Materials Engineering, Queen's university, Kingston, ON

M. Matthew Toews, président du jury
Département de génie des systèmes, École de technologie supérieure, Montréal, QC

M. Conrad Botton, membre du jury
Département de génie de la construction, École de technologie supérieure, Montréal, QC

M. Gabriel Antonius, membre du jury
Département de biochimie, chimie, physique et science forensique, Université du Québec à
Trois-Rivières, QC

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 16 AVRIL 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je souhaite exprimer ma sincère gratitude à mes directeurs de thèse, Claudiane Ouellet-Plamondon et Laurent Karim Béland, pour leur soutien indéfectible et leurs échanges précieux, même à des heures tardives. Leur accompagnement a été essentiel à la réalisation de cette thèse, et je leur en suis profondément reconnaissant.

Je tiens également à remercier chaleureusement les membres du group de Laurent Karim Béland à Kingston, en Ontario, pour leur accueil bienveillant lors des deux stages que j'y ai effectués. Leur expertise et leur soutien ont été déterminants dans l'avancement de mes travaux.

Je suis profondément reconnaissant envers les membres du jury qui ont accepté d'évaluer ma thèse malgré leurs emplois du temps chargés. Leur contribution a été précieuse et est grandement appréciée.

Un grand merci à ma famille, mes amis et mes collègues pour leur soutien inestimable, leurs conseils éclairés et leurs encouragements tout au long de cette aventure. Leur présence a été une source de motivation constante.

Enfin, je remercie le personnel de l'école de technologie supérieure pour leur aide et leur soutien, qui ont grandement facilité mon parcours.

Modélisation conjointe du silicium, de l'oxygène et de la silice

Karim ZONGO

RÉSUMÉ

Cette thèse décrit le développement de potentiels interatomiques pour le silicium et la silice, suivis d'une étude numérique du silicium amorphe basée sur le dit potentiel. Le silicium et la silice sont les constituants fondamentaux de plusieurs matériaux de construction tels que les argiles, les matériaux cimentaires et les polymères activés par alcalins. Ce travail propose une description conjointe par potentiel interatomique d'un composé de la silice (SiO_2) et de ses composants élémentaires, le silicium (Si) et l'oxygène (O), intégrant le concept de transfert de charge. Cependant, cette description combinée et les complexités liées au transfert de charge posent des défis aux méthodes traditionnelles, qui rencontrent souvent des limitations spatiotemporelles et des problèmes de transférabilité.

Dans cette recherche, la silice et ses composants élémentaires — le silicium et l'oxygène — ont été utilisés comme prototypes pour évaluer la capacité du Potentiel par Tenseur de Moments (MTP) à décrire simultanément et avec précision des espaces configurationnels disjoints, tout en intégrant les transferts de charge. Le MTP est un potentiel basé sur l'apprentissage automatique qui a montré des résultats prometteurs dans la description non conjointe de matériaux tels que les semi-conducteurs, l'eau et les alliages. Dans un premier temps, une base de données complète pour la silice, le silicium et l'oxygène a été générée à partir de milliers de calculs basés sur la théorie de la fonctionnelle de la densité électronique (DFT). Par la suite, le MTP a été paramétrisé en utilisant ce jeu de données, adoptant une approche unifiée pour le système disjoint Si-SiO₂-O. Le potentiel obtenu a ensuite été utilisé pour des calculs statiques et des simulations de dynamique moléculaire, avec des résultats comparés aux données DFT, aux résultats expérimentaux et aux modèles semi-empiriques. Le potentiel a été testé en profondeur en le couplant à la technique d'activation et de relaxation *nouveau* (Activation Relaxation Technique *nouveau* - ARTn) pour modéliser le silicium amorphe.

Les prédictions du potentiel unifié s'alignent bien avec les données expérimentales. De plus, il a été conclu que les coordonnées de réaction chimiques sont suffisantes pour décrire les transferts de charge, tandis qu'un choix judicieux du nombre de paramètres du modèle permet de décrire simultanément les régions disjointes de l'espace configurationnel et ce de façon efficace. Suite à ces résultats, une base de données sur l'eau, ainsi que sur les molécules et structures liées au gel de silice, a été conçue et générée. L'étape de l'eau constitue un point de départ pour la modélisation de systèmes plus complexes en lien avec les matériaux cimentaires et argileux.

Le coût computationnel associé à la modélisation dans cette thèse reste modeste grâce à la taille de la base de données générée et aux simulations effectuées. Il existe un potentiel d'optimisation supplémentaire des coûts en élaborant un plan de travail concis et en affinant les codes impliqués dans le développement du potentiel.

Mots-clés: Modèle, simulation numérique, dynamique moléculaire, base de données, potentiel interatomique, apprentissage automatique, paramétrisation, regression, optimisation, coût computationnel

Joint modeling of silicon, oxygen, and silica.

Karim ZONGO

ABSTRACT

This thesis focuses on the development of interatomic potentials for silicon and silica, followed by a numerical study of amorphous silicon, using said potential. Silicon and silica are fundamental constituents of many construction materials, such as clays, cement materials, and alkali-activated polymers. This work proposes a joint interatomic potential description of the silica compound (SiO_2) and its elemental components, silicon (Si) and oxygen (O), implicitly handling charge transfer. This combined description and the complexities associated with charge transfer present challenges for traditional methods, which often encounter spatiotemporal limitations and transferability issues.

Herein, silica and its elemental components—silicon and oxygen—were used as prototypes to assess the ability of the Moment Tensor Potential (MTP) to simultaneously and accurately describe disjoint configuration spaces while integrating charge transfers. The MTP is based on machine learning. It has shown promise in the non-joint description of materials such as semiconductors, water, and alloys. Initially, a comprehensive database for silica, silicon, and oxygen was generated from thousands of calculations based on density functional theory (DFT). Subsequently, the Moment Tensor Potential was parameterized using this extensive dataset, adopting a unified approach for the disjoint system Si-SiO₂-O. The resulting potential was then employed for static calculations and molecular dynamics simulations, with results compared to DFT data, experimental outcomes, and semi-empirical models. The potential was further tested by coupling it with the Activation Relaxation Technique *nouveau* (ARTn) to model amorphous silicon.

The predictions of the unified potential align well with experimental data, demonstrating competitiveness with semi-empirical results for certain properties while surpassing them for others. Furthermore, reaction coordinates are sufficient to describe charge transfers, while a careful selection of the number of model parameters enables the effective simultaneous description of disjoint regions of the configurational space. Following these results, a database was developed and generated for water, as well as for molecules and structures related to silica gel. The water phase serves as a starting point for modeling more complex systems associated with cementitious and clay materials.

The computational cost associated with the modeling in this thesis remains modest due to the size of the generated database and the simulations conducted. There exists further potential for cost optimization by developing a concise work plan and refining the codes involved in the potential's development.

Keywords: Model, numerical simulation, molecular dynamics, database, interatomic potential, machine learning, parameterization, regression, optimization, computational cost

TABLE DES MATIÈRES

	Page
INTRODUCTION	1
0.1 Énoncé du problème	4
0.2 Objectifs de recherche	5
0.3 Organisation de la thèse	6
0.4 Contributions principales	6
0.5 Limitations de la thèse	9
 CHAPITRE 1 REVUE DE LITERATURE	 11
1.1 Matériaux	11
1.2 Processus de simulation numérique	15
1.2.1 Modélisation multi-échelle	15
1.2.2 Simulations atomistiques et moléculaires	19
1.2.2.1 Simulation par dynamique moléculaire	20
1.2.2.2 Intégration des équations du mouvement	22
1.2.2.3 Ensembles thermodynamiques	23
1.2.2.4 Simulation par les méthodes statistiques	25
1.2.3 Conditions aux limites périodiques	28
1.2.4 Méthodes ab initio	28
1.2.5 Champ de force classique	34
1.2.6 Apprentissage automatique	42
1.2.7 Application au potentiel interatomique : Potentiel interatomique par apprentissage automatique	45
1.2.8 Propriété de partitionnement issue de la localité des interactions interatomiques	48
1.2.9 Régresseur à Réseau de Neurones (NN)	53
1.2.9.1 Potentiel par réseau de neurones (NNP : Neural Network Potential)	53
1.2.10 Régression par processus gaussien (GPR)	60
1.2.10.1 Potentiel par approximation gaussienne (GAP : Gaussian Approximation Potential)	60
1.2.11 Régression ridge noyau (KRR : Kernel Ridge Regression)	61
1.2.11.1 Champ de force adaptatif, généralisable et informé par le voisinage (AGNI : Adaptive, Generalizable, and Neighborhood Informed)	61
1.2.12 Régression linéaire	62
1.2.12.1 Le potentiel d'analyse spectrale des voisins (SNAP : Spectral Neighbor Analysis Potential)	62
1.2.12.2 Potentiel par tenseur de moments (MTP : Moment Tensor Potential)	63
1.2.13 Labellisation des données	76

1.2.14	Optimisation numérique	82
1.2.14.1	Propriétés générales de l'optimisation numérique	87
CHAPITRE 2 FIRST-PRINCIPLES DATABASE FOR FITTING A MACHINE- LEARNING SILICON INTER-ATOMIC FORCE-FIELD		
		89
2.1	Introduction	89
2.2	Methods	90
2.2.1	Database generation	90
2.2.2	Computational details	94
2.2.3	Reliability of the database	99
2.2.4	Training details	100
2.3	Results and discussion	101
2.4	Conclusion and future work	105
CHAPITRE 3 A UNIFIED MOMENT TENSOR POTENTIAL FOR SILICON, OXYGEN, AND SILICA		
		107
3.1	Abstract	107
3.2	Introduction	107
3.3	Results	109
3.3.1	Cohesive energy, elastic constant, point defects and extended defects	109
3.3.2	Silicon slab energy	120
3.3.3	Disordered structures	122
3.3.4	Liquid and amorphous silicon	123
3.3.5	Liquid and amorphous silica	123
3.3.6	Phonon dispersion	128
3.3.7	Si - SiO ₂ Interface	128
3.4	Discussion	131
3.5	Methods	135
3.5.1	<i>Ab initio</i> calculations	135
3.5.2	Machine learning model : The moment tensor potential	135
3.5.3	Data curation and optimization	137
3.5.4	Training set	139
3.5.5	Training and validation	140
3.5.6	Static calculations	141
3.5.7	Si - SiO ₂ Interface construction	143
3.5.8	Molecular dynamics simulation	145
3.6	Supplementary Information for : A unified moment tensor potential for silicon, oxygen, and silica	146
3.6.1	Database preparation	146
3.6.2	Silicon	147
3.6.3	Silica	148
3.6.4	Oxygen	149
3.6.5	Active learning	150
3.6.6	Training mode	151

3.6.7	Amorphous silicon	151
3.6.8	Amorphous silica	152
CHAPITRE 4 AMORPHOUS SILICON STRUCTURES GENERATED USING A MOMENT TENSOR POTENTIAL AND THE ACTIVATION RELAXATION TECHNIQUE <i>NOUVEAU</i>		
4.1	Abstract	175
4.2	Introduction	176
4.3	Methods	178
4.3.1	Activation Relaxation Technique <i>nouveau</i>	178
4.3.2	A moment tensor potential for silicon	179
4.3.3	Protocol for input models preparation and simulation details	182
4.3.4	Structural analysis	183
4.4	Results and discussion	184
4.4.1	Short-range order structural analysis of the models	188
4.4.2	Medium-range order structural analysis of the models	191
4.4.3	Refinement of ART-SW optimized model : 1000-aSi-SW	193
4.4.4	Stiffness	196
4.4.5	Potential energy	197
4.4.6	Density	198
4.4.7	Nature of the relaxation	199
4.4.8	Computational cost and nonlinear optimization process	201
4.5	Conclusions	202
4.6	Acknowledgements	203
4.7	Competing interests	204
CHAPITRE 5 THE COST OF MODELLING SILICON-BASED MATERIALS WITH MACHINE LEARNING POTENTIAL		
5.1	Introduction	205
5.2	Computational methods	207
5.2.1	Moment tensor potential	207
5.2.2	Generation of a database on parallel computing	208
5.3	Results	208
5.4	Conclusion	212
CHAPITRE 6 BASE DE DONNÉES DE L'EAU		
6.1	Introduction	215
6.2	Méthodes	219
6.2.1	Génération de la base de données	219
6.3	Résultats	220
6.4	Conclusion	222
CONCLUSION		225

RECOMMANDATIONS	229
DÉPÔT DES DONNÉES ATOMIQUES	235
ANNEXE I ARTICLE DE CONFÉRENCE	237
LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES	243

LISTE DES TABLEAUX

	Page
Tableau 1.1	Paramètres de BKS, $q_{Si} = 2.4$ et $q_O = -1.2$ 40
Tableau 1.2	Liste de descripteurs Adapté de Wang <i>et al.</i> (2024, p. 4) 46
Tableau 1.3	La liste des logiciels implémentés et des dates de publication pour différents modèles MLIP Tiré de Wang <i>et al.</i> (2024, p. 7) 51
Tableau 2.1	Content of the DFT database used for fitting the MTP 95
Tableau 2.2	Comparison between our DFT computed lattice parameters and experimental values 99
Tableau 2.3	Average training and validation errors 100
Tableau 2.4	Point defects diffusion energies barriers (eV) 101
Tableau 3.1	Comparison of lattice constants predicted by MTP, ReaxFF, and COMB models against experimental data and DFT calculations 110
Tableau 3.2	Comparison of force fields prediction of elastic constants for silicon crystal and α -quartz 112
Tableau 3.3	Exploring bulk modulus in silicon and silica polymorphs 114
Tableau 3.4	Comparative analysis of defect formation energies in silicon crystals and silica polymorphs using DFT, MTP, and ReaxFF results 115
Tableau 3.5	Final refined and optimized training dataset derived from extensive uncured database 140
Tableau 3.6	Examples of fitting parameters for levels 26 and 28 of MTP potentials, consisting of 1657 and 2638 parameters respectively, are provided for a two-component system 152
Tableau 3.7	The computational expenses 153
Tableau 3.13	Compositon of silica database 153
Tableau 3.8	Illustrations of training errors (RMSE) on defect subsets 162
Tableau 3.9	Illustrations of training errors (RMSE) on cohesive energy and elastic constant subsets 163

Tableau 3.10	Illustrations of training errors (RMSE) on disordered structures subsets	164
Tableau 3.11	First refined and optimized training set (set 1)	165
Tableau 3.12	The first unification process that led to training set 1	166
Tableau 4.1	The models utilized in this study, the preparation methods of the input configurations, the number of atoms in each model, and the preparation methods for the a-Si models	181
Tableau 4.2	Structural characteristics of ARTn-MTP generated amorphous models	185
Tableau 4.3	Structural and elastic properties for 1000-aSi-SW initial and 1000-aSi-SW final configurations	195
Tableau 4.4	Initial (ρ_i) and final (ρ_f) density for the ARTn-MTP generated a-Si model	198
Tableau 5.1	Total number of configurations in the database and fraction utilized for the implementation of the unified Si-SiO ₂ -O potential (Zongo <i>et al.</i> , 2024)	209
Tableau 6.1	Base de données	221

LISTE DES FIGURES

	Page
Figure 0.1	Illustration du temps de simulation et de la précision des différentes méthodes de simulation numérique 3
Figure 1.1	Bloc de silicium pur et de dioxyde de silicium Tirée respectivement de Wikipedia contributors (2023, p. 1) et de Indiamart (2024, p. 1) 12
Figure 1.2	Modèle mésoscopique du ciment hydraté et cellule unitaire de la montmorillonite obtenue par simulation MD classique Tirée respectivement de Zhou <i>et al.</i> (2019, p. 2) et de Morteza Amirani (2019, non publiée) 13
Figure 1.3	Structure moléculaire du C-S-H à l'échelle nanométrique, incluant l'effet du rapport C/S sur la structure moléculaire du C-S-H 14
Figure 1.4	La cellule unité conventionnelle du silicium et de l' α -quartz 15
Figure 1.5	Les étapes du développement d'un modèle Tirée de Ashby (1992, p. 2) . 16
Figure 1.6	Modèle d'énergie à différentes échelles de longueur et de temps pour la simulation computationnelle 18
Figure 1.7	Boite de simulation et fonctionnement des conditions aux limites périodiques Tirée de Wikipedia contributors (2024, p. 1) 29
Figure 1.8	Schéma de calcul de l'équation de Kohn-Sham pour la détermination de la densité électronique 33
Figure 1.9	Les principaux types d'apprentissage automatique 44
Figure 1.10	Diagramme de processus et les quatre étapes essentielles des MLIPs Tirée Wang <i>et al.</i> (2024, p. 3) 47
Figure 1.11	Illustration d'un environnement atomique local au sein d'une configuration donnée x_q avec N_{tot} atomes, d'énergie totale E^{qm} donnée par la mécanique quantique 49
Figure 1.12	Disposition du réseau de neurones artificiels 54
Figure 1.13	Construction du potentiel de réseau de neurones 55
Figure 1.14	Angle du triplet d'atomes 56

Figure 1.15	Transformation du voisinage atomique local	57
Figure 1.16	Délimitation illustrative de l'environnement atomique local pour une configuration donnée (x)	64
Figure 1.17	Schéma de fonctionnement du potentiel par tenseur de moment	66
Figure 1.18	Rotation de système de coordonnées	70
Figure 1.19	Les six premières fonctions de base radiales définies dans l'équation 1.72	74
Figure 1.20	Principe de génération des données d'entraînement à partir des configurations atomiques	77
Figure 1.21	Illustration de la relation entre deux ensembles de données	78
Figure 1.22	Étiquetage des données générées par la DFT à partir des configurations atomiques	79
Figure 1.23	Étiquetage des données calculées par le MTP à partir des configurations atomiques	79
Figure 1.24	Algorithme du gradient sur une fonction de type bol (vue en plan avec courbes de niveaux) Tirée de Wikipédia (2024, p. 1)	84
Figure 2.1	DFT-computed and MTP-predicted total energies (a) training set- 1062 configurations (b) validation set - 365 configurations	102
Figure 2.2	DFT-computed and MTP-predicted force components (a) training set - 1062 configurations(b) validation set - 365 configurations	102
Figure 2.3	DFT-computed and MTP-predicted stress components (a) training set - 1062 configurations(b) validation set - 365 configurations	103
Figure 2.4	Generalized stacking fault curve	103
Figure 2.5	Generalized stacking fault curve	104
Figure 3.1	Bond and angle energy of oxygen molecules, and equation of state of silicon and silica polymorphs	113
Figure 3.2	Point defects in silicon crystal	117
Figure 3.3	Diffusion of point defects in silicon crystal	118

Figure 3.4	Planar defect in silicon crystal	121
Figure 3.5	Line defect in silicon crystal	122
Figure 3.6	Liquid and amorphous silicon	124
Figure 3.7	Liquid silica	125
Figure 3.8	Amorphous silica	127
Figure 3.9	Phonon dispersions	129
Figure 3.10	Interface structure of Si - SiO ₂	132
Figure 3.11	Interface energy of Si-SiO ₂	133
Figure 3.12	Distribution of configuration sizes in the unified training sets	161
Figure 3.13	Distribution of configuration sizes in the validation set V_1 picked from the curated data such that $V_1 \cap T_{unified} = \emptyset$	162
Figure 3.14	Distribution of configuration sizes in the validation set V_2 selected from AIMD cooling trajectories	163
Figure 3.15	Mean square displacement	164
Figure 3.16	Comparison of RMSE in energy, forces, and stress	167
Figure 3.17	Comparison of RMSE in energy, forces, and stress	168
Figure 3.18	Comparison of RMSE in energy, forces, and stress	169
Figure 3.19	Comparison of RMSE in energy, forces, and stress	170
Figure 3.20	Initial state of the simulation block containing both the solid and liquid phases used for the melting point simulation	171
Figure 3.21	Illustrations of screw dislocation core structures obtained using the unified potential	172
Figure 3.22	Illustrations interface configurations	173
Figure 3.23	Illustrations interface configurations	173
Figure 4.1	Key elements characterizing the structure of amorphous silicon	182

Figure 4.2	Coordination numbers of input configurations and ARTn-MTP optimized configurations (final configurations)184
Figure 4.3	Comparison of the local structural properties of input configurations (top panels : a, b, and c) with the final configuration (100% coordinated) (bottom panels : d, e, and f) generated using the ARTn-MTP coupling scheme186
Figure 4.4	Local structural properties of ARTn-MTP generated a-Si models189
Figure 4.5	Proportion of coordination defects in various a-Si models generated by different methods. 216-R1 has no floating bond and 216-R2 is 100 % fully 4-fold coordinated. No atoms with fewer than 3 neighbors or more than 5 neighbors have been recorded in any model190
Figure 4.6	FWHM of the bond-angle distribution. The FWHM values are computed by fitting a Gaussian distribution to the bond angle distributions. The red line denotes the experimental value of FWHM (Fortner <i>et al.</i> , 1989)191
Figure 4.7	Structure factor for ARTn-MTP generated models compared with that obtained from experimental data and 100k-GAP18 atom simulations ...192
Figure 4.8	Ring statistics for ARTn-MTP generated models as well as those derived from the 512-WWW model and 100k-GAP18 atom models193
Figure 4.9	Local structural properties of ARTn-MTP refined configurations (1000-aSi-SW final) compared to ART-SW optimized configurations (1000-aSi-SW initial)194
Figure 4.10	Bulk modulus for ARTn-MTP generated models as well as models from the literature196
Figure 4.11	Excess energy between ART-MTP configurations (a-Si) and crystalline silicon (c-Si) as a function of the number of accepted events. A total of 4 events per atom was required to achieve 100% 4-fold coordination in 216-R2 a-Si model197
Figure 4.12	Analysis of accepted Events in ARTn-MTP optimization : a) Barrier distribution, b) square root of squared atomic displacement versus energy barrier, c) energy barrier as a function of energy decrease, and d) energy barrier versus active atoms during events200

Figure 4.13	Nonlinear optimization effects characterized by excess energy per atom and number of events computed from independent runs with identical input Configuration201
Figure 5.1	Creating a database of Si and SiO ₂ Configurations : A Blueprint for database creation by combining two or more modelling software such as Xcrysden, VESTA, AtomsK, OVITO, and Quantum ESPRESSO209
Figure 5.2	Migration of computation files and management, monitoring of computations on the Digital Research Alliance of Canada cluster210
Figure 5.3	Atomic structures in the database : Liquid, amorphous, defects of silicon and silica, oxygen molecules, and water molecules. The configurations are labeled with energies, forces, and stress values211
Figure 5.4	Evolution of the cost function during training on Digital Research Alliance of Canada Cluster211
Figure 5.5	Resource consumption for our modelling projects over four calendar years212
Figure 6.1	Modèles moléculaires de l'eau, du peroxyde d'hydrogène et du dihydrogène221
Figure 6.2	Structures moléculaires de l'acide orthosiliciques et de l'acide pyrosiliciques222
Figure 6.3	Eau liquide dans une boîte de simulation cubique et vue de dessus d'une structure de glace hexagonale222

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ETS	École de technologie supérieure
e.g.	Exempli gratia (For example)
i.e.	Id est (That is)
et. al.	Et Alia (And others)
DFT	Density functional theory
LAMMPS	Large-scale Atomic/Molecular Massively Parallel Simulator
PBC	Periodic boundary condition
CPL	Conditions périodiques aux limites
PES	Potential energy surface
SEP	Surface d'énergie potentielle
RMC	Reverse Monte Carlo
HRMC	Hybrid Reverse Monte Carlo
ECSR	Experimentally Constrained Structural Relaxation
ECMR	Experimentally Constrained Molecular Relaxation
WWW	Wooten-Winer-Weaire
CRN	Continuous random network
ART	Activation relaxation technique
ARTn	Activation relaxation technique nouveau
FEAR	Force Enhanced Atomic Refinement
FF	Force fields
KRR	Kernel Ridge Regression
GPR	Gaussian Process Regression
GAP	Gaussian Approximation Potential
NN	Neural Network
NNP	Neural Network Potential

SNAP	Spectral Neighbor Analysis Potential
AGNI	Adaptive, Generalizable, and Neighborhood Informed
SOAP	Smooth Overlap of Atomic Positions
MD	Molecular dynamic
NVE	Constant number of particles (N), a fixed volume (V), and a conserved total energy (E).
NVT	Constant number of particles (N), a fixed volume (V), and a maintained temperature (T)
NPT	Constant number of particles (N), a maintained pressure (P), and a controlled temperature (T)
CPMD	Car-Parrinello Molecular Dynamics
AIMD	<i>Ab initio</i> molecular dynamics simulation
RDF	Radial distribution fuction
ADF	Angular distribution function
MMT	Montmorillonite
SiO ₂	Silicon dioxide
Si	Silicon
O	Oxygen
O ₂	Dioxygen molecule
O ₃	Ozone molecule
H	Hydrogène
H ₂	Dihydrogène
Ca	Calcium
Al	Aluminum
MLIP	Machine learning interatomic potential
MTP	Moment tensor potential
HD	Hexagonal diamond

HCP	Hexagonal closed packek
SH	Simple hexagonal
C24	Clathrate 24
C46	Clathrate 46
C136	Clathrate 136
FCC	Face-Centered Cubic
BCC	Body-Centered Cubic
α -Q	α -quartz
β -Q	β -quartz
α -C	α -cristobalite
β -C	β -cristobalite
α -T	α -tridymite
β -T	β -tridymite
SE	Seifertite
ST	Stishovite
KE	Keatite
CO	Coesite
MO	Moganite
CH	Chabazite
ZG	Zeolite GUS
ZSM	Zeolite ZSM
α -2D	Two-dimensional silica α
β -2D	Two-dimensional silica β
δ -2D	Two-dimensional silica δ
γ -2D	Two-dimensional silica γ
GSF	Generalized stacking fault
RRMSE	Relative Root Mean Square Error

FWHM	Full Width at Half Maximum
cpu	Central processing unit
XC	Exchange-Correlation
NEB	Nudge Elastic Band
GGA	Generalized Gradient Approximation
PBE	Perdew, Burke, and Ernzerhof
SCF	Self-Consistent Field
PAW	Projector Augmented Wave Pseudopotential
BFGS	Broyden-Fletcher-Goldfarb-Shanno
NWMO	Nuclear Waste Management Organization
CNS	Canadian Nuclear Society

LISTE DES SYMBOLES ET UNITÉS DE MESURE

ρ	Density
$\text{\AA}$	Angstrom
$^{\circ}\text{C}$	Degree Celsius
K	Kelvin
eV	Electronvolt
meV	Millielectronvolt
J	Joule
J/m^2	Joules per square meter
GPa	Gigapascal
P	Probability
S	Static Structure Factor
$\mathbf{q}$	Scattering Vector
fs	femtosecond
ps	Picosecond
K/ps	Kelvin per picosecond
$\text{eV/\AA}$	Electronvolts per angstrom

INTRODUCTION

Le silicium et ses oxydes jouent un rôle crucial dans l'industrie et la technologie modernes. Ils sont utilisés dans la fabrication de composants électroniques (Streetman *et al.*, 2000; Laskowski *et al.*, 2019; Sze *et al.*, 2021) et dans divers matériaux de construction (Aggarwal *et al.*, 2015; Siddique, 2011; Shi *et al.*, 2011). Par exemple, dans le domaine de la construction, la fumée de silice est utilisée dans des formulations spécialisées de béton, un des matériaux les plus largement employés sur notre planète (Siddique, 2011; Mehta *et al.*, 2020; Toutanji *et al.*, 1995; Ding *et al.*, 2002). Les chaînes de silice sont également des composantes clés de la structure moléculaire du ciment, formée lors de son hydratation (Flores *et al.*, 2017; Zhou *et al.*, 2019; Abdolhosseini Qomi *et al.*, 2014). Le silicium (Si) et la silice (SiO₂) constituent des éléments essentiels dans les argiles, comme l'argile bentonite, qui est envisagée pour le stockage souterrain des déchets radioactifs (Ma *et al.*, 2019). En outre, les zéolithes sont une famille de matériaux aluminosilicatés cristallins microporeux, constitués principalement de silice, et sont couramment utilisés commercialement comme adsorbants et catalyseur (Kellouai, 2022; Reungoat, 2007; Gottardi *et al.*, 2012). Les oxydes de silicium jouent aussi un rôle important dans la fabrication de polymères inorganiques activés par alcalins. En effet, ces polymères se forment par une réaction de polycondensation des oxydes d'aluminium et de silicium dissous en présence d'un activateur à pH élevé (Di Mare *et al.*, 2022; Van Deventer *et al.*, 2007). Ces matériaux complexes, tels que les matériaux cimentaires, les systèmes argileux, les zéolithes et les polymères inorganiques activés par alcalins, qui renferment une proportion significative de silicium et de silice, jouent un rôle déterminant dans plusieurs secteurs clés, y compris la protection de l'environnement. Une compréhension approfondie de ces matériaux est essentielle, car elle peut contribuer, de manière directe ou indirecte, à la réduction de l'empreinte carbone (Worrell *et al.*, 2001) et à l'adsorption des polluants (Ghayaza, 2012; Coulombe, 2023). La multifonctionnalité de ces matériaux complexes nécessite une compréhension approfondie de leur structure atomique à toutes les échelles, ainsi que de la manière dont ils réagissent

et évoluent en réponse aux conditions externes. Cependant, d'une part, ces matériaux sont complexes, et d'autre part, les applications envisagées nécessitent des performances sur des centaines d'années ou des milliers d'années. Cela rend les expériences longues, difficiles à réaliser, et parfois même impraticables. En effet, la durée de prédiction nécessaire pour évaluer le comportement à long terme de ces matériaux dépasse souvent les capacités des expérimentations réalisables. De ce fait, la simulation numérique s'avère être un atout précieux pour remédier les expériences physiques complexes et difficilement réalisables. Non seulement elle permet de sonder et d'explorer des zones inaccessibles aux tests expérimentaux, mais elle offre également la possibilité de valider et de confirmer les résultats expérimentaux. Les simulations numérique à l'échelle nanométrique peuvent offrir des aperçus qualitatifs et quantitatifs sur les mécanismes importants sur des échelles temporelles très longues, qui pourraient échapper aux expériences à l'échelle macroscopique et microscopique. Elles peuvent également aider à interpréter les expériences, y compris les mesures de transport et de sorption, les tests mécaniques, ainsi que les analyses d'imagerie basées sur des techniques de microscopie avancées et de diffraction.

En général, les simulations à l'échelle atomique en science des matériaux se divisent en deux catégories : les méthodes quantiques et les approches basées sur les champs de force classiques. Pour un système atomique ou moléculaire donné, ces méthodes calculent l'énergie totale en utilisant des approximations différentes. Les approximations de la mécanique quantique, telles que la théorie de la fonctionnelle de la densité (DFT) (Kohn *et al.*, 1965; Sholl *et al.*, 2022; Goedecker, 1999; Schommers, 2019), offrent une grande précision et transférabilité, mais elles sont coûteuses en termes de calcul et sont limitées aux systèmes de petite taille contenant quelques milliers d'atomes, où l'échelle temporelle des phénomènes ou événements est d'environ quelques centaines de picosecondes (Zuo *et al.*, 2020; Behler *et al.*, 2007). En revanche, les champs de force semi-empiriques (Schommers, 2019; Torrens, 2012; Mao *et al.*, 2023; Van Duin *et al.*, 2001; Cygan *et al.*, 2004), bien qu'ils dépassent les limites des approches *ab initio* en terme d'échelle de temps et d'espace, restent inflexibles et leur extensibilité se fait au détriment

de la précision (Zuo *et al.*, 2020; Behler & Parrinello, 2007). Autrement dit comme illustrer à la Figure 0.1, une méthode présente des limites spatiotemporelles, tandis que l'autre souffre de limitations en termes de transférabilité et de précision. Ainsi, nous faisons face à deux problèmes majeurs : d'une part, les limitations des méthodes traditionnelles de modélisation, et d'autre part, la grande complexité des matériaux d'intérêt. Pour remédier à ces enjeux, nous adoptons une approche intégrée, articulée autour de deux volets principaux : le premier concerne la modélisation par étapes ou modélisation multi-échelle et le second, l'utilisation de potentiels interatomiques fondés sur l'apprentissage automatique.

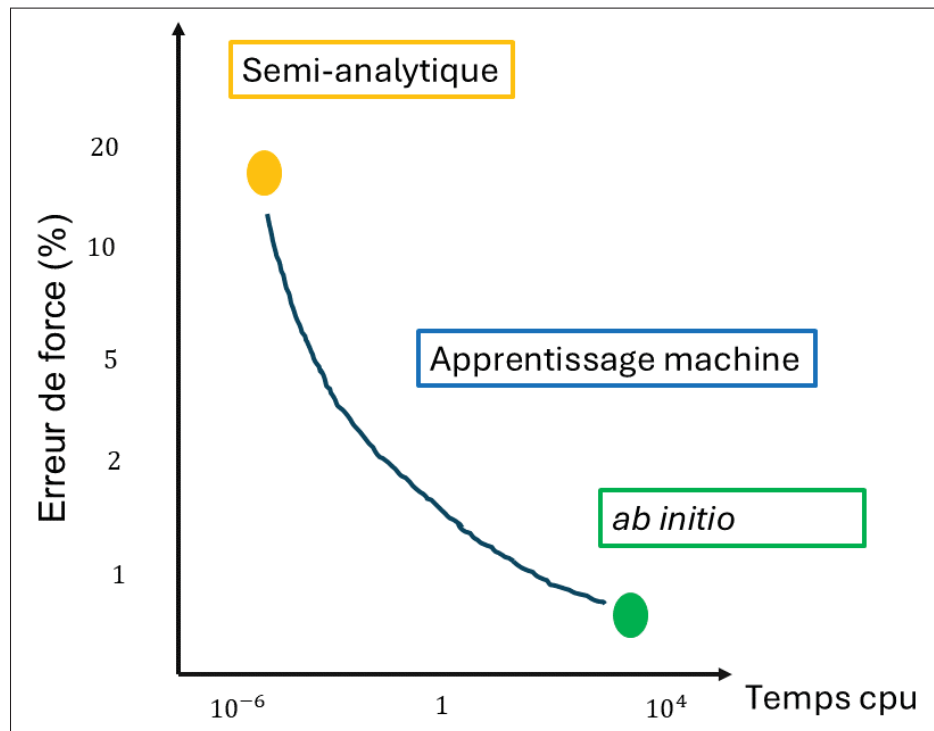


Figure 0.1 Illustration du temps de simulation et de la précision des différentes méthodes de simulation numérique. Pour chaque méthode, la figure montre approximativement l'erreur de prédiction des forces ainsi que le temps nécessaire pour les calculer

La modélisation par étapes consiste à initier l'étude par les constituants élémentaires des matériaux d'intérêt en l'occurrence, le silicium et la silice que nous avons déjà présentés. En principe, la modélisation par étapes se fait progressivement, en partant du silicium et en ajoutant

graduellement des éléments chimiques tels que l'oxygène, l'hydrogène, le calcium, l'aluminium, et d'autres éléments présents dans les matériaux complexes mentionnés. En effet, chacune des structures de ces matériaux complexes qu'il soit ordonnée ou désordonnée peut être représentée par des blocs moléculaire du système (Si, SiO₂, SiO₂-H₂O, SiO₂-H₂O-Ca/Al). À chaque étape de la modélisation, une base de données est conçue et générée, et l'apprentissage automatique est sollicité. Pour ce faire, nous déployons un modèle de potentiel interatomique basé sur les tenseurs de moments (MTP) (Shapeev, 2016). Le MTP est un modèle basé sur l'apprentissage automatique qui a démontré son efficacité dans la modélisation de matériaux tels que les alliages métalliques (Gubaev *et al.*, 2019), les semi-conducteurs (Zuo *et al.*, 2020) et l'eau (Lomaka *et al.*, 2020). Dans cette thèse, ce modèle est développé pour le silicium et la silice, qui constituent les éléments fondamentaux de matériaux complexes tels que les argiles, les matériaux cimentaires et les zéolithes. Le silicium et la silice, en tant que matériaux moins complexes, jouent également le rôle de prototypes pour comprendre le fonctionnement du potentiel par tenseur de moment. Cela a pour objectif de faciliter la transition de l'application du modèle vers une modélisation de matériaux plus complexes, dont les blocs moléculaires constitutifs sont inclus dans le système (Si, SiO₂, SiO₂-H₂O, SiO₂-H₂O-Ca/Al). Le travail réalisé dans cette thèse nous permet donc de poser les bases pour la modélisation de systèmes plus complexes, tels que les systèmes aluminosilicates.

0.1 Énoncé du problème

Les interfaces du silicium et de la silice sont de grand intérêt pour l'industrie de la microélectronique et des semi-conducteurs. La formation de ces interfaces implique des transferts de charge. De plus, ces deux entités constituent des composants essentiels de matériaux multifonctionnels, tels que les argiles, le ciment et le béton, où les transferts de charge sont omniprésents. Ces matériaux jouent un rôle crucial tant dans l'industrie que dans les solutions environnementales.

Notre recherche se concentrera sur l'amélioration de la description conjointe des forces interatomiques au sein de la silice et du silicium, en tenant compte des transferts de charge associés. Nous examinerons également le coût computationnel lié à cette description conjointe des forces pour un composé et ses constituants élémentaires. En plus, il est question de comprendre le fonctionnement du modèle de potentiel basé sur le tenseur de moment à un niveau moins complexe, afin de préparer son application à des matériaux plus complexes. Dans ce contexte, notre recherche explorera également la modélisation introductive de l'eau (H_2O), qui constitue un passage essentiel entre les matériaux moins complexes, tels que le Si et le SiO_2 , et ceux beaucoup plus complexes, comme les argiles et les matériaux cimentaires.

0.2 Objectifs de recherche

L'objectif principal de cette thèse est de modéliser les interactions interatomiques au sein du silicium et de la silice, qui constituent des éléments fondamentaux des matériaux complexes tels que les argiles, les matériaux cimentaires et les zéolithes. Cette thèse s'intéresse également à la modélisation introductive de l'eau, qui représente un point de départ crucial pour la modélisation des matériaux complexes.

Objectif spécifique 1 : Développer une base de données *ab initio* pour silicium.

Objectif spécifique 2 : Développer un modèle d'apprentissage-automatique pour le silicium basé sur cette base de donnée.

Objectif spécifique 3 : Validation approfondie du potentiel pour le silicium en l'utilisant pour modéliser le silicium amorphe.

Objectif spécifique 4 : Étendre la base de données pour y inclure l'oxygène et la silice.

Objectif spécifique 5 : Étendre le modèle d'apprentissage-automatique pour y inclure la silice et l'oxygène.

Objectif spécifique 6 : Validation du potentiel Si/SiO₂/O₂ contre d'autres calculs *ab initio* et mesures expérimentales.

Objectif spécifique 7 : Étendre la base de données pour y inclure l'hydrogène et l'eau.

0.3 Organisation de la thèse

Cette thèse se compose d'une INTRODUCTION, de six chapitres, d'une CONCLUSION, suivie des RECOMMANDATIONS, ainsi que des ANNEXES. Le chapitre 1 offre une revue de la littérature. Les chapitres 2, 3, 4 et 6 sont des articles scientifiques. Trois ont été révisés par les pairs et publiés. L'un d'entre eux, le chapitre 4, a récemment été soumis à *Physical Review B* pour révision. Le chapitre 2 aborde la conception et le calcul d'une base de données atomiques pour le silicium destinée aux potentiels d'apprentissage automatique. Ce chapitre a été publié dans le journal *MRS Advances*. Dans le chapitre 3, nous proposons un potentiel par apprentissage automatique unifié pour le silicium, la silice et l'oxygène. Ce chapitre a également été publié dans le journal *npj Computational Materials*. Le chapitre 4 est dédié à l'application du potentiel développé dans le chapitre 3 pour modéliser la structure amorphe du silicium. Comme nous l'avons déjà mentionné, les résultats ont été publiés dans le journal *Physical Review B*. Le chapitre 5, quant à lui, est plus concis et décrit la préparation de la base de données de l'eau. Enfin, le chapitre 6 résume l'utilisation des ressources de l'Alliance de recherche numérique du Canada pour l'élaboration des travaux de cette thèse. Ce chapitre a été publié dans le journal *Springer Nature Switzerland*. La thèse donne également des informations sur le dépôt de la base de données dans Borealis.

0.4 Contributions principales

Les travaux de cette thèse se sont principalement concentrés sur la résolution des problèmes précédemment évoqués. Cette démarche se traduit par une série de contributions significatives

pour les projets futurs et pour la communauté scientifique œuvrant dans le domaine de la modélisation. Ces contributions se déclinent en trois composantes essentielles : une vaste base de données, un potentiel unifié pour le système (Si, O, SiO₂), le développement de modèles de très haute qualité pour le silicium amorphe et une composante d'ordre technique. Notons que quatre des chapitres de la présente thèse sont des articles scientifiques révisés par les pairs, signe que les travaux qui y sont décrits sont pertinents pour l'avancement du savoir.

Contribution 1 : Base de données du silicium (Si), oxygène (O), de la silice (SiO₂) et de l'eau (H₂O)

La première contribution majeure de cette thèse est la création d'une base de données exhaustive, qui regroupe des informations cruciales sur les systèmes étudiés. Cette base de données est essentielle pour la modélisation et l'analyse des interactions atomiques, permettant ainsi de nourrir les futurs travaux de recherche dans le domaine. Une base de données complète sur le silicium, l'oxygène, la silice, l'hydrogène et l'eau est désormais disponible et accessible à la communauté scientifique. Une partie de cette base peut également servir aux travaux de projets connexes concernant l'un des matériaux inclus. Elle constitue un point de départ essentiel pour modéliser des systèmes plus complexes, tels que les argiles ou les matériaux cimentaires, qui sont indispensables pour de nombreuses applications dans notre société.

Contribution 2 : Potentiel unifié Si - O - SiO₂

Un potentiel unifié pour le système (Si, O, SiO₂) est désormais disponible. Ce potentiel offre de nouvelles perspectives pour mener des recherches variées, en tant que potentiel d'interaction. Par exemple, il permet désormais de générer du silicium amorphe de haute qualité en le couplant à la technique d'activation et de relaxation *nouveau* (ARTn). De plus, d'autres polymorphes du silicium ou de la silice peuvent désormais être explorés à l'aide de ce potentiel. Cette avancée peut ouvrir la voie à une compréhension approfondie des défauts dans le silicium, tels que les

dislocations, y compris les paires de kink, grâce aux capacités offertes par ce potentiel déjà développé.

Contribution 3 : Composante d'ordre technique

Du point de vue technique, cette thèse a permis de repousser les limites de la compréhension de la modélisation des matériaux par apprentissage automatique. Nous avons démontré qu'un seul potentiel peut être paramétrisé pour décrire un système disjoint tel que (Si, O, SiO₂), tout en tenant compte des transferts de charge. Ce système disjoint constitue en réalité un composé, ainsi que ses éléments constitutifs. Il est désormais évident que les coordonnées des réactions chimiques suffisent à représenter ces transferts de charge de manière indirecte. De plus, l'ajustement du nombre de paramètres peut grandement faciliter la description des régions disjointes de l'espace configurationnel. Dans l'espace configurationnel, les configurations du silicium pur forment une vaste région partagée par ses différentes structures cristallines, ainsi que par ses défauts et ses états désordonnés. L'oxygène, quant à lui, occupe des niveaux d'énergie distincts, et ses molécules occupent une région différente de l'espace configurationnel. La silice présente également des niveaux d'énergie spécifiques. Ses polymorphes, ses formes désordonnées et les structures contenant des défauts occupent une grande région de l'espace configurationnel. Ces trois régions ne se chevauchent pas, ce qui est illustré par les niveaux d'énergie de ces entités ; les minima associés au silicium, à l'oxygène et au dioxyde de silicium (SiO₂) sont différents. Chacune de ces trois grandes régions de l'espace configurationnel peut être représentée de manière alternative par sa surface d'énergie potentielle. L'unification consiste alors à décrire simultanément l'union de ces trois surfaces d'énergie potentielle. Pour rappel, le silicium, l'oxygène et la silice présentent des niveaux d'énergie distincts, ce qui se traduit par trois surfaces d'énergie potentielle qui ne se chevauchent pas. En outre, la description unifiée inclut non seulement les transitions d'état ou de phase propre à chaque surface d'énergie potentielle, mais elle prend également en compte les réactions chimiques entre les surfaces

d'énergie potentielle différentes impliquant des transfert de charges, en utilisant des coordonnées de réaction appropriées.

0.5 Limitations de la thèse

Les résultats et les diverses contributions sont significatifs ; toutefois, cette thèse a été réalisée dans le cadre de certaines contraintes et approximations susceptibles de limiter la portée et la validité des résultats obtenus.

- Les données du système (Si, O, SiO₂) ont été générées en utilisant une fonctionnelle d'échange-corrélation de type approximation par gradient généralisé (GGA), spécifiquement celle de Perdew, Burke et Ernzerhof (PBE) (Perdew *et al.*, 1996). D'autres types de fonctionnelles XC pourraient entraîner des résultats légèrement différents. En ce qui concerne les données de l'eau, la fonctionnelle SCAN (Strongly Constrained and Appropriately Normed) a été utilisée (Sun *et al.*, 2015).
- Le potentiel par tenseur de moments, ainsi que d'autres modèles d'apprentissage automatique, repose sur une approximation qui stipule que les interactions interatomiques sont locales dans des systèmes atomiques non chargés ou non polarisés. Cette approche permet d'exploiter la propriété de partitionnement de l'énergie totale en contributions d'énergie individuelle pour chaque atome. L'approximation permet alors d'introduire un rayon de coupure (cut-off radius R_C) dans le calcul des énergies atomiques. En principe, ce rayon de coupure R_C est optimisé pour établir un équilibre entre la précision et le coût computationnel.
- La base de données a été optimisée, et une partie a été utilisée pour l'entraînement. En raison de la taille finie du modèle, les configurations de silice à haute énergie n'ont pas été incluses dans les données d'entraînement, ce qui pourrait restreindre la capacité du potentiel à décrire certaines phases liquides de la silice hors équilibre à des températures extrêmes. Quelles que soient les données d'entraînement, il existe toujours des limites à la précision et à l'efficacité des prédictions.

- La base de données du système (Si, O, SiO₂, H, H₂O) ne couvre pas l'intégralité de l'espace configurationnel réel unifié. Pour une entité donnée et ses polymorphes, il existe une infinité de dispositions possibles des coordonnées atomiques qui ne peuvent pas être intégralement couvertes ou incluse dans la base de donnée avant l'entraînement du potentiel. En d'autres termes, il n'existe pas de théorie capable de définir de manière exhaustive l'espace configurationnel d'un matériau donné. Ainsi, la conception de la base de données et l'entraînement du potentiel se concentrent généralement sur des régions d'intérêt spécifiques de l'espace configurationnel.
- Certaines phases solides de la silice, telles que la silice fibreuse, ainsi que l'oxygène solide, un solide moléculaire présentant des interactions covalentes et de Van der Waals, n'ont pas pu être décrites par la fonctionnelle PBE. Cela a conduit à leur exclusion de la base de données.

CHAPITRE 1

REVUE DE LITERATURE

Introduction

La modélisation de matériaux, basée sur les données de la mécanique quantiques, est devenue un outil puissant pour surmonter les limitations des méthodes traditionnelles et approfondir notre compréhension de la fonctionnalité des matériaux. Le domaine de la modélisation des matériaux a été particulièrement séduit par le model du potentiel par apprentissage automatique. Parmi les divers aspects de ce domaine en évolution, nous nous concentrons sur les potentiels interatomiques basés sur l'apprentissage automatique (MLIP : *Machine Learning Interatomic Potential*). Ces modèles utilisent des techniques d'apprentissage automatique pour prédire les interactions atomiques avec une grande précision, offrant des améliorations significatives par rapport aux potentiels conventionnels et renforçant notre capacité à simuler et à comprendre les comportements complexes des matériaux. Dans ce chapitre, nous proposons une revue critique de la littérature sur les simulations numériques des matériaux.

1.1 Matériaux

Le silicium (Si) et l'oxygène (O) sont les éléments chimiques les plus abondants dans la croûte terrestre. Leur combinaison chimique donne naissance aux minéraux silicatés, dont le dioxyde de silicium (SiO_2) est l'un des oxydes les plus importants sur terre, utilisé pour fabriquer une grande variété de matériaux. En général, les matériaux à base de silicium, d'oxygène et de silice sont largement utilisés dans l'industrie et la technologie. Ils sont indispensables pour de nombreux secteurs industriels, notamment l'industrie du verre et de la céramique, ainsi que les industries de la microélectronique.

Le silicium et la silice existent sous diverses formes cristallines et amorphes. Ils peuvent être classés respectivement comme matériaux covalents forts et ioniques, avec une large gamme de géométrie de liaison et de coordination. En particulier, le silicium se présente sous différentes

formes cristallines correspondant à une pression et/ou une température données (Li *et al.*, 2018). Le nombre de coordination et les angles de liaison varient largement selon les polymorphes (Lenosky *et al.*, 2000). Certains polymorphes du silicium présentent plusieurs angles de liaison (Kumagai *et al.*, 2007; Kasper *et al.*, 1964). De manière similaire au silicium, on observe une grande polymorphie dans la silice en fonction de la pression et/ou de la température (Gibbs, 1926; Li *et al.*, 1994). Cependant, la coordination tétraédrique du silicium et la coordination double de l'oxygène dans la phase ordinaire (α -quartz) restent les mêmes dans la plupart des polymorphes, à l'exception des structures de stishovite et de seifertite, où le silicium et l'oxygène sont respectivement coordonnés de manière octaédrique (Smyth *et al.*, 1995) et trigonal (Dera *et al.*, 2002).



Figure 1.1 Bloc de silicium pur et de dioxyde de silicium

Tirée respectivement de Wikipedia contributors (2023, p. 1) et de Indiamart (2024, p. 1)

Le silicium et la silice jouent un rôle essentiel dans les matériaux de construction (Siddique, 2011). Le ciment, un composant clé du béton, contient principalement du dioxyde de silicium (SiO_2) sous forme de silicate de calcium. Comme le montre les Figures 1.2a et 1.3, la silice contribue à la formation des silicates de calcium, qui sont responsables de la résistance et de la durabilité du béton. Le béton, largement utilisé dans la construction de bâtiments, ponts et infrastructures, bénéficie de la robustesse que la silice lui confère (Neville *et al.*, 1987; Siddique,

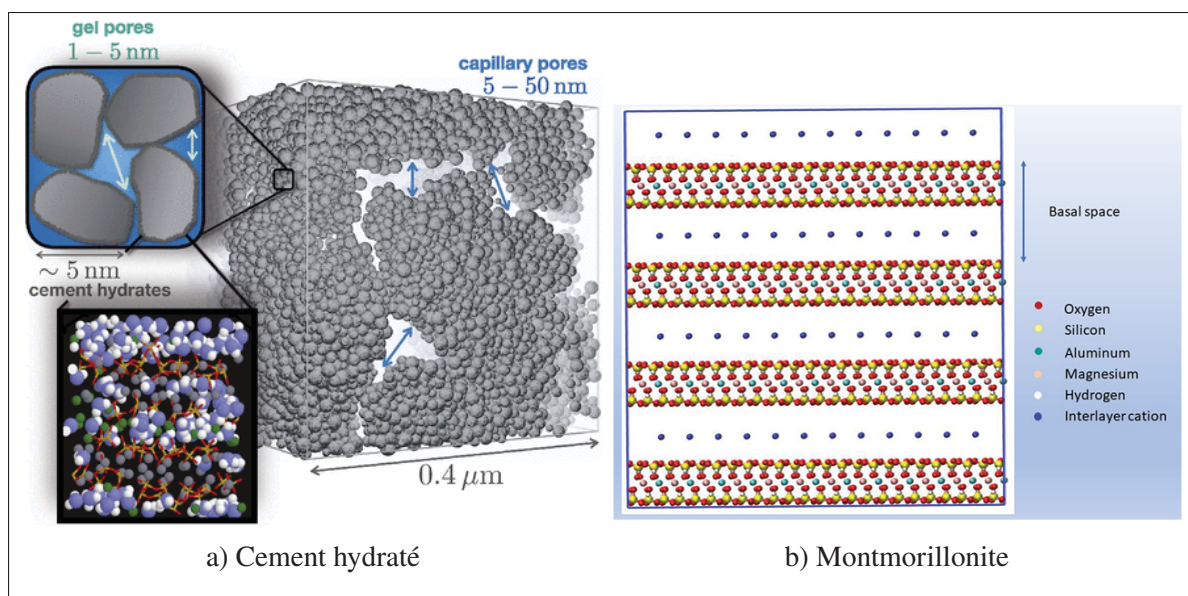


Figure 1.2 Modèle mésoscopique du ciment hydraté et cellule unitaire de la montmorillonite obtenue par simulation MD classique

Tirée respectivement de Zhou *et al.* (2019, p. 2) et de Morteza Amirani (2019, non publiée)

2011; Mazloom *et al.*, 2004). Dans le même ordre d'idée, le verre est principalement constitué de silice fondue (Shelby, 2020). Le dioxyde de silicium est chauffé à haute température pour former du verre, qui est utilisé dans une variété d'applications, y compris les fenêtres, les bouteilles, et les matériaux de vitrage (Shelby, 2020). La pureté et la qualité de la silice sont cruciales pour obtenir un verre transparent et durable. Les briques et les tuiles de construction sont souvent fabriquées à partir d'argiles contenant des minéraux silicatés (Singh, 2022). L'argile bentonite, par exemple, est utilisée dans la fabrication de briques pour sa capacité à se transformer en un matériau solide et durable après cuisson (Singh, 2022). Cette argile particulière, qui se distingue par ses propriétés absorbantes et gonflantes, est largement utilisée dans divers domaines, de la construction à l'industrie pétrolière, en passant par la protection de l'environnement (Coulombe, 2023; Ahmed, 2019; Papin, 1956). Par exemple, l'Organisation de Gestion des Déchets Nucléaires (OGDN) (Nuclear Waste Management Organization, 2002) considère la bentonite comme le principal candidat pour servir de matériau de buffer dans le cadre du plan de stockage souterrain des déchets nucléaires. Cela est dû au fait que la bentonite possède de nombreuses propriétés physico-chimiques, telles qu'une capacité de gonflement, une grande surface spécifique, une

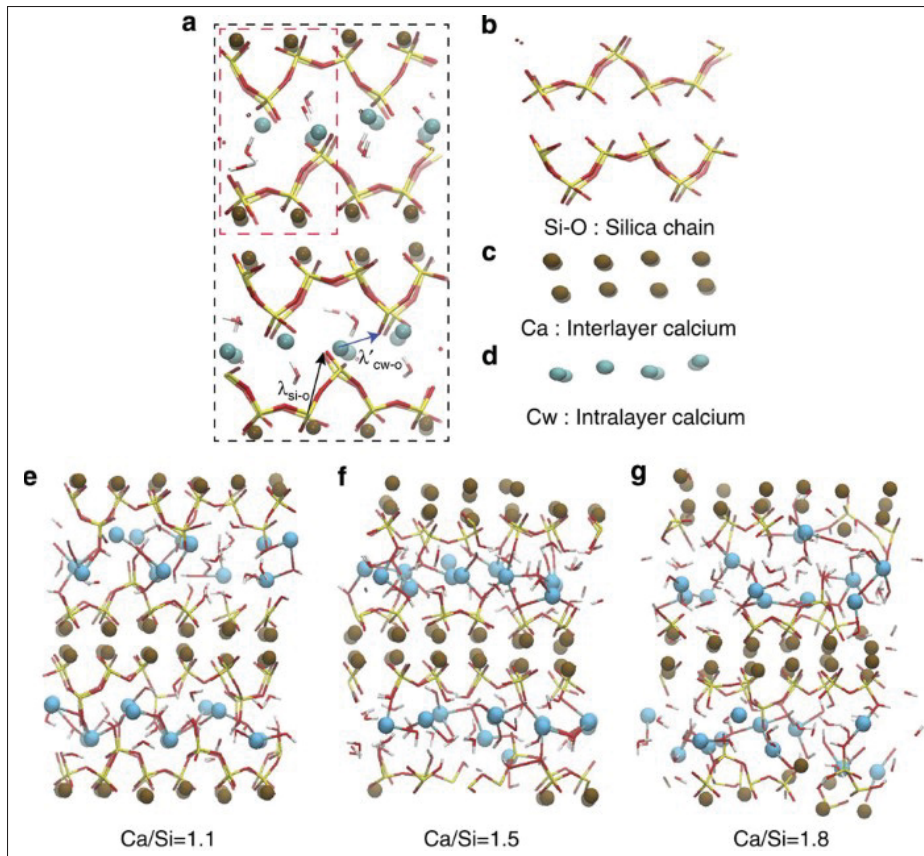


Figure 1.3 Structure moléculaire du C-S-H à l'échelle nanométrique, incluant l'effet du rapport C/S sur la structure moléculaire du C-S-H. Les chaînes de silice (Si-O) sont également indiquées

Tirée de Abdolhosseini Qomi *et al.* (2014, p. 3)

bonne capacité de sorption et une forte capacité d'échange cationique (Galamboš *et al.*, 2010; Murray, 2006; Christidis, 2011). La bentonite est principalement composée de montmorillonite, dont la cellule unitaire est présentée dans la figure ci-dessous 1.2b. La présence de silicium et de silice dans ces matériaux confère des caractéristiques spécifiques qui sont cruciales pour leur performance et leur utilité dans de nombreuses applications industrielles et environnementales.

Du point de vue cristallographique, le silicium cristallise dans la structure du diamant (Troler-McKinstry *et al.*, 2017). La structure du diamant (Figure 1.4a) est un type de structure cristalline très spécifique et importante en science des matériaux. La structure du diamant est également une

structure de type cubique à faces centrées, aussi appelée structure FCC (face-centered cubic). En revanche, la forme la plus commune de la silice, le quartz, cristallise dans le système de réseau trigonale (Lager *et al.*, 1982). Sa cellule unité est représentée dans la figure 1.4b ci-dessous.

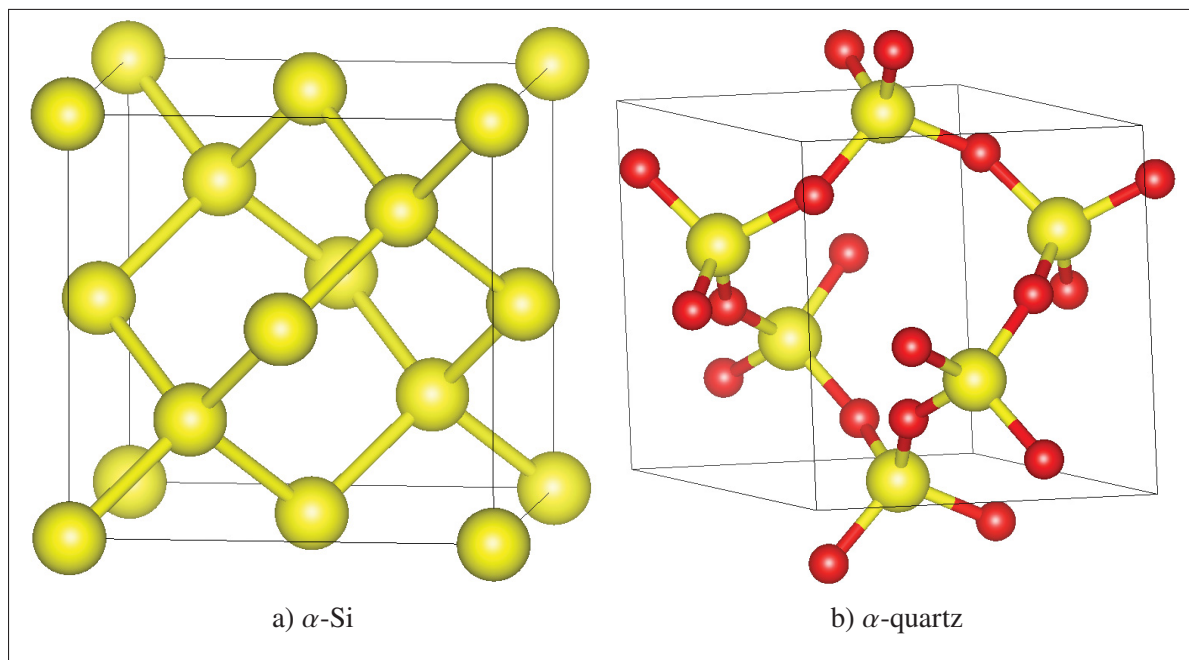


Figure 1.4 La cellule unité conventionnelle du silicium et de l' α -quartz

1.2 Processus de simulation numérique

1.2.1 Modélisation multi-échelle

La simulation numérique est un ensemble de calculs informatiques utilisés pour prédire, imiter ou suivre un événement ou un phénomène réel, principalement dans les sciences naturelles, y compris les sciences sociales. Ces calculs, souvent désignés sous le terme exécution du programme, nécessitent des équations mathématiques ou des algorithmes connus sous le nom de modèle. Le modèle est donc une condition préalable à la simulation numérique dans une grande variété de disciplines scientifiques, telles que la chimie, la physique et les sciences appliquées. Par exemple, nous ne simulons pas directement un matériau, mais plutôt un modèle de ce

matériau (LeSar, 2013; Ashby, 1992). En ce sens, la simulation numérique consiste à étudier la réponse d'un système modélisé face à des forces extérieures et à des contraintes (LeSar, 2013). Comme on peut le voir à la Figure 1.5, la construction du modèle est généralement basée sur la

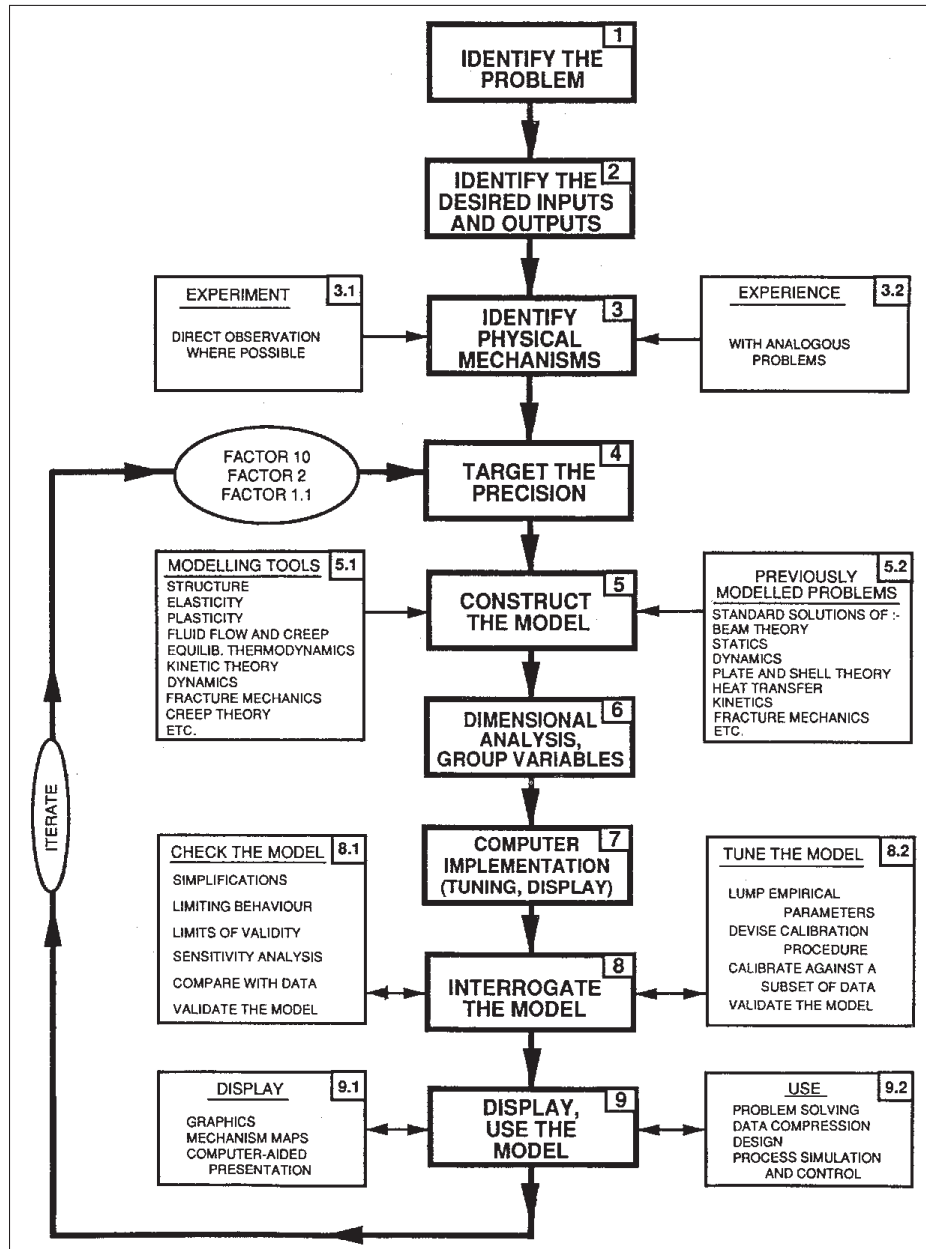


Figure 1.5 Les étapes du développement d'un modèle
Tirée de Ashby (1992, p. 2)

nature du système, ses composants, leur géométrie ainsi que les processus physico-chimiques,

principalement à une échelle de longueur et de temps donnée (LeSar, 2013). La construction du modèle peut également dépendre d'un type d'application spécifique. Par exemple, les ingénieurs étant principalement intéressés par le comportement macroscopique des solides, les modèles continus sont généralement utilisés, et les relations constitutives représentent les effets atomistiques dans les applications d'ingénierie (Wang *et al.*, 2018). En revanche, comme le soulignent les physiciens de l'état solide et les chimistes, ces derniers se concentrent davantage sur le comportement des matériaux à l'échelle électronique et atomique, en supposant que les processus correspondants à l'échelle macroscopique sont homogènes (Wang *et al.*, 2018).

À toute échelle, la qualité d'un modèle est déterminée par la précision de la description des interactions et l'échantillonnage des configurations du système (Tozzini, 2010). Les modèles conventionnels se basent généralement sur une échelle unique en termes de longueur et de temps (Wang *et al.*, 2018). Souvent, un modèle à une seule échelle est plus adapté à un système homogène avec des composants géométriques uniformes internes et où un processus unique a lieu. Cependant, une approche unitaire est limitée pour les systèmes complexes où un processus unique ou plusieurs processus impliquent une cascade d'événements multi-échelles (Tozzini, 2010). De plus, le système physique peut contenir des défauts, des singularités ou des hétérogénéités (Wang *et al.*, 2018). Par conséquent, la simulation numérique de tels systèmes nécessite une combinaison d'approches de modélisation, connue sous le nom de modélisation multi-échelle (Wang *et al.*, 2018; Tozzini, 2010; Ebrahimi, 2013).

La modélisation multi-échelle (Figure 1.6) est une méthode technique progressive qui permet de relier les informations, les phénomènes ainsi que les modèles uniques entre différentes échelles de systèmes complexes (Fermeglia, 2008). Cela permet d'étudier un système à différents niveaux avec des résolutions variées en termes de temps et de longueur, et de réaliser un passage d'échelle entre les modèles, facilitant ainsi la transmission d'informations d'une échelle plus petite à un modèle à une échelle plus large avec un degré de liberté interne réduit (Wang *et al.*, 2018; Ebrahimi, 2013; Fermeglia, 2008; Ebrahimi *et al.*, 2016; Lawrimore *et al.*, 2016). Par exemple, dans une étude (Ebrahimi *et al.*, 2016), l'agrégation des plaquettes d'argile et les propriétés mécaniques correspondantes ont été examinées avec succès à l'échelle mésoscopique

grâce au potentiel Gay-Berne (GB) calibré avec les interactions à l'échelle atomistique. La même approche de passage d'échelle a été utilisée pour étudier l'agrégation des nanotubes d'imogolite via le potentiel de force moyenne (PMF) (Zhu *et al.*, 2019). Le PMF, basé sur l'énergie libre, est un moyen de faire le lien entre les échelles de longueur et de temps dans la modélisation de systèmes complexes (Ioannidou *et al.*, 2017; Jeanmairet, 2014). Dans cette

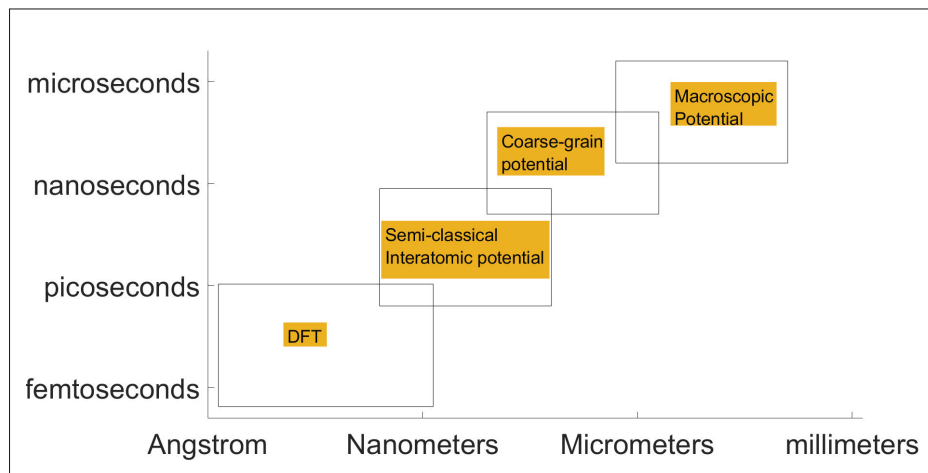


Figure 1.6 Modèle d'énergie à différentes échelles de longueur et de temps pour la simulation computationnelle

perspective, l'approche de modélisation multi-échelle ascendante est souvent utilisée pour aborder des problèmes à une échelle plus grande à partir d'une échelle fondamentale (Ebrahimi, 2013; Fermeglia, 2008; Lawrimore *et al.*, 2016). Les paramètres de champs de force non empiriques utilisés à l'échelle atomistique dans les simulations de dynamique moléculaire sont généralement déterminés à l'aide de modèles mécaniques quantiques, qui diffèrent par la manière dont l'interaction électron-électron est traitée (Tozzini, 2010). Ainsi, à cette échelle la plus basse (échelle quantique), le modèle DFT est largement adopté pour définir les champs de force pour les simulations de dynamique moléculaire (Cygan *et al.*, 2004; Xu *et al.*, 2018; Seppälä *et al.*, 2016; Greathouse *et al.*, 2015) et pour préparer des structures optimisées pour des simulations à différentes échelles (Tavares *et al.*, 2018).

En résumé, comme détaillé ci-dessus, tout système peut être modélisé à une échelle plus grande grâce au concept de transmission ou de liaison des informations pertinentes entre les

modèles, de l'échelle quantique à l'échelle atomistique, puis de l'échelle atomistique à l'échelle mésoscopique, et ainsi de suite jusqu'au niveau macroscopique ou de processus (Fermeglia, 2008). La description des interactions à une échelle donnée résulte de l'échantillonnage du système et des informations physico-chimiques à une échelle inférieure. Cela est en accord avec notre tâche principale, qui consiste à développer un modèle d'interaction pour le dioxyde de silicium et ses constituants principaux.

1.2.2 Simulations atomistiques et moléculaires

L'utilisation des ordinateurs pour les simulations atomistiques et moléculaires est désormais omniprésente dans presque tous les domaines de la recherche scientifique. Grâce aux avancées en matière de puissance de calcul et de logiciels efficaces, il est désormais possible d'explorer les matériaux à différents niveaux. De l'échelle électronique à l'analyse par éléments finis, en passant par l'échelle atomique et moléculaire, la modélisation computationnelle constitue un outil puissant pour prédire et extraire diverses propriétés dans un large éventail de disciplines scientifiques, telles que la physique, la chimie, la biologie et la science des matériaux. En plus de prédire les propriétés et de planifier des expériences, la modélisation est aujourd'hui une méthode alternative pour éviter des expériences coûteuses, dangereuses ou impossibles (Ma *et al.*, 2019; Tavares *et al.*, 2018; Paquet *et al.*, 2018). Par exemple, divers matériaux, y compris les minéraux argileux, sont étudiés à l'aide de techniques de modélisation computationnelle (Ma *et al.*, 2019; Tavares *et al.*, 2018).

Les principales méthodes de simulation numérique se classifient en deux grandes catégories : les méthodes dynamiques et les méthodes statiques (LeSar, 2013; McGreevy *et al.*, 1988). Les méthodes dynamiques suivent l'évolution du système atomique au fil du temps en réponse à des conditions externes, en utilisant les équations du mouvement de Newton (LeSar, 2013). Elles sont particulièrement utiles pour explorer les propriétés des matériaux et des molécules à l'échelle atomique, en permettant d'observer leur comportement dynamique et leurs interactions au cours du temps. En revanche, les méthodes statiques sont particulièrement utiles pour explorer la surface d'énergie potentielle d'un système, permettant ainsi de localiser les états de transition

et les configurations d'équilibre (McGreevy & Pusztai, 1988; Wooten *et al.*, 1985; Pandey *et al.*, 2016b; Barkema *et al.*, 1996). Ces méthodes sont essentielles pour identifier les minima locaux de l'énergie, qui correspondent souvent aux états stables ou métastables du système. Elles permettent également d'analyser les barrières énergétiques entre différents états, fournissant des informations cruciales sur les mécanismes de transition et les processus de diffusion.

1.2.2.1 Simulation par dynamique moléculaire

La dynamique moléculaire est une méthode computationnelle qui permet de calculer l'évolution temporelle du comportement d'un système atomique donné. Les propriétés à l'équilibre ainsi que les propriétés dépendant du temps sont calculées à partir des mouvements des atomes du système considéré (LeSar, 2013; Becquart *et al.*, 2016; Hansson *et al.*, 2002; Allen *et al.*, 2004). En général, on impose des conditions externes et des contraintes au système étudié, puis on suit son évolution au cours du temps à l'aide des équations du mouvement de Newton.

Pour illustrer le concept de simulation par dynamique moléculaire, considérons un système composé de N atomes. Soit $\vec{r}_i$ et $\vec{v}_i$ la position et la vitesse de la i -ème particule à un instant donné. La seconde loi de Newton permet d'écrire l'équation du mouvement d'une particule comme suit (LeSar, 2013) :

$$m_i \frac{d^2 \vec{r}_i}{dt^2} = \vec{f}_i = m_i \vec{a}_i \quad (1.1)$$

Dans cette équation, m_i représente la masse de la particule i , $\vec{a}_i$ est son accélération, et $\vec{f}_i$ correspond à la somme de toutes les forces agissant sur la particule i . Sur la base de la somme des forces, l'équation 1.1 peut être réécrite comme suit :

$$m_i \frac{d^2 \vec{r}_i}{dt^2} = \vec{f}_i = m_i \vec{a}_i = \sum_{j \neq i} \vec{f}_{ij}(\vec{r}_{ij}) \quad (1.2)$$

où $\vec{f}_{ij}$ est la force d'interaction entre les atomes i et j . En particulier, la forme de l'équation 1.2 est propre à la modélisation à l'aide de potentiels interatomiques dans le cadre de la dynamique moléculaire classique. Étant donné qu'un système atomique est régi par sa surface d'énergie potentielle, il est possible de déterminer les forces $\vec{f}_{ij}$ entre les atomes ainsi que d'autres propriétés du système, à condition que les coordonnées cartésiennes des atomes soient connues. La force peut être déterminée en prenant le gradient négatif de l'énergie potentielle (LeSar, 2013; Allen *et al.*, 2004).

$$\vec{f}_i = -\Delta_i U(r_1, r_2, \dots, r_N) = -\Delta_i U(r^N) \quad (1.3)$$

Le gradient est évalué en fonction des coordonnées atomiques de l'atome i considéré.

$$\Delta_i U(r^N) = \frac{\partial U(r^N)}{\partial x_i} \vec{u}_x + \frac{\partial U(r^N)}{\partial y_i} \vec{u}_y + \frac{\partial U(r^N)}{\partial z_i} \vec{u}_z \quad (1.4)$$

Le potentiel $U(r^N)$ peut être obtenu soit à partir des premiers principes de la mécanique quantique, soit en utilisant un potentiel semi-empirique. Selon la méthode utilisée pour déterminer les forces d'interaction, on distingue la dynamique moléculaire *ab initio* de la dynamique moléculaire classique (LeSar, 2013; Becquart & Perez, 2016; Hansson *et al.*, 2002; Allen *et al.*, 2004). La dynamique moléculaire classique repose sur des équations de mouvement classiques, telles que celles de Newton, pour simuler le comportement des molécules. Elle utilise des potentiels empiriques ou semi-empiriques pour modéliser les interactions entre les particules. Bien que moins précise que l'approche *ab initio*, sa rapidité permet de simuler des systèmes plus vastes sur des périodes allant jusqu'à la microseconde. En revanche, la dynamique moléculaire *ab initio* utilise également les équations du mouvement classique, mais l'énergie et les forces agissant sur les atomes sont calculées à l'aide de la mécanique quantique. Basée sur des méthodes de calcul de la structure électronique, comme la théorie de la fonctionnelle de la densité (DFT) ou la méthode de Hartree-Fock (Fischer, 1977; Slater, 1963), elle offre des résultats très précis en tenant compte des effets quantiques. Cependant, en intégrant les orbitales électroniques dans ses

calculs, elle se révèle coûteuse et est limitée à de petits systèmes atomiques et à des échelles de temps très courtes, de l'ordre de la picoseconde.

1.2.2.2 Intégration des équations du mouvement

Le principe clé est de résoudre correctement et efficacement l'équation différentielle (1.2). Pour un système composé de N atomes, nous avons à résoudre $3N$ équations couplées, pour lesquelles il n'existe pas de solution analytique continue en fonction du temps (LeSar, 2013; Allen *et al.*, 2004). Il est donc nécessaire de discrétiser le temps en intervalles discrets. L'intégration des équations du mouvement devient ainsi plus simple à réaliser numériquement. On choisit un intervalle de temps, également appelé pas de temps (time step), et on approxime les équations différentielles par des équations aux différences finies (LeSar, 2013; Becquart & Perez, 2016; Hansson *et al.*, 2002; Allen *et al.*, 2004). La résolution de ces équations aux différences finies de manière computationnelle nécessite un algorithme capable de s'exécuter rapidement, avec un pas d'intégration δt optimal pour garantir la conservation de l'énergie totale du système. L'algorithme d'intégration des équations du mouvement le plus couramment utilisé est celui de Verlet (Verlet, 1967). En connaissant les informations à l'instant t , cet algorithme permet d'évaluer les quantités nécessaires à l'instant $t + \delta t$.

$$r_i(t + \delta t) = r_i(t) + v_i(t)\delta t + \frac{1}{2m}f_i(t)\delta t^2 \quad (1.5)$$

$$v_i(t + \delta t) = v_i(t) + \frac{1}{2m}(f_i(t) + f_i(t + \delta t))\delta t \quad (1.6)$$

Par exemple, dans l'équation (1.5), les positions des atomes à l'instant $t + \delta t$ peuvent être déterminées en connaissant celles à l'instant t . Ensuite, les forces sont calculées à partir du potentiel d'interaction en utilisant les positions $r_i(t + \delta t)$. Enfin, les vitesses $v_i(t + \delta t)$, décrites par l'équation (1.6), sont déterminées. En pratique, la simulation commence avec les positions atomiques connues, c'est-à-dire les coordonnées cartésiennes des atomes $r_i(0)$. Les vitesses

initiales à $t = 0$ sont généralement générées selon une distribution gaussienne ou uniforme. Ainsi, le calcul est répété en suivant l'algorithme jusqu'à ce que la simulation soit terminée. La connaissance des vitesses à l'instant t permet également de déterminer la température instantanée du système.

$$T = \frac{1}{3(N-1)k_B} \sum_{i=1} m_i |\vec{v}_i|^2 \quad (1.7)$$

Dans l'équation (1.7), le k_B est la constante de Boltzmann, et $N-1$ tient compte de la conservation de l'impulsion, souvent appelée "momentum conservation"; T représente la température du système, m_i désigne la masse de la particule i , et v_i correspond à la vitesse de cette même particule au sein du système atomique.

En résumé, la dynamique moléculaire est un outil puissant pour étudier et prédire le comportement des matériaux sous diverses contraintes. Elle nécessite un potentiel d'interaction adapté au système atomique considéré. Un des principaux défis est le choix du pas de temps pour l'intégration des équations du mouvement, qui doit être ajusté en fonction des processus atomiques dans les solides. Par exemple, les vibrations atomiques se déroulent généralement sur une échelle de temps de quelques fractions de picoseconde, ce qui impose un pas de temps de l'ordre de la femtoseconde (Dove, 2008).

1.2.2.3 Ensembles thermodynamiques

Pour suivre le comportement d'un système à l'aide de la dynamique moléculaire, il est nécessaire de définir un ensemble thermodynamique approprié. Un ensemble thermodynamique est un modèle conceptuel utilisé pour décrire un système physique en fonction de certaines variables et conditions (Jaynes, 1957; Lustig, 1994; Hünenberger, 2005). Chaque ensemble représente une manière différente de caractériser les systèmes et leurs interactions. Nous pouvons citer parmi les ensembles les plus utilisés dans les simulations numériques :

- Ensemble microcanonique (NVE - nombre de particules, volume et énergie constant) : Dans cet ensemble, le système est isolé, avec un volume constant et une énergie totale fixe, et il n'y a aucun échange d'énergie ni de particules avec l'extérieur (Hünenberger, 2005).
- Ensemble canonique (NVT - Nombre de particules, Volume et Température constant) : Ici, le système échange de l'énergie avec un réservoir thermique tout en maintenant une température constante, un volume fixe, et un nombre de particules constant (Hünenberger, 2005). Dans ce modèle, l'équation du mouvement est modifiée par l'introduction d'un thermostat, qui permet de maintenir la température du système à une valeur constante en ajustant les vitesses des particules ou en simulant des échanges d'énergie avec un réservoir thermique (Hünenberger, 2005). Il existe trois grandes catégories de thermostats couramment utilisés dans les simulations moléculaires, dont l'ajustement de vitesse (Velocity Rescaling) (Bussi *et al.*, 2007), le thermostat d'Andersen (Weinan *et al.*, 2008) et le thermostat de Nosé-Hoover (Evans *et al.*, 1985). Dans le thermostat d'ajustement de vitesse, les vitesses des particules sont ajustées à chaque étape de la simulation pour maintenir la température du système à une valeur constante. Bien que cette méthode soit simple à mettre en œuvre, elle peut introduire des artefacts dans les propriétés dynamiques du système. Le thermostat d'Andersen introduit des collisions aléatoires entre les particules du système et des particules fictives à une température donnée. En simulant des échanges énergétiques avec un réservoir thermique, il permet de maintenir la température constante tout en préservant les propriétés dynamiques du système. Le thermostat de Nosé-Hoover est conçu pour réguler la température d'un système en introduisant des variables supplémentaires qui permettent de contrôler l'énergie cinétique moyenne des particules. Il modifie le mouvement des particules de manière à ajuster la température du système vers une valeur cible. Les variable supplémentaire introduites sont souvent appelée variable de contrôle ou variable de relaxation (ξ), qui est liée à un oscillateur harmonique. Cette variable contrôle le « couplage » entre le système principal et un réservoir thermique fictif.

$$m_i \vec{a}_i = -\xi \frac{d\vec{r}_i}{dt} - \frac{\partial U(r^N)}{\partial r_i} \quad (1.8)$$

Où les termes sont les mêmes que ceux définis précédemment.

- Ensemble Isobarique-Isotherme (NPT - nombre de particules, pression et température constant) : Le modèle NPT est celui dans lequel le nombre de particules, la température et la pression sont maintenus constants (Hünenberger, 2005). Ce modèle est particulièrement adapté pour étudier les systèmes soumis à des contraintes fixes et reflète de manière plus précise les conditions expérimentales. Ici, on a employé souvent un thermostat de type Nose-Hoover pour réguler la température et le barostat pour contrôler la pression. (Ke *et al.*, 2022; Berendsen *et al.*, 1984).

1.2.2.4 Simulation par les méthodes statistiques

Les méthodes statiques de simulation reposent souvent sur les techniques de Monte Carlo pour étudier des systèmes atomiques ou des processus physico-chimiques se déroulant dans ces systèmes (Hayoun, 2002; Landau *et al.*, 2021; McGreevy & Pusztai, 1988; McGreevy *et al.*, 1992; Gereben *et al.*, 2007; Wooten *et al.*, 1985). Contrairement aux méthodes dynamiques, qui suivent l'évolution temporelle des systèmes, les méthodes statiques se concentrent sur l'exploration de l'espace des configurations du système à l'aide d'échantillonnages aléatoires. Ces méthodes permettent d'analyser des propriétés thermodynamiques et d'équilibre en générant et en évaluant diverses configurations possibles sans tenir compte de la dynamique temporelle des particules. La plupart des méthodes statiques reposent sur la méthode de Métropolis qui fonctionne sur des probabilités (Metropolis *et al.*, 1953; Hayoun, 2002). On commence avec une configuration initiale du système. Cette configuration peut être générée aléatoirement ou choisie en fonction des conditions de départ spécifiques. Un petit déplacement ou une modification est proposé pour une partie du système. Par exemple, on pourrait déplacer une particule, changer une coordonnée, ou ajuster une variable dans le système. On calcule l'énergie du système avant et après le déplacement proposé. Cela permet d'évaluer le changement d'énergie ΔE associé à la modification. Le déplacement est accepté ou rejeté en fonction d'un critère probabiliste. Si l'énergie du système diminue ($\Delta E < 0$), le déplacement est toujours accepté. Si l'énergie

du système augmente ($\Delta E > 0$), le déplacement est accepté avec la probabilité (P) suivante (McGreevy & Howe, 1992; Vink, 2002; Hayoun, 2002; N'Tsouaglo, 2015) :

$$P = \exp\left(\frac{-\Delta E}{k_B T}\right) \quad (1.9)$$

Où ΔE est la différence d'énergie entre la configuration initiale et la configuration finale, T est la température, et k_B est la constante de Boltzmann. On répète ce processus un grand nombre de fois pour explorer l'espace des configurations et obtenir des échantillons statistiques représentatifs du système. Les configurations acceptées sont utilisées pour calculer les propriétés thermodynamiques et les autres propriétés d'intérêt du système. Beaucoup de méthodes statiques reposent sur des principes de fonctionnement régis par l'équation (1.8) et nécessitent uniquement un potentiel d'interaction. On peut citer, parmi ces méthodes, la méthode Monte Carlo Metropolis (MMC) (Metropolis *et al.*, 1953) standard, l'algorithme de commutation des liaisons de Wooten-Winer-Weaire (WWW) (Wooten *et al.*, 1985), la technique d'activation et relaxation *nouveau* (ARTn) (Barkema *et al.*, 2000; Mousseau *et al.*, 2012). La méthode WWW est efficace mais est limitée à un seul type de potentiel d'interaction, celui de Keating. En revanche, l'ARTn peut être couplée à n'importe quel type de potentiel, y compris les potentiels *ab initio*. ARTn est une méthode de recherche des états de transitions développée par Normand Mousseau et ses collaborateurs à l'Université de Montréal. Techniquement, la recherche du chemin de transition se fait en utilisant la méthode des vecteurs propres, combinée avec l'algorithme de Lanczos (Barkema & Mousseau, 2000; Mousseau *et al.*, 2012; Jay *et al.*, 2022). La technique d'activation-relaxation se déroule en trois étapes principales : l'activation, la convergence vers le point selle, et enfin la relaxation dans un nouveau minimum. Ces trois étapes forment ce que l'on appelle un événement, constitué d'un ensemble de trois configurations : la configuration du minimum initial, celle de l'état de transition, et enfin la configuration du minimum final. Une autre méthode populaire est la méthode Monte Carlo inverse (RMC). Il s'agit d'une variante des techniques traditionnelles de la méthode Monte Carlo de Metropolis (MMC) standard (Metropolis *et al.*, 1953; McGreevy & Pusztai, 1988; McGreevy & Howe, 1992; Gereben *et al.*, 2007). Les méthodes de Monte Carlo inverse (RMC) reposent fondamentalement sur des données

expérimentales, utilisant des informations telles que les facteurs de structure et les fonctions de corrélation de paires (Keen, 2001). Cette technique inverse les données expérimentales sans tenir compte des interactions atomiques, permettant ainsi une approche purement basée sur les données (McGreevy & Pusztai, 1988; McGreevy & Howe, 1992; Gereben *et al.*, 2007). Aucun potentiel d'interaction n'est requis pour la méthode Monte Carlo inverse (RMC). Le processus débute avec une configuration initiale du système, qui peut être créée de manière aléatoire ou sélectionnée selon des conditions de départ spécifiques. Au départ, la fonction de distribution radiale est calculée, puis transformée en un facteur de structure (Evrard *et al.*, 2005; McGreevy & Howe, 1992). Ensuite, la différence χ_0^2 entre le facteur de structure total calculé de la configuration initiale et le facteur de structure total mesuré expérimentalement est déterminée. Par la suite, un atome est déplacé au hasard, et une nouvelle fonction de distribution radiale ainsi qu'un nouveau facteur de structure total sont calculés. La nouvelle différence χ_{new}^2 entre le facteur de structure total calculé et le facteur de structure expérimental est ensuite déterminée. Si $\chi_n^2 < \chi_0^2$, le déplacement est accepté. Si $\chi_n^2 > \chi_0^2$, le déplacement est accepté avec une probabilité de :

$$P = \exp \left[-\frac{(\chi_n^2 - \chi_0^2)}{2} \right] \quad (1.10)$$

Sinon, le déplacement est rejeté (McGreevy & Pusztai, 1988; Gereben *et al.*, 2007; Evrard & Pusztai, 2005). Ce processus est répété jusqu'à ce qu'une configuration d'énergie minimale acceptable soit atteinte. Il est important de noter que les configurations générées ne sont pas uniques (McGreevy & Pusztai, 1988; McGreevy & Howe, 1992). La méthode RMC est conçue pour produire des configurations conformes à une fonction de distribution radiale spécifiée $g(r)$. Notamment, cette méthode génère une variété de configurations distinctes, toutes cohérentes avec la même $g(r)$. Pour résoudre ce problème, des méthodes hybrides ont été développées. En plus des données expérimentales, elles utilisent un potentiel d'interaction classique ou *ab initio* comme contrainte additionnelle. La méthode FEAR (Force-Enhanced Atomic Refinement) (Pandey *et al.*, 2015) ainsi que d'autres méthodes (Borisenko *et al.*, 2012; Biswas *et al.*, 2004; Limbu *et al.*, 2018) sont des exemples de ce type hybride.

1.2.3 Conditions aux limites périodiques

En pratique, les systèmes simulés sont placés dans des boîtes de simulation pour limiter leur taille et mieux gérer les contraintes computationnelles (Senac, 2018; Makov *et al.*, 1995). Dans la réalité, les atomes ne sont pas toujours distribués périodiquement comme dans un cristal ; il est donc nécessaire de définir une boîte de simulation. Les conditions physico-chimiques des atomes ou molécules situés aux bords de la boîte diffèrent de celles observées à l'intérieur, ce qui entraîne un effet de bord et pose un problème pour la simulation. Pour contourner ce problème et simuler un système quasi-infini à partir d'une boîte de dimensions finies, on applique des conditions aux limites périodiques (CLP, en anglais "Periodic Boundary Conditions", PBC). Les boîtes de simulation sont reproduites dans toutes les directions, de sorte que les bords se connectent directement à leurs bords opposés. Par exemple, comme le montre la Figure 1.7, une particule (en vert) avec une quantité de mouvement sort de la boîte rouge. Grâce à la périodicité du système, une particule identique entre dans la boîte par la face opposée avec la même quantité de mouvement. La taille de la boîte de simulation doit être choisie en fonction des interactions au sein du système considéré (Senac, 2018; Makov & Payne, 1995). Les dimensions de la boîte doivent être suffisamment grandes pour que la distance entre une particule et son image dans les boîtes voisines dépasse la distance à partir de laquelle les interactions non liées sont significatives (Senac, 2018; Makov & Payne, 1995). Cela permet d'éviter que les particules interagissent avec leurs images dans les boîtes voisines. Dans la pratique, il est courant d'effectuer un test de convergence des propriétés désirées en fonction de la taille de la boîte de simulation.

1.2.4 Méthodes *ab initio*

Les calculs en mécanique quantique reposent sur l'équation de Schrödinger. Ils estiment l'énergie totale d'un système donné en résolvant cette équation à l'aide de diverses approximations, parmi lesquelles on trouve la méthode Hartree-Fock (HF), la théorie de perturbation de Møller-Plesset (MP), l'interaction de configuration (CI), et la théorie de la fonctionnelle de la densité (DFT), entre autres (Tavares *et al.*, 2018). Parmi ces approximations, la DFT est particulièrement populaire et largement utilisée en science des matériaux. Cette approche utile découle de

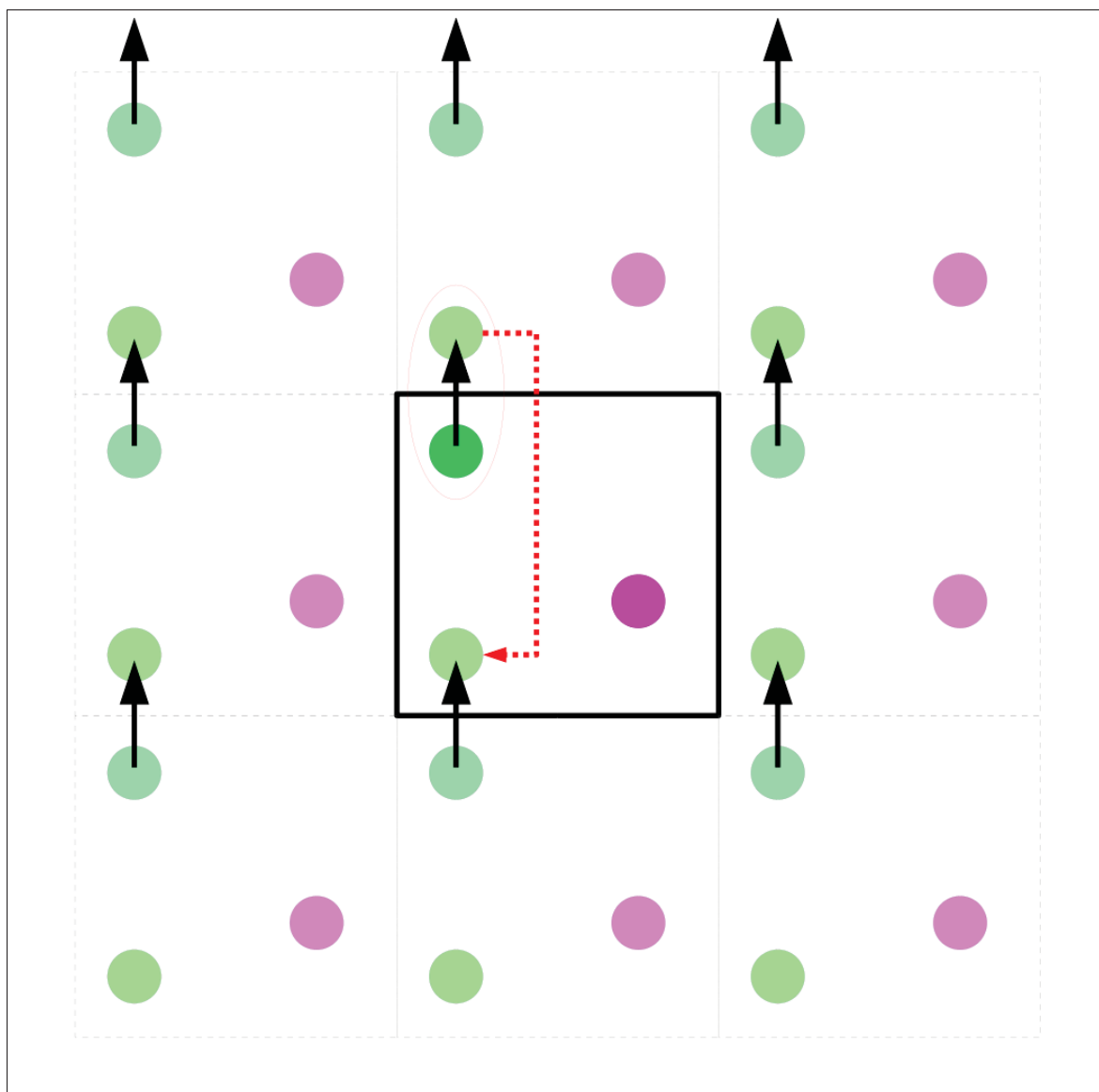


Figure 1.7 Boîte de simulation et fonctionnement des conditions aux limites périodiques
Tirée de Wikipedia contributors (2024, p. 1)

plusieurs approximations dans la résolution de l'équation de Schrödinger, qui est un problème à nombreux corps (Sholl & Steckel, 2022). La première approximation, connue sous le nom d'approximation de Born-Oppenheimer, repose sur le fait que les électrons sont beaucoup plus légers que les noyaux, ce qui les rend beaucoup plus réactifs aux changements que les noyaux (Pisana *et al.*, 2007; Sholl & Steckel, 2022). En considérant cette approximation, l'équation de

Schrödinger traite du mouvement des électrons tandis que les noyaux sont considérés comme fixes.

$$\left[\frac{\hbar^2}{2m} \sum_{i=1}^N \nabla_i^2 + \sum_{i=1}^N V(r_i) + \sum_{i=1}^N \sum_{j<i}^N U(r_i, r_j) \right] \psi = E\psi \quad (1.11)$$

- **Premier terme** : $-\frac{\hbar^2}{2m} \sum_{i=1}^N \nabla_i^2$: énergie cinétique des N particules, où m est la masse d'une particule, ∇^2 est l'opérateur Laplacien et $\hbar$ est la constante de Planck réduite.
- **Deuxième terme** : $\sum_{i=1}^N V(\mathbf{r}_i)$: potentiel externe agissant sur chaque particule i à la position $\mathbf{r}_i$.
- **Troisième terme** : $\sum_{i=1}^N \sum_{j<i}^N U(\mathbf{r}_i, \mathbf{r}_j)$: interaction entre les particules i et j où U est le potentiel d'interaction.
- ψ : fonction d'onde du système, qui contient toutes les informations quantiques.
- E : énergie totale du système.

Tous les termes entre crochets représentent l'Hamiltonien, qui comprend l'énergie cinétique des électrons, l'énergie d'interaction électron-noyau et l'énergie d'interaction électron-électron. ψ est la fonction d'onde électronique qui donne la probabilité de trouver les électrons à un ensemble particulier de coordonnées. En réalité, dans ce système complexe, la quantité d'intérêt physique mesurable est la probabilité (Sholl & Steckel, 2022; Schommers, 2019). Puisque ψ peut être approximée par le produit des fonctions d'onde individuelles des électrons ψ_i , la densité des électrons $n(\vec{r})$ à une coordonnée spatiale donnée peut s'écrire comme suit (Car *et al.*, 1985a; Sholl & Steckel, 2022; Schommers, 2019) :

$$n(\vec{r}) = 2 \sum_i \psi_i^*(\vec{r}) \psi_i(\vec{r}) \quad (1.12)$$

Dans cette équation, le symbole mathématique $*$ désigne le conjugué complexe. Le concept de densité électronique permet de réduire considérablement le nombre de variables dans l'équation 1.12. Alors que la fonction d'onde complète pour un système à N électrons dépend de $3N$ coordonnées spatiales, la fonction de densité électronique se caractérise par une dépendance uniquement en 3 coordonnées spatiales, comme le montre l'équation. La densité électronique,

telle qu'exprimée dans l'équation 1.12, est d'un intérêt crucial pour simplifier l'équation de Schrödinger complexe, car le premier théorème formulé par Kohn et Hohenberg repose sur cette quantité. Le premier théorème stipule que l'énergie de l'état fondamental dans l'équation 1.11 est une fonction unique de la densité électronique. En conséquence, ce théorème implique une correspondance biunivoque entre la fonction d'onde de l'état fondamental et la densité électronique de l'état fondamental (Sholl & Steckel, 2022; Schommers, 2019). Étant donné que l'état fondamental correspond à l'état d'énergie le plus bas des électrons, on sait d'un point de vue mathématique que la fonctionnelle globale doit posséder certaines propriétés importantes. Comme rapporté dans (Sholl & Steckel, 2022; Schommers, 2019), le deuxième théorème fournit une propriété importante de la fonctionnelle, car le premier théorème, bien qu'il prouve l'existence d'une fonctionnelle pouvant être utilisée pour résoudre l'équation de Schrödinger, ne fournit pas de détails mathématiques sur la fonctionnelle. Ainsi, le deuxième théorème de Kohn et Hohenberg stipule que la densité électronique qui minimise l'énergie de la fonctionnelle globale est la véritable densité électronique correspondant à la solution complète de l'équation de Schrödinger (Sholl & Steckel, 2022; Schommers, 2019). La forme mathématique exacte de la fonctionnelle est inconnue, c'est pourquoi le principe variationnel est utilisé avec des formes fonctionnelles approximatives dans l'équation mono-électronique de Kohn et Sham, dont la résolution repose sur l'approximation de Born-Oppenheimer et les théorèmes de Kohn et Hohenberg mentionnés ci-dessus. L'équation mono-électronique de Kohn et Sham est donnée comme suit (Mohapatra *et al.*, 2018; Karakasidis *et al.*, 2007; Sholl & Steckel, 2022; Schommers, 2019) :

$$\left[\frac{\hbar^2}{2m} \nabla^2 + V(r) + V_H(r) + V_{XC}(r) \right] \psi_i(r) = \varepsilon_i \psi_i \quad (1.13)$$

- **Premier terme :** $-\frac{\hbar^2}{2m} \nabla^2$: énergie cinétique de l'électron dont les termes sont identiques à ceux de l'équation 1.11.
- **Deuxième terme :** $V(\mathbf{r})$: potentiel externe agissant sur l'électron, tel que celui dû à un noyau.
- **Troisième terme :** $V_H(\mathbf{r})$: potentiel de Hartree, représentant l'interaction électrostatique entre l'électron et la densité électronique.

- **Quatrième terme : $V_{XC}(\mathbf{r})$** : Potentiel d'échange-corrélation, qui intègre les effets quantiques ainsi que les interactions d'échange et de corrélation entre les électrons. L'échange découle de la contrainte d'antisymétrie imposée à la fonction d'onde électronique, conformément au principe d'exclusion de Pauli, et se traduit par une interaction répulsive entre les électrons. La corrélation, quant à elle, résulte de la nature non séparable de la fonction d'onde à plusieurs corps. Le potentiel de corrélation vient corriger l'approximation de séparabilité et induit une interaction attractive entre les électrons.
- $\psi_i(\mathbf{r})$: fonction d'onde de l'électron i , décrivant son état quantique.
- ϵ_i : énergie associée à l'état quantique $\psi_i(\mathbf{r})$.

Dans l'équation 1.13, $V(r)$ représente l'énergie potentielle d'interaction de l'électron avec l'ensemble des noyaux. $V_H(r)$ est le potentiel de Hartree qui décrit la répulsion coulombienne entre l'électron considéré et la densité électronique totale. Comme le potentiel de Hartree inclut une auto-interaction, ce qui est non physique, un troisième terme, noté $V_{XC}(r)$, est introduit à des fins de correction. Ainsi, $V_{XC}(r)$ définit la contribution d'échange et de corrélation dans l'équation de l'électron unique (Karakasidis & Charitidis, 2007; Sholl & Steckel, 2022). En raison de plusieurs fonctions inconnues, le calcul se fait généralement en supposant une densité électronique initiale. La figure 1.8 présente les détails du processus de calcul (Mohapatra *et al.*, 2018; Sholl & Steckel, 2022).

On peut citer, par exemple, l'approximation de densité locale (LDA) classique Ziesche *et al.* (1998); Perdew *et al.* (2001), le modèle Perdew, Burke et Ernzerhof (PBE) (Perdew *et al.*, 1996; Perdew & Schmidt, 2001), ainsi que la fonctionnelle Strongly Constrained and Appropriately Normed (SCAN) (Sun *et al.*, 2015) comme exemples de fonctionnelles d'échange-corrélation (XC : *Exchange-correlation*). En général, il existe des fonctionnelles XC semi-empiriques et non-empiriques. Lorsque les fonctionnelles XC dérivent uniquement des contraintes découlant des premiers principes de la mécanique quantique, la DFT est considérée comme appartenant aux méthodes *ab initio*. Cependant, de manière générale, les fonctionnelles XC nécessitent des paramètres ajustables, ce qui conduit généralement à classer la DFT parmi les méthodes *ab*

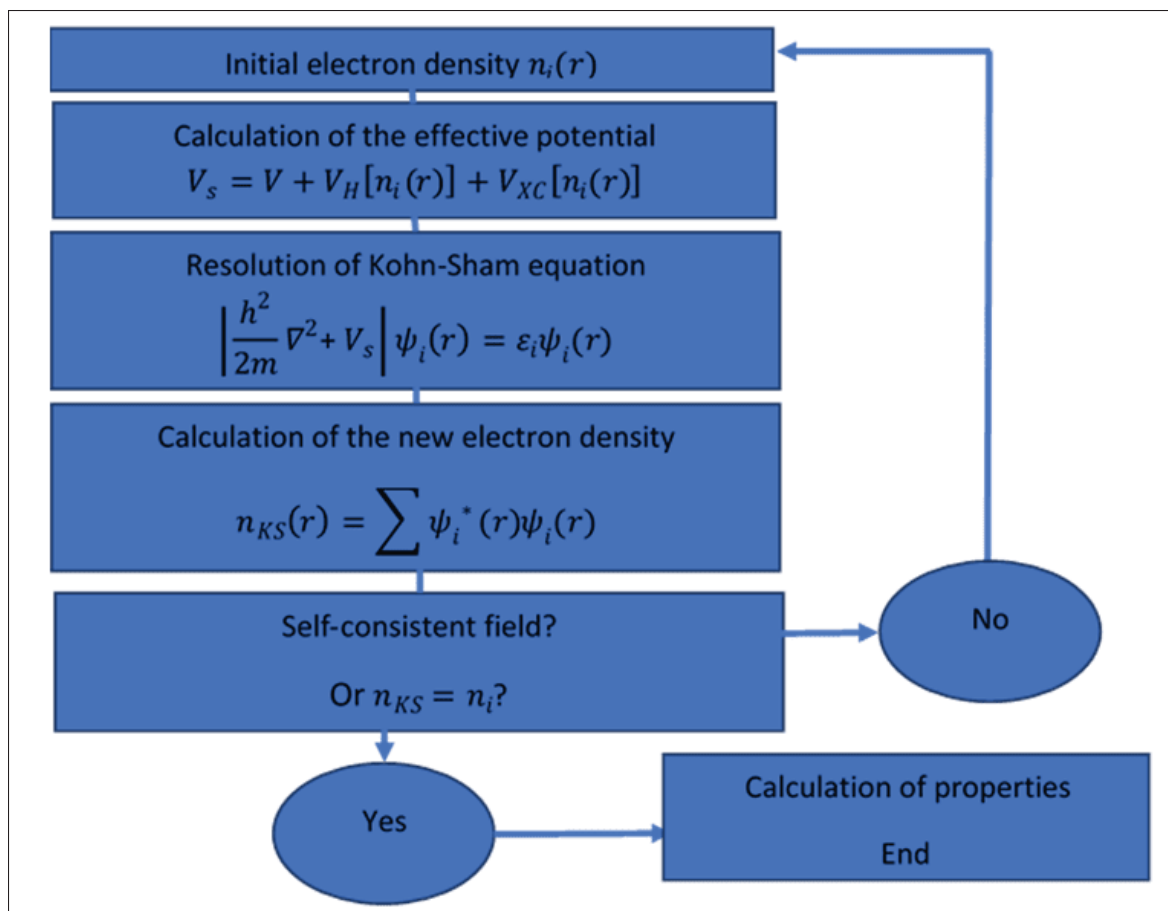


Figure 1.8 Schéma de calcul de l'équation de Kohn-Sham pour la détermination de la densité électronique

initio, même si les fonctionnelles elles-mêmes ne sont pas strictement *ab initio*, mais plutôt semi-empiriques.

La DFT a joué un rôle clé et continue d'être employée dans le développement du champ de force semi-empirique. Dans nos sous-objectifs, les calculs DFT seront utilisés pour construire une base de données *ab initio* afin d'ajuster le modèle de potentiel de tenseur de moment et pour effectuer d'autres calculs à des fins de validation. Malgré sa précision, la simulation par ordinateur basée sur des calculs en mécanique quantique tels que la DFT est coûteuse et difficilement améliorable (Shapeev, 2016; Xu *et al.*, 2018; Gubaev *et al.*, 2019; Zuo *et al.*, 2020); en particulier pour des systèmes complexes tels que les minéraux d'argile, certains calculs peuvent être impraticables.

Cependant, la description classique peut être utilisée pour approcher de près l'énergie totale au lieu de résoudre l'équation de Schrödinger à plusieurs corps. Les structures et propriétés des matériaux sont reproductibles avec des approximations classiques de l'énergie totale de tout système sous l'équation de Schrödinger. Cette approximation provient de l'examen des divers modes et façons dont les atomes peuvent interagir selon l'équation de Schrödinger. La description classique repose sur un champ de force ou une fonction d'énergie potentielle pour approximer la surface d'énergie de l'état fondamental en fonction des positions des noyaux. Cette approche alternative sera détaillée dans la section suivante.

1.2.5 Champ de force classique

Alors que la mécanique quantique traite des électrons, le principe des champ de force reposent sur les interactions atomiques (Tavares *et al.*, 2018; Cygan *et al.*, 2004). Contrairement à la description en mécanique quantique où le mouvement des électrons est crucial, les méthodes de champ de force se concentrent sur les mouvements des noyaux. En effet, pour un ensemble d'atomes ou de systèmes moléculaires, divers modes d'interaction tels que l'allongement des liaisons, la déformation des angles, l'interaction de Van der Waals, l'interaction coulombienne et d'autres interactions sont attendus. Par conséquent, l'énergie associée à chaque mode d'interaction est exprimée par des fonctions analytiques (Cygan *et al.*, 1998; Senac, 2018). Ainsi, l'énergie totale du système est décrite par l'ensemble de ces fonctions analytiques. Le champ de force est donc un ensemble de fonctions analytiques qui prennent en compte l'ensemble des positions atomiques et renvoient l'énergie totale due aux interactions interatomiques dans le système. L'objectif du champ de force est de décrire classiquement tous les faits de la mécanique quantique en partitionnant l'énergie électronique totale en contributions séparées d'atomes-atomes (interactions) (González, 2011). En réalité, les fonctions de champ de force sont généralement paramétrées de sorte que les paramètres puissent reproduire les propriétés physiques, chimiques et électroniques ainsi que d'autres propriétés expérimentales des matériaux (N'Tsouaglo, 2015). Ainsi, comme rapporté dans (Tavares *et al.*, 2018; Cygan *et al.*, 2004; Dufresne *et al.*, 2018), les approches de champ de force calculent l'énergie totale de tout système

atomique ou moléculaire donné en utilisant des fonctions paramétrées. Ces fonctions paramétrées sont adaptées au type d'interaction atomique. Les interactions liées, également connues sous le nom d'interactions covalentes, sont principalement décrites par des fonctions harmoniques (Tavares *et al.*, 2018; Cygan *et al.*, 2004). Les interactions non liées ou non covalentes sont généralement régies par des fonctions de potentiel de Lennard-Jones et coulombiennes (Tavares *et al.*, 2018; Cygan *et al.*, 1998; Dufresne *et al.*, 2018). L'énergie totale peut être écrite comme suit (Senac, 2018; Cygan *et al.*, 1998; Dufresne *et al.*, 2018) :

$$E_{TOT} = E_{\text{lié}} + E_{\text{non-lié}} \quad (1.14)$$

$$E_{\text{lié}} = E_{\text{étirement-liaison}} + E_{\text{angle - coudée}} + E_{\text{dièdre}} \quad (1.15)$$

$$E_{\text{non-lié}} = E_{\text{vdw}} + E_{\text{Coulomb}} \quad (1.16)$$

L'énergie totale E_{TOT} du système est la somme de l'énergie des interactions liées $E_{\text{lié}}$ et de l'énergie des interactions non liées $E_{\text{non-lié}}$, où $E_{\text{lié}}$ est composée de l'énergie d'étirement des liaisons $E_{\text{étirement-liaison}}$, de l'énergie des angles $E_{\text{angle-coudée}}$, et de l'énergie diédrale $E_{\text{dièdre}}$, tandis que $E_{\text{non-lié}}$ inclut l'énergie de van der Waals E_{vdW} et l'énergie Coulombienne E_{Coulomb} . Toutes les formules explicites avec les détails des termes du côté droit des équations 1.15 et 1.16 sont fournies dans (Senac, 2018; González, 2011; Cygan *et al.*, 2004). Parfois, d'autres termes sont ajoutés aux interactions non liées pour prendre en compte les interactions de liaison hydrogène (González, 2011; Kholmurodov *et al.*, 2010) et les termes croisés dus à un couplage entre torsion, étirement et flexion (González, 2011). Comme mentionné précédemment, ces formules explicite (fonctions analytiques) contiennent des paramètres qui peuvent être obtenus en ajustant les fonctions aux calculs quantiques *ab initio* et aux données expérimentales (González, 2011). Les champs de force basés uniquement sur des données expérimentales sont appelés champs de force empiriques. Cependant, la plupart des champs de force sont semi-empiriques ; ils combinent à la fois des données quantiques et expérimentales. De même que les logiciels informatiques qui possèdent plusieurs versions en raison des améliorations périodiques, le champ de force est également en évolution (González, 2011). Parmi les exemples de champs de

force largement utilisés dans les simulations de dynamique moléculaire, on trouve REAXFF (Van Duin *et al.*, 2001), IFF (Winetroun *et al.*, 2021), CLAYFF (Cygan *et al.*, 2004) et d'autres (Wang *et al.*, 2004; Mishra *et al.*, 2017).

En considérant la représentation fonctionnelle, ces potentiels peuvent être classifiés en trois grandes catégories : les potentiels d'ordre de liaison, les potentiels de cluster et les potentiels de méthode d'atome enchâssé (EAM) (Daw *et al.*, 1984). Parmi eux, les potentiels d'ordre de liaison sont de loin les plus utilisés dans la modélisation et la simulation computationnelle des structures de silicium. Parmi les exemples de potentiels d'ordre de liaison, on trouve le potentiel de Tersoff (Tersoff, 1986), le potentiel de Tersoff modifié (MOD) (Kumagai *et al.*, 2007), le potentiel de Dodson (DOD) (Dodson, 1987) et d'autres.

Le concept de ces potentiels repose sur le fait que la force d'une liaison chimique dépend de l'environnement de liaison, y compris le nombre de liaisons ainsi que les angles et longueurs des liaisons. Le modèle de Tersoff (TS), qui est un potentiel à plusieurs corps, est le plus largement utilisé. Ce modèle décrit par les équations 1.17 et 1.18, a réussi à étudier une large gamme de propriétés, y compris la propagation des phonons (Li *et al.*, 1988), les réponses thermo-mécaniques (Zhao *et al.*, 2006), la stabilité des défauts ponctuels et les phases non cristallines (Tersoff, 1988a), pour n'en nommer que quelques-unes.

$$E = \sum_i E_i = \frac{1}{2} \sum_{i,j \neq i} V_{ij} \quad (1.17)$$

$$V_{ij} = f_c(r_{ij}) [A \exp(-\lambda_1 r_{ij}) - B_{ij} \exp(-\lambda_2 r_{ij})] \quad (1.18)$$

où E est l'énergie totale du système, E_i est l'énergie de l'atome i , V_{ij} est l'énergie d'interaction entre les atomes i et j , r_{ij} est la distance entre eux, A , B , λ_1 et λ_2 sont tous des paramètres positifs et f_c est la fonction de coupure bien détaillé dans l'article de Tersoff (Tersoff, 1986). On remarque qu'en imposant, par exemple, $\lambda_1 = 2\lambda_2$, on retrouve le potentiel de Morse. Le potentiel de Morse est utilisé principalement pour modéliser les interactions entre deux atomes, en particulier les liaisons covalentes.

Le modèle de Tersoff a également été étendu avec succès aux systèmes hétéro-nucléaires tels que le carbure de silicium (SiC) et le silicium-germanium (SiGe). D'autres modèles de potentiels d'ordre de liaison, comme le potentiel de Tersoff modifié (Pastewka *et al.*, 2008), le potentiel interatomique dépendant de l'environnement (EDIP) (Justo *et al.*, 1998) et le ReaxFF (Van Duin *et al.*, 2001), ont été introduits pour incorporer les effets de plusieurs corps dans la modélisation du silicium. Plus récemment, dans (Pun *et al.*, 2017), le MOD (Kumagai *et al.*, 2007) a été optimisé pour modéliser une large gamme de propriétés du silicium, y compris les énergies de formation des défauts ponctuels, les énergies de formation des surfaces et de reconstructions, les facteurs d'expansion thermique, la température de fusion et quelques autres propriétés.

La deuxième classe importante de potentiels semi-empiriques pour le silicium est celle des potentiels de cluster (Balamane *et al.*, 1992), tels que Stillinger-Weber (Stillinger *et al.*, 1985), le potentiel de Biswas Harnann (BH) (Biswas *et al.*, 1985) et le potentiel de P. Carson, Takai, Halicioglu et Tiller (PTHT) (Pearson *et al.*, 1984). Le modèle de potentiel de cluster comprend des contributions d'interactions à deux et trois corps. Le terme à deux corps rend compte des interactions répulsives, tandis que le terme à trois corps traite des interactions attractives et des contributions angulaires. Comme le potentiel de Tersoff, le potentiel Stillinger-Weber (Stillinger & Weber, 1985) est également largement utilisé pour le silicium. Sa forme analytique s'écrit :

$$E = \epsilon A \sum_{i < j} \left[f_{ij}^{(2)}(r_{ij}) + \frac{\lambda}{A} \sum_{k \neq i, j} f_{ijk}^{(3)}(r_{ij}, r_{ik}, \theta_{ijk}) \right] \quad (1.19)$$

Le potentiel à deux corps $f_{ij}^{(2)}(r_{ij})$ peut s'écrire :

$$f_{ij}^{(2)}(r_{ij}) = A \exp\left(-\frac{r_{ij}}{\rho}\right) - B \exp\left(-\frac{r_{ij}}{\sigma}\right) \quad (1.20)$$

Le potentiel à trois corps $f_{ijk}^{(3)}(r_{ij}, r_{ik}, \theta_{ijk})$ peut s'écrire :

$$f_{ijk}^{(3)}(r_{ij}, r_{ik}, \theta_{ijk}) = C \left[\exp\left(\frac{r_{ijk}}{d}\right) - \exp\left(\frac{r_{ijk}}{e}\right) \right] (\cos \theta_{ijk} - \cos \theta_0)^2 \quad (1.21)$$

Où E est l'énergie totale du système. Les paramètres du modèle incluent ϵ , A , λ , B , ρ , σ , C , θ_0 , d , et e . Les distances de séparation interatomiques sont notées r_{ij} et r_{ik} , tandis que θ_{ijk} désigne l'angle formé par l'atome central i et ses atomes voisins j et k .

D'autres formes plus explicites des fonctions à deux et à trois corps sont détaillées dans la référence (Vink *et al.*, 2001b).

Le modèle SW est bien connu pour son succès dans la modélisation des propriétés thermiques des nanofils de silicium et du matériau massif de silicium (Yang *et al.*, 2008; Volz *et al.*, 1999; Feng, 2003; Maiti *et al.*, 1997), ainsi que des propriétés du silicium liquide (Tavares *et al.*, 2018).

La méthode d'atome incorporé modifiée (MEAM) (Lenosky *et al.*, 2000; Cai, 1999; Baskes, 1987) est la troisième classe de potentiels pour le silicium. Le modèle original connu sous le nom de méthode d'atome incorporé (EAM) (Daw *et al.*, 1983; Daw & Baskes, 1984; Schommers, 2019) pour les métaux a été modifié pour tenir compte des contributions angulaires dans l'énergie d'interaction au sein des systèmes covalents.

$$E = \frac{1}{2} \sum_{i \neq j} \phi(r_{ij}) + \sum_i E_{\text{embed}}(\rho_i) \quad (1.22)$$

$$\rho_i = \sum_{j \neq i} \rho(r_{ij}) \quad (1.23)$$

Comme on peut le voir, l'énergie totale E du système est décomposée en un potentiel d'interaction à deux corps, qui dépend uniquement de la distance r_{ij} entre deux atomes i et j différents, et en une énergie d'incorporation $E_{\text{embed}}(\rho_i)$, qui représente l'énergie requise pour placer l'atome i dans la densité électronique de fond ρ_i à la position i . La fonction ρ_i désigne la densité électronique de l'atome i , tandis que $\rho(r_{ij})$ est la densité électronique de l'atome j à une distance r_{ij} du noyau de l'atome i .

Dans le concept EAM, chaque atome est enchâssé dans un gaz électronique avec une densité, affectée par les autres atomes. En résumé, tous les potentiels mentionnés pour le silicium souffrent de limitations en termes de polyvalence, et l'amélioration de n'importe quel potentiel

semi-empirique pour décrire des propriétés spécifiques est généralement accompagnée de sacrifices sur d'autres propriétés du matériau. Les potentiels semi-classiques sont évidemment trop simples pour couvrir la large gamme de propriétés à décrire. De plus, ils sont peu adaptables en raison du grand nombre de polymorphes qui sont stables à différentes températures et pressions avec différents nombres de coordination et types de liaison allant de fortement covalents à métalliques (Li *et al.*, 2018; Lenosky *et al.*, 2000; Kumagai *et al.*, 2007; Kasper & Richards, 1964; Gibbs, 1926; Li *et al.*, 1994). Parfois, on souhaite modéliser l'interaction du matériau avec d'autres substances chimiques. Par exemple, les oxydes de silicium résultent de l'interaction du silicium avec l'oxygène. Ces oxydes sont des matériaux importants pour l'industrie du verre (Shelby, 2020). De plus, les agrégats des oxydes de silicium, principalement le quartz, se trouvent souvent dans le béton, qui est une source d'empreinte carbone (Siddique, 2011; Neville & Brooks, 1987; Siddique, 2011; Mazloom *et al.*, 2004). En outre, les oxydes de silicium, y compris le silicium, sont les éléments constitutifs des systèmes d'argile (Singh, 2022). Ainsi, disposer d'un potentiel capable de décrire avec précision le silicium et ses oxydes est très important. Comme pour le silicium, des modèles de potentiels interatomiques simples ont été proposés pour les systèmes basés sur la silice et les systèmes mixtes silice-silicium. La plupart de ces potentiels, à l'exception de certains (Yasukawa, 1996; Van Beest *et al.*, 1990; Tsuneyuki *et al.*, 1988), ont été paramétrés pour les systèmes Si-O (Munetoh *et al.*, 2007; Lee *et al.*, 2016; Yu *et al.*, 2007; Van Duin *et al.*, 2003; Watanabe *et al.*, 1999; Fischer *et al.*, 2006; Ganster *et al.*, 2010b; France-Lanord *et al.*, 2016; Billeter *et al.*, 2006). Ils incluent des modèles simples tels que le potentiel Stillinger-Weber (SW) (Stillinger & Weber, 1985) et le potentiel de Tersoff (Tersoff, 1986), qui sont les plus adaptés aux systèmes où les interactions interatomiques à courte portée prédominent. Comme les interactions interatomiques dans le SiO₂ sont de nature ionique, des potentiels à charge fixe ont été introduits pour capturer à la fois les interactions coulombiennes à courte et longue portée. Parmi ces potentiels, on trouve le potentiel BKS (Van Beest *et al.*, 1990) et le potentiel TTAM (Tsuneyuki *et al.*, 1988). Le potentiel BKS est donné par l'équation 1.24.

$$\Phi_{ij}(r_{ij}) = \left[A_{ij} \exp(-B_{ij}r_{ij}) - \frac{C_{ij}}{r_{ij}^6} \right] + \frac{q_i q_j}{r_{ij}} \quad (1.24)$$

ou A_{ij} , B_{ij} et C_{ij} sont des paramètres déterminés par régression du modèle à l'aide des données *ab initio* et expérimentales (Van Beest *et al.*, 1990). Les paramètres A_{ij} , B_{ij} et C_{ij} sont fixes pour une paire d'atomes donnée. Ils sont fournis dans le tableau 1.1 ci-dessous. Les q_i et q_j sont les charges partielles. En plus des caractéristiques de charge fixe, ces potentiels

Tableau 1.1 Paramètres de BKS, $q_{Si} = 2.4$ et $q_O = -1.2$

i-j	A_{ij} (eV)	B_{ij} ($\text{\AA}^{-1}$)	C_{ij} (eV $\cdot \text{\AA}^6$)
O - O	1388	2.760	175
O - Si	18003	4.873	133
Si - Si	0	0.000	0
Al - O	16008	4.797	130
Al - Al	0	0.000	0
P - O	9034	5.191	19
P - P	0	0.000	0

sont insensibles aux variations des angles de liaison ou à la sur/sous-coordination pendant la simulation. Comme l'oxygène est connu pour être très déformable et polarisable en fonction de son environnement de coordination immédiat, l'idée de charge fixe a été étendue pour tenir compte des effets de la polarisation des dipôles de l'oxygène (Tangney *et al.*, 2002; Kermode *et al.*, 2010). Bien que la charge soit optimisée, il est crucial de permettre à la charge atomique de répondre à l'environnement électrostatique changeant pour la précision des prévisions du potentiel. En effet, la capacité du potentiel à capturer la distribution dynamique des charges au sein des atomes est cruciale pour la détermination précise de la contribution de l'énergie électrostatique à l'énergie totale d'interaction. Ainsi, des modèles interatomiques semi-classiques plus sophistiqués incorporant le modèle d'équilibrage de charge (QEq) (Rappe *et al.*, 1991) ont été développés. Ces potentiels incluent le Tersoff modifié (Yasukawa, 1996), le potentiel à plusieurs corps optimisé pour les charges (COMB) (Yu *et al.*, 2007), le potentiel de méthode d'atome enchâssé modifié pour les systèmes ioniques 2NNMEAM + Qeq (Lee *et al.*, 2016), et le champ de force réactif (ReaxFF) pour les systèmes de silicium et d'oxyde de silicium (Van Duin

et al., 2003). Le formalisme général du champ de force ReaxFF est présenté par l'équation 1.25.

$$E_{\text{total}} = \sum_{i < j} [V_{\text{bond}}(r_{ij}) + V_{\text{angle}}(\theta_{ijk}) + V_{\text{torsion}}(\phi_{ijkl}) + V_{\text{overlap}}(r_{ij}, r_{ik}, \theta_{ijk})] + \sum_{i < j} V_{\text{coulomb}}(r_{ij}) \quad (1.25)$$

Dans l'équation 1.25, chaque terme représente une énergie d'interaction, expliqué comme suit :

- $V_{\text{bond}}(r_{ij})$: Ce terme représente l'énergie potentielle de liaison entre deux atomes i et j , qui dépend de la distance r_{ij} les séparant. Il quantifie l'énergie associée à la liaison formée entre ces atomes, généralement modélisée par des fonctions telles que le potentiel harmonique.
- $V_{\text{angle}}(\theta_{ijk})$: Ce terme prend en compte l'énergie potentielle angulaire impliquant trois atomes : i , j et k . L'angle θ_{ijk} est formé par les liaisons reliant ces atomes. Ce potentiel décrit le coût énergétique associé à la déformation ou au changement de l'angle entre les liaisons, essentiel pour maintenir la géométrie de la molécule.
- $V_{\text{torsion}}(\phi_{ijkl})$: Ce terme décrit l'énergie potentielle de torsion liée à quatre atomes : i , j , k et l . L'angle de torsion ϕ_{ijkl} représente la rotation autour de la liaison reliant les atomes j et k . Ce potentiel capture les variations d'énergie qui se produisent lorsque ces atomes sont tournés, influençant la conformation globale de la molécule.
- $V_{\text{overlap}}(r_{ij}, r_{ik}, \theta_{ijk})$: Ce terme fait référence à l'énergie potentielle d'overlap, qui découle des interactions entre les nuages électroniques se chevauchant des atomes i , j et k . Il est fonction des distances r_{ij} et r_{ik} ainsi que de l'angle θ_{ijk} . Ce terme prend en compte les interactions répulsives lorsque les densités électroniques se chevauchent trop, contribuant ainsi à stabiliser la structure moléculaire.
- $V_{\text{coulomb}}(r_{ij})$: Ce terme représente l'énergie potentielle coulombienne entre deux particules chargées (ou atomes) i et j , en fonction de leur distance de séparation r_{ij} . Il capture les interactions électrostatiques entre atomes chargés, qui peuvent influencer significativement la stabilité et la dynamique moléculaires.

L'incorporation du modèle Qeq (Rappe & Goddard III, 1991) dans ces potentiels permet aux charges atomiques de répondre de manière dynamique à l'environnement atomique changeant pendant la simulation de dynamique moléculaire. Cependant, la variation de la charge est connue

pour avoir un impact négatif sur le coût de calcul, principalement pour les simulations à grande échelle.

1.2.6 Apprentissage automatique

L'apprentissage automatique, l'un des sous-domaines de l'intelligence artificielle, est une méthode computationnelle qui utilise des informations passées pour améliorer les performances et/ou faire des prédictions précises à l'aide d'algorithmes (Mohri *et al.*, 2018). Ainsi, grâce aux algorithmes implémentés, la machine apprend à partir des informations passées, des expériences ou des exemples, appelés données d'échantillon ou données d'entraînement, afin de prendre de meilleures décisions sans être explicitement programmée. Cette approche basée sur les données permet à l'ordinateur d'apprendre automatiquement et de s'ajuster en fonction de la situation ou des problèmes rencontrés, limitant ainsi l'intervention humaine et l'assistance. En général, tout sous-domaine de l'informatique dépend de la qualité de ses algorithmes, principalement déterminée par leur complexité en temps et en espace (Mohri *et al.*, 2018). Cependant, dans le domaine de l'apprentissage automatique, la complexité des échantillons est également un paramètre clé pour le succès de l'apprentissage. L'ensemble du processus technique de l'apprentissage automatique découle des idées et du formalisme de la statistique, de la probabilité et de l'optimisation, combinés aux concepts de l'informatique (Mohri *et al.*, 2018). L'apprentissage automatique est généralement divisé en trois grandes catégories (Figure 1.9) : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement (Sutton *et al.*, 2018). Contrairement aux autres, l'apprentissage supervisé est similaire à la procédure de régression standard, qui consiste à trouver une fonction inconnue reliant des données connues (entrées) à une sortie inconnue (Schmidt *et al.*, 2019). Comme rapporté dans (Schmidt *et al.*, 2019), les sorties pour les valeurs aberrantes sont estimées par extrapolation des motifs dans le jeu de données d'entraînement étiqueté. L'étiquetage des données est généralement réalisé par des humains et est plus coûteux que les données brutes non étiquetées. L'apprentissage non supervisé traite des données non étiquetées. Il consiste à trouver des motifs au sein des données non étiquetées. Par exemple, les problèmes de regroupement et d'association sont généralement

traités avec des approches d'apprentissage non supervisé. Parfois, les deux méthodes sont combinées pour former un apprentissage semi-supervisé, idéal pour une situation où un petit nombre de données étiquetées et un grand nombre de données non étiquetées sont disponibles (Mohri *et al.*, 2018). Enfin, la méthode d'apprentissage par renforcement est basée sur les interactions (Mohri *et al.*, 2018; Sutton & Barto, 2018). Cette approche vise à trouver la meilleure façon de mapper les situations aux actions afin de maximiser un signal de récompense numérique (Mohri *et al.*, 2018; Sutton & Barto, 2018).

L'apprentissage automatique, avec ses nombreuses applications pratiques telles que la classification de texte ou de documents, le traitement du langage naturel, le traitement de la parole, les applications de vision par ordinateur, et bien d'autres (Mohri *et al.*, 2018; Schmidt *et al.*, 2019), ainsi que les sciences des matériaux computationnelles (Mohri *et al.*, 2018; Schmidt *et al.*, 2019; Butler *et al.*, 2018a; Liu *et al.*, 2017), constitue un domaine en pleine expansion. Nous nous concentrerons principalement sur les techniques d'apprentissage automatique appliquées au domaine des matériaux. Introduit au cours du dernier siècle, l'apprentissage automatique s'est répandu dans tous les domaines des sciences des matériaux en raison des limitations des approches traditionnelles telles que l'essai-erreur et la théorie de la fonctionnelle de la densité (DFT) (Wei *et al.*, 2019). De la prédiction des propriétés des matériaux et des molécules à la conception et à la découverte de nouveaux matériaux, l'apprentissage automatique est devenu un outil de substitution puissant permettant de contourner la consommation de temps et de ressources, ainsi que le coût élevé des approches traditionnelles (Wei *et al.*, 2019; Liu *et al.*, 2017). Ainsi, l'apprentissage automatique peut considérablement réduire le cycle de développement de nouveaux matériaux impliquant des expériences de laboratoire répétitives et prédire efficacement les propriétés des matériaux avec des ressources computationnelles moindres et beaucoup plus rapidement que la DFT.

Dans le domaine des sciences des matériaux, les techniques d'apprentissage automatique ont été utilisées avec succès pour déterminer le diagramme de phases, les propriétés des matériaux basées sur des calculs mécaniques quantiques, et le développement du potentiel interatomique (Mueller *et al.*, 2020). Ce dernier est d'un grand intérêt car de nombreuses propriétés physico-chimiques en

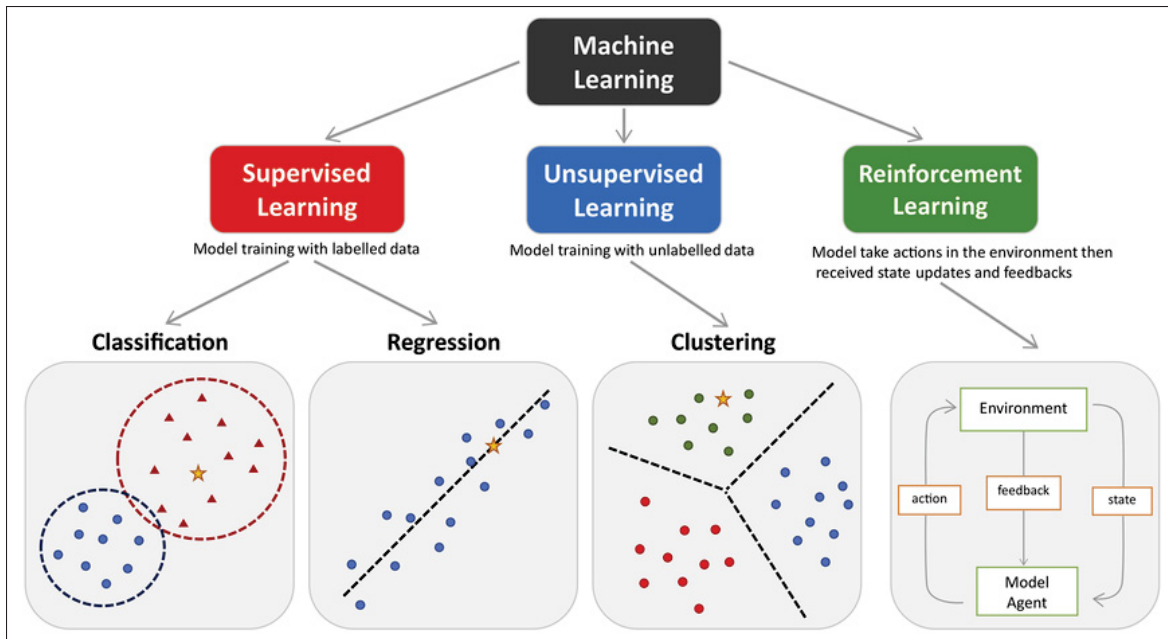


Figure 1.9 Les principales approches de l'apprentissage automatique incluent la classification et la régression dans l'apprentissage supervisé, ainsi que le clustering dans l'apprentissage non supervisé. L'apprentissage par renforcement améliore les performances du modèle en interagissant avec l'environnement. Les points colorés et les triangles représentent les données d'entraînement. Les étoiles jaunes représentent les nouvelles données qui peuvent être prédites par le modèle entraîné
Tirée de Peng *et al.* (2021, p. 2)

dépendent. En effet, la surface d'énergie potentielle (SEP) contient de nombreuses informations et propriétés des matériaux. Par exemple, les structures stables, les fréquences vibratoires et phononiques, la constante de force, le module élastique, les défauts et l'énergie d'activation des réactions chimiques, ainsi que la diffusion et les transitions de phase sont cryptés dans la SEP (Mueller *et al.*, 2020; Deringer *et al.*, 2019). Ainsi, la détermination précise de la surface d'énergie potentielle a été l'objet de plusieurs investigations computationnelles basées sur des approches bien connues telles que la théorie de la fonctionnelle de la densité et les champs de forces semi-empiriques, jusqu'à la percée avec les approches d'apprentissage automatique. L'apprentissage automatique a résolu les défis posés par les deux premières approches : d'une

part, les limitations spatiotemporelles de la DFT, et d'autre part, le manque de versatilités des champs de forces semi-empiriques (Mueller *et al.*, 2020; Mueller *et al.*, 2016).

Le concept de potentiel interatomique basé sur l'apprentissage automatique est similaire à celui des champs de forces semi-empiriques qui reposent sur des hypothèses physiques telles que la localité et la lissité de la SEP (Deringer *et al.*, 2019). Cependant, contrairement au champ de forces semi-empiriques, qui nécessite une forme fonctionnelle fermée et une paramétrisation appropriée de la SEP, les MLIP reposent sur un ensemble de fonctions invariantes qui nécessitent uniquement des données de référence pour extraire la SEP avec une précision comparable à celle de la DFT, tout en étant plus polyvalents que les champs de forces semi-classiques (Botu *et al.*, 2017). Bien que l'idée de l'apprentissage automatique dans le sous-domaine des interactions atomiques ait été brièvement introduite au début des années 90 (Schmidt *et al.*, 2019), elle n'a été développée et mise en œuvre qu'en 2007 par Behler et Parrinello (Deringer *et al.*, 2019; Schmidt *et al.*, 2019; Behler & Parrinello, 2007). Ils ont développé un potentiel interatomique basé sur l'apprentissage automatique utilisant un algorithme de réseau de neurones, qui imite le fonctionnement du cerveau humain. Depuis lors, il y a eu un développement continu de différents potentiels interatomiques basés sur l'apprentissage automatique, distinguables par leurs méthodes de paramétrisation et de descripteur (Wang *et al.*, 2024; Deringer *et al.*, 2019; Musil *et al.*, 2021; Schütt *et al.*, 2020). La plupart d'entre eux reposent sur un schéma de partitionnement de l'environnement atomique et s'appuient sur des données d'entraînement principalement issues des calculs quantiques basés sur la DFT.

1.2.7 Application au potentiel interatomique : Potentiel interatomique par apprentissage automatique

La modélisation moléculaire, pour rivaliser avec les expériences ou constituer une alternative significative, repose sur sa capacité à calculer efficacement les forces interatomiques d'une configuration atomique donnée. Cela nécessite un bon descripteur (Tableau 1.2) de l'environnement atomique local qui prend en compte tous les effets physico-chimiques, y compris tous les types

d'interactions pertinents, et sa capacité à intégrer les transformations isométriques telles que la translation, la permutation des atomes, la rotation, ainsi que la lissité des fonctions.

Tableau 1.2 Liste de descripteurs
Adapté de Wang *et al.* (2024, p. 4)

Descripteur	Année
Atom-centered symmetry functions (ACSF) (Behler, 2011)	2011
Coulomb Matrix (Rupp <i>et al.</i> , 2012)	2012
Smooth Overlap of Atomic Positions (SOAP) (Bartók <i>et al.</i> , 2013)	2013
Ewald sum and Sine Matrix (Faber <i>et al.</i> , 2015)	2015
Spectral neighbor analysis method (SNAP) (Thompson <i>et al.</i> , 2015)	2015
Moment Tensor Potentials (MTP) (Shapeev, 2016)	2016
Weighted atom-centered symmetry functions (wACSF) (Gastegger <i>et al.</i> , 2018)	2018
Automatic selection of fingerprints (Imbalzano <i>et al.</i> , 2018)	2018
Atomic-position independent material descriptor (U-api) (Drautz, 2019)	2018
Deep Potential Dynamic Descriptor (DP) (Zhang <i>et al.</i> , 2018)	2018
Graph Descriptor (Xie <i>et al.</i> , 2018)	2018
Optimizing SOAP (Caro, 2019)	2019
Atomic Cluster Expansion (ACE) (Drautz, 2019)	2019
Many-body Tensor Representation (MBTR) (Huo <i>et al.</i> , 2022)	2022

Étant donné que l'apprentissage automatique nécessite finalement l'existence de données passées, un problème d'apprentissage peut se résumer comme suit : étant donné un ensemble de données {matériaux, propriétés}, où les propriétés (énergie, forces, contraintes) sont calculées avec des méthodes mécaniques quantiques, comment pouvons-nous estimer les propriétés de nouveaux matériaux non présents dans l'ensemble de données, à condition que le nouveau matériau appartienne à la même classe chimique-structurale (Botu *et al.*, 2017) que celle de l'ensemble de données ? Mathématiquement, cela s'apparente à une tâche d'interpolation et nécessite a priori une fonction, plus précisément un modèle, qui associe les matériaux aux propriétés dans l'ensemble de données {matériaux, propriétés}. Ainsi, il convient de construire un modèle F pour adapter les données mécaniques quantiques. Étant donné une configuration x_q avec une

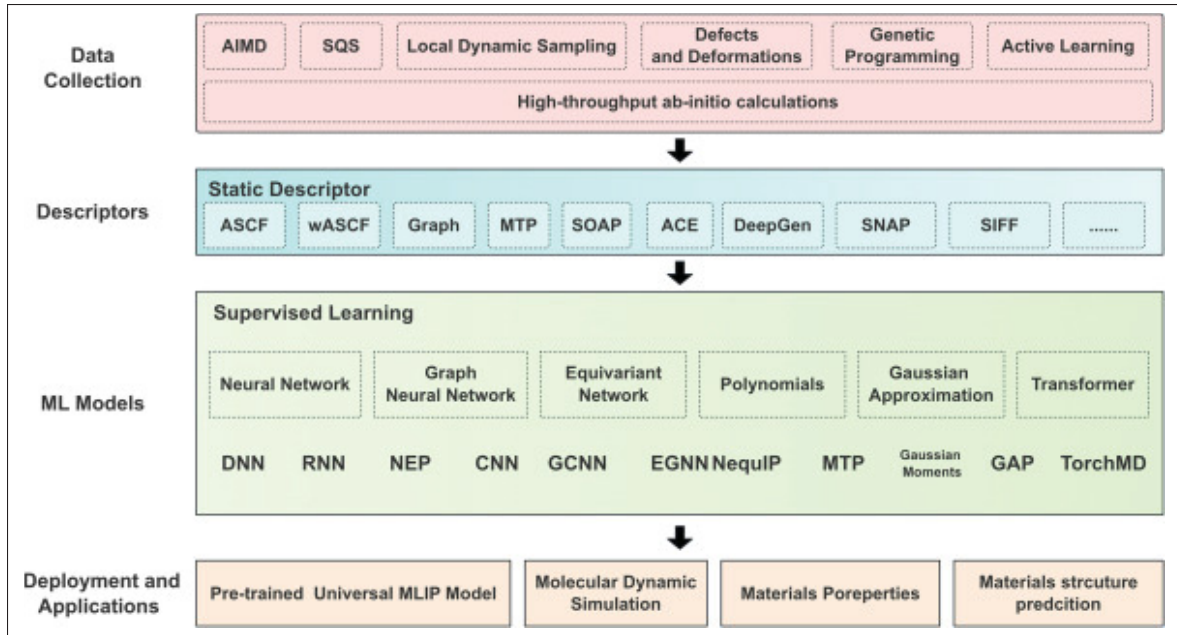


Figure 1.10 Diagramme de processus et les quatre étapes essentielles des MLIPs

Tirée Wang *et al.* (2024, p. 3)

énergie de la mécanique quantique $E^{qm}(x_q)$, l'énergie $E(x_q)$ fournie par le modèle F doit être aussi proche que possible de l'énergie de la mécanique quantique. On peut trouver les paramètres optimaux de F en minimisant la fonction (fonction de perte) ci-dessous.

$$f = \sum_q |E(x_q) - E^{qm}|^2 + (forces) + \dots + \quad (1.26)$$

La fonction f sera décrite plus en détail dans les sections suivantes. Le processus global nécessite un modèle de régression, des données de référence et un algorithme d'optimisation. Maintenant que nous savons qu'il est possible de construire un ensemble de fonctions mathématiques, plus précisément un modèle, pour prédire les données mécaniques quantiques, comment ces fonctions agissent-elles sur une configuration atomique donnée dans le contexte de l'apprentissage automatique ? Comment ces fonctions peuvent-elles approximer l'énergie totale d'une configuration donnée sans agir directement sur la structure électronique comme le fait l'équation de Schrödinger ? Pour répondre à ces questions, nous allons donner plus en

details dans les sections ci-dessous ce qui a permis l'application de l'approche d'apprentissage automatique au domaine des potentiels interatomiques. Le processus de développement d'un potentiel interatomique par apprentissage automatique est illustré à la Figure 1.10.

1.2.8 Propriété de partitionnement issue de la localité des interactions interatomiques

Comme mentionné ci-dessus, les potentiels interatomiques basés sur l'apprentissage automatique (MLIP) traitent la surface d'énergie potentielle (SEP) comme une fonction des descripteurs de l'environnement atomique local, avec des propriétés de symétrie telles que la rotation, la translation et la permutation (Bartók *et al.*, 2013; Zuo *et al.*, 2020). Dans l'approche MLIP, la SEP est exprimée comme une somme d'énergies atomiques individuelles qui dépendent uniquement de leur environnement local (Zuo *et al.*, 2020), comme illustré par la figure 1.11 ci-dessous. Autrement dit, l'énergie totale d'une configuration donnée est la somme des contributions des énergies atomiques individuelles, calculées dans un rayon de coupure définissant les limites de l'environnement atomique local.

$$SEP = E_{Total}(x_q) = \sum_i E_i \quad (1.27)$$

E_i is the energy of atom i given by its neighborhood.

Soit maintenant u_i un tuple, une collection des distances relatives $r_i - r_j$ des atomes j par rapport à l'atome i et ce dans un cerlce ou une sphere de rayon R_c :

$$u_i = (r_i - r_j)_{1 \leq i \leq N_{tot}, 0 \leq |r_i - r_j| \leq R_c} \quad (1.28)$$

En fait, u_i désigne l'environnement de l'atome i . Le tuple est constitué de cette manière pour chaque atome i parmi le total N_{tot} d'atomes. L'équation 1.27 peut être réécrite comme suit :

$$E_{Total} = \sum_i E_i = \sum_i V(u_i) \quad (1.29)$$

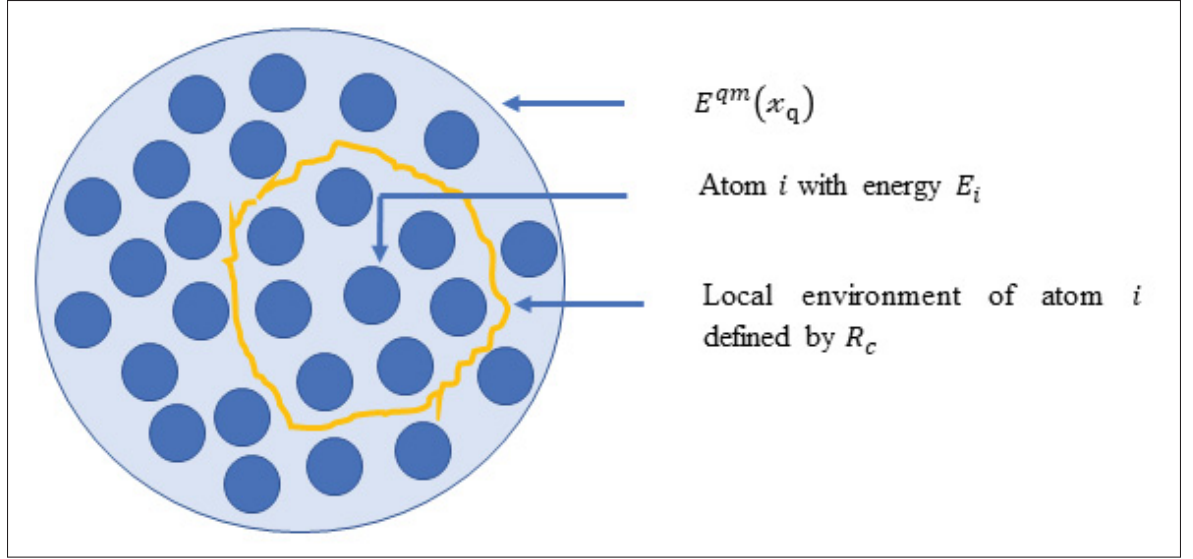


Figure 1.11 Illustration d'un environnement atomique local au sein d'une configuration donnée x_q avec N_{tot} atomes, d'énergie totale E^{qm} donnée par la mécanique quantique

La force agissant sur un atome k à la position r_k dans la configuration x_q est déterminée par le gradient ∇ de l'énergie totale E_{Total} . Cette énergie totale peut également être exprimée comme la somme des dérivées de l'énergie atomique V de chaque atome i par rapport à la position r_k :

$$F_k(x_q) = -\nabla E_{\text{Total}}(x_q) = -\sum_i \frac{\partial V(u_i)}{\partial r_k(x_q)} \quad (1.30)$$

Maintenant, en connaissant l'énergie du système et le volume de la maille, nous pouvons calculer les contraintes à partir du concept de *stress virial*. Ce dernier représente une mesure du stress mécanique à l'échelle atomique pour les systèmes homogènes. Ainsi, le stress virial dans la configuration atomique à zéro kelvin (où toutes les vitesses sont nulles) est défini comme suit :

$$\tau_k(x_q) = \frac{1}{2\Omega} \sum_{k \in \Omega} \sum_{l \in \Omega} (r_l^{(i)} - r_k^{(i)}) F_{kl}^{(i)} \quad (1.31)$$

où Ω est le volume de la maille de la configuration et $F_{kl}^{(i)}$ est le j^{me} composant des forces appliquées sur l'atome k par l'atome l .

Notez que les calculs par la théorie de la fonctionnelle de densité (DFT) sont généralement effectués à zéro Kelvin et fournissent un résultat numérique pour les équations 1.29 à 1.31 en fonction de la taille du système.

La fonction V dans l'équation 1.29 est appelée potentiel interatomique, et la somme de $V(u_i)$ sur tout l'environnement atomique de la configuration x_q est supposée approximativement égale à l'énergie totale $E^{qm}(x_q)$ donnée par l'approche mécanique quantique. V doit être construit de telle sorte que $E_{Total}(x_q)$ soit aussi proche que possible de $E^{qm}(x_q)$.

$$E_{Total} = \sum_i V(u_i) \approx E^{qm}(x_q) \quad (1.32)$$

Cette hypothèse est basée sur la localité des interactions interatomiques et fonctionne dans la plupart des systèmes avec des forces à courte portée où les interactions à longue portée, telles que les interactions coulombiennes dans un système chargé ou polarisé, peuvent être négligées.

La taille du tuple u_i est déterminée par la densité des atomes dans le voisinage et peut varier, donc la fonction V est en pratique une famille de fonctions mais nous la référons encore comme une fonction pour des raisons de commodité. La fonction V est par définition symétrique par translation et doit satisfaire les symétries suivantes : (R1) invariance par permutation, (R2) invariance par rotation et réflexion, et (R3) lissité.

Tous les MLIP sont développés à partir de l'idée de partitionnement de l'énergie totale en énergies atomiques individuelles comme illustré ci-dessus. Ils diffèrent par le type de descripteurs de l'environnement atomique local (Zuo *et al.*, 2020) et de régresseurs (Deringer *et al.*, 2019). À ce jour, il existe six principaux régresseurs pour le potentiel interatomique en apprentissage automatique. Le tableau 1.3 résume les méthodes d'apprentissage automatique et les implémentations disponibles pour les potentiels interatomiques (Deringer *et al.*, 2019; Wang *et al.*, 2024).

Tableau 1.3 La liste des logiciels implémentés et des dates de publication pour différents modèles MLIP

Tiré de Wang *et al.* (2024, p. 7)

Model	Type	Date	Reference
NNP	NN	1995	Blank (Blank <i>et al.</i> , 1995)
BPNN	NN	2007	Behler (Behler & Parrinello, 2007)
ænet-Fortran	NN	2016	Artrith (Artrith <i>et al.</i> , 2016)
DTNN	NN	2017	Schutt (Schütt <i>et al.</i> , 2017b)
ANI-1	NN	2017	Smith (Smith <i>et al.</i> , 2017)
HIP-NN	NN	2018	Nebgen (Nebgen <i>et al.</i> , 2018)
SchNet	NN	2018	Schutt (Schütt <i>et al.</i> , 2017a)
DeepPotential	NN	2018	Zhang (Zhang <i>et al.</i> , 2018)
PhysNet	NN	2019	Unke (Unke <i>et al.</i> , 2019)
ACE	NN	2019	Drautz (Drautz, 2019)
LieConv	NN	2020	Finzi (Finzi <i>et al.</i> , 2020)
4G-HDNNP	NN	2021	KO (Ko <i>et al.</i> , 2021)
NEP	NN	2021	FAN (Fan <i>et al.</i> , 2021)
ænet-Pytorch	NN	2021	Artrith (López-Zorrilla <i>et al.</i> , 2023)
GNNFF	GraphNN	2021	Pa (Park <i>et al.</i> , 2021)
MDGNN	GraphNN	2021	Wang (Wang <i>et al.</i> , 2021)
NoisyNodes	GraphNN	2022	Godwin (Godwin <i>et al.</i> , 2021)
M3GNet	GraphNN	2022	Chen (Chen <i>et al.</i> , 2022)
TeaNet	GraphNN	2022	Takamoto (Takamoto <i>et al.</i> , 2022)
PotentialMind	GraphNN	2023	Wang (Wang <i>et al.</i> , 2023)
CHGNet	GraphNN	2023	Deng (Deng <i>et al.</i> , 2023)
TensorMol 0.1	GraphNN	2018	Yao (Yao <i>et al.</i> , 2018)
DimeNet++	GraphNN	2020	Gasteiger (Gasteiger <i>et al.</i> , 2020a)

Tableau 1.3 La liste des logiciels implémentés et des dates de publication pour différents modèles MLIP

Tiré de Wang *et al.* (2024, p. 7) (suite)

Model	Type	Date	Reference
DimeNet	GraphNN	2020	Gasteiger (Gasteiger <i>et al.</i> , 2020b)
NewtonNet	GraphNN	2021	Haghighatlari (Haghighatlari <i>et al.</i> , 2022)
SphereNet	GraphNN	2021	Liu (Liu <i>et al.</i> , 2021a)
PAINN	GraphNN	2021	Schutt (Schütt <i>et al.</i> , 2021)
GemNet	GraphNN	2022	Gasteiger (Gasteiger <i>et al.</i> , 2021)
HermNet	GraphNN	2022	Wang (Wang <i>et al.</i> , 2022)
MACE	GraphNN	2022	Batatia (Batatia <i>et al.</i> , 2022)
Allegro	GraphNN	2023	Musaelian (Musaelian <i>et al.</i> , 2023)
Tensor-Field	EquivariantNetwork	2018	Thomas (Thomas <i>et al.</i> , 2018)
Cormorant	EquivariantNetwork	2019	Anderson (Anderson <i>et al.</i> , 2019)
UNET	EquivariantNetwork	2021	Qiao (Qiao <i>et al.</i> , 2022)
EGNN	EquivariantNetwork	2021	Satorras (Satorras <i>et al.</i> , 2021)
NequIP	EquivariantNetwork	2022	Batzner (Batzner <i>et al.</i> , 2022)
MTP	Polynomials	2016	Shapeev (Shapeev, 2016)
IPMLs	Polynomials	2018	Bereau (Bereau <i>et al.</i> , 2018)
aPIPs	Polynomials	2021	Allen (Allen <i>et al.</i> , 2021)
Bartok	Gaussian	2010	Bartok (Bartok, 2010)
FCHL19	Gaussian	2020	Christensen (Christensen <i>et al.</i> , 2020)
GMNN	Gaussian	2020	Zaverkin (Zaverkin <i>et al.</i> , 2020)
SE(3)-Transformers	Transformer	2020	Fuchs (Fuchs <i>et al.</i> , 2020)
DeepMoleNet	Transformer	2021	Liu (Liu <i>et al.</i> , 2021b)
TorchMD-NET	Transformer	2022	Thölke (Tholke <i>et al.</i> , 2022)

1.2.9 Régresseur à Réseau de Neurones (NN)

L'idée de cette approche provient des fonctions cérébrales. La Figure 1.12 montre le groupe interconnecté de nœuds inspiré par la simplification du cerveau. La sortie illustrée à la Figure 1.12 peut être exprimée mathématiquement comme suit :

$$y_k(x_1, \dots, x_n) = f_a(w_{k0} + \sum_{j=1}^n w_{kj} g_a(w_{j0} + \sum_{i=1}^n w_{ji} x_i)) \quad (1.33)$$

où y_k et $x_1, \dots, x_n$ sont respectivement la sortie et les entrées, w_{j0} est le biais, et w_{pq} sont les poids reliant les nœuds p dans la couche m aux nœuds q dans la couche $m - 1$ (plus précisément, on peut écrire w_{pq}^m où m désigne le numéro de la couche); f_a et f_b désignent les fonctions d'activation. Les entrées x_i sont le plus souvent des coordonnées cartésiennes atomiques transformées, parfois appelées coordonnées généralisées. Cette transformation est généralement réalisée avec des descripteurs de l'environnement atomique local qui sont des fonctions de symétrie.

1.2.9.1 Potentiel par réseau de neurones (NNP : Neural Network Potential)

Plusieurs types de descripteurs atomiques locaux sont utilisés dans l'approche du potentiel par réseau de neurones (Wang *et al.*, 2024). Le plus courant est celui introduit par Behler (Behler & Parrinello, 2007), connu sous le nom de fonctions de symétrie de type gaussien. La fonction de symétrie radiale et la fonction de symétrie angulaire sont données ci-dessous.

$$G_i^{rad} = \sum_j^{N_c} e^{-\eta(r_{ij}-R_s)^2} \cdot f_c(r_{ij}) \quad (1.34)$$

$$G_i^{ang} = 2^{1-\xi} \sum_{j,k \neq i}^{N_c} (1 + \lambda \cos \theta_{ijk})^\xi \cdot e^{-\eta(r_{ij}^2 + r_{ik}^2 + r_{jk}^2)} \cdot f_c(r_{ij}) \cdot f_c(r_{ik}) \cdot f_c(r_{jk}) \quad (1.35)$$

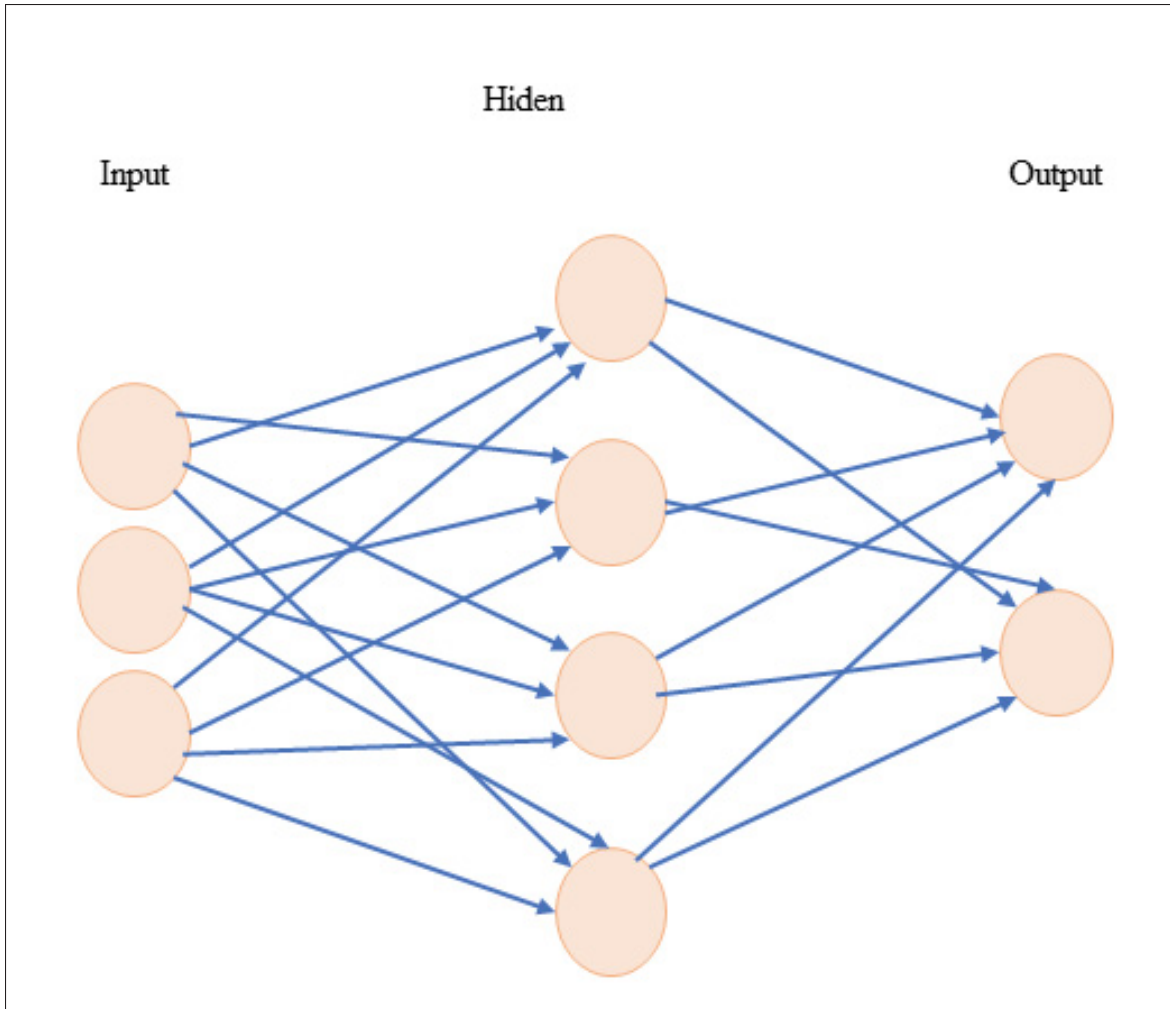


Figure 1.12 Disposition du réseau de neurones artificiels

où $r_{ik}=r_i-r_j$ est la distance relative de l'atome j à l'atome central i ; η et R_s sont respectivement la largeur de la gaussienne et le décalage de position pour tous les atomes voisins N_c dans le rayon de coupure R_c . Le paramètre ξ contrôle la résolution angulaire et est parfois pris comme $(-1,1)$. La contribution de l'angle θ_{ijk} comme présenté dans la figure 1.14 est calculée pour tous les triplets (i, j, k) dans le système à partir des coordonnées relatives. En effectuant le produit scalaire avec les composantes cartésiennes des vecteurs relatifs, ainsi qu'avec les normes de ces vecteurs, il est possible de déterminer aisément la valeur du cosinus comme suit :

$$\cos \theta_{ijk} = \frac{\vec{r}_{ij} \cdot \vec{r}_{ik}}{r_{ij} \cdot r_{ik}} \quad (1.36)$$

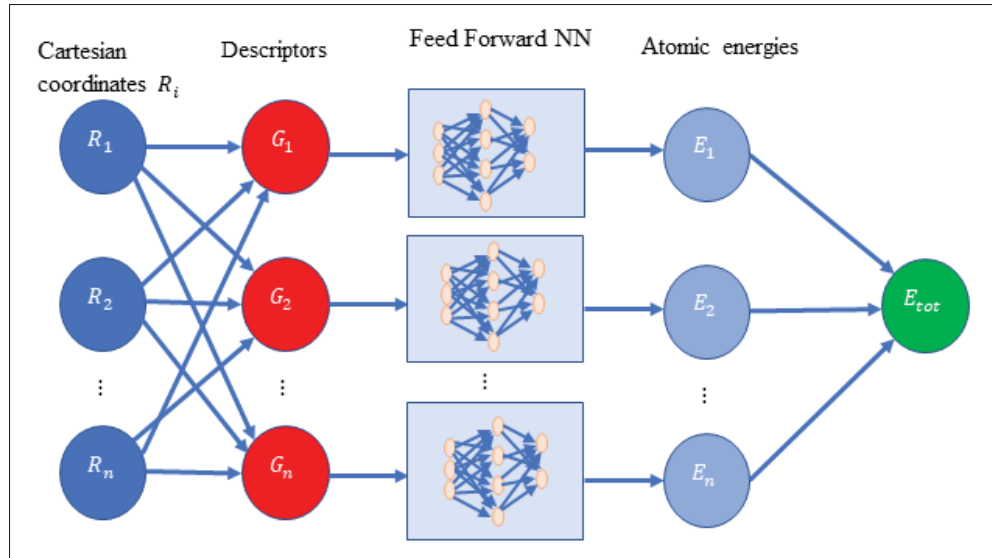


Figure 1.13 Construction du potentiel de réseau de neurones

La fonction f_c est la fonction de coupure qui permet de définir le voisinage d'interaction des atomes. Une autre fonction de coupure de type polynomial permet aux utilisateurs de spécifier un paramètre mesurant le taux de décroissance de la fonction de coupure dans le voisinage atomique local. Cette fonction de coupure, utilisée dans (Khorshidi *et al.*, 2016), est donnée comme suit :

$$f_c(r_{ij}) = \begin{cases} 1 + \gamma \cdot \left(\frac{r_{ij}}{R_c}\right)^{\gamma+1} - (1 + \gamma) \cdot \left(\frac{r_{ij}}{R_c}\right)^{\gamma} & \text{pour } r_{ij} \leq R_c \\ 0 & \text{pour } r_{ij} \geq R_c \end{cases} \quad (1.37)$$

Le paramètre γ peut être ajusté pour retarder la décroissance et permettre à l'utilisateur de contrôler le comportement de la fonction de coupure. Le réseau de neurones artificiels (ANN) et AENET sont implémentés avec ces types de descripteurs, tandis que le *atomistic machine learning potential* (AMP) est construit à partir de trois types de descripteurs, et son package correspondant permet aux utilisateurs de choisir le descripteur adapté à leur situation. Les descripteurs dans le package AMP incluent des fonctions radiales et angulaires gaussiennes détaillées ci-dessus, les moments de Zernike et les fonctions de bispectre. Les composants du descripteur 3D-Zernike G_i sont calculés à partir du moment 3D-Zernike de la fonction de densité

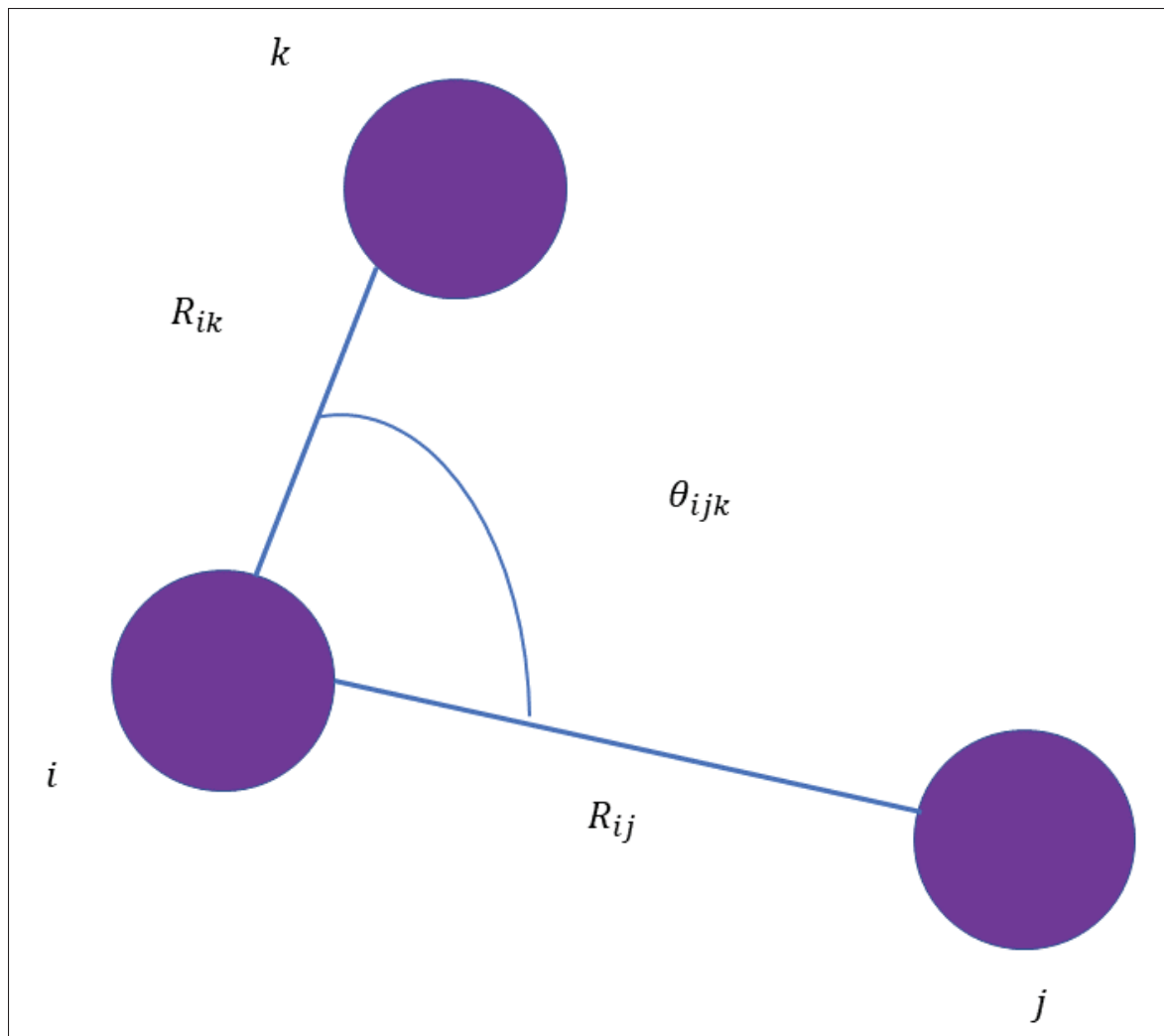


Figure 1.14 Angle du triplet d'atomes

atomique $\rho_i(r)$ du voisinage local de chaque atome i (Khorshidi & Peterson, 2016) :

$$\rho_i(r) = \sum_{j \neq i} \eta_j \delta(r - r_{ij}) f_c(|r_{ij}|) \quad (1.38)$$

où η_j est un paramètre d'invariance de permutation qui permet à l'utilisateur de spécifier le même η_j pour tous les atomes chimiquement équivalents, δ est la distribution de Dirac et r désigne la coordonnée sphérique de tout point à l'intérieur de la sphère unité du voisinage local. Le voisinage atomique local est transformé en une 2-sphère (S^2) de rayon unité centrée sur

l'atome central i , de sorte que pour tout point du voisinage local, $\sum_{i=1}^3 x_i^2 = 1$. En rappel, une 2-sphère est une sphère ordinaire à deux dimensions dans l'espace euclidien tridimensionnel. La

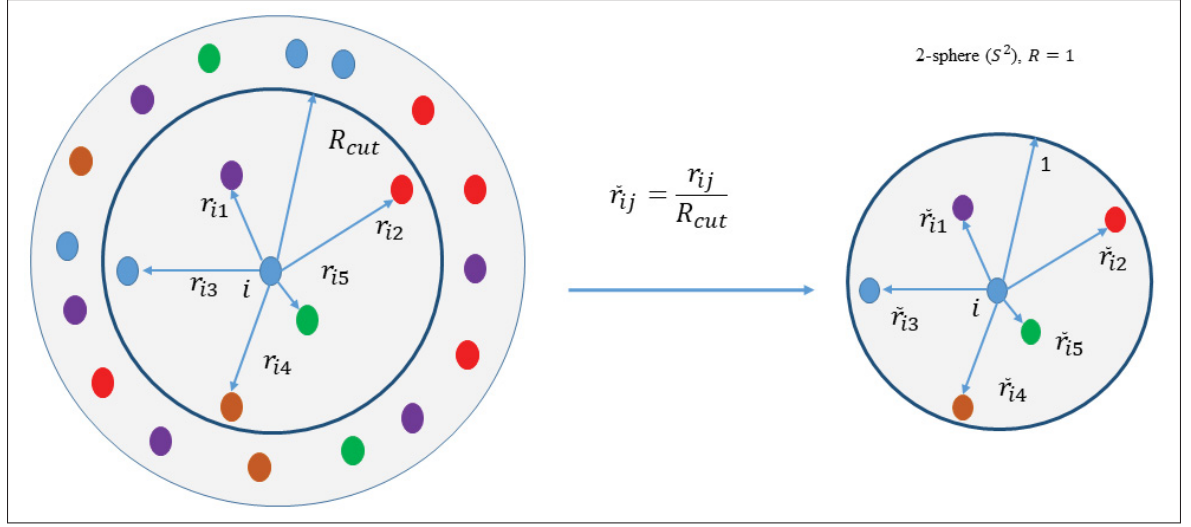


Figure 1.15 Transformation du voisinage atomique local

sphère de rayon unité, qui est maintenant le nouvel environnement atomique local de l'atome i , définit l'espace à l'intérieur du rayon de coupure R_{cut} et est sous le système de coordonnées sphériques (r, θ, φ) . La fonction de densité atomique $\rho_i(r)$ est ensuite mise à l'échelle pour s'adapter à la sphère unité, donnant ainsi une nouvelle densité atomique $\rho_i(r, \theta, \varphi)$ qui est développée en une base de fonctions de Zernike 3D (Canterakis, 1999; Novotni *et al.*, 2004; Khorshidi & Peterson, 2016) comme suit :

$$\rho_i(\tilde{r}, \theta, \varphi) = \sum_{i=0}^N \sum_{l=0}^n \sum_{m=-l}^l C_{nl}^m Z_{nl}^m(\tilde{r}, \theta, \varphi) \quad (1.39)$$

Les sommes sur l sont restreintes à des $n - l$ pairs avec n tronqué à une valeur N pour des raisons de calcul, et C_{nl}^m sont les moments de Zernike 3D :

$$C_{nl}^m = \langle Z_{nl}^m(\tilde{r}, \theta, \varphi), \rho_i(\tilde{r}, \theta, \varphi) \rangle \quad (1.40)$$

Les fonctions de base Z_{nl}^m formant l'ensemble de bases de Zernike 3D (Z) sont définies à l'intérieur et sur la surface de S^2 ; elles sont le produit du polynôme de Zernike $R_n^l(\tilde{r})$, qui est complet et orthogonal à l'intérieur de la sphère unité, et des harmoniques sphériques $Y_l^m(\theta, \varphi)$, qui sont complets et orthogonaux sur la surface de la sphère unité.

$$Z_{nl}^m(\tilde{r}, \theta, \varphi) = R_n^l(\tilde{r})Y_l^m(\theta, \varphi) \quad (1.41)$$

Les composants $||C_{nl}||$ du descripteur de Zernike G_i (G_i est le descripteur pour chaque atome i en partant de la densité donnée par l'équation 1.39) sont calculés à partir des moments de Zernike C_{nl}^m . Contrairement au descripteur gaussien donné ci-dessus, l'empreinte de Zernike G_i est sensible aux interactions entre quatre atomes.

Le troisième descripteur utilisé dans le package AMP est le descripteur de bispectrum (Khorshidi & Peterson, 2016). Tout d'abord, l'environnement atomique local est transformé en une 3-sphère (S^3) de rayon unité (pour tout point de l'environnement local, $X \in \mathbb{R}^4$, $\sum_{i=1}^4 x_i^2 = 1$) avec le système de coordonnées hypersphériques ($\tilde{r} = 1, \psi, \theta, \phi$). Ensuite, la fonction de densité atomique est projetée sur la sphère S^3 de rayon unité et est ensuite développée en une base d'harmoniques sphériques.

$$\rho_i(\psi, \theta, \phi) = \sum_{j,m',m} C_{m',m}^j U_{m',m}^j(\psi, \theta, \phi) \quad (1.42)$$

avec

$$C_{m',m}^j = \langle U_{m',m}^j(\psi, \theta, \phi), \rho_i(\psi, \theta, \phi) \rangle \quad (1.43)$$

À partir des coefficients $C_{m',m}^j$, le bispectrum est déduit :

$$B_{j,j_1,j_2} = \sum_{m'_1, m_1 = -j_1}^{j_1} C_{m'_1, m_1}^{j_1} \sum_{m'_2, m_2 = -j_2}^{j_2} C_{m'_2, m_2}^{j_2} \sum_{m', m = -j}^j C_{m, m_1, m_2}^{j, j_1, j_2} C_{m', m'_1, m'_2}^{j, j_1, j_2} (C_{m', m}^j) \quad (1.44)$$

où $C_{m, m_1, m_2}^{j, j_1, j_2}$ sont les coefficients de Lebesgue–Gordan. Les composants de B_{j, j_1, j_2} sont invariants par permutation des atomes et rotation de l'espace 4D ainsi que de l'espace 3D original. Les

éléments du descripteur G_i (le descripteur G_i est calculé pour chaque atome i individuellement en partant de la densité donnée par l'équation 1.42) sont les composants du bispectrum calculés à partir de l'équation 1.44. Ce vecteur descripteur G_i implémenté dans le package AMP est également sensible aux interactions entre quatre atomes.

Un autre potentiel interatomique basé sur l'apprentissage automatique récemment développé en 2018 utilisant les réseaux de neurones comme régressors est le DeePMD (Wang *et al.*, 2018). Contrairement à tous les autres MLIP qui reposent sur des fonctions spéciales comme des descripteurs d'environnement atomique, l'approche DeePMD utilise une transformation du système de coordonnées pour mapper la position atomique sur le descripteur. Une base orthonormale $[e_{i1}, e_{i2}, e_{i3}]$ centrée sur l'atome i est construite à partir de la position relative ($r_{ij} = x_{0ij}e_{01} + y_{0ij}e_{02} + z_{0ij}e_{03}$) de deux atomes différents tirés du voisinage de l'atome, où $[e_{01}, e_{02}, e_{03}]$ est le système de coordonnées cartésiennes. Ensuite, la position relative dans la base orthonormale est donnée comme suit :

$$(x_{ij}, y_{ij}, z_{ij}) = (x_{0ij}, y_{0ij}, z_{0ij})[e_{i1}, e_{i2}, e_{i3}] \quad (1.45)$$

En fait, $[e_{i1}, e_{i2}, e_{i3}]$ est la matrice de rotation construite à partir de deux atomes indexés $a(i)$ et $b(i)$ de l'environnement local de l'atome i .

$$\mathcal{R}(r_{ia(i)}, r_{ib(i)}) = [e_{i1}, e_{i2}, e_{i3}] \quad (1.46)$$

Le descripteur D_i de l'environnement local de l'atome i est le suivant :

$$D_i = D_i(r_i, \{r_j \mid j \in N_{R_c(i)}\}) \quad (1.47)$$

où $N_{R_c(i)}$ est l'ensemble des indices de tous les voisins de l'atome i dans le rayon de coupure R_c . Les composants de D_i sont donnés par :

$$\{D_{ij}^\alpha\} = \{a\{D_{ij}^\alpha\}\} \quad (1.48)$$

Parties radiales et angulaires :

$$\{D_{ij}^\alpha\} = \begin{cases} \left\{ \frac{1}{r_{ij}}, \frac{x_{ij}}{r_{ij}}, \frac{y_{ij}}{r_{ij}}, \frac{z_{ij}}{r_{ij}} \right\} & \text{Partie radiale et angulaire} \\ \left\{ \frac{1}{r_{ij}} \right\} & \text{Partie radiale} \end{cases} \quad (1.49)$$

Le paramètre $\alpha = 0, 1, 2, 3$ permet de choisir entre les parties radiales, angulaires, ou les deux. Par exemple, lorsque α vaut 0, seule la partie radiale est fournie. La $\{D_{ij}^\alpha\}$ est construite selon l'ordre de l'indice j en fonction des espèces chimiques, de sorte que la symétrie de permutation est ainsi préservée. Par conséquent, le descripteur D_i est invariant par rotation, translation et permutation des atomes chimiquement équivalents.

1.2.10 Régression par processus gaussien (GPR)

1.2.10.1 Potentiel par approximation gaussienne (GAP : Gaussian Approximation Potential)

Le potentiel par approximation gaussienne (Gaussian Approximation Potential) utilise la régression par processus gaussien (GPR) comme régresseur. Le modèle GAP exprime l'énergie totale d'un système donné comme la somme du potentiel de paire et des termes à plusieurs corps. Les termes à plusieurs corps sont donnés par une somme linéaire sur les fonctions de base Kernel.

$$E_{tot} = \sum_{i < j} V_{pair}(r_{ij}) + \sum_i \sum_s^M \alpha_s K(r_i, r_s) \quad (1.50)$$

où M désigne la taille des atomes représentatifs, r_i est la collection des distances relatives à l'atome central i du voisinage, et K est le noyau SOAP calculé à partir de l'expansion de la fonction de densité atomique $\rho_i(r)$ dans les fonctions de base des harmoniques sphériques $Y_{lm}(\tilde{r})$ et les fonctions de base radiales $g_n(r)$.

$$\rho_i(r) = \sum_{i'} f_{cut}(r_{ii'}) e^{-(r-r_{ii'})/2\sigma_{atom}^2} \quad (1.51)$$

$$\rho_i(r) = \sum_{nlm} Y_{lm}(\tilde{r}) g_n(r) \quad (1.52)$$

Les coefficients α_s sont généralement calculés à partir des données DFT et peuvent être compris comme des paramètres d'ajustement.

1.2.11 Régression ridge noyau (KRR : Kernel Ridge Regression)

1.2.11.1 Champ de force adaptatif, généralisable et informé par le voisinage (AGNI : Adaptive, Generalizable, and Neighborhood Informed)

La KRR non linéaire est celle implémentée dans les champs de forces adaptatifs, généralisables et informés par le voisinage (AGNI). Contrairement aux autres MLIP, le processus KRR apprend d'abord les forces et déduit les énergies atomiques. La KRR est une mesure de similarité entre une configuration atomique donnée et les données de référence. Une interpolation est effectuée pour déterminer les composantes des forces atomiques en comparant le descripteur de l'atome avec ceux des données de référence. Le descripteur directionnel de l'environnement atomique local est donné comme suit :

$$V_i^u(\eta) = \sum_{j \neq i} \frac{r_{ij}^u}{r_{ij}} e^{-(r_{ij}/\eta)^2} f_d(r_{ij}) \quad (1.53)$$

où η est la largeur de la fonction gaussienne, u est la direction le long de laquelle les coordonnées relatives r_{ij} sont projetées pour calculer le descripteur, et f_d est une fonction d'atténuation similaire à la fonction de coupure f_c détaillée ci-dessus. En raison de la sommation, le descripteur est insensible à la permutation des atomes dans le voisinage ou à la translation de l'environnement local. La symétrie de rotation est préservée par une matrice de rotation qui agit de manière identique sur les descripteurs et les forces. La composante de la force atomique le long de la direction u est interpolée comme suit :

$$F_i^u = \sum_t \alpha_t e^{-\frac{(d_{i,t}^u)^2}{2t^2}} \quad (1.54)$$

où N_t désigne le nombre total d'environnements atomiques de référence, chacun étiqueté par t , $d_{i,t}^u$ est la distance euclidienne entre deux descripteurs dans la direction u , et les paramètres α et l sont déterminés lors de l'entraînement.

1.2.12 Régression linéaire

À ce jour, quelques MLIP ont été développés avec les méthodes de régression linéaire. Le Potentiel par analyse du spectre du voisinage (spectral neighbor analysis potential - SNAP) et le potentiel par tenseur de moment (moment tensor potential - MTP) sont des exemples.

1.2.12.1 Le potentiel d'analyse spectrale des voisins (SNAP : Spectral Neighbor Analysis Potential)

Dans ce modèle, les descripteurs sont les composants du bispectre mentionnés ci-dessus. Ces composants sont calculés à partir des fonctions de densité atomique locale de l'atome central i , similaires à celles exprimées précédemment.

$$\rho_i(r) = \delta(r) + \sum_{r_{ij} < R_{cut}} f_c(r_{ij}) \omega_j \delta(r - r_{ij}) \quad (1.55)$$

L'énergie atomique est ensuite développée linéairement comme suit :

$$E_{SNAP}^i(B^i) = \beta_0^{\alpha_i} + \sum_{k=1}^K \beta_k^{\alpha_i} B_k^i \quad (1.56)$$

où α_i et $\beta_k^{\alpha_i}$ représentent respectivement le type chimique et le coefficient linéaire de l'atome i . Les B_k^i sont les composants du bispectre $B^i = \{B_1^i, B_2^i, \dots, B_K^i\}$. Comme mentionné précédemment, les composants du bispectre satisfont aux transformations isométriques de l'environnement local. La plupart de ces potentiels interatomiques basés sur l'apprentissage automatique, y compris le potentiel de moment que nous visons à développer pour les systèmes d'argile, traitent des systèmes simples. Ils ne sont pas adaptés aux systèmes très complexes tels que les argiles. Compte tenu du coût computationnel et de l'efficacité du descripteur, certains de ces potentiels

sont meilleurs que d'autres, mais tous doivent être redessinés ou modifiés pour des matériaux complexes.

1.2.12.2 Potentiel par tenseur de moments (MTP : Moment Tensor Potential)

L'idée provient du fait que toute fonction potentielle lisse V peut être approximée par des polynômes $p(u)$.

$$V(u) \approx p(u) \quad (1.57)$$

Le polynôme $p(u)$ n'est pas initialement un polynôme symétrisé, mais il peut être symétrisé pour satisfaire toutes les symétries requises. Le polynôme symétrisé peut être développé sur une base de polynômes. Ainsi, si nous notons $p(u)^*$ comme le polynôme symétrisé, nous pouvons l'expanser sur une base de polynômes $B_\beta(u)$:

$$p(u)^* = \sum_{\beta} c_{\beta} B_{\beta}(u) \quad (1.58)$$

L'ensemble des polynômes doit être lisse et invariant par permutation, rotation et réflexion. Ainsi, toute fonction V de l'environnement atomique local peut être approximée par ces polynômes invariants. Par conséquent, la fonction potentielle locale V peut être exprimée comme une combinaison linéaire de fonctions de base polynomiales :

$$V(u) \approx \sum_{\beta} c_{\beta} B_{\beta}(u) \quad (1.59)$$

Les coefficients c_{β} sont des paramètres linéaires généralement déterminés au cours de la phase d'entraînement à partir des données DFT, et u est un tuple, la collection des coordonnées relatives à l'atome central i de l'environnement local défini dans R_{cut} , comme illustré dans la figure 1.16 ci-dessous.

En considérant l'environnement atomique de la figure 1.16 ci-dessus, l'énergie totale, telle que donnée dans l'équation 1.27, peut être réécrite en suivant la propriété de partitionnement.

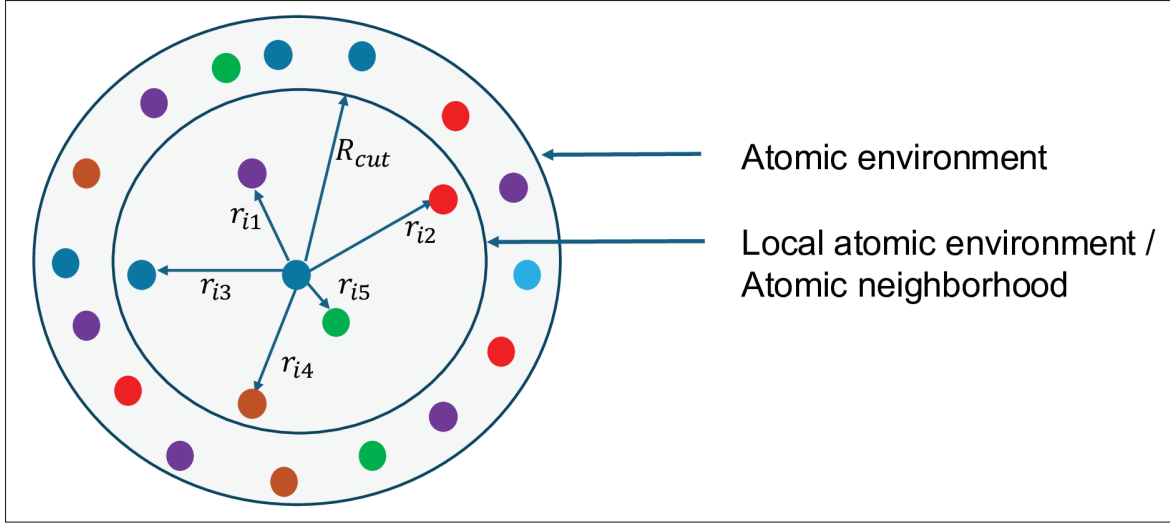


Figure 1.16 Délimitation illustrative de l'environnement atomique local pour une configuration donnée (x)

$$E(x) = \sum_i V_{local}(\xi_i) \quad (1.60)$$

À partir de la forme générale de l'équation (1.41), nous pouvons exprimer V_{local} comme suit :

$$V_{local}(\xi_i) = \sum_{\beta} c_{\beta} B_{\beta}(\xi_i) \quad (1.61)$$

où $\xi_i = \{(r_{ij}, \tau_i, \tau_j)\}$ contient la position relative de l'atome j par rapport à l'atome i , τ_i et τ_j définissent les types d'atomes représentés par différentes couleurs dans la figure 1.16. Les fonctions de base polynomiales B_{β} résultent de la contraction du tenseur de moments qui décrit l'environnement atomique local. Les descripteurs de tenseur de moments ont les formes suivantes respectivement pour un système à composant unique ($\tau_i = \tau_j$), $u_i = \{r_{ij}\}$ et un système multicomposant $\{\xi_i\}$ (Shapeev, 2016; Gubaev *et al.*, 2018) :

$$M_{\mu,\nu}(u_i) = \sum_j f_{\mu}(|r_{ij}|) \underbrace{r_{ij} \otimes \cdots \otimes r_{ij}}_{\nu \text{ fois}} \quad (1.62)$$

et,

$$M_{\mu,\nu}(\xi_i) = \sum_j f_{\mu}(|r_{ij}|, \tau_i, \tau_j) \underbrace{r_{ij} \otimes \cdots \otimes r_{ij}}_{\nu \text{ fois}} \quad (1.63)$$

où $f_{\mu}(|r_{ij}|, \tau_i, \tau_j)$, qui dépend uniquement de la distance interatomique et du type d'atome, représente la distribution radiale, et le terme $r_{ij} \otimes \cdots \otimes r_{ij}$ est un tenseur de rang ν incluant l'information angulaire du voisinage local. Ce terme ressemble à un moment d'inertie du voisinage de l'atome i , d'où le nom de "Moment Tensor Potential" pour cette approche. Par construction, $M_{\mu,\nu}(\xi_i)$ est invariant par translation du système et permutation des atomes chimiquement équivalents (Gubaev *et al.*, 2019, 2018). Les invariances de rotation et de réflexion résultent de la contraction de (Gubaev *et al.*, 2019). Par conséquent, le potentiel local résultant V_{local} possède toutes les symétries nécessaires. De plus, à partir de l'équation 1.63, l'expression de V_{local} dépend des coordonnées relatives r_{ij} et met en évidence uniquement les interactions à deux corps, mais la contraction de $M_{\mu,\nu}(\xi_i)$ peut contenir des termes d'interactions à plusieurs corps d'ordre supérieur (Gubaev *et al.*, 2019; Mueller *et al.*, 2020). Le modèle de la méthode du potentiel à tenseur de moments est illustré dans la figure 1.17 ci-dessous. Le modèle local est une fonction analytique résultant de la construction de la base polynomiale. Le paramètre linéaire c_{β} , ainsi que d'autres paramètres issus de la construction des fonctions de base, forment l'ensemble des paramètres du modèle local qui peuvent être notés par θ . Notez que θ est déterminé au cours du processus d'ajustement avec les données DFT. En résumé, la tâche principale consiste à construire une fonction analytique par tenseur de moments contenant des paramètres inconnus qui peuvent être déterminés par ajustement linéaire aux données de référence.

- Les polynômes de base B_{β}

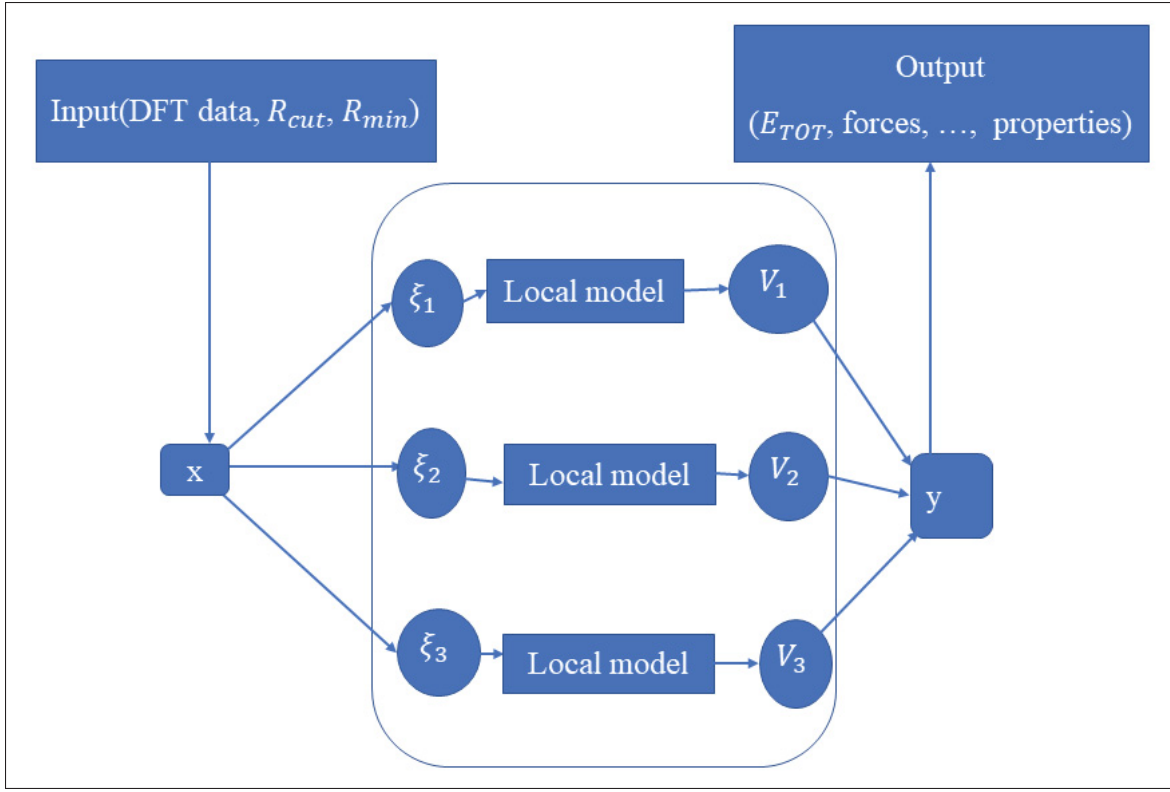


Figure 1.17 Schéma de fonctionnement du potentiel par tenseur de moment. Les V ($V_1, V_2, \dots$) et ξ ($\xi_1, \xi_2, \dots$) représentent les potentiels et les paramètres locaux respectivement

Les fonctions de base sont déterminées par la contraction de $M_{\nu,\mu}$ définie ci-dessus.

$$B_{\beta}(u) = \beta \prod_{i=1}^k M_{\beta_{ii}\beta'_i}(u) \quad (1.64)$$

où $\prod$ est un opérateur de contraction agissant sur les tenseurs : pour un tenseur $T^{(i)} \in (\mathbb{R}^d)^{\beta'_i}$ avec d représentant la dimension de l'espace et β'_i étant la somme des éléments hors diagonale d'une matrice symétrique $k \times k$ de nombres entiers positifs ($\beta_{ij} \geq 0, k \in \mathbb{N} = 0, 1, \dots, n$).

$$\beta'_i = \sum_{\substack{j=1 \\ j \neq i}}^k \beta_{ij} \quad (1.65)$$

L'application de l'opérateur $\prod$ sur le tenseur $T^{(i)}$ donne :

$$\beta \prod_{i=1}^k T^{(i)} = \sum_{\alpha \in \mathcal{A}} \prod_{i=1}^k T^{(i)}_{\alpha_1^{(1,i)} \dots \alpha_{\beta_{1i}}^{(1,i)} \dots \alpha_1^{(i-1,1)} \dots \alpha_{\beta_{i-1,i}}^{(i-1,i)} \alpha_1^{(i,i+1)} \dots \alpha_{\beta_{i+1,i}}^{(i+1,i)} \dots \alpha_1^{(i,k)} \dots \alpha_{\beta_{i,k}}^{(i,k)}} \quad (1.66)$$

où $\mathcal{A} = (\{1 \dots d\}^{\beta_{ij}})$, $(\alpha_1^{(i,j)} \dots \alpha_{\beta_{ij}}^{(i,j)}) \in \mathcal{A}$, $(i, j) \in \mathcal{M}$ et $\mathcal{M} = \{(i, j) \in \{1, \dots, m\}^2, i < j\}$.

L'opérateur $\prod$ a deux effets sur les tenseurs : d'une part, il multiplie leurs composants, et d'autre part, il les additionne. Comme le montre l'équation 1.66, les indices $\alpha_{\beta_{i,k}}^{(i,k)}$ sont les indices des éléments du tenseur, formé à partir d'une collection de multi-indices α et β est la matrice symétrique $k \times k$ mentionnée ci-dessus, permettant de déterminer les fonctions de base B_β .

Illustration

Prenons $k = 2$ et β comme suit :

$$\beta = \begin{pmatrix} \mu_1 & 1 \\ 1 & \mu_2 \end{pmatrix}$$

avec $\beta'_1 = \beta'_2 = 1$, $\mathcal{A} = (\{1 \dots 2\}^1) = (1, 2)$, et $\mathcal{M} = (1, 2)$, où $1 \leq i < j \leq k$.

- Pour $i = 1$, $\alpha_1^{(1,2)} \dots \alpha_1^{(1,2)}$, un composant du tenseur $M_{\beta_{11}1}$.
- Pour $i = 2$, $\alpha_1^{(1,2)} \dots \alpha_{\beta_{12}}^{(1,2)} \dots \alpha_1^{(1,2)} \alpha_{\beta_{12}}^{(1,2)} \dots = \alpha_1^{(1,2)}$ puisque $\beta_{12} = 1$, ce qui donne à nouveau un composant unique.

En utilisant les équations 1.64 et 1.66, d'abord :

$$\prod_{i=1}^2 M_{\beta_{ii}1}(u)_{\alpha_1^{(1,2)} \alpha_1^{(1,2)}} = (M_{\beta_{11}1}(u))_{\alpha_1^{(1,2)}} \cdot (M_{\beta_{22}1}(u))_{\alpha_1^{(1,2)}}$$

Puis :

$$B_\beta(u) = \sum_{\alpha_1^{(1,2)}=1}^{d=2} (M_{\beta_{11}1}(u))_{\alpha_1^{(1,2)}} \cdot (M_{\beta_{22}1}(u))_{\alpha_1^{(1,2)}}$$

$$B_\beta(u) = (M_{\beta_{11}}(u))_1 \cdot (M_{\beta_{22}}(u))_1 + (M_{\beta_{11}}(u))_2 \cdot (M_{\beta_{22}}(u))_2$$

$$B_\beta(u) = (M_{\mu_1}(u))_1 \cdot (M_{\mu_2}(u))_1 + (M_{\mu_1}(u))_2 \cdot (M_{\mu_2}(u))_2$$

Il s'agit du produit scalaire de $M_{\mu_1}(u)$ et $M_{\mu_2}(u)$ en 2 dimensions, donc :

$$B_\beta(u) = M_{\mu_1}(u) \cdot M_{\mu_2}(u)$$

Lorsque $\nu = 1$, la partie angulaire du descripteur est un vecteur ($\vec{r}_{ij}$) et l'équation 1.63 devient :

$$M_{\mu,\nu}(\xi_i) = \sum_j f_\mu(|r_{ij}|, \tau_i, \tau_j) \vec{r}_{ij} \quad (1.67)$$

Gardons la même valeur de k avec un β différent :

$\begin{pmatrix} \mu_1 & 2 \\ 2 & \mu_2 \end{pmatrix}$, $\beta'_1 = \beta'_2 = 2$, $\mathcal{A} = (\{1 \cdots 2\}^2) = \{(1, 1), (1, 2), (2, 1), (2, 2)\}$ Puisque $i < j$, nous avons toujours $\mathcal{A} = (1, 2)$.

- $i = 1$, $\alpha_1^{(1,2)} \cdots \alpha_{\beta_{12}}^{(1,2)} = \alpha_1^{(1,2)}, \alpha_{\beta_{12}}^{(1,2)}$; nous avons deux composants du tenseur $M_{\beta_{ii}2}$.
- $i = 2$, $\alpha_1^{(1,2)} \cdots \alpha_{\beta_{12}}^{(1,2)} \cdots \alpha_1^{(1,2)} \cdots \alpha_{\beta_{12}}^{(1,2)} \cdots$, $1 \leq i < j \leq k$ ($k = 2$), cela conduit à deux composants :

$$\alpha_1^{(1,2)} \cdots \alpha_{\beta_{12}}^{(1,2)} \cdots \alpha_1^{(1,2)} \cdots \alpha_{\beta_{12}}^{(1,2)} \cdots = \alpha_1^{(1,2)}, \alpha_2^{(1,2)}$$

$$\prod_{i=1}^2 M_{\beta_{ii}2}(u)_{\alpha_1^{(1,2)} \alpha_2^{(1,2)}} = (M_{\mu_1 2}(u))_{\alpha_1^{(1,2)} \alpha_2^{(1,2)}} \cdot (M_{\mu_2 2}(u))_{\alpha_1^{(1,2)} \alpha_2^{(1,2)}}$$

$$B_{\beta}(u) = \sum_{\alpha_2^{(1,2)}=1}^{d=2} \sum_{\alpha_1^{(1,2)}=1}^{d=2} (M_{\mu_1 2}(u))_{\alpha_1^{(1,2)} \alpha_2^{(1,2)}} \cdot (M_{\mu_2 2}(u))_{\alpha_1^{(1,2)} \alpha_2^{(1,2)}}$$

$$B_{\beta}(u) = ((M_{\mu_1 2}(u))_{(1,1)} \cdot (M_{\mu_2 2}(u))_{(1,1)} + (M_{\mu_1 2}(u))_{(2,1)} \cdot (M_{\mu_2 2}(u))_{(2,1)}) + ((M_{\mu_1 2}(u))_{(1,2)} \cdot (M_{\mu_2 2}(u))_{(1,2)} + (M_{\mu_1 2}(u))_{(2,2)} \cdot (M_{\mu_2 2}(u))_{(2,2)})$$

$B_{\beta}(u)$ est la somme du produit scalaire composant par composant de deux matrices $(M_{\mu_1 2}(u))$ et $(M_{\mu_2 2}(u))$, c'est-à-dire la somme du produit des composants correspondants de deux matrices, ce qui correspond au produit scalaire de Frobenius (Horn *et al.*, 2012). Par conséquent,

$$B_{\beta}(u) = (M_{\mu_1 2}(u)) : (M_{\mu_2 2}(u))$$

La partie angulaire est maintenant $r_{ij} \otimes r_{ij}$ ($\nu = 2$),

$$M_{\mu, \nu}(\xi_i) = \sum_j f_{\mu}(|r_{ij}|, \tau_i, \tau_j) r_{ij} \otimes r_{ij} \quad (1.68)$$

En 2D, $r_{ij} = x_{ij}e_1 + y_{ij}e_2$, ainsi $r_{ij} \otimes r_{ij} = (x_{ij}e_1 + y_{ij}e_2) \otimes (x_{ij}e_1 + y_{ij}e_2) = r_{ij}r_{ij}^T$

$$r_{ij} \otimes r_{ij} = \begin{pmatrix} x_{ij}^2 & x_{ij}y_{ij} \\ x_{ij}y_{ij} & y_{ij}^2 \end{pmatrix}_{e_p \otimes e_q} \quad (1.69)$$

La partie angulaire est maintenant

avec $p = 1, 2$ pour l'indice de ligne et $q = 1, 2$ pour l'indice de colonne de la matrice ; $\{e_p\}$ est une base de $\mathbb{R}^2$ et $e_p \otimes e_q$ une base de $\mathbb{R} \otimes \mathbb{R}$.

Passant aux coordonnées polaires

$$\begin{cases} x = r \cos \theta \\ y = r \sin \theta \end{cases}$$

cela conduit à

$$r_{ij} \otimes r_{ij} = r_{ij}^2 \begin{pmatrix} \cos^2 \theta & \cos \theta \sin \theta \\ \cos \theta \sin \theta & \sin^2 \theta \end{pmatrix}$$

Maintenant, faisons une rotation de l'environnement atomique local (rotations de la configuration) par un angle ϕ .

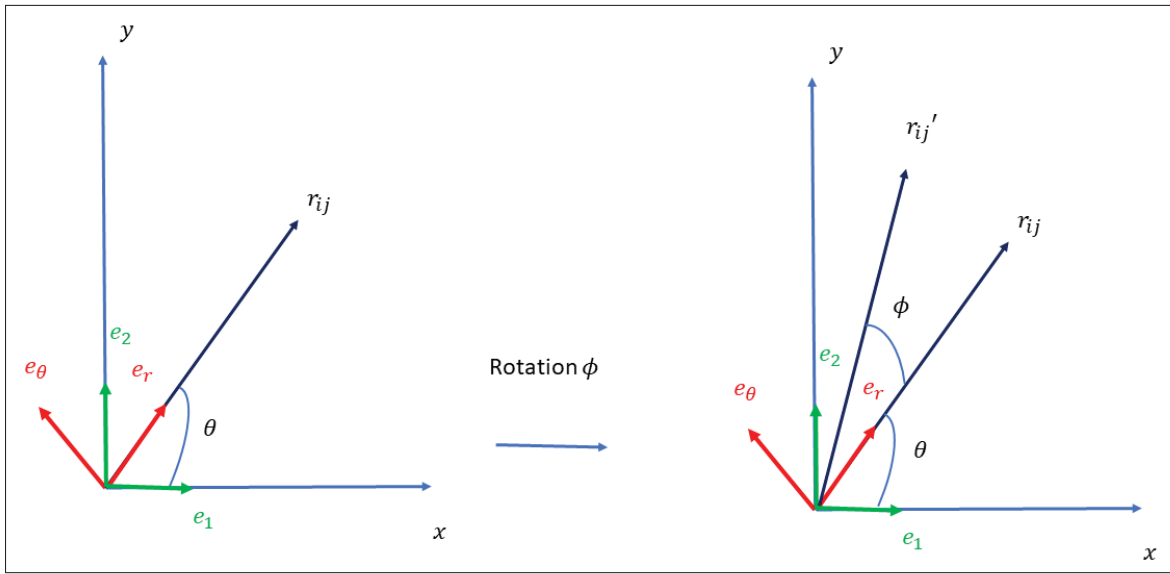


Figure 1.18 Rotation de système de coordonnées. Les r_{ij} représente des vecteurs issus des coordonnées relatives, les e_i sont les vecteurs de la base et θ l'angle entre les r_{ij} et e_i .
L'angle de rotation est donné par ϕ

Les vecteurs r'_{ij} sont calculés avec la matrice de rotation $\mathcal{R}$.

$$r'_{ij} = \mathcal{R} r_{ij} = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} r_{ij} \text{ et } r'_{ij} \otimes r'_{ij} = r_{ij} \otimes r_{ij} \text{ conservent leur invariance par rotation.}$$

$r'_{ij}{}^T = (\mathcal{R} \times r_{ij})^T = r_{ij}^T \mathcal{R}^T$, donc $r'_{ij} \otimes r'_{ij} = \mathcal{R} r_{ij} r_{ij}^T \mathcal{R}^T = \mathcal{R} r_{ij} \otimes r_{ij} \mathcal{R}^T$. Le terme $r_{ij} \otimes r_{ij}$ est un tenseur et $\mathcal{R}$ une matrice scalaire, nous pouvons commuter les opérations. Ainsi,

$$r'_{ij} \otimes r'_{ij} = \mathcal{R} \mathcal{R}^T r_{ij} \otimes r_{ij} \text{ et } \mathcal{R} \mathcal{R}^T = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} \cdot \begin{pmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$$

Par conséquent,

$$r'_{ij} \otimes r'_{ij} = r_{ij} \otimes r_{ij} = r_{ij}^2 \begin{pmatrix} \cos^2 \theta & \cos \theta \sin \theta \\ \cos \theta \sin \theta & \sin^2 \theta \end{pmatrix}$$

Comme dernière illustration, considérons une matrice en 3D, $k = 3$, $\beta = \begin{pmatrix} \mu_1 & 1 & 1 \\ 1 & \mu_2 & 0 \\ 1 & 0 & \mu_3 \end{pmatrix}$
 $\beta'_1 = 2, \beta'_2 = 1, \beta'_3 = 1, \mathcal{A} = (\{1 \cdots 3\}^1, \beta_{12} = 1, \{1 \cdots 3\}^1, \beta_{13} = 1, \{1 \cdots 3\}^0, \beta_{23} = 0) = ((1, 2, 3), \beta_{12}, (1, 2, 3), \beta_{13}, (1), \beta_{23}), i < j$

• $i = 1, \alpha_1^{(1,2)} \cdots \alpha_{\beta_{12}}^{(1,2)} \cdots \alpha_1^{(1,3)} \cdots \alpha_{\beta_{13}}^{(1,3)} = \alpha_1^{(1,2)}, \alpha_1^{(1,3)}; \beta_{12} = 1 \text{ and } \beta_{13} = 1$ Cela signifie que nous avons deux composants.

• $i = 2, \alpha_1^{(1,2)} \cdots \alpha_{\beta_{12}}^{(1,2)} \cdots \alpha_1^{(1,2)} \cdots \alpha_{\beta_{12}}^{(1,2)} \alpha_1^{(2,3)} \cdots \alpha_{\beta_{23}}^{(2,3)} \cdots \alpha_1^{(2,3)} \cdots \alpha_{\beta_{23}}^{(2,3)}$ et $\forall (i, j), \alpha_1^{(i,j)}, \alpha_{\beta_{i,j}}^{(i,j)} \in \mathcal{A} = (\{1 \cdots d\}^{\beta_{ij}})$; cependant $\beta_{23} = 0 \rightarrow \mathcal{A} = (\{1 \cdots d\}^0) = \emptyset$, Ainsi, nous avons un composant $\alpha_1^{(1,2)} \cdots \alpha_{\beta_{12}}^{(1,2)} \cdots \alpha_1^{(1,2)} \cdots \alpha_{\beta_{12}}^{(1,2)} = \alpha_1^{(1,2)}$.

• $i = 3$, en suivant la même procédure que pour $i = 2$, nous obtenons un composant $\alpha_1^{(1,3)}$

$$\prod_{i=1}^3 M_{\beta_{ii}\beta_i'}(u)_{\alpha_1^{(1,2)} \cdots \alpha_1^{(1,3)}} = (M_{\beta_{11}\beta_1'}(u))_{\alpha_1^{(1,2)} \alpha_1^{(1,3)}} \cdot (M_{\beta_{22}\beta_2'}(u))_{\alpha_1^{(1,2)}} \cdot (M_{\beta_{33}\beta_3'}(u))_{\alpha_1^{(1,3)}}$$

$$B_\beta(u) = \sum_{\alpha_1^{(1,2)}=1}^{d=3} \sum_{\alpha_1^{(1,3)}=1}^{d=3} (M_{\mu_1 2}(u))_{\alpha_1^{(1,2)} \alpha_1^{(1,3)}} \cdot (M_{\mu_2 1}(u))_{\alpha_1^{(1,2)}} \cdot (M_{\mu_3 1}(u))_{\alpha_1^{(1,3)}}$$

Soit $m = \alpha_1^{(1,2)}$ et $n = \alpha_1^{(1,3)}$,

$$B_\beta(u) = \sum_{m=1}^{d=3} \sum_{n=1}^{d=3} (M_{\mu_1 2}(u))_{mn} \cdot (M_{\mu_2 1}(u))_m \cdot (M_{\mu_3 1}(u))_n$$

Nous savons que pour $\nu = 2$, $M_{\mu,\nu}$ est un tenseur de rang 2 et un vecteur si $\nu = 1$, comme le montre l'équation 1.62 ou 1.63. Donc, $B_\beta(u)$ est une contraction du tenseur $M_{\mu_1,2}(u)$ et de deux vecteurs $M_{\mu_2,1}(u)$ et $M_{\mu_3,1}(u)$.

Demonstration

Soit $A = A_{mn}e_m \otimes e_n$ un tenseur et $B = B_p e_p$ et $C = C_q e_q$ deux vecteurs. Notez que lorsque l'indice est répété, la convention de sommation s'applique.

$$\begin{aligned} (A \cdot B) \cdot C &= [(A_{mn}e_m \otimes e_n) \cdot B_p e_p] \cdot C_q e_q \\ &= [A_{mn}B_p \delta_{np} e_m] \cdot C_q e_q \\ &= A_{mn}B_n C_q (e_m \cdot e_q) \\ &= A_{mn}B_n C_q \delta_{mq} \\ &= A_{mn}B_n C_m \\ &= A_{nm}B_n C_m \end{aligned}$$

Cela est vrai en général puisque nous avons traité une double sommation avec les mêmes bornes inférieures et supérieures, l'ordre des indices de A n'a pas d'importance (les indices de A sont interchangeables). De plus, la nature symétrique des tenseurs $r_{ij} \otimes r_{ij}$ (voir équations 1.69 ou 1.70) permet d'écrire :

$$A_{mn}B_n C_m = A_{nm}B_n C_m$$

Donc,

$$B_\beta(u) = (M_{\mu_1 2}(u) \cdot M_{\mu_2 1}(u)) \cdot M_{\mu_3 1}(u)$$

- Limitation des fonctions de base $B_\beta(u)$

Reconsidérons l'équation 1.63 à nouveau :

$$M_{\mu,\nu}(\xi_i) = \sum_j f_\mu(|r_{ij}|, \tau_i, \tau_j) \underbrace{r_{ij} \otimes \dots \otimes r_{ij}}_{\nu \text{ times}}$$

Le vecteur de position relative r_{ij} en trois dimensions est $r_{ij} = (x_{ij}, y_{ij}, z_{ij})$. Si nous prenons $\nu = 2$, en étendant l'équation 1.69 à trois dimensions, la partie angulaire est :

$$r_{ij} \otimes r_{ij} = \begin{pmatrix} x_{ij}^2 & x_{ij}y_{ij} & x_{ij}z_{ij} \\ x_{ij}y_{ij} & y_{ij}^2 & y_{ij}z_{ij} \\ x_{ij}z_{ij} & y_{ij}z_{ij} & z_{ij}^2 \end{pmatrix}_{e_p \otimes e_q} \quad (1.70)$$

La partie radiale du descripteur est développée en fonctions de base radiales $Q^{(\alpha)}(|r_{ij}|)$.

$$f_\mu(|r_{ij}|, \tau_i, \tau_j) = \sum_{\alpha}^{N_Q} c_{\mu,\tau_i,\tau_j}^{(\alpha)} Q^{(\alpha)}(|r_{ij}|) \quad (1.71)$$

avec,

$$Q^{(\alpha)}(|r_{ij}|) = \begin{cases} \varphi^{(\alpha)}(|r_{ij}|)(R_{cut} - |r_{ij}|)^2 & \text{pour } r_{ij} < R_{cut} \\ 0 & \text{pour } r_{ij} \geq R_{cut} \end{cases} \quad (1.72)$$

où $Q^{(\alpha)}(|r_{ij}|)$ est l'ensemble des paramètres radiaux qui sont calculés durant l'ajustement, et $\varphi^{(\alpha)}$ sont des fonctions polynômiales. Le MTP est implementé avec des polynômes de Chebyshev $\varphi^{(\alpha)}$ définis sur l'intervalle $[R_{min}, R_{max}]$. R_{min} est introduit pour refléter le comportement répulsif des atomes ; il y a toujours une distance minimale entre les atomes dans un système donné. Pour illustrer certaines fonctions de base, nous utilisons les polynômes de Chebyshev du premier genre définis sur l'intervalle $[-1, 1]$ comme suit :

$$\varphi^{(0)}(r) = 1$$

$$\varphi^{(1)}(r) = r$$

$$\varphi^{(\alpha+1)}(r) = 2r\varphi^{(\alpha)}(r) - \varphi^{(\alpha-1)}(r) \text{ pour } \alpha \geq 1$$

Soit $R_{min} = 2 \text{ \AA}$ et $R_{max} = 5 \text{ \AA}$; l'intervalle $[R_{min}, R_{max}]$ est donc $[2, 5]$. Nous devons transformer cet intervalle en $[-1, 1]$. Cela peut être fait en trouvant une fonction g telle que $g([2, 5]) = [-1, 1]$. Il y a deux paires de points $(2, -1)$ et $(5, 1)$. Par conséquent, g est une fonction linéaire de la forme $g(r) = ar + b$. En résolvant pour $g(2) = -1$ et $g(5) = 1$, nous obtenons $a = \frac{2}{3}$ et $b = \frac{7}{3}$. Nous notons $\varphi_*^{(\alpha)}$ comme la transformation de $\varphi^{(\alpha)}$,

$$\varphi_*^{(\alpha)}(r) = \varphi^{(\alpha)}\left(\frac{2}{3}r - \frac{7}{3}\right), \text{ ainsi}$$

$$\varphi_*^{(0)}(r) = 1$$

$$\varphi_*^{(1)}(r) = \left(\frac{2}{3}r - \frac{7}{3}\right)$$

$$\varphi_*^{(\alpha+1)}(r) = 2\left(\frac{2}{3}r - \frac{7}{3}\right)\varphi_*^{(\alpha)}\left(\frac{2}{3}r - \frac{7}{3}\right) - \varphi_*^{(\alpha-1)}\left(\frac{2}{3}r - \frac{7}{3}\right) \text{ pour } \alpha \geq 1$$

Maintenant, nous pouvons tracer $Q^{(\alpha)}$ comme illustré dans la figure 1.19 ci-dessous. La mise

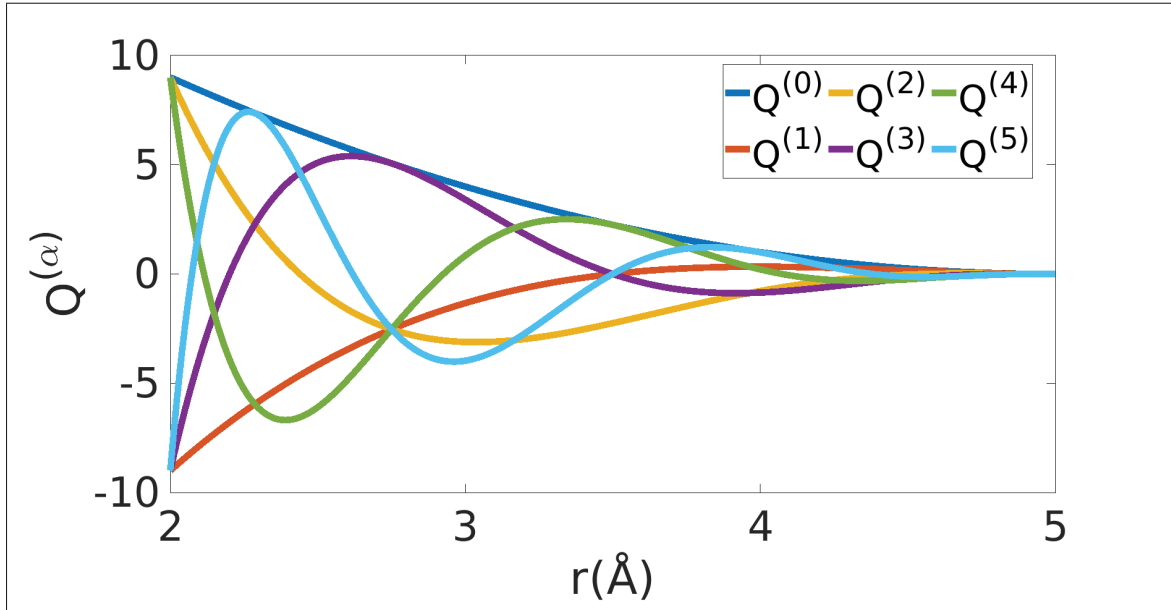


Figure 1.19 Les six premières fonctions de base radiales définies dans l'équation 1.72

en œuvre pratique nécessite que le nombre de fonctions de base soit fini. Ainsi, un niveau du

moment est défini comme suit :

$$\text{lev}M_{\mu,\nu} = 2 + 4\mu + \nu \quad (1.73)$$

où 2 et 4 sont des coefficients empiriques optimaux trouvés lors de tests effectués avant la mise en œuvre finale. Si la contraction implique de nombreux moments $M_{\mu_1,\nu_1}, M_{\mu_2,\nu_2}, \dots, M_{\mu_n,\nu_n}$, alors,

$$\text{lev}B_\beta = (2 + 4\mu_1 + \nu_1) + (2 + 4\mu_2 + \nu_2) \cdots (2 + 4\mu_n + \nu_n) \quad (1.74)$$

Ce nombre permet de limiter la taille de la base. Par exemple, $\text{lev}M_{0,0} = 2$; $\text{lev}M_{1,0} = 6$; $\text{lev}M_{0,1} = 3$; $\text{lev}M_{1,1} = 7$. En pratique, un niveau maximum $\text{lev}_{\max}$ est généralement choisi, et toutes les B_β avec un niveau inférieur ou égal à $\text{lev}_{\max}$ sont incluses dans le calcul.

$$\text{lev}B_\beta \leq \text{lev}_{\max} \quad (1.75)$$

À des fins d'illustration, fixons $\text{lev}_{\max}$ à 8 ($\text{lev}_{\max} = 8$). Nous allons ici trouver les fonctions de base satisfaisant la condition donnée par l'équation 1.75. B_β devrait être une fonction scalaire ; comme vous pouvez le voir à partir de l'équation 1.63 lorsque $\nu = 0$, $M_{\mu,0}$ est une fonction scalaire, on peut construire B_β à partir de n'importe quel nombre de $M_{\mu,0}$ par un simple produit ($B_\beta = M_{\mu,0} \times M_{\mu,0} \cdots$ ou $B_\beta = M_{\mu_1,0} \times M_{\mu_2,0} \cdots$) ; pour $\nu = 1$, $M_{\mu,1}$ est un vecteur et un produit scalaire doit être appliqué ($B_\beta = M_{\mu,1} \cdot M_{\mu,1} \cdots$ ou $B_\beta = M_{\mu_1,1} \cdot M_{\mu_2,1} \cdots$) et pour $\nu = 2, \dots$, la base du tenseur $M_{\mu,\nu}$ est $e_1 \otimes e_2 \otimes \cdots$ et au moins une double contraction est nécessaire. ν indique le nombre de contractions, $\nu = n$, impliquant n contractions, c'est-à-dire n produits scalaires de vecteurs de base. Ainsi, pour un $\text{lev}_{\max}$ donné, on doit combiner toutes les opérations possibles telles que le produit simple, le produit scalaire, les n -contractions pour obtenir des fonctions scalaires B_β dont le niveau satisfait l'équation 1.75. Notez que l'ordre des B_β n'a pas d'importance tant que tous les B_β satisfaisant l'équation 1.75 sont inclus (on peut prendre par exemple $B_5 = M_{0,0}$ et $B_1 = M_{0,2} : M_{0,2}$ et ainsi de suite $\cdots$).

$$B_1 = M_{0,0} \quad \text{lev}B_1 = 2$$

$$B_2 = M_{1,0} \quad \text{lev}B_2 = 6$$

$$\begin{array}{ll}
B_3 = M_{0,0} \times M_{0,0} = M_{0,0}^2 & lev B_3 = 4 \\
B_4 = M_{0,1} \cdot M_{0,1} & lev B_4 = 6 \\
B_5 = M_{0,2} : M_{0,2} & lev B_5 = 8 \\
B_6 = M_{0,0} \times M_{0,1} & lev B_6 = 8 \\
B_7 = M_{0,0}^3 & lev B_7 = 6 \\
B_8 = M_{0,0}^4 & lev B_8 = 8 \\
B_9 = M_{0,0} \times (M_{0,1} \cdot M_{0,1}) & lev B_9 = 8
\end{array}$$

Par exemple, de manière similaire à l'opération de contraction effectuée ci-dessus, B_5 peut être calculé en trois dimensions en utilisant l'opérateur de contraction avec

$$\beta = \begin{pmatrix} 0 & 2 & 0 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}$$

, $\nu = \sum_{i \neq j}^3 \beta_{ij} = 2, \mu = 0$ ce qui correspond à $M_{0,2}$.

En règle générale, β sert à composer les composants pour l'opérateur de contraction. En pratique, la contraction de $M_{\mu,\nu}$ en un scalaire est formulée par une matrice β symétrique $n \times n$ dont les éléments hors-diagonale indiquent le nombre de dimensions à contracter entre les tenseurs M_{lm} et M_{pq} . Les indices μ et ν permettent de limiter la taille de la base pour l'implémentation pratique et aussi d'estimer la complexité et l'efficacité computationnelles comme démontré dans (Shapeev, 2016; Novikov *et al.*, 2020a).

1.2.13 Labellisation des données

Comme nous l'avons mentionné précédemment, l'apprentissage automatique repose sur des informations historiques. Mathématiquement, cela entre dans le domaine de l'interpolation, où des données connues sont finalement nécessaires. Pour appliquer ce concept dans le domaine des potentiels interatomiques, il est nécessaire de préparer des données de référence à partir des calculs de la Théorie de la Fonctionnelle de Densité (DFT). La préparation de la base de

données consiste à identifier toutes les configurations possibles dans lesquelles le matériau peut exister, puis à calculer les propriétés correspondantes, principalement l'énergie, les forces et les contraintes, en utilisant la DFT. Soit $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ l'ensemble des configurations dans lesquelles le matériau peut exister, et $\{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}$ les propriétés correspondantes ($y^{(i)} = \{\text{Énergie, Forces, Contraintes}\}$) qui ne sont pas connues a priori. Une correspondance un à un de $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ avec les propriétés connues $\{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}$ est généralement effectuée par des calculs DFT. Pour chaque structure $x^{(i)}$ dans $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$, ses propriétés $y^{(i)}$ sont calculées à l'aide de la DFT comme l'illustre le diagramme 1.20 ci-dessous.

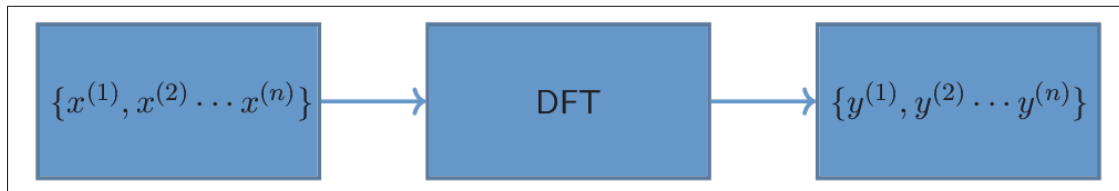


Figure 1.20 Principe de génération des données d'entraînement à partir des configurations atomiques

L'ensemble $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})\}$ est appelé données étiquetées, ce qui constitue l'une des caractéristiques principales de l'apprentissage supervisé.

Procédure d'ajustement

Tout d'abord, les données étiquetées doivent être séparées en deux parties. Une partie pour l'entraînement et l'autre pour la validation.

- Ensemble d'entraînement $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$
- Ensemble de validation $\{(x^{(m+1)}, y^{(m+1)}), (x^{(m+2)}, y^{(m+2)}), \dots, (x^{(n)}, y^{(n)})\}$, avec $m < n$.

Dans la majorité des modèles d'apprentissage automatique, en particulier pour les réseaux de neurones, les données sont divisées de manière disjointe en trois ensembles : un ensemble d'entraînement, un ensemble de validation et un ensemble de test.

L'ensemble d'entraînement est utilisé pour ajuster les paramètres du modèle. L'ensemble de validation, quant à lui, sert de régulation durant l'entraînement. Il permet d'ajuster les

hyperparamètres et de surveiller le modèle afin de détecter et stopper le surapprentissage lorsqu'il commence à se produire. Enfin, l'ensemble de test est réservé pour évaluer la performance finale du modèle sur des données qu'il n'a pas encore été rencontrées. L'ensemble de validation dans ces modèles joue un rôle essentiel en aidant à prévenir le surapprentissage et en garantissant que le modèle généralise bien à de nouvelles données.

Maintenant, la tâche consiste à trouver une fonction $\mathcal{F}$ qui associe $x^{(i)}$ à $y^{(i)}$.

$$\mathcal{F}(\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}) = \{y^{(1)}, y^{(2)}, \dots, y^{(m)}\}$$

Dans d'autres domaines, la forme mathématique et/ou la nature de la fonction $\mathcal{F}$ n'est pas connue a priori, mais la détermination complète de $\mathcal{F}$ dépendra du comportement des données. Un simple tracé est généralement effectué pour capturer la forme de $\mathcal{F}$. Comme illustré dans la figure 1.21 ci-dessous, en (a) nous avons une relation linéaire entre la sortie et l'entrée, ce qui signifie que $\mathcal{F}$ est une fonction linéaire ; en (b) $\mathcal{F}$ est une fonction quadratique. Contrairement

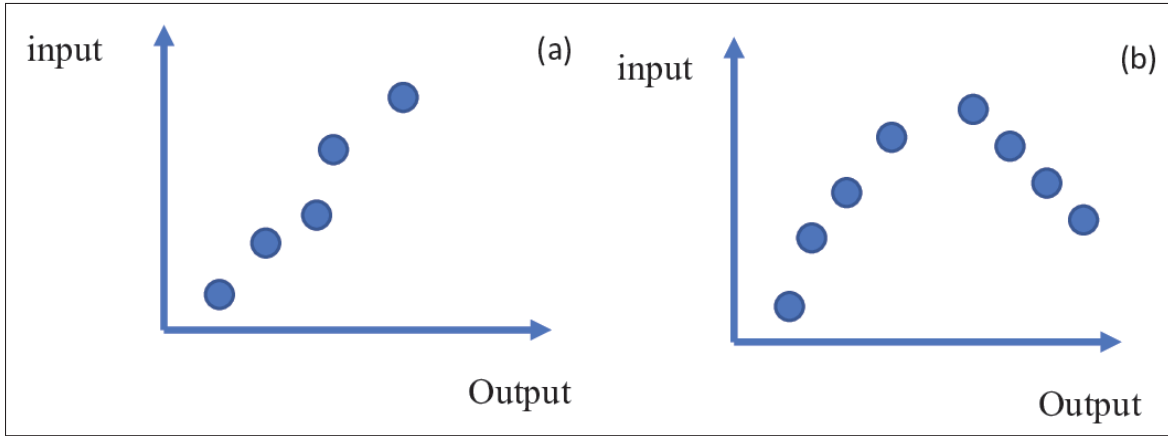


Figure 1.21 Illustration de la relation entre deux ensembles de données

aux cas où la forme de $\mathcal{F}$ n'est pas connue, dans le sous-domaine des potentiels interatomiques, la forme mathématique de la fonction $\mathcal{F}$ est connue mais avec des paramètres inconnus. La forme générale de $\mathcal{F}$ est le résultat d'une combinaison des paramètres introduits par le type de régression et le descripteur, généralement construit en fonction de la symétrie des transformations isométriques et de la lissité. Rigorousement, $\mathcal{F}$ est une famille de fonctions ; soit $\{\zeta\}$ l'ensemble des paramètres de $\mathcal{F}$, également appelé modèle, $\{\zeta\} = \{\lambda, \gamma\} = \{\{\lambda\}, \{\gamma\}\}$; les ensembles

$\{\lambda\}$ et $\{\gamma\}$ sont respectivement les paramètres du type de régression et les paramètres des descripteurs. Comme vous pouvez le voir dans l'équation 1.61 et l'équation 1.71, l'ensemble des paramètres du modèle MTP est $\Theta = \{c_\beta, c\} = \{\{c_\beta\}, \{c_{\mu, \tau_i, \tau_j}^{(\alpha)}\}\}$ où $\{c_\beta\}$ est le paramètre de régression linéaire et $\{c_{\mu, \tau_i, \tau_j}^{(\alpha)}\}$ est le paramètre du descripteur, les deux étant déterminés lors de l'entraînement avec les données mécaniques quantiques. Ainsi, l'ajustement consiste à trouver les paramètres optimaux d'un modèle connu plutôt qu'à trouver un modèle (fonction) incluant ses paramètres. Cela peut être fait en résolvant le problème de minimisation :

$$L(\zeta) = \sum_{i=1}^n (y^{(i)} - \mathcal{F}(\zeta, x^{(i)}))^2 \rightarrow \min \quad (1.76)$$

Pour être plus illustratif (figure 1.22), réécrivons notre ensemble de données étiqueté et remplaçons $y^{(i)}$ par $y_{qm}^{(i)}$, qui représente les énergies, forces et contraintes d'une configuration $x^{(i)}$ calculé par la mécanique quantique, $y_{qm}^{(i)} = \{E^{qm}(x^{(i)}), F^{qm}(x^{(i)}), \sigma^{qm}(x^{(i)})\}$

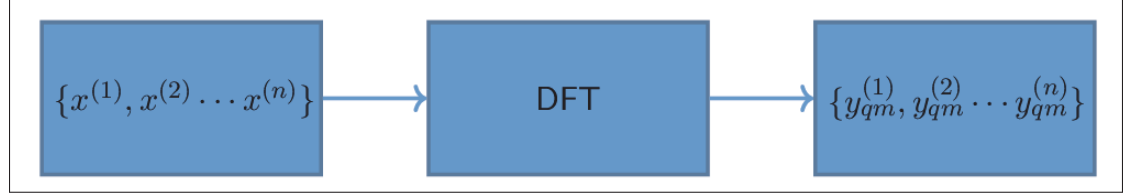


Figure 1.22 Étiquetage des données générées par la DFT à partir des configurations atomiques

Comme le montre la figure 1.23, les propriétés peuvent également être recalculées avec le modèle (MTP),

$$y^{(i)} = \{E^{mtp}(x^{(i)}), F^{mtp}(x^{(i)}), \sigma^{mtp}(x^{(i)})\}$$

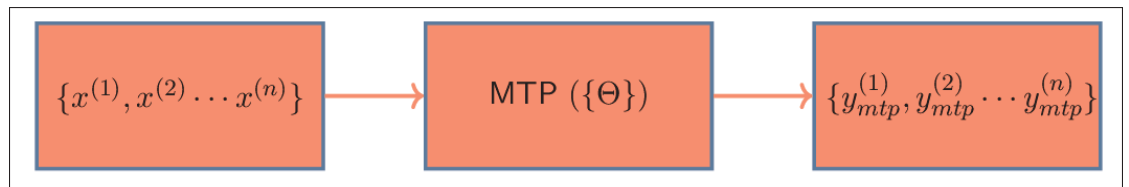


Figure 1.23 Étiquetage des données calculées par le MTP à partir des configurations atomiques

On souhaite que $y_{mtp}^{(i)}$ soit aussi proche que possible de $y_{qm}^{(i)}$. Comment trouver l'ensemble des paramètres $\{\Theta\}$ tel que $y_{mtp}^{(i)} \approx y_{qm}^{(i)}$? Cette question peut être résolue en résolvant le problème d'optimisation suivant (Novikov *et al.*, 2020a) :

$$\begin{aligned} \sum_{i=1}^n \left[w_e \left(E^{mtp}(x^{(i)}, \Theta) - E^{qm}(x^{(i)}) \right)^2 \right. \\ + w_f \sum_{j=1}^{N_a(x^{(i)})} \left| F_j^{mtp}(x^{(i)}, \Theta) - F_j^{qm}(x^{(i)}) \right|^2 \\ \left. + w_\sigma \left| \sigma^{mtp}(x^{(i)}, \Theta) - \sigma^{qm}(x^{(i)}) \right|^2 \right] \rightarrow \min \end{aligned} \quad (1.77)$$

où w_e , w_f et w_σ sont des poids non négatifs indiquant l'importance de chaque propriété (énergie, force ou contrainte) dans le problème d'optimisation. $N_a(x^{(i)})$ est le nombre d'atomes dans la configuration $x^{(i)}$. La norme de Frobenius d'une matrice, qui s'étend aux tenseurs ($\sigma = \sigma_{pq} e_p \otimes e_q$), est donnée comme suit :

$$\|\sigma\| = \sqrt{\sum_{p,q=1}^3 |\sigma_{pq}|^2} \quad (1.78)$$

L'erreur sur les énergies, les forces et les contraintes peut être calculée à partir de l'erreur quadratique moyen (RMSE : Root Mean Square Error) (Novikov *et al.*, 2020a; Novikov *et al.*, 2019).

$$RMSE(E)^2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{E^{mtp}(x^{(i)}, \Theta)}{N^{(i)}} - \frac{E^{qm}(x^{(i)})}{N^{(i)}} \right)^2 \quad (1.79)$$

$$RMSE(F)^2 = \frac{1}{n} \sum_{i=1}^n \frac{1}{3N^{(i)}} \sum_{j=1}^{N_a(x^{(i)})} \left| F_j^{mtp}(x^{(i)}, \Theta) - F_j^{qm}(x^{(i)}) \right|^2 \quad (1.80)$$

$$RMSE(\sigma)^2 = \frac{1}{n} \sum_{i=1}^n \frac{1}{9} |\sigma^{mtp}(x^{(i)}, \Theta) - \sigma^{qm}(x^{(i)})|^2 \quad (1.81)$$

où $N^{(i)}$ désigne le nombre de configurations $x^{(i)}$ dans l'ensemble d'apprentissage. L'ensemble d'apprentissage peut contenir différentes tailles de supercellules de la même structure. Notez qu'en 3 dimensions, il y a 3 composantes de forces et que le tenseur σ de rang 2 a 3^2 (9) composantes. Comme une partie des données disponibles n'est pas incluse dans l'apprentissage (ensemble de validation), les erreurs de prédiction peuvent être calculées en utilisant les équations 1.79 à 1.81 et aideront à estimer de manière impartiale l'exactitude du potentiel.

Cas pratique d'entraînement du model

Après l'installation du paquet MLIP, l'entraînement du MTP est effectué en résolvant le problème de minimisation souligné dans l'équation 1.77 ci-dessus. Tout d'abord, nous commençons par vérifier la distance minimale dans l'ensemble d'apprentissage avec la commande suivante *mindist*. Cette commande peut s'exécuter sur un terminal local ou dans un terminal des grappes de l'Alliance canada.

```
.../mlp mindist Train.cfg
.../mlp mindist Validation.cfg
mpirun -n (#cores) .../mlp train (level#).mtp Train.cfg --trained-pot-name
=pot.mtp --valid-cfgs=Validation.cfg
```

L'itération par défaut est fixée à 1000, mais on peut choisir n'importe quel nombre d'itérations en fonction du niveau MTP non entraîné avec l'option *max - iter*. Notez que la méthode d'optimisation BFGS (Broyden–Fletcher–Goldfarb–Shanno) est celle implémentée dans le paquet MTP. Par exemple, si nous choisissons 2000 itérations, nous devrions écrire :

```
mpirun -n (#cores) .../mlp train (level#).mtp Train.cfg --max-iter=2000 --trained-pot-name=pot.mtp
--valid-cfgs=Validation.cfg
```

Notez que l'entraînement et la validation peuvent être effectués séparément, c'est-à-dire :

```
mpirun -n (#cores) .../mlp train (level#).mtp Train.cfg --trained-pot-name
= pot.mtp
```

et après l'entraînement

```
mpirun -n (#cores) ... /mlp calc-errors pot.mtp Validation.cfg
```

L'énergie, les forces et les contraintes pour toute configuration donnée sont calculées avec la commande *calc - efs*. D'autres commandes telles que *relax* et l'interface avec LAMMPS peuvent aider à calculer des propriétés importantes telles que le paramètre de réseau, le module de compression, l'équation d'état, les constantes élastiques ainsi que les énergies de formation des défauts.

1.2.14 Optimisation numérique

Dans la plupart des problèmes de modélisation et de simulation numériques, on rencontre fréquemment des fonctions mathématiques complexes issues des problèmes à résoudre, pour lesquelles des solutions analytiques sont pratiquement impossibles à obtenir. Ces fonctions présentent des minima dont la détermination permet d'approcher les solutions des problèmes en question. Par exemple, la surface d'énergie potentielle d'un matériau peut être considérée comme une fonction complexe des coordonnées atomiques. Il existe alors un ensemble de coordonnées qui minimise cette fonction d'énergie potentielle. Trouver ces coordonnées atomiques minimisantes constitue un problème d'optimisation numérique (Sholl & Steckel, 2022). L'optimisation vise donc à identifier un ensemble de variables ou de paramètres qui minimise une fonction donnée. Dans le contexte spécifique de l'apprentissage supervisé, il s'agit de minimiser une fonction de coût ou une fonction de perte, dont les paramètres sont inconnus.

L'optimisation se déroule généralement en plusieurs étapes. Contrairement à d'autres domaines où la fonction de coût est choisie librement, dans le domaine des potentiels interatomiques par apprentissage automatique, la forme de la fonction de coût est déterminée par le modèle d'interaction utilisé. Les étapes clés du processus sont :

- Définition de la fonction de coût : Dans le contexte des potentiels interatomiques, la fonction de coût est déterminée par le modèle d'interaction. Ce modèle définit comment les erreurs de prédiction sont calculées en fonction des paramètres du modèle.

- Calcul du gradient de la fonction de coût : Une fois la fonction de coût définie, il est nécessaire de calculer son gradient. Ce gradient informe la méthode d'optimisation sur la direction et l'ampleur du changement des paramètres pour réduire la fonction de coût.
- Choix de la méthode d'optimisation : Enfin, on sélectionne une méthode d'optimisation itérative appropriée pour ajuster les paramètres du modèle. Ce choix dépend des caractéristiques du problème, telles que la forme de la fonction de coût et les contraintes associées.

Les méthodes itératives pour l'optimisation sont entre autres l'algorithme de gradient (Cauchy *et al.*, 1847), l'algorithme de Newton, l'algorithme de quasi-Newton et d'autres (Méthode de simulated annealing, algorithme génétique, méthodes de lagrange (multiplicateurs de Lagrange)). Ces méthodes peuvent être adaptées à divers types de problèmes en fonction de leur structure et des exigences spécifiques. La mise à jour des paramètres dans le cadre de la descente de gradient s'écrit :

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha \nabla f(\mathbf{x}_n) \quad (1.82)$$

ou α est le taux d'apprentissage et $\nabla f(\mathbf{x}_n)$ est le gradient de la fonction f au point $\mathbf{x}_n$. L'itération de la descente de gradient commence avec les conditions initiales des paramètres, notées $\mathbf{x}_0$, qui sont spécifiées avant le début de l'algorithme. Elles sont fixées avant le début de l'algorithme et influencent le point de départ pour l'optimisation. Il existe d'autres variantes telles que la Descente de Gradient Stochastique (SGD) et Descente de Gradient par Mini-Lots. La SGD utilise un sous-ensemble aléatoire de données pour chaque mise à jour. L'algorithme choisit aléatoirement une petite partie des données pour chaque mise à jour des paramètres, ce qui permet de traiter les données en ligne ou par petits lots.

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha \nabla f(\mathbf{x}_n; \mathbf{z}_i) \quad (1.83)$$

ou $\mathbf{z}_i$ représente un point de données ou une observation utilisée pour calculer le gradient de la fonction de coût.

Quand à la descente de gradient par mini-lots, de petits lots de données sont utilisés pour chaque mise à jour. Cette méthode utilise des groupes de données plus petits que l'ensemble complet

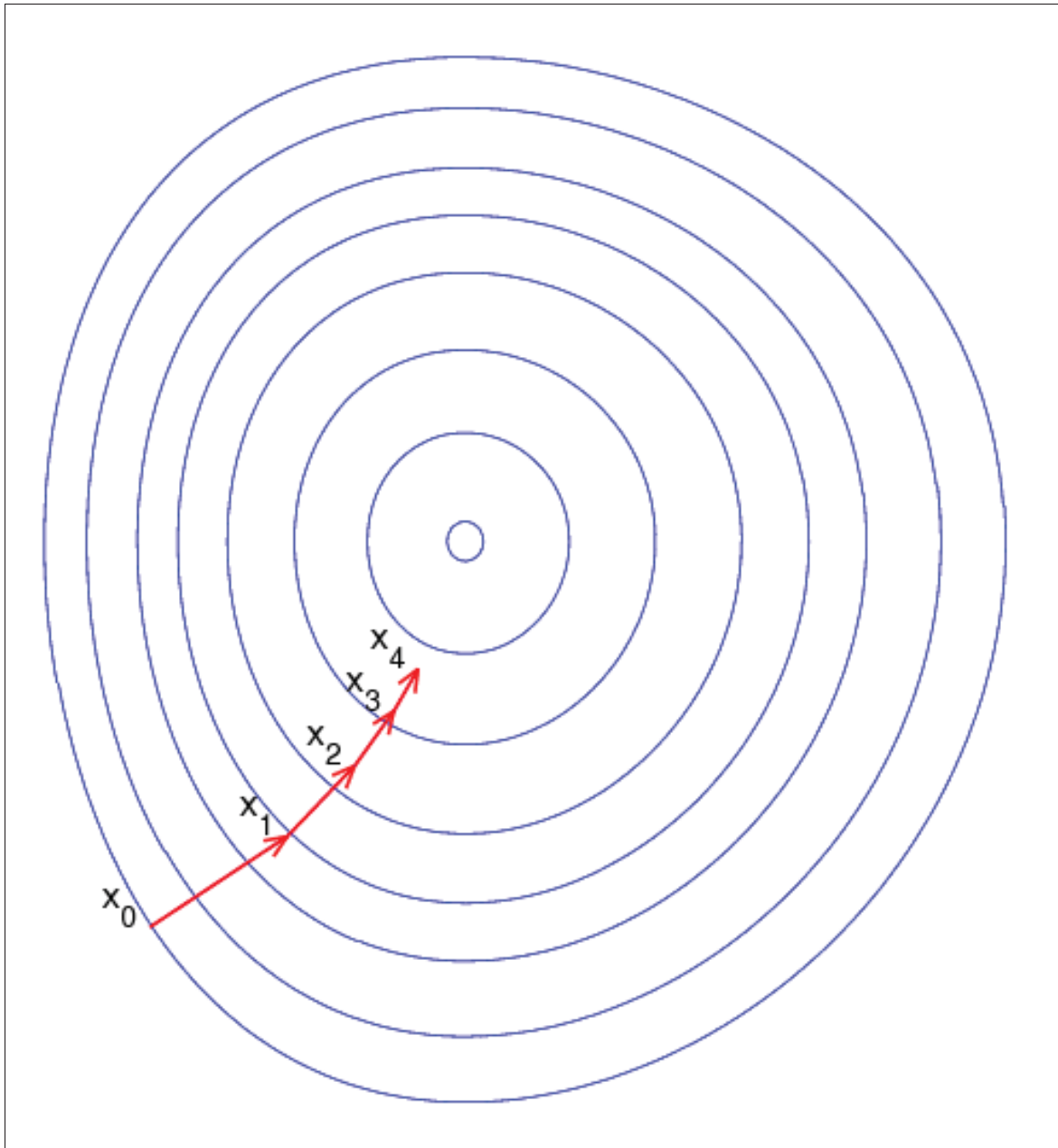


Figure 1.24 Algorithme du gradient sur une fonction de type bol (vue en plan avec courbes de niveaux)
Tirée de Wikipédia (2024, p. 1)

mais plus grands que ceux utilisés en SGD. Cela permet de bénéficier d'une meilleure estimation

du gradient tout en conservant une certaine rapidité de calcul.

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha \frac{1}{m} \sum_{i=1}^m \nabla f(\mathbf{x}_n; \mathbf{z}_i) \quad (1.84)$$

Dans l'équation 1.84, m est la taille du mini-lot (le nombre d'observations dans le mini-lot).

Pour la méthode de Newton, la mise à jour des paramètres s'écrit généralement comme suit :

$$\mathbf{x}_{n+1} = \mathbf{x}_n - [H_f(\mathbf{x}_n)]^{-1} \nabla f(\mathbf{x}_n) \quad (1.85)$$

$[H_f(\mathbf{x}_n)]^{-1}$ ou représente l'inverse de la matrice Hessienne de la fonction f au point x_n (Bofill, 2003). La matrice Hessienne $H_f(\mathbf{x})$ d'une fonction f au point $\mathbf{x}$ est définie comme :

$$H_f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad (1.86)$$

où $\frac{\partial^2 f}{\partial x_i \partial x_j}$ est la dérivée partielle seconde de la fonction f par rapport aux variables x_i et x_j . Dans la méthode quasi-Newton, la mise à jour des paramètres est similaire à celle de la méthode de Newton standard, mais elle utilise une approximation de la matrice Hessienne au lieu de la matrice Hessienne exacte :

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha_n B_n^{-1} \nabla f(\mathbf{x}_n) \quad (1.87)$$

B_n^{-1} est l'inverse d'une approximation de la matrice Hessienne ; $\mathbf{x}_n$ et α_n désignent respectivement la position actuelle et la taille du pas, déterminée par une recherche linéaire. La mise à jour de la matrice Hessienne approximative dans la méthode BFGS est donnée par :

$$B_{n+1} = B_n + \frac{y_n y_n^T}{y_n^T s_n} - \frac{B_n s_n s_n^T B_n}{s_n^T B_n s_n} \quad (1.88)$$

où $s_n = \mathbf{x}_{n+1} - \mathbf{x}_n$ et $y_n = \nabla f(\mathbf{x}_{n+1}) - \nabla f(\mathbf{x}_n)$.

La méthode BFGS (Broyden-Fletcher-Goldfarb-Shanno) (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) est une méthode d'optimisation itérative qui se situe dans la catégorie des méthodes quasi-Newton. C'est une technique très utilisée pour résoudre des problèmes d'optimisation non linéaire, notamment pour les fonctions non linéaires différentiables. La méthode BFGS est spécifiquement conçue pour approximativement inverser la matrice hessienne sans calculer directement cette matrice, ce qui est particulièrement utile pour les grands problèmes d'optimisation.

L'objectif est de trouver un minimum local d'une fonction f en utilisant des approximations de la matrice hessienne pour guider la direction de recherche. De façon général on procède de la façon suivante :

- Initialisation : Commence avec une approximation initiale de la matrice hessienne (souvent la matrice identité)
- Mise à jour : Met à jour cette approximation à chaque itération en fonction des gradients et des directions de recherche précédente
- Direction de recherche : La direction de recherche à chaque itération est obtenue en multipliant l'inverse approximé de la hessienne par le gradient

La méthode BFGS offre plusieurs avantages notables. Elle ne nécessite pas le calcul explicite de la matrice hessienne, ce qui réduit considérablement le coût computationnel par rapport à la méthode de Newton. De plus, elle converge généralement plus rapidement que les méthodes de descente du gradient simple grâce à l'utilisation d'une approximation de la matrice hessienne. Cependant, cette méthode présente également certains inconvénients. Elle peut être moins efficace pour les fonctions comportant des contraintes ou pour les problèmes très mal conditionnés. De plus, obtenir des résultats optimaux peut nécessiter un ajustement approprié des paramètres de l'algorithme.

1.2.14.1 Propriétés générales de l'optimisation numérique

L'optimisation numérique présente des caractéristiques et des propriétés essentielles pour résoudre efficacement les problèmes numériques. Comprendre ces propriétés est indispensable pour réussir l'optimisation numérique. En effet, une bonne maîtrise de ces éléments peut contribuer à réduire les coûts computationnels et à trouver des solutions optimales. Les propriétés de l'optimisation numérique incluent, entre autres (Sholl & Steckel, 2022) :

- Les algorithmes sont itératifs : Ils ne fournissent pas de solution exacte, mais plutôt une série d'approximations de la solution exacte.
- Estimation initiale requise : Il est nécessaire de fournir une estimation initiale pour utiliser les algorithmes. Ces algorithmes ne donnent pas de conseils sur le choix de cette estimation initiale.
- Contrôle par un paramètre de tolérance : Le nombre d'itérations effectuées est contrôlé par un paramètre de tolérance qui évalue la proximité de la solution actuelle avec la solution exacte.
- Variabilité des solutions : La répétition d'un algorithme avec un paramètre de tolérance différent ou une estimation initiale différente génère plusieurs solutions approximatives finales, qui sont généralement similaires mais pas identiques.
- Taux de convergence : Le taux de convergence des différents algorithmes peut varier considérablement ; choisir un algorithme approprié peut réduire considérablement le nombre d'itérations nécessaires.
- Identification des minima multiples : Les méthodes ne permettent pas de déterminer si la fonction a plusieurs minima si la méthode est appliquée une seule fois. L'application répétée avec différentes estimations initiales peut révéler plusieurs minima, mais ne garantit pas la découverte de tous les minima possibles.
- Garantie de convergence : Pour la plupart des méthodes, il n'y a pas de garantie que l'algorithme convergera vers une solution pour une estimation initiale arbitraire.

En conclusion, cette revue de littérature nous a permis de prendre connaissance des travaux réalisés dans le passé. Les informations recueillies dans ce chapitre ont mis en évidence les

méthodes de simulation actuelles et leurs limites. Ce chapitre nous a aidés à planifier notre recherche, à identifier les limites et les problèmes des méthodes standards et traditionnelles de modélisation des matériaux. Les résultats de la mise en pratique des concepts de bases de données, de potentiel par apprentissage automatique, de paramétrisation de modèles et de simulation de dynamique moléculaire seront détaillés dans les prochains chapitres.

CHAPITRE 2

FIRST-PRINCIPLES DATABASE FOR FITTING A MACHINE-LEARNING SILICON INTER-ATOMIC FORCE-FIELD

Karim Zongo¹, Laurent Karim Béland², Claudiane Ouellet-Plamondon¹

¹ Department of Construction Engineering, École de Technologie Supérieure,
1100 Notre-Dame St W, Montreal (Quebec) H3C 1K3, Canada

² Department of Mechanical and Materials Engineering, Queen's University
99 University Ave, Kingston, Ontario K7L 3N6

Paper published in Journal of MRS Advances, vol. 7, no 2, p. 39-47, February 2022

Abstract

Data-driven machine learning has emerged to address the limitations of traditional methods when modelling interatomic interactions in materials, such as electronic density functional theory (DFT) and semi-empirical potentials. These machine learning frameworks involve mathematical models coupled to quantum mechanical data. In the present article, we focus on the moment tensor potential (MTP) machine learning framework. More specifically, we provide an account of the development of a preliminary MTP for silicon, including details pertaining to the construction of a DFT database.

2.1 Introduction

Material modeling has become a powerful tool for gaining insight into materials properties and to circumvent obstacles met by experiments. It can also help to guide experimentalists and accelerate materials discovery by providing computationally predicted compounds of potential interest ((Hautier *et al.*, 2011; Liu *et al.*, 2017; Podryabinkin *et al.*, 2019)). Models span different time and length scales, allowing one to choose a scale based on desired properties or available computational means. Current materials modeling methods include quantum mechanical approaches such as density functional theory (DFT), semi-empirical methods and machine learning. The latter is a data driven approach introduced to address the limitations of

the first two methods such as spatio-temporal and transferability limitations respectively (Zuo *et al.*, 2020; Behler, 2014). Machine learning approaches allow for flexible and adaptive force fields for material research and simulation (Mueller *et al.*, 2016). Indeed, their computational cost are typically several orders of magnitude lower than that of DFT (Mueller *et al.*, 2016). They provide interpolative predictions of properties of a new atomic configuration using reference data (atomic configuration, properties) usually carefully prepared with DFT (Mueller *et al.*, 2016; Mohri *et al.*, 2018). The approach maintains near-chemical accuracy and is more versatile than semi-empirical potentials and faster than both DFT and even than some semi-empirical potentials (Mohri *et al.*, 2018; Botu *et al.*, 2017). The machine learning method is made of three part, namely its underlying mathematical model, its database and its implementation (Deringer *et al.*, 2019). This work is devoted to the second part, the database. Thus, in the present paper we describe the quantum mechanical database preparation in the context of clay minerals modeling using machine learning. Our long-term goal is to develop a machine learning force field that will enable a better description of interatomic forces within clay systems in the context of radioactive waste sealing and environmental applications. Our strategy is to build the database in a step-wise manner starting by one of the main constituent elements of clay systems. Namely, we start at a simple level with a single chemical element (silicon) and we will keep adding other chemical elements.

In this article, we will describe how our silicon atomic configurations database was constructed. The database is assessed by carrying out a preliminary implementation of a machine learning interatomic potential for silicon based on the moment tensor potential model (Shapeev, 2016).

2.2 Methods

2.2.1 Database generation

Databases are one of the main components that enable machine learning potential to address the challenges faced by DFT and semi-empirical potentials. Recall that the ultimate goal of machine learning interatomic potential (ML-IAP) is to accurately handle previously unknown

scenarios and structures (Deringer *et al.*, 2019). In fact different atomic environments may occur during manufacturing and practical applications of materials (Deringer *et al.*, 2019). Thus, truly predictive machine learning potential not only depend on the ML model but also need to be designed based on a highly diversified database either from experiments or *ab initio* calculations (Ramprasad *et al.*, 2017; Gubaev *et al.*, 2018). Although we are limited to a certain number of *ab initio* calculations that can be conducted, our database is designed by sampling most of the possible configurations and phases in which the material may chemically exist. First, we include all the possible phases including solid and liquid of the material of interest. Second for solid phase we take into account all the reported crystallographic and amorphous phases. Up to 13 different crystal structures of silicon with qualitatively different bonding have been reported in the literature (Zhao *et al.*, 2016). Thus, we build most of the configurations from experimental parameters such as lattice constant, bond length, bond angle and atomic coordinates found in the literature (Persson, 2014; Mehl *et al.*, 2017; Hicks *et al.*, 2019; Gadzhiev *et al.*, 2013; Downs *et al.*, 2003; Zheng *et al.*, 2020; Tran *et al.*, 2016). However all these experimental parameters were not available for some polymorphs including face centered cubic (FCC), body centered cubic silicon(BCC) and ST12 as well as non crystalline structures such as grain boundary. In these particular situations we use lattice constant and atomic coordinates from other DFT calculations (Lee, 2007; Li *et al.*, 2018; Crain *et al.*, 1994; Zheng *et al.*, 2020) as starting point and then compute new properties with our own optimal DFT parameters such as kinetic cut-off energy and k points. Regardless of theses parameters the re-computation is necessary as most of the configurations from others DFT calculations in the literature except the lattice constant and atomic coordinates usually lack useful properties such as energies, forces and stress which are the main ingredient of our machine learning model. We begin by computing the ground state for each crystallographic phase and then apply strain in six different modes following the Voigt notation as detailed in (De Jong *et al.*, 2015). In other words, we apply shear, tensile and compression from the ground state configuration with strain value ranging from -10% to 10% at 2% intervals. This strain increment will ensure to get different configurations with respect to atomic positions, bonding and possibly coordination numbers. In the other words, we ensure that atomic coordinates and bonding information are highly sensitive to the applied strain and

will be significantly different between two consecutive deformations.

In addition, the ground state configuration of each solid phase is replicated sufficiently and oriented before we introduce defects including point, line, planar and 3-dimensional defects. As interactions such as defect wave-function overlap, magnetic interactions, and strain field may affect the defect calculation in the supercell method using periodic boundary conditions (Freysoldt *et al.*, 2014), we employ replicated supercell of the diamond structure, 2x2x2 for vacancy, 3x3x3 for divacancy and interstitial and 4x4x4 for vacancy clusters as detailed in table 1 below. Configurations containing line defects such as dislocation were generated using three different replicated supercell 4x3x1, 5x4x1 and 6x5x1 in which quadrupole was introduced in order to cancel out long-range elastic strain field (Clouet, 2020). Both edge and screw dislocations as well as related planar defects such as generalized stacking faults(GSF) were considered in our database. The GSF is associated with dislocation properties that govern plastic deformation and fracture of crystalline solids. Here, the slab method (Goyal *et al.*, 2021) was used in two directions $[\bar{1} 10]$ and $[11\bar{2}]$ representing respectively full dislocation and partial dislocation directions in the (111) plane which is the natural cleavage plane of silicon. As stacking faults have very short-range interactions (one or two atomic layer distance) (Denteneer *et al.*, 1987; Tian *et al.*, 2012; Li *et al.*, 2014), we considered an orthorhombic supercell of the diamond structure with lengths $a\frac{\sqrt{2}}{2}$, $3a\sqrt{3}$ and $a\frac{\sqrt{6}}{2}$ (a is the lattice parameter) corresponding to $\frac{1}{2}[\bar{1} 10] \times 3[111] \times \frac{1}{2}[11\bar{2}]$ respectively. In total the supercell consist of 36 atoms and containing nine bilayers in the $[111]$ direction. We cut the crystal following the well known two distinct ways namely the shuffle and glide cut (Juan *et al.*, 1996) as illustrated in figure 5 in the supplemental information. The GSF were introduced within a bilayer and between bilayers for glide and shuffle cut respectively. A vacuum of 20 Å is created on each side of the (111) faces to avoid spurious interaction between the slab and its periodic images in the calculation. During the calculation, atoms were only allowed to relaxed perpendicularly to the plane of the cut. Another important planar defect that we considered in our database is the external interface defects (surfaces). Surfaces are important in a wide range of technological field including interfaces, catalysis, semiconductor fabrication and many others (Sun *et al.*, 2013; Sholl *et al.*, 2011). They are also inherent in the study and the understanding of physico-chemical processes at heterophase

interfaces such as gas-liquid (Ghoufi *et al.*, 2016; Neyt *et al.*, 2011), solid-liquid (Geysermans *et al.*, 2000; Šolc *et al.*, 2011; Ulian *et al.*, 2021) and solid-gas interfaces (Liang *et al.*, 2013). Thus, we include (100), (110) and (111) of diamond surfaces along with (1x1), (2x1) and (2x2) reconstructions of (100), (110) and (111). Here again, we used the slab model with 8 atomic layers for (100) and (110) and 12 atomic layers for (111) in both relaxation and reconstruction. A vacuum of 20 Å is then added in the direction normal to the considered surface.

Liquid and amorphous configurations are also generated using replicated unit cells. We run an NVT *ab initio* molecular dynamics simulation (AIMD) at 300 K and 1x, 1.5x, and 2x of the melting point with a time step of 1 fs and equilibration of 20000 time steps. Because of compute resource limitations, AIMD amorphization via the melting-quenching method was not possible. Instead, we generate amorphous configurations in two steps. First, we use the semi-classical Stillinger-Weber (Stillinger & Weber, 1985) and Tersoff (Tersoff, 1986) potentials to melt and quench the silicon crystal. Second, the resulting structures are refined via DFT geometry optimization thereby improving local geometry distortions and bringing each configuration to the nearest energy minimum. Note that only atomic coordinates are optimized so that the density is not altered. In addition we add some amorphous structures from (Barkema & Mousseau, 2000; Deringer *et al.*, 2018b). These structures only have atomic coordinates and cell parameters so we recalculate associated energies, forces and stress with our own kinetic cut-off energy and k points. We also add disorders structures to the database through random displacements of the optimized bulk unit cell of each crystallographic phase. We randomly displace atoms of up to 0.6 Å in certain cases in two different modes. In the first mode, atoms are randomly displaced in each of the x, y, and z Cartesian directions while their bulk positions are maintained in the other two directions. The second mode correspond to an isotropic random displacement in which all the atoms are displaced with same magnitude in three Cartesian directions. The configurations including different phases of silicon are detailed in the table 1 in below.

2.2.2 Computational details

We perform all the calculations using the Quantum Espresso package (Giannozzi *et al.*, 2009). The interactions between valence electrons and ionic cores are described with the projector augmented wave pseudopotential (PAW) (Blöchl, 1994) while the exchange-correlation interaction energy is calculated with the generalized gradient approximation (GGA) of Perdew, Burke, and Ernzerhof (PBE) (Perdew *et al.*, 1996). The energy criterion for convergence of self-consistent field (SCF) iterations was set to 10^{-10} eV. Convergence criteria for geometry optimization including atomic relaxation were set to 10^{-6} eV and 10^{-5} eV/Å for energy and forces, respectively. Before calculation on each crystallographic phase, we performed a convergence test for both kinetic cut-off energy and k-point. Each of these parameters is varied until the total energy become stable and the change in total energy become less than 10^{-4} eV. Thus, a value of 884 eV was set as the kinetic cut-off energy for the expansion of the wave function into plane waves. We employ Monkhorst-Pack grid (Monkhorst *et al.*, 1976) for Brillouin zone sampling and various k-point are used for the database generation including 8x8x8 for the ordinary phase. In this work, we describe the potential energy surface (PES) by means of the moment tensor potential (MTP) (Shapeev, 2016). Similarly to others ML-IAP, this interatomic interaction model assume that the PES is partitioned into a sum of individual atomic energies that only depend on their local environment delimited by some cut-off radius R_c . Thus, given an atomic configuration in which long-range interaction can be neglected, the total energy is expressed as a sum of the contributions of each of the n atoms :

$$E_{Total} = \sum_{i=1}^n E_i = \sum_{i=1}^n V_{local}(u_i) \quad (2.1)$$

in which u_i is a collection of relative distance r_{ij} of atoms j to atom i within the sphere or circle of radius R_c . The potential energy V_{local} of an atom i is then expanded as a linear combination of basis function B_β .

$$V_{local} = \sum_{\beta} c_{\beta} B_{\beta} \quad (2.2)$$

The basis functions B_β depend on the atomic environment of the atom i and are determined from the scalar tensorial contraction of the local atomic environment descriptors $M_{\mu,\nu}$ designed to account for all the physical symmetries.

$$M_{\mu,\nu}(r_{ij}, \tau_i, \tau_j) = \sum_j f_\mu(|r_{ij}|, \tau_i, \tau_j) \underbrace{r_{ij} \otimes \dots \otimes r_{ij}}_{\nu \text{ times}} \quad (2.3)$$

where τ_i and τ_j depict the type of atoms i and j respectively. $M_{\mu,\nu}$ is a tensor of rank ν enclosing the radial distribution f_μ and the angular information $r_{ij} \otimes \dots \otimes r_{ij}$ of the neighborhood of the central atom i . The radial part of the descriptor is further expanded in the radial basis functions $Q^{(\alpha)}$ which are expressed through the Chebyshev polynomials in the current implementation of MTP (Novikov *et al.*, 2020a).

$$f_\mu(|r_{ij}|, \tau_i, \tau_j) = \sum_\alpha c_{\mu,\tau_i,\tau_j}^{(\alpha)} Q^{(\alpha)}(|r_{ij}|) \quad (4)$$

$c_{\mu,\tau_i,\tau_j}^{(\alpha)}$ along with c_β are the MTP parameters that are computed during the fitting. The model has shown good performance and computational efficiency in various problems involving different materials (Zuo *et al.*, 2020; Mortazavi *et al.*, 2020; Lomaka & Tamm, 2020; Novikov *et al.*, 2018; Gubaev *et al.*, 2019; Novoselov *et al.*, 2019).

Tableau 2.1 Content of the DFT database used for fitting the MTP. The first column indicates which structure types were included in the database. The second column indicates the dimension of the simulation box, in terms of unit cell length—note that a mix of primitive and conventional unit were used. The third column depicts the number of atoms in the unit cell. The fourth column gives the total number of configurations for each structure type that were included in the database. Note that 25% of the structures were reserved as a validation set—they were not included in the training set

Content	Replication	Atom/cell	Total
Diamond structure			
Bulk	1x1x1	2	1
Bulk deformations (tensile, compression, shear)	1x1x1	2	177

Tableau 2.1 Content of the DFT database used for fitting the MTP. The first column indicates which structure types were included in the database. The second column indicates the dimension of the simulation box, in terms of unit cell length—note that a mix of primitive and conventional unit were used. The third column depicts the number of atoms in the unit cell. The fourth column gives the total number of configurations for each structure type that were included in the database. Note that 25% of the structures were reserved as a validation set—they were not included in the training set (continued)

Content	Replication	Atom/cell	Total
Vacancies	2x2x2 & 3x3x3	63, 213, 214, 215	58
Divacancies	3x3x3	214	44
Vacancy cluster (V_3, V_4, V_5, V_{17})	4x4x4	507, 508, 509, 495	20
Interstitials $I_1 - I_4$	3x3x3	217, 218, 219, 220	81
Diamond surfaces (100), (110), (111)	$\frac{\sqrt{2}}{2} \times \frac{\sqrt{2}}{2} \times 4, \frac{\sqrt{2}}{2} \times 1 \times 4\sqrt{2}$ $\frac{\sqrt{2}}{2} \times \frac{\sqrt{2}}{2} \times 4\sqrt{3}$	8, 12, 16	41
Surface reconstruction (100), (110), (111)	$\sqrt{2} \times \frac{\sqrt{2}}{2} \times 4, \frac{\sqrt{2}}{2} \times 1 \times 4\sqrt{2}$ $\frac{\sqrt{2}}{2} \times 1 \times 4\sqrt{3}$	16, 24	9
Stacking fault	$\frac{\sqrt{2}}{2} \times 3\sqrt{3} \times \frac{\sqrt{6}}{2}$	36	168
Dislocation (screw)	$5\frac{\sqrt{6}}{2} \times 4\sqrt{3} \times \frac{\sqrt{2}}{2}$	240	6
Grain boundaries (tilt & twist)	$\sqrt{5} \times 3\sqrt{5} \times 1, 2 \times 4 \times 2$ $\sqrt{2} \times 4\sqrt{2} \times \frac{\sqrt{2}}{2}$	44, 86, 119, 130, 160	8
MD (300 K)	2x2x2	64	25
Random displacement	2x2x2, 3x3x3, 4x4x4	32, 64, 192, 216, 512	19
Disordered phases			
Amorphous	2x2x2, 3x3x3, 4x4x4 5x5x5	64, 216, 512, 1000	42
MD-liquids (1770 K, 2530 K, 3370 K)	2x2x2	64	48
β - S_n			
Bulk	1x1x1	2	1

Tableau 2.1 Content of the DFT database used for fitting the MTP. The first column indicates which structure types were included in the database. The second column indicates the dimension of the simulation box, in terms of unit cell length—note that a mix of primitive and conventional unit were used. The third column depicts the number of atoms in the unit cell. The fourth column gives the total number of configurations for each structure type that were included in the database. Note that 25% of the structures were reserved as a validation set—they were not included in the training set (continued)

Content	Replication	Atom/cell	Total
Bulk deformations	1x1x1	2	105
Random displacement	2x2x2	54	2
Simple hexagonal			
Bulk	1x1x1	1	1
Bulk deformations	1x1x1	1	105
Random displacement	2x2x2	64	2
Hexagonal closed packed			
Bulk	1x1x1	2	1
Bulk deformations	1x1x1	2	105
Random displacement	2x2x2	54	2
Hexagonal diamond			
Bulk	1x1x1	4	1
Bulk deformations	1x1x1	4	105
Random displacement	2x2x2	108	2
FCC			
Bulk	1x1x1	4	1
Bulk deformations	1x1x1	1	9
Random displacement	2x2x2	108	2
BCC			
Bulk	1x1x1	2	1
Bulk deformations	1x1x1	1	9
Random displacement	2x2x2	54	2

Tableau 2.1 Content of the DFT database used for fitting the MTP. The first column indicates which structure types were included in the database. The second column indicates the dimension of the simulation box, in terms of unit cell length—note that a mix of primitive and conventional unit were used. The third column depicts the number of atoms in the unit cell. The fourth column gives the total number of configurations for each structure type that were included in the database. Note that 25% of the structures were reserved as a validation set—they were not included in the training set (continued)

Content	Replication	Atom/cell	Total
BC8			
Bulk	1x1x1	8	1
Bulk deformations	1x1x1	8	66
Random displacement	2x2x2	64	2
ST12			
Bulk	1x1x1	12	1
Bulk deformations	1x1x1	12	66
Random displacement	2x2x2	96	2
Clathrate Structure Si ₂₄			
Bulk	1x1x1	24	1
Bulk deformations	1x1x1	24	56
Random displacement	2x2x2	192	2
Clathrate Structure Si ₄₆			
Bulk	1x1x1	46	1
Bulk deformations	1x1x1	46	56
Random displacement	2x2x2	46	2
Clathrate Structure Si ₁₃₆			
Bulk	1x1x1	136	1
Bulk deformations	1x1x1	136	56
Random displacement	2x2x2	136	2
Planar and puckered silicene			
Bulk	1x1x1	2	2

Tableau 2.1 Content of the DFT database used for fitting the MTP. The first column indicates which structure types were included in the database. The second column indicates the dimension of the simulation box, in terms of unit cell length—note that a mix of primitive and conventional unit were used. The third column depicts the number of atoms in the unit cell. The fourth column gives the total number of configurations for each structure type that were included in the database. Note that 25% of the structures were reserved as a validation set—they were not included in the training set (continued)

Content	Replication	Atom/cell	Total
Bulk deformations	1x1x1	2	15
Grand total			1427

2.2.3 Reliability of the database

Tableau 2.2 Comparison between our DFT computed lattice parameters and experimental values. In the second column, the first value is a and the second value is c

	DFT	Experimental	Deviation (%)
Diamond structure	5.469	5.435 (Cannon, 1974)	0.6256
β - S_n	4.809 2.655	4.69 2.579 (Hu <i>et al.</i> , 1986)	2.53 2.94

In the context of interatomic potential, machine learning consist of several key steps including the generation of reference data, training of the machine learning model and the using of the implemented potential to run molecular dynamic simulation for the targeted property. At each step, a numerical noise is expected. Thus, one need to handle carefully the data involve in each process mainly at the calculation step of the reference data as numerical noise are intrinsic to atomistic calculations (Butler *et al.*, 2018b). Therefore a large number of basic properties such as lattice constant, bulk modulus, elastic constants, bond length and bond angle are computed and compared with experimental or other DFT calculation taking into account that experimental properties are usually measured at ambient conditions and others DFT calculations may use different parameters and functionals. Point defects formation energies are also compared with the literature. When these properties deviate more than 3% from experimental or theoretical

results, we first recheck the structure and if necessary we change our calculations parameters such as kinetic energy cut-off and k-points. As we deal with more than 1000 configurations for each material, we do not record every deviation but table 2 below provide an example on how we do a quick check of the reliability of our data.

2.2.4 Training details

Tableau 2.3 Average training and validation errors

	Energy error meV atom ⁻¹	Force error meV Å ⁻¹ (%)	Stress error GPa (%)
Training	1.4	10 (1.82)	0.076 (0.9)
Validation	1.9	22 (0.15)	0.095 (0.17)

The training of MTP using n atomic configurations consist of finding the parameters $\{\Theta\}$ of the model by solving the minimization problem as given below.

$$\sum_{i=1}^n \left[w_e (E^{mtp}(x^{(i)}, \Theta) - E^{qm}(x^{(i)}))^2 + w_f \sum_{j=1}^{N_a(x^{(i)})} |F_j^{mtp}(x^{(i)}, \Theta) - F_j^{qm}(x^{(i)})|^2 + w_\sigma |\sigma^{mtp}(x^{(i)}, \Theta) - \sigma^{qm}(x^{(i)})|^2 \right] \rightarrow \min \quad (2.4)$$

where w_e , w_f and w_σ are non-negative weight indicating the importance of each property (energy, force, or stress) in the optimization problem; $N_a(x^i)$ is the number of atoms in the configuration $x^{(i)}$. $\{E^{qm}(x^{(i)}), F^{qm}(x^{(i)}), \sigma^{qm}(x^{(i)})\}$ and $\{E^{mtp}(x^{(i)}), F^{mtp}(x^{(i)}), \sigma^{mtp}(x^{(i)})\}$ denote the properties calculated with DFT and MTP respectively. Here we choose the silicon as a prototype material for our first implementation of moment tensor potential model. A total number of 1427 configurations has been used with a split of 68 :32 i.e 68% for training and 32% for validation. We use the default setting of the MTP code such as scaling and weighting. Thus a weight of 1, 0.01 and 0.001 were used for energy, forces and stress respectively. We use a MTP functional form of level 26 with four radial basis functions including a total number of 889 free parameters. The cut off and minimum distance were set to $R_{cut} = 5.0$ Å and $R_{min} = 1.5$ Å

respectively. We set the number of iterations to 2000. The average fitting errors are detailed in table 2 below.

As can be seen from table 3 above, except the energy for which the training and validation errors are closed to each other ($1.4 \text{ meV atom}^{-1}$ versus $1.9 \text{ meV atom}^{-1}$), there are significant difference between those errors for forces and stress that can be a sign of underfitting or overfitting. Possible reasons for this include the large number of configurations used in the training and validation and the lack of the variation of MTP level. For instance, we used MTP level 26 in our first test which is not necessary an optimal level. Thus, optimal level including cut off distance as well as radial basis size were supposed to be found by varying simultaneously the MTP level like 18, 20, 22, 24, and the associated parameters (R_{cut} and basis size). Another important point is that MTP model parameters are randomly initialized at the beginning of the training so the training need to be repeated several times in order to quantify the uncertainty of the predictions and to circumvent overfitting issues.

2.3 Results and discussion

Tableau 2.4 Point defects diffusion energies barriers (eV)

Vacancy		
	Exp	DFT
E_v	3.25	3.33
E_m	0.21	0.17
E_a	3.46	3.50
Divacancy (movement)		
	Exp	DFT
E_v	5.45	4.98
E_m	1.36	1.23
E_a	-	6.21
Divacancy(dissociation)		
	Exp	DFT
E_v	5.45	4.98
E_m	2.00	2.07
E_a	-	7.05

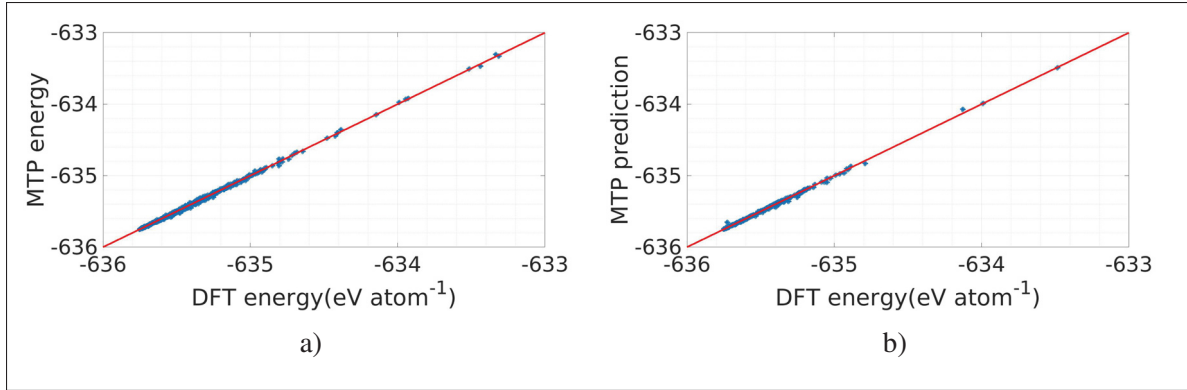


Figure 2.1 DFT-computed and MTP-predicted total energies (a) training set- 1062 configurations (b) validation set - 365 configurations

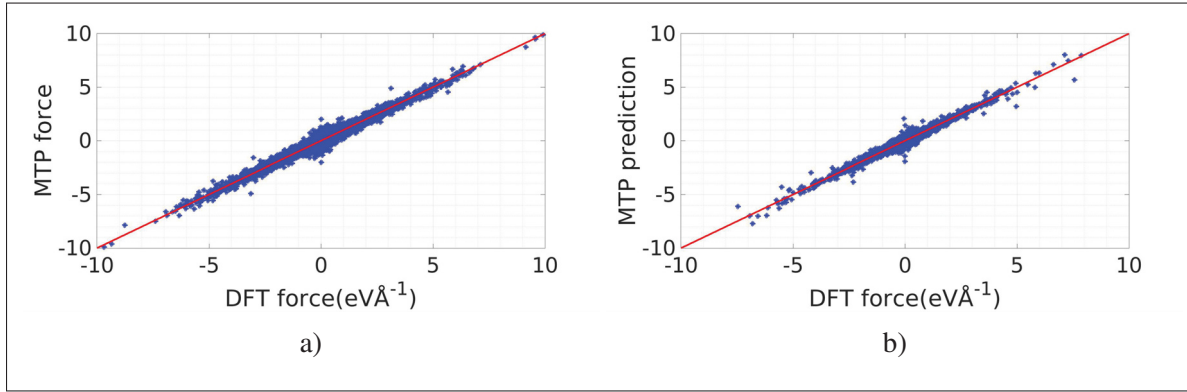


Figure 2.2 DFT-computed and MTP-predicted force components (a) training set - 1062 configurations (b) validation set - 365 configurations

In this section, we present the first set of results obtained by testing the MTP model on silicon. Fig. 1 (a), (b), Fig. 2 (a), (b) and Fig. 3 (a), (b) illustrates the comparison of the DFT-calculated energies, atomic forces, and stress and those calculated by the moment tensor potential. These plots of energies, force components, and stress components were obtained using training set and validation set. As can be seen, all of them show good correlation between DFT and MTP models. The excellent agreement between the MTP models and DFT from the validation set indicates that MTP can make accurate predictions within configuration space that was not explicitly employed in the training sets. We investigate generalized stacking fault (GSF) and point defects diffusion in silicon crystal both relevant to the mechanical properties and structural evolution of solids.

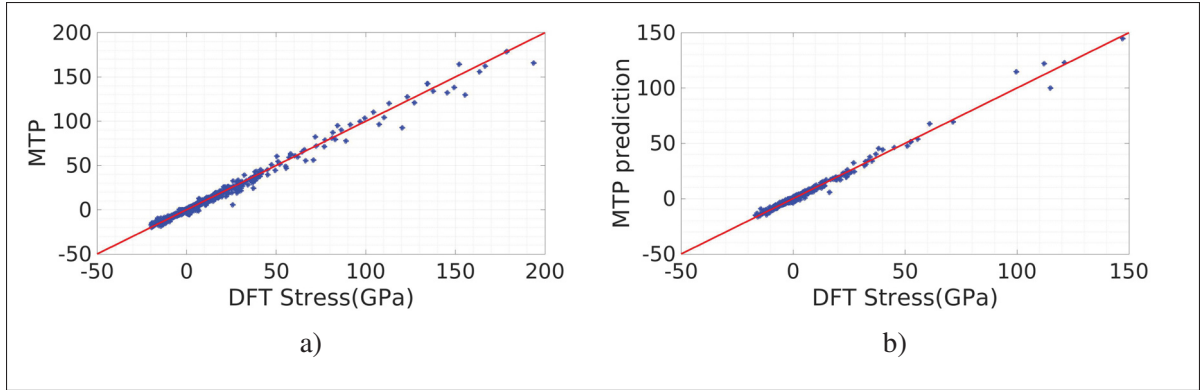


Figure 2.3 DFT-computed and MTP-predicted stress components (a) training set - 1062 configurations(b) validation set - 365 configurations

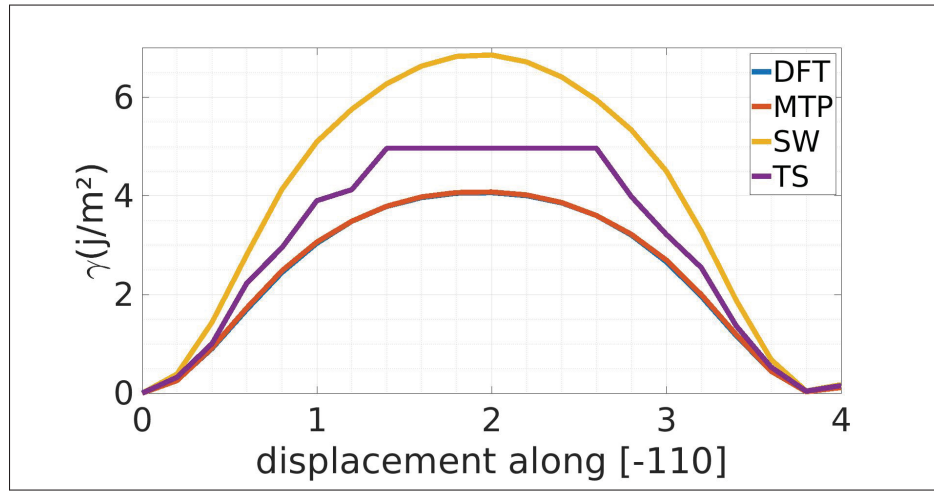


Figure 2.4 Generalized stacking fault curve

As mentioned earlier we considered shuffle and glide cut of diamond crystal. The shuffle cut result was already reported in conference article of the Canadian nuclear society (CNS) (Zongo *et al.*, 2021). So here, we only report the result of the glide cut of diamond lattice. As shown in figure 4 and figure 5, first our MTP stacking fault energies are in good agreement with our DFT energies. The first peak in figure 5 indicate the unstable stacking fault energy (γ_{USFE}) in the (111) planes along the $[112]$ direction. The DFT calculated value of γ_{USFE} is 1.564 J/m^2 which is comparable to that of figure 22 in (Bartók *et al.*, 2018). Second, MTP results compare better to DFT than semi-empirical Stillinger-Weber potential (Stillinger & Weber, 1985) and Tersoff

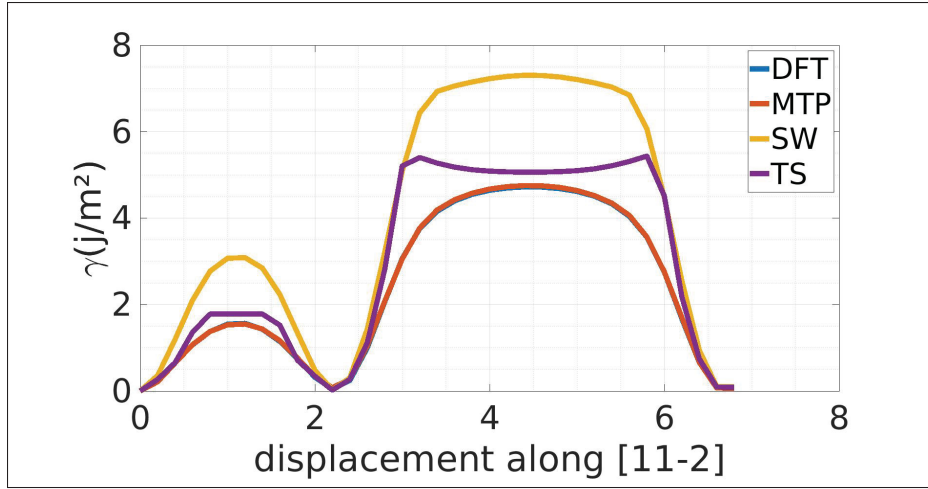


Figure 2.5 Generalized stacking fault curve

potential (Tersoff, 1986).

The second properties we investigate was the point defect diffusion such as vacancy-, divacancy- and interstitial-type defects. As such calculations are time consuming and very expensive using regular method like nudged elastic band(neb) (Henkelman *et al.*, 2000) coupled with density functional theory (DFT), we first manually moved the targeted atom along the considered path and compute the corresponding energies with DFT. Since this is not necessary a minimum energy path, the resulting energy barriers are higher, however the energy profile compare well with those found in the literature (Hwang *et al.*, 2002; Du *et al.*, 2005; El-Mellouhi *et al.*, 2004) as can be seen in the supporting information. Here the MTP reproduce with accuracy DFT energy profiles. Second, we used our implemented silicon MTP potential in LAMMPS to calculated energy barrier associated with the aforementioned point defect diffusion. The computed formation energy E_v , migration energy E_m and activation energy barrier E_a are reported in table 4 below. They compare reasonably to the DFT energies barrier that were obtained using different calculation parameters such as kinetic energy cut off, k points as well as exchange correlation functional and pseudo-potentials.

2.4 Conclusion and future work

The preliminary results described herein are very encouraging. Our implementation of a MTP-based model of silicon is a good demonstration of the versatility of MTP models. Benchmarks against 476 validation DFT calculations indicate that the MTP potential can correctly describe energies, forces and stresses in a wide variety of scenarios where semi-empirical potentials fail. Future work will include the extension of our database to silica and oxygen as well as the revision of the MTP training methods. We will also investigate point defects diffusion in silicon crystal, the amorphization of silica and the simulation of gas-liquid interface of oxygen using a unified fully implemented MTP potential of the silicon, oxygen and silica.

Acknowledgement

We thank Dr. Yaoting Zhang for insightful discussions. The authors wish to thank Compute Canada for generous allocation of computer resources, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Nuclear Waste Management Organization (NWMO) for financial support. There is no conflict of interest. Data will be made available upon request.

CHAPITRE 3

A UNIFIED MOMENT TENSOR POTENTIAL FOR SILICON, OXYGEN, AND SILICA

Karim Zongo¹, Hao Sun², Claudiane Ouellet-Plamondon¹, Laurent Karim Béland²

¹ Department of Construction Engineering, École de Technologie Supérieure,
1100 Notre-Dame St W, Montreal (Quebec) H3C 1K3, Canada

² Department of Mechanical and Materials Engineering, Queen's University
99 University Ave, Kingston, Ontario K7L 3N6

Paper published in Journal of npj Computational Materials, vol. 10, no 218, p. 1-16, September 2024

3.1 Abstract

Si and its oxides have been extensively explored in theoretical research due to their technological importance. Simultaneously describing interatomic interactions within both Si and SiO₂ without the use of *ab initio* methods is considered challenging, given the charge transfers involved. Herein, this challenge is overcome by developing a unified machine learning interatomic potentials describing the Si/ SiO₂/ O system, based on the moment tensor potential (MTP) framework. This MTP is trained using a comprehensive database generated using density functional theory simulations, encompassing diverse crystal structures, point defects, extended defects, and disordered structure. Extensive testing of the MTP is performed, indicating it can describe static and dynamic features of very diverse Si, O, and SiO₂ atomic structures with a degree of fidelity approaching that of DFT.

3.2 Introduction

Si/SiO₂ interfaces are ubiquitous in semiconductor manufacturing, which includes metal—oxide—semiconductor field-effect transistors (Arns, 1998), nanowire- and nanodot-transistors. The formation of SiO₂ layers involves charge transfer during the oxidation of Si substrates (Bauza, 2001). Additionally, siliceous materials—including clay minerals and cement, which comprise Si, O, and SiO₂ components—are governed by interactions that entail similar charge transfer.

Abundant past theoretical research has focused on understanding Si, its oxides, the formation of SiO₂ multilayered structures, early oxidation rates, and amorphization of oxide layers (Pasquarello *et al.*, 1998; Ganster *et al.*, 2010a, 2012; Cvitkovich *et al.*, 2023, 2022; Salles *et al.*, 2017; Takahashi *et al.*, 2014). These studies predominantly relied on electronic density functional theory (DFT) (Kohn & Sham, 1965; Sholl & Steckel, 2022; Goedecker, 1999) and nonflexible classical potentials (Torrens, 2012; Pizzagalli, 2004). However, simulating large systems with multiple components, charge transfer, and hetero-interfacial systems poses challenges within these frameworks. An ideal modeling approach should explicitly or implicitly capture charge transfer without compromising accuracy or incurring prohibitively large computational costs. Existing charge equilibration potentials like ReaxFF (Van Duin *et al.*, 2001, 2003) and COMB (Phillpot *et al.*, 2018; Yu *et al.*, 2007; Liang *et al.*, 2012), while being capable of describing chemical interactions during MD simulations, tend to have a limited ability to describe mechanical properties of materials (Van Duin *et al.*, 2003; Yu *et al.*, 2007; Shan *et al.*, 2010) unless special reparametrization is applied.

Recent advances, such as linear-scaling DFT (Goedecker, 1999; Hine *et al.*, 2009) and machine learning (ML) force fields—e.g. the Gaussian approximation potentials (GAP) (Bartok, 2010) and artificial neural networks (Behler & Parrinello, 2007; Behler, 2021)—lift the limitations of traditional methods. ML force fields have demonstrated high accuracy in modeling Si (Bartók *et al.*, 2018; Li *et al.*, 2020; Hu *et al.*, 2019; Babaei *et al.*, 2019) and many other elements (Dragoni *et al.*, 2018; Szlachta *et al.*, 2014; Rowe *et al.*, 2020; Smith *et al.*, 2021; Deringer *et al.*, 2020). Similar progress has been made in improving interatomic interaction descriptions in Si oxides (Novikov & Shapeev, 2019; Erhard *et al.*, 2022; Li *et al.*, 2018; Balyakin *et al.*, 2020) and metal oxides (Artrith *et al.*, 2011; Artrith & Urban, 2016; Kandy *et al.*, 2023; Kobayashi *et al.*, 2022). However, jointly describing compounds and their constituents using ML force fields presents challenges due to the disjointed configurational space of multi-phase forms and the need to handle charge transfer. ML force fields (Behler, 2016; Das Sarma *et al.*, 2019; Mishin, 2021; Onat *et al.*, 2020; Friederich *et al.*, 2021) combine a descriptor to a regression procedure to encode geometry and *ab initio* properties, usually omitting explicit electronic structures. A

previous study focusing on modeling SiO₂ using the moment tensor potential (MTP) suggests incorporating additional reference data is preferable to adding explicit charge equilibration for long-range interactions (Novikov & Shapeev, 2019).

The novelty of this article is a MTP (Shapeev, 2016; Novikov *et al.*, 2020b) that jointly describes interatomic interactions in SiO₂ and its constituents (Si and O), enabling the representation of multiple charge states. The developed MTP for Si/ O/ SiO₂ systems is parameterized using an *ab initio* database containing diverse crystal structures, point defects, extended defects, and disordered structures. This MTP is then utilized for molecular statics (MS) and molecular dynamics (MD) simulations to investigate crystalline, interfaces, amorphous, and liquid states of Si and SiO₂. These test simulations indicate that the MTP can provide a unified description of these disjoint systems.

3.3 Results

Our analysis encompasses both MS and MD simulations. MS results include cohesive energy, lattice constant, elastic constant, defect formation energies, interface relaxation, as well as linear and planar defect. The MD results bracket vacancy diffusion coefficients, melting point, interfaces, as well as the liquid and amorphous structures of Si and SiO₂. The outcome of these runs are compared against the reference method (DFT) and those derived from the semi-classical potential.

3.3.1 Cohesive energy, elastic constant, point defects and extended defects

First, the state equations of various Si and SiO₂ polymorphs are presented as cohesive energy vs lattice parameter. Please see the methodology section for calculation details. Additionally, the cohesive energy values for molecular oxygen, both O₂ and O₃, are provided. As shown in Fig. 1, the MTP replicated the cohesive energies of the references states with remarkable accuracy. The Tab. 3.1 compares lattice parameters predicted by the MTP, COMB, and ReaxFF models with

benchmark and experimental data. Remarkably, the MTP predictions show excellent agreement with both the benchmark and experimental data.

Tableau 3.1 Comparison of lattice constants predicted by MTP, ReaxFF, and COMB models against experimental data and DFT calculations

		Exp	DFT	MTP	COMB	ReaxFF
α Si	a	5.4300 (Brandes <i>et al.</i> , 2013)	5.4690	5.4682	5.4399	5.4635
β -Si	a	4.665 (McMahon <i>et al.</i> , 1994)	4.8090	4.8170	4.8242	4.7650
	c	2.565	2.6554	2.6493	2.6533	2.6207
HD	a		3.8508	3.8601	3.8411	3.8635
	c		6.2706	6.3028	6.2718	6.3084
HCP	a		2.6961	2.8038	2.8065	2.7192
	c		4.5577	4.6013	4.6057	4.4624
SH	a	2.5530(McMahon <i>et al.</i> , 1994)	2.6458	2.6478	2.7232	2.9162
	c	2.3820	2.4840	2.4861	2.5569	2.7382
ST12	a		5.6763	5.6639	5.7680	5.65236
	c		6.8308	6.7607	6.8849	6.7469
BC8	a		5.7657	5.7662	5.7539	5.6405
	c		4.7077	4.7081	4.6980	4.6054
C24	a	3.82(Kim <i>et al.</i> , 2015)	3.8510	3.8533	3.8482	3.7800
	b	10.70	10.2298	10.7874	10.7732	10.5821
	c	12.63	12.7483	12.7281	12.7113	12.4859
C46	a	10.3550(Adams <i>et al.</i> , 1994)	10.2298	10.2777	10.1513	10.1761
C136	a	14.864(Adams <i>et al.</i> , 1994)	14.7417	14.7989	14.8817	14.6751
α -Q	a	4.9160 (Levien <i>et al.</i> , 1980)	5.0339	5.0280	4.9105	4.8872
	c	5.4054	5.5216	5.5314	5.4022	5.3765
β -Q	a	4.9977 (Wright <i>et al.</i> , 1981)	5.1060	5.1171	5.05321	5.0529
	c	5.4601	5.5869	5.5905	5.5207	5.52045
α -C	a	4.9720 (Downs <i>et al.</i> , 1994)	5.1026	5.1004	4.7418	4.8774

Tableau 3.1 Comparison of lattice constants predicted by MTP, ReaxFF, and COMB models against experimental data and DFT calculations (continued)

		Exp	DFT	MTP	COMB	ReaxFF
β -C	c	6.922	7.1364	7.1205	6.6199	6.8092
	a	7.1590 (Barth, 1932)	7.4608	7.4625	7.5429	7.3724
α -T	a	5.0100 (Cellai <i>et al.</i> , 1995)	5.0616	5.1275	4.8502	4.9462
	b	8.600	8.9187	8.8062	8.3299	8.4948
β -T	c	8.2200	8.4166	8.4159	7.9607	8.1183
	a	5.0350 (Villars, 1985)	5.2756	5.2731	5.2957	5.2255
	c	8.220	8.6107	8.6069	8.6438	8.5292
CO	a	7.1356 (Levien <i>et al.</i> , 1981)	7.2653	7.2354	7.2648	7.6759
	b	12.3692	12.5368	12.4253	11.9582	12.3520
SE	c	7.1736	6.2766	6.2687	6.2943	6.6504
	a	4.0730 (Grocholski <i>et al.</i> , 2013)	4.1145	4.1155	4.2222	4.4444
	b	5.0260	4.5263	4.5287	4.6461	4.8906
ST	c	4.4770	5.0814	5.0830	5.2147	5.4892
	a	4.1797 (Ross <i>et al.</i> , 1990)	4.2345	4.2285	4.2858	4.8368
KE	c	2.6669	2.69170	2.6905	2.7269	3.0776
	a	7.464 (Shropshire <i>et al.</i> , 1959)	7.5581	7.6446	7.4497	7.5001
MO	c	8.62	9.1695	8.8285	8.6035	8.6617
	a	7.3317 (Miehe <i>et al.</i> , 1992)	7.5199	7.4262	7.7782	7.4709
	b	4.876	5.0728	4.9944	5.2312	5.0245
CH	c	8.758	6.9780	6.8935	7.2203	6.9350
	a	13.6046 (Díaz-Cabañas <i>et al.</i> , 1998)	13.70564	13.7061	13.6346	13.5269
	c	14.8290	14.9563	14.9569	14.8789	14.7614
ZG	a	16.4110 (Plévert <i>et al.</i> , 2000)	16.7106	16.9151	16.3306	16.6907
	b	20.0440	20.4473	20.3998	19.6949	20.1292
	c	5.0427	5.1559	5.2060	5.0261	5.1370

Tableau 3.1 Comparison of lattice constants predicted by MTP, ReaxFF, and COMB models against experimental data and DFT calculations (continued)

		Exp	DFT	MTP	COMB	ReaxFF
ZSM	a	20.0511 (Artioli <i>et al.</i> , 2000)	20.4927	20.6177	20.2347	20.1282
	b	19.8757	20.1528	20.2357	19.8598	19.7552
	c	13.3682	13.5828	13.7043	13.4498	13.3789
RRMSE (%)				0.17	0.39	0.37

Tableau 3.2 Comparison of force fields prediction of elastic constants for silicon crystal and α -quartz. The RRMSE values calculated against DFT data are provided in parentheses. Note that the BKS potential was implemented by optimizing bulk parameters based on the predictions of experimental elastic constants of α -quartz (Van Beest *et al.*, 1990)

Si elastic constants (GPa)					
	Exp (Brandes & Brook, 2013)	DFT	MTP	COMB	ReaxFF
C ₁₁	165	153.1	168.1	142.5	109.3
C ₁₂	65	57.0	51.8	75.4	46.8
C ₄₄	79.2	74.3	65.5	69.0	37.0
RRMSE (%)			5.7 (5.8)	8.0 (7.0)	21.5 (18.8)
α -quartz elastic constants (GPa)					
	Exp (McSkimin <i>et al.</i> , 1965)	DFT	MTP	BKS	
C ₁₁	86.8	84.2	83.1	90.6	-
C ₁₂	7	3.3	10.6	8.1	-
C ₁₃	19.1	8.26	17.8	15.2	-
C ₁₄	-18	17.1	-21.8	-17.6	-
C ₃₃	105.8	89.5	100.1	107.0	-
C ₄₄	58.2	52.0	34.7	50.2	-
C ₆₆	39.9	42.8	36.2	41.2	-
RRMSE (%)			6.1 (12.3)	2.4 (10.8)	-

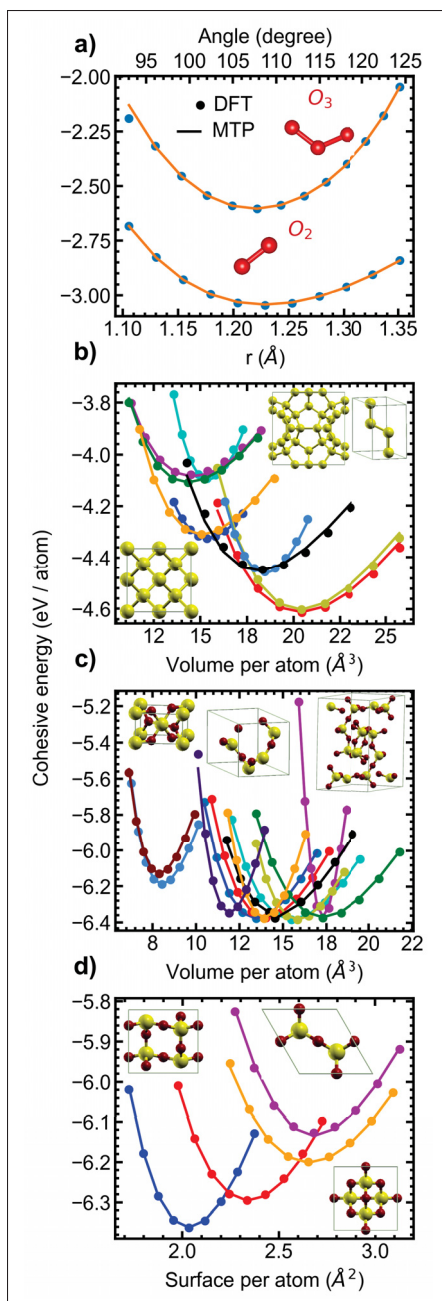


Figure 3.1 Bond and angle energy of oxygen molecules, and equation of state of silicon and silica polymorphs. **a** Bond and angle energy of dioxygen and ozone as calculated using MTP (lines) and DFT (dots). **b-d** Energy-volume relationships in crystalline SiO₂ and Si polymorphs calculated using the MTP (lines) and DFT (dots). 3D Si polymorphs **b**, 3D **c** and 2D **d** SiO₂ structures were considered. Details pertaining to these crystal structures are available as S.I. Agreement between MTP and DFT is excellent. Insets : illustration of selected polymorphs' crystal structures

Tableau 3.3 Exploring bulk modulus in silicon and silica polymorphs : Calculations using experimental lattice parameters input in lammps code. The units are expressed in GPa. The RRMSE value (calculated against DFT data) is provided in parentheses. The value of the bulk modulus for amorphous structures may vary depending on the cooling rate. In our study, we used cooling rates of 5 K/ps for silicon and 10 K/ps for silica

Structure	Exp	DFT	MTP	COMB	ReaxFF
Diamond silicon	98.3 (Brandes & Brook, 2013)	89.0	90.5	97.8	67.6
β -tin silicon		105	110.0	180.3	98.2
HD		87.3	91.2	97.8	67
HCP silicon		86.5	99.8	601.936	-
BC8 silicon		83.4	94	95.8	291.1
ST12 silicon		75.0	99.5	56.9	397
α -quartz	38 (Levien <i>et al.</i> , 1980)	32.8	39.9	33.7	-
α -tridymite		26.1	22.6	20.8	-
Keatite		66.7	70.9	12.3	-
Coesite	96.3 (Levien & Prewitt, 1981)	114	97.4	-	-
Moganite	32.2 (Léger <i>et al.</i> , 2001)	37.2	28.9	-	-
Seifertite	290 (Pabst <i>et al.</i> , 2013)	278	320	112.8	-
Zeolite (Chabazite)	54 (Leardini <i>et al.</i> , 2013)	62.0	56.3	-	-
Zeolite (GUS-1)		24.4	24.8	9.7	-
Zeolite (ZSM)		32.9	47.1	11.5	-
a-Si	-	82.5 (Durandurdu <i>et al.</i> , 2001)	70.9	-	-
a-SiO2	36.4 (Guerette <i>et al.</i> , 2012)	-	37.9	-	-
RRMSE			3.3	51.4	44.7
			(3.5)	(24.1)	(33.6)

Tableau 3.4 Comparative analysis of defect formation energies in silicon crystals and silica polymorphs using DFT, MTP, and ReaxFF results. Our comparison excludes the results from BKS and COMB as they are not parameterized for a single oxygen system. In the training set, only the silicon vacancy defect is present, while other defects are used to test the portability of the potential. For Si-I₂, Si-I₃, Si-I_{4C1}, and Si-I_{4C2}, the formation energy is given in eV per atom

	DFT	MTP	ReaxFF
Si-V	3.7	2.3	3.0
Si-V ₂	5.5	3.7	6.7
Si-I ₂	2.9	3.0	1.9
Si-I ₃	2.1	2.5	2.1
Si-I _{4C1}	2.2	2.5	2.0
Si-I _{4C2}	2.1	2.4	1.1
α -AQ	6.1	5.8	6.6
β -BQ	5.4	5.8	5.1
α -AC	5.1	5.1	4.9
α -AT	5.0	4.8	1.3
KE	5.2	5.7	0.5
CO	5.40	6.20	4.4
CO	5.9	6.78	4.3
MO	5.2	5.8	5.5
SE	6.2	4.3	0.9
ST	6.2	5.3	2.3
RRMSE (%)		4.5	12.0

The second-order elastic constants and bulk modulus are determined using finite difference as detailed in the methods section. Tab. 3.2 provides the relative root mean square error (RRMSE) on elastic constant with respect to the DFT benchmark and experimental data. The MTP model demonstrates lower RRMSE when compared to those of ReaxFF and COMB potentials, it competes closely with the Beest Kramer van Santen (BKS) (Van Beest *et al.*, 1990) potential. Moreover, other semi-empirical models reported in Ref. (Bartók *et al.*, 2018) demonstrate higher errors compared to the predictions made by the MTP model. The bulk modulus values for various silicon and silica polymorphs can also be found in Tab. 3.3. As evident, the MTP predictions closely align with the reference methods and experimental values, although it is worth noting that the training set did not encompass the deformation of certain polymorphs. Our potential accurately predicts the elastic constant of amorphous silica, even though amorphous

configurations were not included in the training set. In our testing of the MTP potential, we have also considered point defects like vacancies, divacancies, and self-interstitials. The RRMSE values for these defects are reported in Tab. 3.4. Once again, using the MTP leads to smaller relative errors in comparison to ReaxFF potential. Since the majority of potentials were not specifically parameterized for the oxygen system alone, our comparison was limited solely to ReaxFF. In our study, we examined a specific case involving the I4 compact cluster (Arai *et al.*, 1997) within the Si crystal, which was not included in our training set. The atoms within the cluster exhibit a harmonious four-coordinated arrangement. Notably, the cluster boasts the presence of five-, six-, and seven-membered atomic rings. The bond lengths and bond angles of this cluster were calculated based on relaxed structures obtained from DFT, MTP, SW, ReaxFF and COMB calculations, as illustrated in Fig. 2. Notably, no dangling bonds were observed for all potentials, except for the COMB potential, which failed to reproduce the I4 structure. When analyzing the formation energy of the cluster, the MTP model exhibited a prediction within 14% of the reference value, while the SW and ReaxFF potentials displayed errors reaching up to 27% and 47%, respectively. Among the models assessed, the MTP model exhibited superior agreement with DFT calculations for both bond lengths and bond angles. As this is a perfectly coordinated tetra-interstitial, we also tested 3 and 5-fold coordinated interstitials, namely, di-interstitial, tri-interstitial, and tetra-interstitial, as shown in Fig. 2. While these defects are not included in the training set, the MTP exhibits better agreement with the benchmark than ReaxFF, as detailed in Tab. 3.4. The MTP also outperforms ReaxFF in describing the vacancy formation energy in silica polymorphs, as demonstrated in Tab. 3.4. Again, no SiO₂ point defects were incorporated to our training set.

The static migration barrier energy of the vacancy was determined using the nudged elastic band (NEB) (Henkelman *et al.*, 2000). The migration barrier profiles, including the DFT-based profiles as well as the MTP- and SW-based profiles, are depicted in Fig. 3b. The MTP migration barrier profile shows excellent agreement with the DFT reference profile. In contrast, the SW potential does not capture the reference profiles with similar accuracy. Examining the barrier for vacancy migration reveals a relative error in barrier height of 15.0% for MTP, while for SW, it is 73.1%.

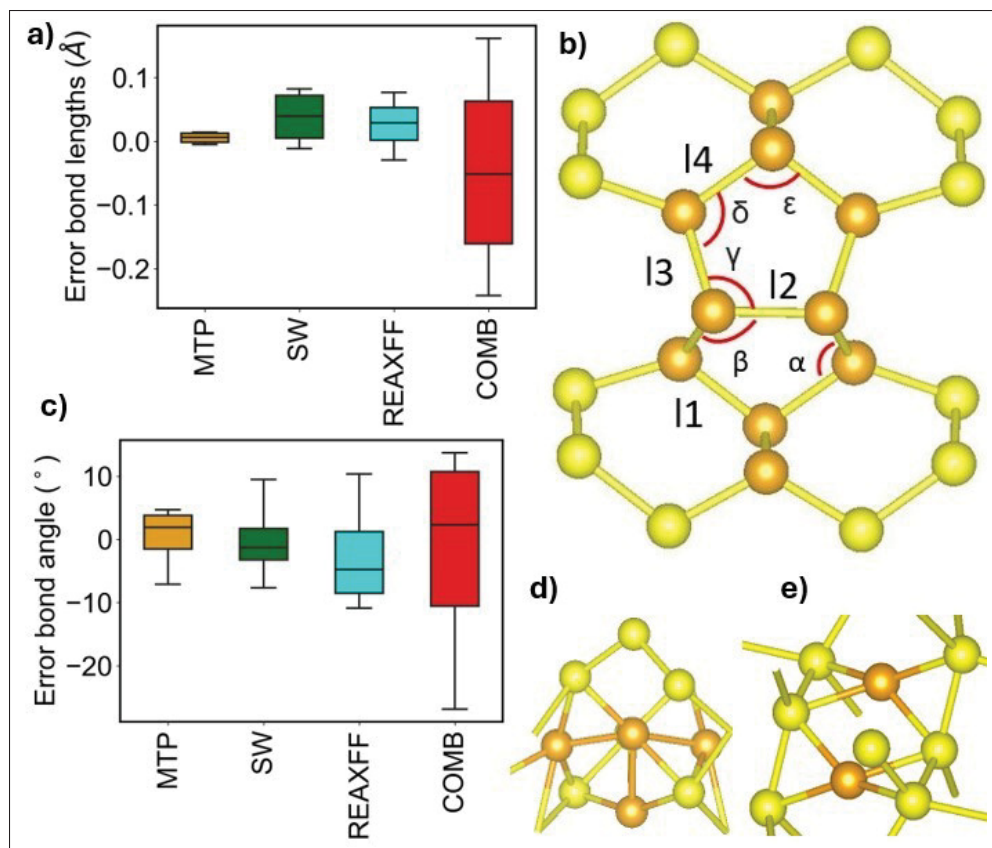


Figure 3.2 Point defects in silicon crystal. In silicon crystal, a relaxed perfectly coordinated four-interstitial cluster is depicted. left **a** and **c** : the MTP leads to bond lengths and angles (as illustrated in the panel **b**) in better agreement with DFT as compared to SW, ReaxFF and comb models. Additionally, two other interstitial clusters with coordination defects are presented, namely a four-interstitial cluster **d** and a di-interstitial **e**. All the interstitial atoms are colored orange . In **d** and **e**, the interstitial bonds are also colored orange. Their formation energies are indicated in Tab. 3.4, where configurations **b**, **d**, and **e** correspond to Si-I4C2, Si-I4C1, and Si-I2, respectively. The notations l_1 , l_2 , l_3 , and l_4 , as well as α , β , γ , δ , and ϵ , represent bond lengths and bond angles within the cluster, respectively. These interstitial configurations were not included in the training set

These results demonstrate that the MTP model can better mimic the reference method when studying point defects within larger systems, as demonstrated in references (Ohbitsu *et al.*, 2022;

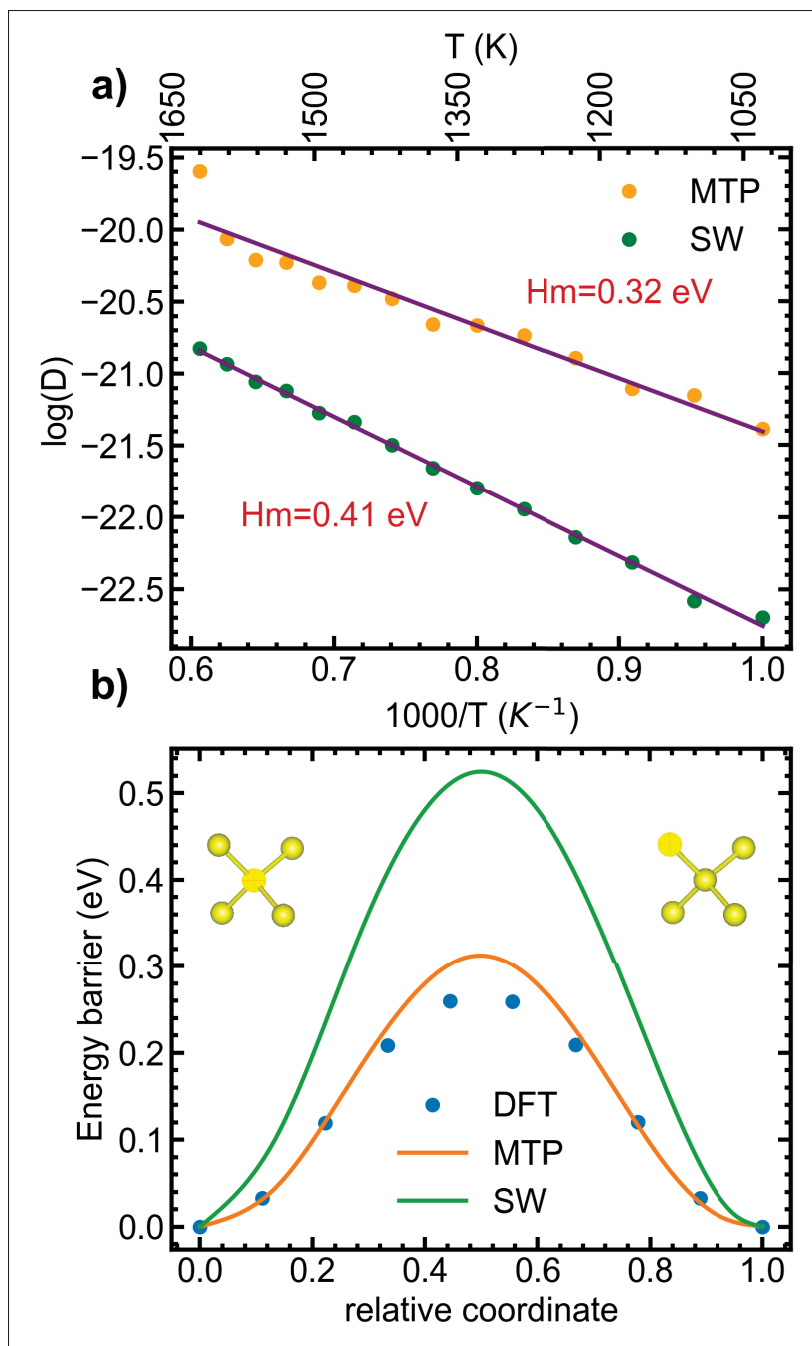


Figure 3.3 Diffusion of point defects in silicon crystal. **a** Temperature dependence of vacancy diffusion coefficients simulated using the MTP and SW. **b** NEB-based mono-vacancy jump barrier. DFT, MTP and SW are compared. AIMD configurations containing vacancies, along with NEB configurations, were excluded from the training set. Insets : Illustration of vacancy position before and after the jump within the silicon crystal

Dragoni *et al.*, 2018; Cheng *et al.*, 2020). The activation barriers for mono-vacancy hopping were further investigated using molecular dynamic simulations, considering temperatures ranging from 1000 K to 1650 K. The simulation details are provided in the method section for reference. The mean square displacements are also provided in the supplementary Fig. S4. As reported in Fig. 3a, the activation energy for mono-vacancy hopping is 0.31 eV for MTP and 0.41 eV for SW. These values are close to the static activation barriers computed at 0 K; the MTP model behaves in a physically plausible fashion. As observed in Fig. 3a-b, the barrier obtained from MD simulation is 0.32 eV, whereas the static migration barrier stands at 0.31 eV. It is noteworthy that NEB configurations, including AIMD configurations containing vacancies, were not included in the training set. We investigated extended defects such as generalized stacking faults and dislocations in the Si crystal. Generalized stacking faults are planar defects closely linked to slip. In turn, the behavior of dislocations and their core properties are particularly important for understanding plasticity. The methods section of our study provides a detailed description of our model for generalized stacking faults and dislocations, as well as the calculations involved. In Fig. 4a-b, the excess energy per unit area, also known as a γ -line, is presented. The γ -line was calculated using the benchmark method, the MTP model, the SW (Stillinger & Weber, 1985) and Tersoff (TS) (Tersoff, 1986, 1988b) semi-classical potentials. As shown in the Fig. 4, the MTP is in good agreement with the benchmark results. In contrast, SW and TS demonstrate lesser agreement with the benchmark. In Fig. 5, the dislocation core structures are presented. The core structures predicted by the MTP model exhibit a nearly perfect agreement with those predicted by DFT, as reported in Ref. (Guénolé *et al.*, 2010). An important and distinctive feature of our potential is the direct relaxation of the C_1 core structure to the C_2 structure, which is commonly referred to as the double-period reconstruction of the C_1 core. In past studies, the C_2 core structure was often manually reconstructed from the relaxed C_1 core (Huang *et al.*, 2021; Guénolé *et al.*, 2010). However, our potential eliminates the need for manual reconstruction by obtaining the C_2 structure directly. To obtain the relaxed structure and energy of the C_1 configuration, a snapshot is selected from the relaxation steps that ultimately lead to the C_2 structure. Our investigation also revealed that the C_2 core is the most stable configuration, a

result consistent with previous reports (Huang *et al.*, 2021; Guénolé *et al.*, 2010). In addition to Fig. 5, the core structures are also depicted in Supplementary Fig. S10.

Coexistence simulation

We determined the solid-liquid coexistence temperature of silicon using the solid-liquid interface method described in Ref. (Morris *et al.*, 1994). Our MTP potential predicts the silicon melting point to be 1485 ± 5 K, which is about 0.5% lower than the benchmark DFT-GGA value of 1492 K (Alfè *et al.*, 2003). Note that both the MTP potential and the benchmark value are approximately 12% lower than the experimental melting point of 1687 K, as reported in Ref. (Lide, 2004). Notably, our database initially lacked solid-liquid interface data, so we integrated a few AIMD configurations gathered around the experimental melting point into our unified training set. Additionally, it is important to acknowledge the influence of the exchange-correlation (XC) functional on melting behavior, as discussed in Ref. (Alfè & Gillan, 2003; Dorner *et al.*, 2018). The capability of MTP to accurately replicate the melting point of the GGA exchange-correlation (XC) functional showcases the high-quality simulation of liquid structure by MTP, as elaborated in the following section. For comprehensive details on our coexistence simulation approach, please refer to the accompanying supplementary Fig. S9.

3.3.2 Silicon slab energy

We compare the surface energy predicted by MTP against experimental data and DFT references, as slab data were not included in the training set. Experimental slab energy values for Si (100), Si(110), and Si(111) are reported to be 2.1, 1.5, and 1.2 J/m² (Jaccodine, 1963), respectively, while DFT values are 2.1, 1.8, and 1.6. The MTP predicts these surface energies to be 2.0, 1.3, and 1.2, respectively. While a large discrepancy between MTP and DFT is observed, especially for Si(110) with relative errors of up to 25%, the MTP-predicted surface energy closely matches the experimental slab energy.

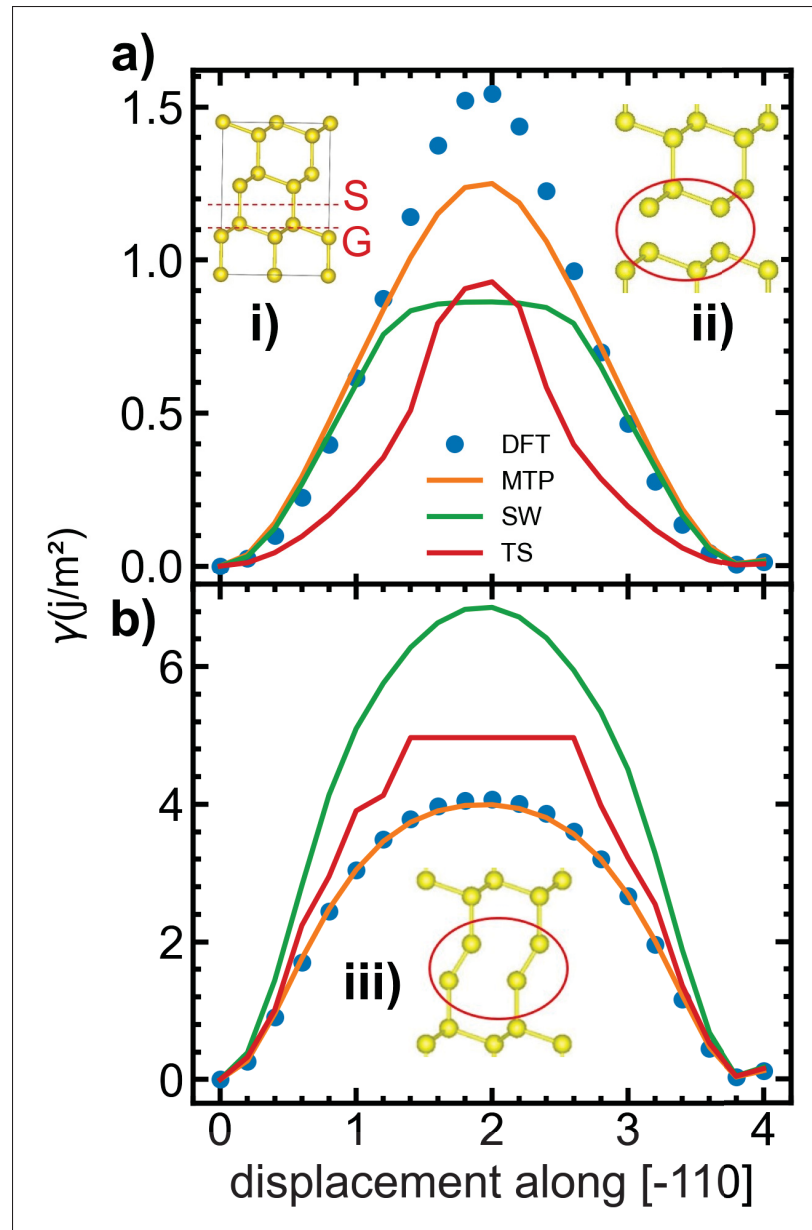


Figure 3.4 Planar defect in silicon crystal. γ -lines on the (111) plane as predicted by DFT, MTP, TS and SW. The shuffle (S) and glide (G) cuts are illustrated in the inset. The MTP provides a good description of the γ -line associated to the shuffle cut (top) and a near-perfect description of the γ -line associated to the glide cut. Insets i), ii), and iii) represent the bulk structure, the relaxed shuffle (S) structure, and the relaxed glide (G) structure, respectively

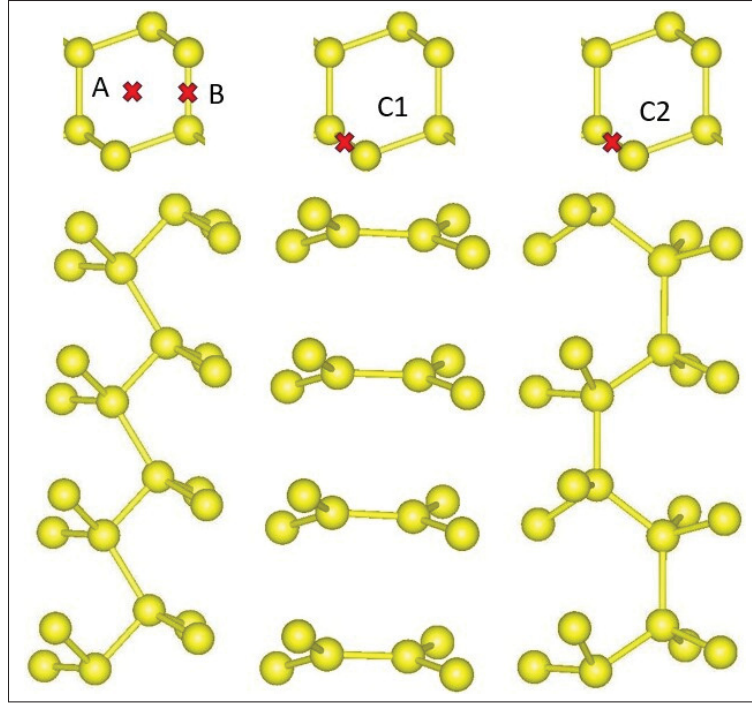


Figure 3.5 Line defect in silicon crystal. Positions and relaxed structures of $[110]$ screw dislocation cores in Si obtained with the MTP potential. The system size was set to 14400 atoms and oriented such that the x, y and z directions coincide to $[11-2]$, $[111]$, $[110]$ respectively. The dislocation core structures are in good agreement with DFT core structures reported in the literature (Guénolé *et al.*, 2010). The core types are represented by A, B, C_1 , and C_2 respectively. The red mark indicates the position of the dislocation core. Dislocation configurations were excluded from the training dataset

3.3.3 Disordered structures

We also considered an extensive set of disordered Si and SiO₂ structures. We compared structures generated using *ab initio* MD, MTP-MD, and semi-empirical-MD. All three cases were subjected to identical MD simulation conditions, except for amorphous silica, where the MTP simulation time was shorter compared to the other potentials. For comparison, we have chosen the Vashista (VA) (Vashishta *et al.*, 1990), Munetoh (TS) (Munetoh *et al.*, 2007), BKS (Van Beest *et al.*, 1990), and Sundararaman (SHK1 and SHK2) (Sundararaman *et al.*, 2018) models. The details

regarding the *ab initio* and classical MD simulations are comprehensively provided in the methods section. To analyze the disordered structures, we utilized both the pair correlation functions and the bond angle distribution functions.

3.3.4 Liquid and amorphous silicon

As observed from the radial distribution function shown in Fig. 6a, the MTP describes the structure of liquid Si with high precision. In contrast, the others potential leads to a shifted position of the first-neighbor peak compared to the results obtained from the reference (DFT). Additionally, there is an overestimation of the peak height, primarily observed with the EDIP and Tersoff potentials. The angular distribution function Fig. 6b also demonstrates excellent agreement between the MTP model and experimental data. When considering amorphous Si, the MTP model accurately describes the structural features in agreement with experimental data. Conversely, semi-classical potentials like SW and Tersoff fail to replicate the experimental radial distribution profile. As shown in Fig. 6c, experimental $g(r)$ and MTP-MD lead to nearly identical first-neighbor peaks (2.36 Å) and second neighbor peaks (3.89 Å). Additionally, the MTP model exhibits better agreement with the experimental bond angle distribution of 108.6° (Fortner & Lannin, 1989). Although the bond angle distributions of semi-empirical models like SW and EDIP are closer to the experimental values, they exhibit angle distributions below 90°, as shown in Fig. 6d. Note that the *ab initio* cooling and equilibration trajectory was not included in the training set of the MTP model, which suggests the MTP is fairly general, which can be attributed to the fact that the training dataset encompasses a variety of configurations within the disordered Si systems. Overall, the MTP model demonstrates a level of accuracy comparable to that of DFT and experiment when describing the structural features and bonding characteristics of disordered Si.

3.3.5 Liquid and amorphous silica

Fig. 7 illustrates results pertaining to liquid SiO₂. It includes pair distribution functions (PDF) and angular distribution functions (ADF). The MTP is in better agreement with DFT as compared

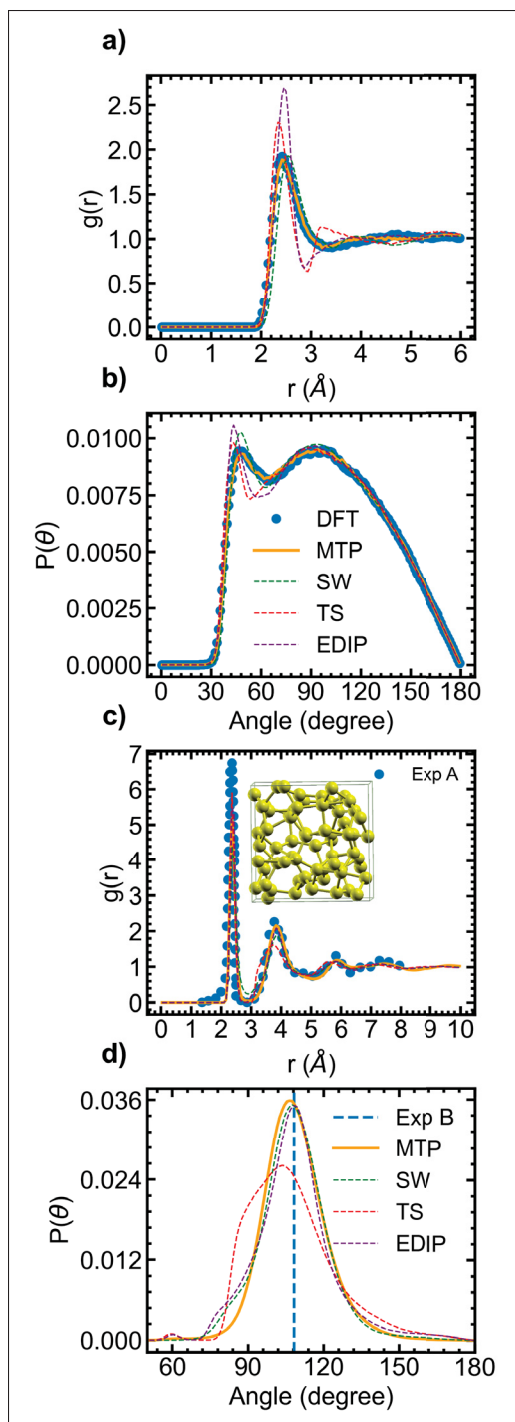


Figure 3.6 Liquid and amorphous silicon. Disordered structures of Si simulated using DFT, MTP, semi-empirical models **a** Radial and **b** angular distribution functions of liquid Si (3370K, 64 atoms); **c** Radial and **d** angular distribution functions of amorphous Si (300 K, 1000 atoms). The amorphous distribution functions are compared against experimental data from Exp A (Laaziri *et al.*, 1999a; Meidanshahi *et al.*, 2019) and Exp B (Fortner & Lannin, 1989)

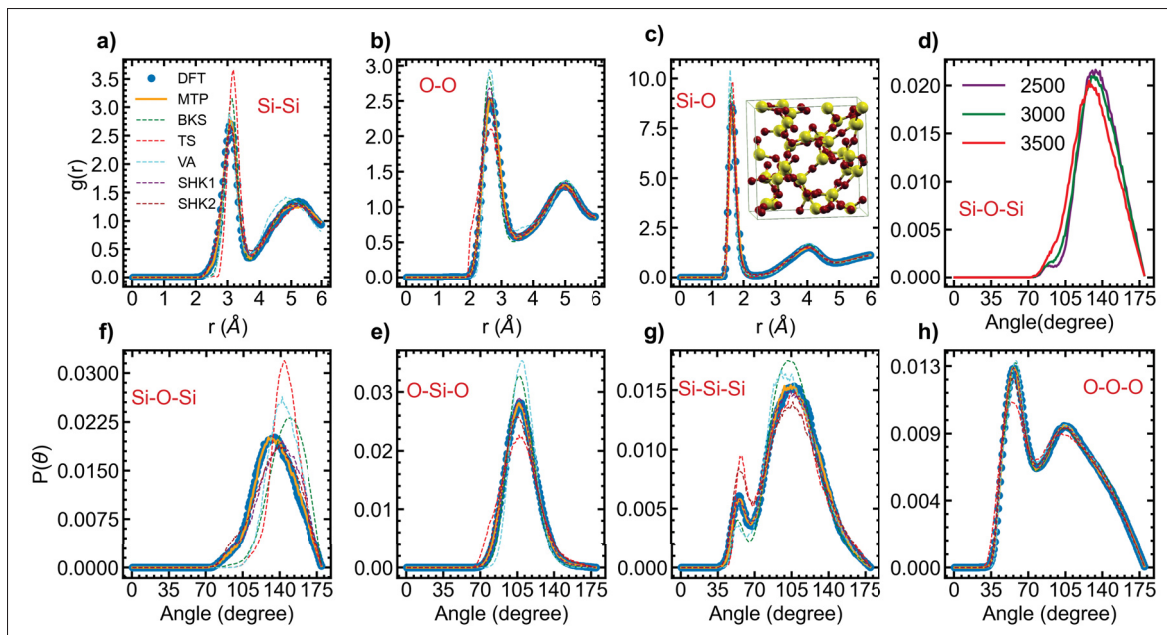


Figure 3.7 Liquid silica. Si-Si, O-O and Si-O correlation functions $g(r)$ and partial bond angle distribution function for liquid SiO_2 at 3600 K simulated using DFT, MTP, BKS, TS, VA, SHK1 and SHK2 with a 96-atom simulation box.
Top-right : Distribution of the Si-O-Si angle against temperature using the MTP potential

to the others potential. We notice both quantitative and qualitative differences between the semi-empirical potentials and DFT, except for the Si-O pair correlation function. In this case, the semi-empirical potentials only overestimates the height of the first peak, which is located around 1.62 Å (Mozzi *et al.*, 1969; Grimley *et al.*, 1990; Mei *et al.*, 2007). This value is consistent with the experimental Si-O bond length observed in liquid SiO_2 , indicating a strong chemical interaction between the Si-O pairs. Both the O-O and Si-Si pair correlation functions exhibit a shift in the first peak as given by the others models, while there is a strong quantitative and qualitative match between the MTP-based and the DFT-based structures. Furthermore, the BKS-based Si-O-Si ADF, along with those of other semi-empirical models, does not match the DFT-based ADF. Conversely, the MTP-based ADF demonstrates a good match. At 3600 K, using our 96-atom simulation box model, the average Si-O-Si angle between two SiO_4 tetrahedra is determined as follows : DFT - 134.5°, MTP - 131.4°, BKS - 146.0°, TS - 143.6°, VA - 141.8°,

SHK1 - 142.6° , and SHK2 - 140.4° . Our DFT value aligns closely with literature, approximately 136.0 (Carré *et al.*, 2016) and 135.0 (Liu *et al.*, 2019) respectively. It's evident that the MTP closely resembles the reference method, whereas other models align with experimental values for the Si-O-Si angle in amorphous silica, ranging between 140° and 152° (Farnan *et al.*, 1992; Da Silva *et al.*, 1975; Coombs *et al.*, 1985). This likely stems from the semi-empirical models being meticulously fitted with consideration of experimental properties. Most of them were optimized based on mixed *ab initio*-experimental data. We then varied the temperature of liquid silica from 2500 K to 3500 K using a 648-atom box and recorded the Si-O-Si angle, as depicted in Fig. 7d. We have found that the Si-O-Si angle in liquid silica changes with temperature, consistent with the findings reported in Ref. (Kobayashi *et al.*, 2021). As the temperature decreases, the angle between tetrahedral interconnections increases, contributing to network relaxation. It is likely that the relaxation of the network at room temperature upon cooling is primarily attributed to variations in bond lengths and angles, given that the network structure of silica liquid does not qualitatively change between 3500 K and 300 K.

Furthermore *Ab initio* MD, MTP-MD, BKS-MD and other models lead to a within-tetrahedra O-Si-O bond angle distribution centered around 109° , which is nearly equal to the experimental bond angle (Mozzi & Warren, 1969; Grimley *et al.*, 1990; Mei *et al.*, 2007). However, the BKS and VA potentials overestimate the average probability at 109° , while the TS potential underestimates this probability. The SHK1, SHK2, and MTP potentials match the benchmark. The DFT, MTP, and all other models except TS lead to very similar O-O-O ADFs. However, when it comes to the Si-Si-Si angle, qualitative and quantitative discrepancies are observed between BKS-generated structures and DFT-generated structures, while MTP-generated structures match the benchmark. There is also a notable discrepancy between the structure predicted by other models and that of the benchmark. The partial pair distribution functions of vitreous SiO_2 are illustrated in Fig. 8a-c. The comparison is made against the pair distribution function computed from experimental data using the reverse Monte Carlo (RMC) method (Tucker *et al.*, 2005). The MTP potential exhibits qualitative agreement with RMC data, as our pair distribution function profiles match those obtained from RMC. For example, only the MTP accurately reproduces the

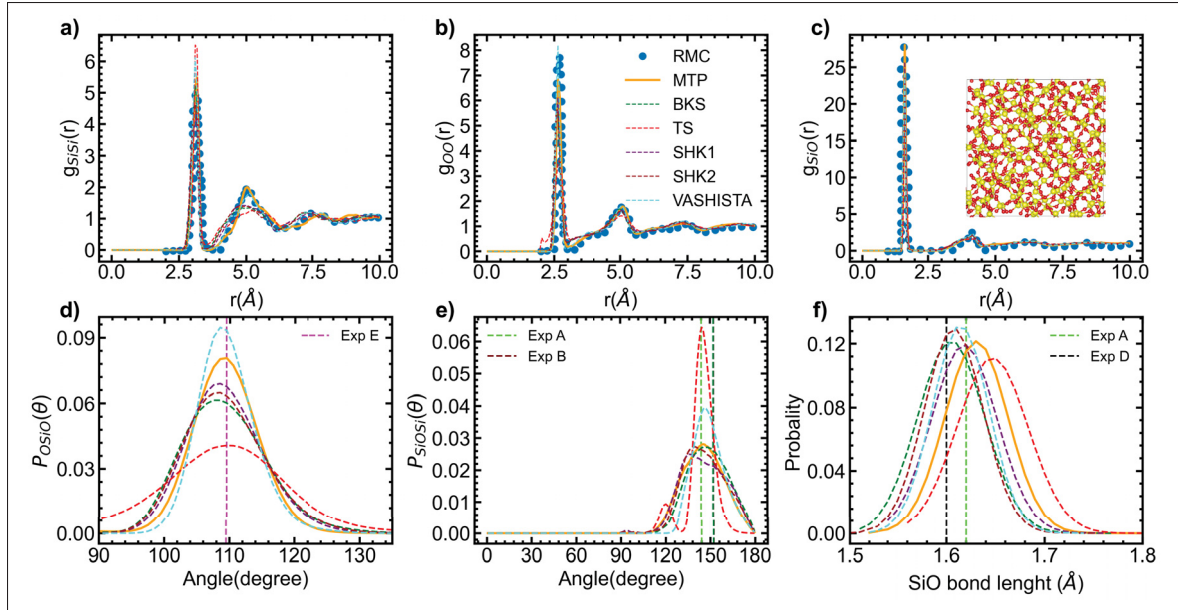


Figure 3.8 Amorphous silica. Partial radial distribution functions in vitreous silica were obtained using various potential including MTP, BKS, TS, VA, SHK1, and SHK2. The vitreous systems were equilibrated at 300 K, consisting of 648 atoms. These partial radial distribution functions are compared with experimental data obtained using Reverse Monte Carlo (RMC) (Tucker *et al.*, 2005). Additionally, the angular distribution function and bond length distribution function were analyzed and compared with experimental data from multiple sources : Exp A (Mozzi & Warren, 1969), Exp B (Da Silva *et al.*, 1975), Exp C (Vukcevic, 1972), and Exp D (Khouchaf *et al.*, 2020)

RMC profiles for Si-Si interactions, as the second peak around 5 Å is not reproduced by the other models, including the BKS model. Additionally, while the other models fail to reproduce the height of the Si-O RMC pair distribution function, the MTP potential shows a good match. To further analyze the structure, we also computed the most important angular distribution function, as well as the Si-O bond length distribution function. For O-Si-O (refer to Fig. 8c, all the profiles are similar but vary in height, centered around the experimental value of 109.0°. When it comes to Si-O-Si angle (Fig. 8d), both the MTP and BKS show similar profiles centered between the experimental values of 144° and 152°. The average values for BKS Si-O-Si angle is 150

and the one of the MTP is 145.5. As the average value for the same angle predicted by MTP in liquid silica at 3600 is around 131.4, this confirms that the Si-O-Si angle varies in liquid silica. Considering all the potential, the the Si-O bond length distribution are centered between 1.60 and 1.66 Å. The average bond length distribution for MTP potential is 1.63 Å which is close to experimental values of 1.62 Å. Note that these structural properties may vary slightly depending on the employed cooling rate. Despite the high cooling rate, no major coordination defects were observed. This indicates that the configuration has been well equilibrated at 3000 K, resulting in the establishment of a strong network. Note that the *ab initio* MD trajectory, which encompasses both the cooling and equilibration stages of the amorphous structure preparation, was not included in the MTP training set, which is an indicator of the MTP's generalization capability. Overall, the MTP potential demonstrates a remarkable improvement in accurately describing the structure of disordered SiO₂ compared to well-established potentials such as the BKS potential and others. The MTP model captures the essential structural features of the disordered systems with greater precision, resulting in a better agreement with experimental observations (Mozzi & Warren, 1969; Grimley *et al.*, 1990; Tucker *et al.*, 2005; Mei *et al.*, 2007).

3.3.6 Phonon dispersion

We calculated the phonon dispersion of c-Si and α -quartz, as illustrated in Fig. 9a-b respectively. The MTP model exhibits very good agreement with the reference method (DFT) except the higher frequencies for α -quartz. Once again, the corresponding frozen phonon configurations were not explicitly included in the training set.

3.3.7 Si - SiO₂ Interface

The Si-SiO₂ interface, a cornerstone in semiconductor physics and material science, plays a fundamental role in device fabrication and significantly impacts device performance. Exploring hetero-structures involving Si - SiO₂ interfaces opens avenues for novel functionalities and applications in microelectronics and beyond. Given its importance, capturing the structure and

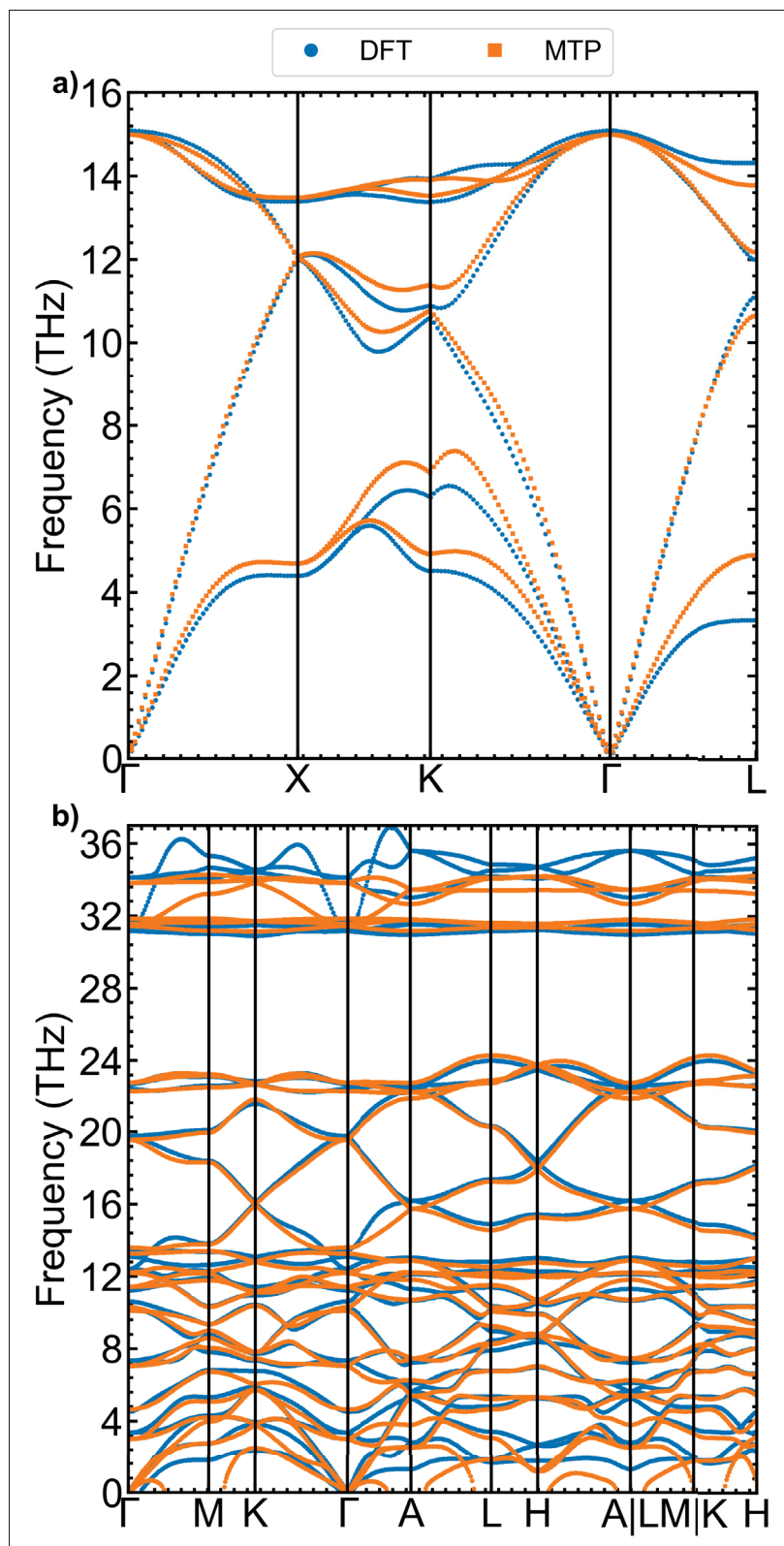


Figure 3.9 Phonon dispersions. Phonon dispersions of **a** *c*-Si and **b** α -quartz computed using DFT and MTP

dynamics of the Si - SiO₂ interface is paramount for a potential model of the Si-O system. To achieve this, we employed various models, encompassing crystalline Si slabs with different orientations, such as Si(100), Si(110), and Si(111). Our approach involved utilizing both α -quartz, β -cristobalite, and other polymorphs in constructing the Si - SiO₂ interface. For a detailed explanation of our construction scheme, please refer to the Method section. To assess the suitability of the potential for modeling the Si/SiO₂ interface, we examined interfaces using both O-terminated SiO₂ slabs and Si-terminated slabs. For each Si/SiO₂ crystalline interface configuration, we perform force and energy minimization, allowing the positions of atoms and the simulation box size to change simultaneously. Following geometry optimization, molecular dynamics simulations are conducted for 50 ps at 300 K in the NPT ensemble. First, our potential stabilized the majority of the interfaces in both static and dynamic runs, with the extent of stabilization depending on the orientation of the Si slab and the termination of the quartz slab. For Si(110) in contact with either a Si- or O-terminated quartz slab, our potential successfully stabilizes and describes the dynamics of the resulting interfaces, whether symmetric or non-symmetric. However, our potential only successfully describes symmetric interfaces for Si(100) and Si(111), which are built from quartz slabs terminated by Si. These combinations of termination and orientation were not considered in the training set, showcasing the remarkable generalization ability of our interaction potential. It's worth noting that defects, such as silicon dangling bonds or over-coordinated oxygen atoms, are observed at the interface following both minimization and dynamic runs, as depicted in Fig. 10. As noted in Ref. (Himpsel *et al.*, 1988), the presence of the dangling bonds is a natural occurrence and constitutes a typical aspect of interface defects. Such anomalies are commonplace and anticipated in these interfaces, owing to the inherent lattice mismatch between the involved materials. Usually, interfacial defects are passivated or special construction schemes are adopted to eliminate them. However, our study does not aim to create defect-free interfaces. Instead, our goal is to evaluate the potential's capability to manage complex heterostructures with varying bonding types not encountered during the training process. To further validate our potential for silicon and silica interfaces, we compare the interfacial energies of small models computed using MTP and DFT. Detailed descriptions of these small models are provided in the Methods section. Our results (shown in

Fig. 11) exhibit good correlation between DFT and MTP models. Importantly, configurations generated by the MTP potential through relaxation and MD simulations converged easily, typically requiring fewer than 50 iterations of single-point energy calculations of DFT. This highlights the reliability of interfacial configurations generated by the MTP potential. These findings demonstrate the capability of MTP potentials to effectively investigate heterosystems containing Si-SiO₂ interfaces. The relaxed small models of the Si-SiO₂ interfaces are presented in the supporting information (Fig. S11 and Fig. S12).

3.4 Discussion

In this work, we have successfully parametrized a machine learning (ML) potential that can implicitly capture and describe different charge states. Additionally, this ML potential has the remarkable ability to describe disjoint zones and hetero-zones of the configurational space of SiO₂ and its constituent elements, Si and oxygen. The potential description of various phenomena—including point defects, diffusion in Si crystals, extended defects, the liquid phase of Si and SiO₂, and the amorphous phase of Si and SiO₂—either rivals or out-performs existing potentials. The potential exhibits very good agreement with experimental data in challenging configurational zones, such as the amorphous state, even though these configurations were not included in the training data. In many scenarios, such as disordered phases (liquid, amorphous), the potential achieved a near-perfect match with the reference method in terms of accuracy, using exactly the same simulation time (very short) and conditions. Even when utilizing longer simulation times and large system with a semi-empirical model, it does not reach a level of accuracy similar to the reference method. Furthermore, the potential displays an intriguing capability regarding dislocation behavior. It autonomously transitioned the C₁ structure to the C₂ structure without the need for manual reconstruction. While there is a growing consensus that machine learning (ML) potentials can effectively serve as surrogates for density functional theory (DFT) in terms of accuracy and speed, generalization including charge state modeling remains a challenging task. This study provides evidence that reaction coordinates are sufficient to implicitly capture charge transfer or charge states involved in chemical reactions.

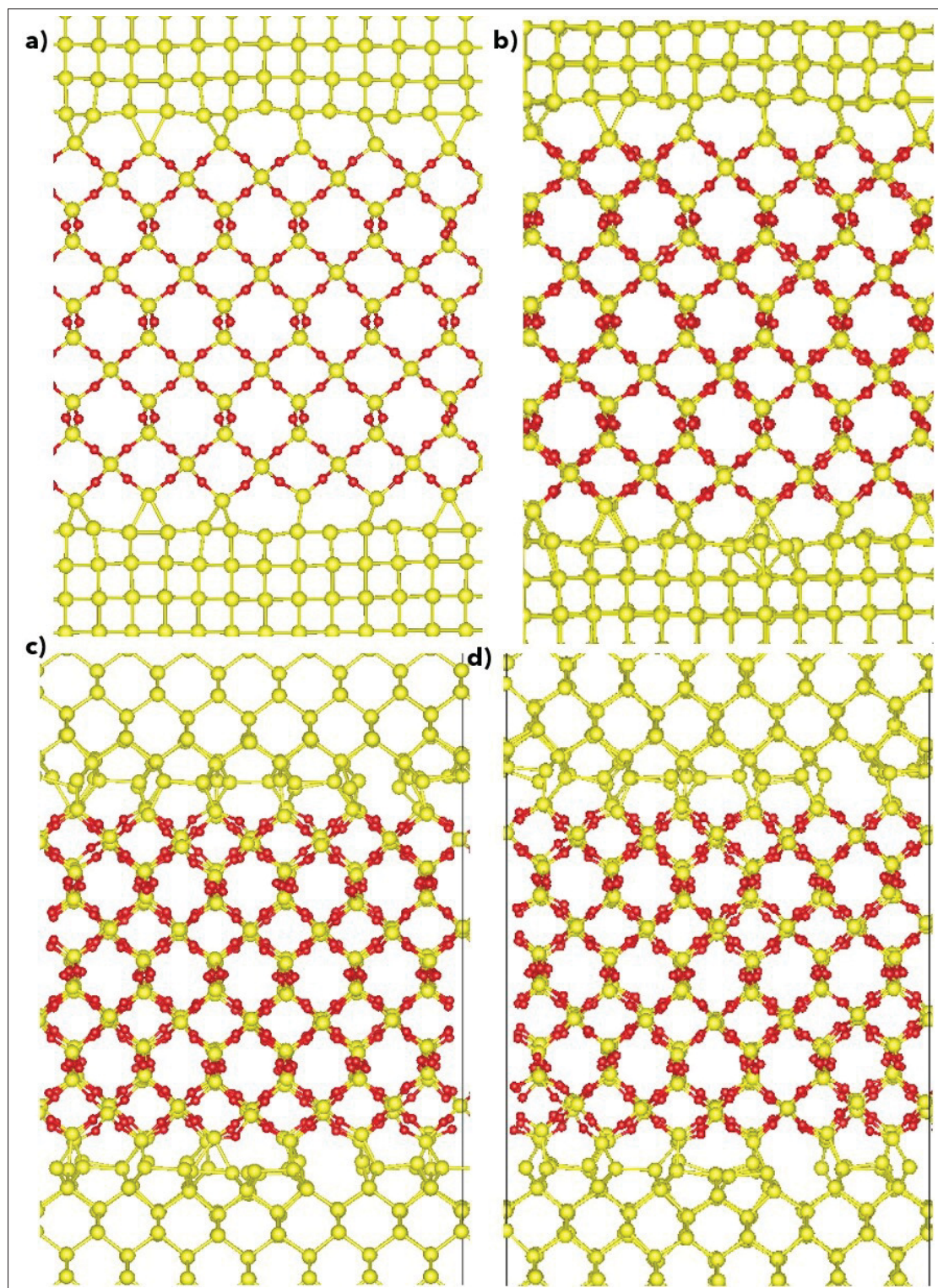


Figure 3.10 Interface structure of Si - SiO₂. Geometry optimization and MD equilibration of 50 ps simulation : Si(010)/ α -Quartz (001) (top) and Si(110)/ α -Quartz (001) (bottom) **a** & **c** Relaxed at 0 K, **b** & **d** Annealed at 300 K

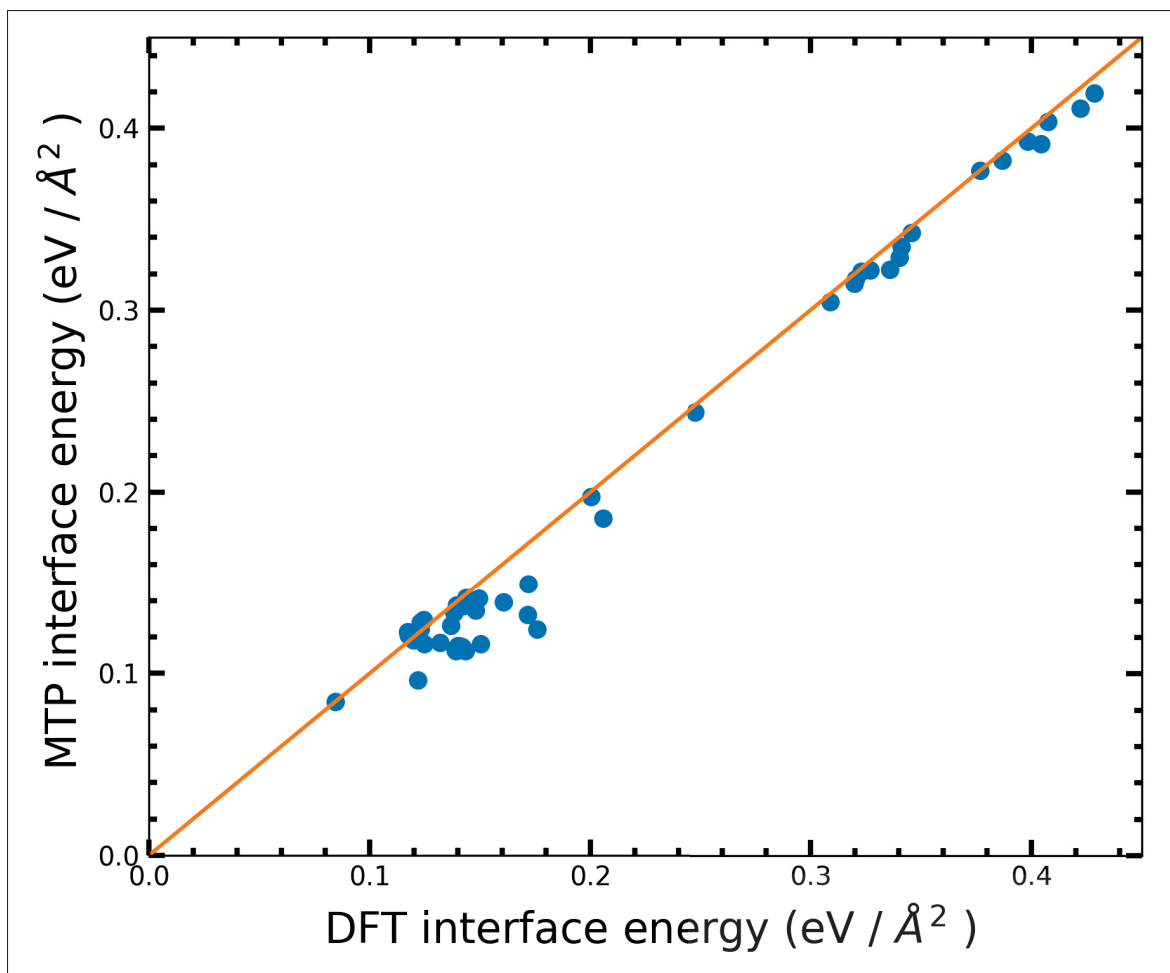


Figure 3.11 Interface energy of Si-SiO₂. Comparing interface energies of silicon and various silica polymorphs, including amorphous silica, using MTP potential and DFT. The orange line indicates the diagonal $y = x$, corresponding to a perfect correlation

Aside from potentials explicitly considering charge-transfer, separate machine learning (ML) potentials are developed for each individual chemical element or compound which database is constructible by DFT. Our approach suggests this is not necessary; compounds and their elemental constituents can be trained jointly. Indeed, joint parametrization, where parameters are derived simultaneously for both silicon, silica and oxygen, offers several advantages. By employing a single potential to describe the interactions between atoms in both silicon and silica, the computational model becomes more streamlined and easier to manage. The unified potential

can save time and resources by avoiding the need to recalibrate parameters of the model for each of the materials involved. The unification idea is also important for some areas of application, such as interfacial modeling for electronic devices, energy storage and conversion, and surface coatings and tribology. Here, the reference data for machine learning potential must include both the individual materials in contact as well as the boundary region. In addition, as chemical reactions can occur in MD simulation, good modeling of a multi-component material or complex systems under certain conditions requires joint parametrization of the considered system and its constituents. For instance, oxygen aggregation in high-temperature molecular dynamics (MD) simulations was observed by researchers, as noted in Ref. (Erhard *et al.*, 2022) (supporting information). This observation led to the inclusion of oxygen molecules in the training set by the researchers. Our preliminary study pertained to a semiconductor and its oxide. However, whether this approach can be generalized to other elements and mixtures—including multi-component alloys and compounds remains to be seen. To ensure a well-implemented unified interaction potential, several other aspects need to be explored in the future. Given that the compound and its constituents are not located within the same zone of the configurational space, achieving an accurate, efficient, low-cost, and general unified potential for both the material and its oxides with limited data may require adjustments to the underlying mathematical model, the fitting procedure, and the database sampling methods (including active learning). These adjustments could help attain the same level of accuracy at a more affordable computational cost, resembling a feature of potentials parameterized for a single compound. Likewise, while this study achieved a joint description of Si and SiO₂ using the MTP framework, it is likely that other currently developed ML potential frameworks would have led to a similar result. In conjunction with these questions, our aim in the future is to extend this work by incorporating the element of hydrogen to model silica gels.

3.5 Methods

3.5.1 *Ab initio* calculations

The database was constructed using DFT, as implemented in the Quantum ESPRESSO (Giannozzi *et al.*, 2009) package. The exchange-correlation potential was treated using the generalized gradient approximation of Perdew-Burke-Ernzerhof (GGA-PBE) (Perdew *et al.*, 1996). Projector augmented waves (PAW) (Blöchl, 1994) were employed. Kinetic energy cutoffs of 884 eV for Si and 1224 eV for both SiO₂ and oxygen were chosen. In all calculations, the Brillouin zone was sampled using the Monkhorst-Pack grid (Monkhorst & Pack, 1976) scheme. Different k-points were used for each polymorph, including an 8x8x8 for the ordinary phases of Si and an 11x11x11 for SiO₂. The gamma point was used for oxygen molecules.

3.5.2 Machine learning model : The moment tensor potential

In this work, the moment tensor potential (MTP) (Shapeev, 2016) was chosen as the machine learning model. The moment tensor potential (MTP) is a multi-component potential. In a previous comparative study (Zuo *et al.*, 2020), it demonstrated a favorable trade-off between accuracy and computational speed across a range of modeling problems. The model derives its name from its use of a tensorial representation of atomic coordinates and utilizes linear regression to determine the local atomic energy. These local atomic energies are subsequently summed to obtain the total energy of the system under consideration. The MTP model considers the total energy of a specific atomic configuration as a sum of individual atomic energy contributions.

$$E_{Total} = \sum_{i=1}^n E_i = \sum_{i=1}^n V_{local}(\zeta_i) \quad (3.1)$$

The argument ζ_i is a tuple $\zeta_i = (r_{ij}, \tau_i, \tau_j)$ containing the relative coordinate r_{ij} and atomic types τ_i, τ_j . Here, V_{local} is approximately computed within the sphere or circle of radius (R_c) of 5.7 Å, beyond which the central atom no longer feels any interaction. Practically, in the MTP framework, the expansion of the atomic energy V_{local} into basis functions B_β serves as the

foundation for linear regression. within the sphere or circle of radius

$$V_{local} = \sum_{\beta} c_{\beta} B_{\beta} \quad (3.2)$$

Since the potential energy function V_{local} is smooth, the force acting on an atom k at position r_k in a given configuration x_q can be calculated by taking the gradient of the total energy.

$$F_k(x_q) = -\nabla E_{Total}(x_q) = -\sum_i \frac{\partial V(\zeta_i)}{\partial r_k(x_q)} \quad (3.3)$$

The virial stress within an atomic configuration x_q of volume Ω can be expressed as follows.

$$\sigma_{ij}(x_q) = \frac{1}{2\Omega} \sum_{k \in \Omega} \sum_{l \in \Omega} (x_i^{(l)} - x_i^{(k)}) F_j^{(kl)} \quad (3.4)$$

The functions B_{β} in equation 2 are obtained through the contraction of the descriptors. In the MTP model, the descriptors are formed by tensors of atomic coordinates weighted by radial functions. These descriptors consider both the radial distribution and the angular distribution of the neighborhood surrounding each atom. By incorporating information from both the radial and angular aspects, the descriptors capture the local atomic environment in a more comprehensive manner, enabling a more accurate representation of the atomic energy within the MTP framework.

$$M_{\mu, \nu}(r_{ij}, \tau_i, \tau_j) = \sum_j f_{\mu}(|r_{ij}|, \tau_i, \tau_j) \underbrace{r_{ij} \otimes \cdots \otimes r_{ij}}_{\nu \text{ times}} \quad (3.5)$$

The radial function f_{μ} , is further expanded using radial basis functions $Q^{(\alpha)}$ and fitting parameter $c_{\mu, \tau_i, \tau_j}^{(\alpha)}$ as expansion coefficients. This expansion allows for a more flexible and accurate representation of the radial dependence of the atomic interactions.

$$f_{\mu}(|r_{ij}|, \tau_i, \tau_j) = \sum_{\alpha} c_{\mu, \tau_i, \tau_j}^{(\alpha)} Q^{(\alpha)}(|r_{ij}|) \quad (3.6)$$

The model parameters $\Theta = (c_\beta, c_{\mu, \tau_i, \tau_j})$ are determined during the minimization of the cost function as given by Equation 3.5.

3.5.3 Data curation and optimization

We acquired *ab initio* data using established methods and databases from prior research. The database construction involved two methodologies, namely manual processing (Bartók *et al.*, 2018; Zuo *et al.*, 2020; Zongo *et al.*, 2022) and active learning (Podryabinkin *et al.*, 2017), as explained deeply in supporting information. Specifically, we referred to the dataset created for the Gaussian approximation potential for silicon (Bartók *et al.*, 2018), the comparative study (Zuo *et al.*, 2020), and active learning techniques detailed in Ref. (Podryabinkin & Shapeev, 2017). For liquid silica, we utilized the temperature range (1000 K - 5000 K) from previous databases specifically designed for neural network inter-atomic potentials (NNIP), which covered temperatures exceeding its boiling point and extended as high as 5000 K, as referenced in (Balyakin *et al.*, 2020). While NNIP potentials involve a very large number of adjustable parameters—typically tens of thousands—allowing to jointly describe a large number of off-equilibrium configurations, accommodating such deviations from equilibrium becomes challenging within the MTP framework due to limited numbers of parameters. Effective MTP training, therefore, relies on carefully selecting the training set. Our final training dataset, herein referred to as the unified training set, was constructed through a two-step process : curation and subsequent optimization.

To enhance the quality of our training set and to properly assess the error of the test set, down selection was applied. To begin, we sorted our full database into smaller subsets as elaborated in Tab. S3 through S6 in the supporting information. Then, within each subset, we utilize a filtering strategy referred to herein as the "train-remove-train" approach. We first train while monitoring for significant reductions in energy and force errors associated to each configuration as we incrementally raise the MTP level by one unit (We focus on levels 08 to 14 for deformation and defects, while levels 16 to 18 are used for disordered structures). Next, we analyze the error

reduction between 2 or 3 consecutive levels. If a significant decrease is not observed, we then eliminate configurations based on factors such as :

- (1) Total energy : configurations with similar total energies yet differing atomic coordinates to other in the training set, as well as those with fluctuating total energies but nearly identical atomic positions to others in the training set are removed. These configurations originated from relaxation of molecules, single-point calculations of unrelaxed defects, manually constructed unrelaxed jump paths, and strained configurations where lattice parameters or vectors were strained without corresponding adjustments to atomic positions and embedded dimer.
- (2) Contributions from smearing : this can be primarily attributed to the extensive use of high-temperature AIMD simulations. Silica configurations generated from AIMD simulations, featuring a number of atoms greater than 36, are excluded if they exhibit smearing contributions to the total energy greater than zero.
- (3) Minimum interatomic distances (this criteria complements the total energy criteria) : if multiple configurations from the same batch exhibit similar minimum distances, some are removed. This technique was mostly used for oxygen molecules. For example, we check the inter-atomic distance in the batch of relaxed O_3 molecules. We also discard embedded dimer configurations by comparing their minimum inter-atomic distances with those of the AIMD configurations. If the minimum inter-atomic distances are equal, we choose AIMD over embedded dimer configuration.
- (4) Polymorphism : These configurations are derived from deformations following thermal expansion. Most of these configurations have been accurately computed and were included in the training database. However, some polymorphs arising from displacive phase transformations and polymorphs sharing the same lattice system with identical coordination numbers were excluded. In the case of displacive phase transformation, we retain the parent crystal and exclude the child crystal. When dealing with two polymorphs that share the same lattice system and coordination number, we typically choose one of them.

After removing these undesired configurations, where we reinitiate the training process incrementally, following a pattern akin to the first stage. This “train-remove-train” process is iterated until we attain a high level of confidence in the cleanliness of the subset—*i.e.* all configurations included in the subset are associated with training errors that decrease as the MTP level increases. In the subsequent curation phase, we examine the possibility of extracting an even smaller subset from each previously cleaned subset. (Podryabinkin & Shapeev, 2017). To achieve this, we use the "select-add" command embedded in the MTP code, as described in Ref. (Novikov *et al.*, 2020b). We applied the "select-add" command to every cleaned subset.

3.5.4 Training set

In implementing the unified potential, we selected a range of configurations from the optimized and curated database, as outlined in the Data Curation and Optimization section. Similarly, the test set was chosen from the same optimized database to eliminate any overlap between the two sets. While the test set encompasses all properties or types of configurations represented in the extensive database, the training set only includes certain configuration types. This strategic approach aims to ensure the portability of the interaction potential. Note that the database contains configurations of substantial size, with up to 1000 atoms. However, we constrained the maximum box size in the training set to 36, except for the interfaces set where a few configurations are of size 80. Consequently, the number of atoms in the training set’s boxes ranges from 1 to 80. Conversely, the test set contains configurations of the maximum size found in the database. The specific types of configurations represented in the training set are detailed in Tab 3.5. By restricting the training set to specific configuration types, we aim to enhance the generalizability of our potential by simulating properties not present in the training set. Within the current implementation of the MTP potential, we do not employ a validation set in the typical manner used to estimate overfitting or underfitting during the training of neural network models. Instead, we opt to utilize an offline test set for this purpose. To gauge the potential’s portability more comprehensively, we perform molecular dynamic simulations for properties that were not included in the training set.

Tableau 3.5 Final refined and optimized training dataset derived from extensive uncured database. Please refer to the supporting information for specific details about the number of atoms per configuration

Cleaned and optimized subset	Total
Vacancy Si	50
Single interstitial Si	63
Stacking fault Si	40
Isolated atom Si and O	2
Deformation Si	99
Deformation SiO ₂	139
Liquid Si	217
Liquid SiO ₂	353
Oxygen molecules	21
Interface Si/SiO ₂	164
Embedded dimer (Si, SiO ₂)	2
SiO ₂ slab	9
Unified training set	1159

3.5.5 Training and validation

The cost function, as described by Equation 3.5, was minimized using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, a quasi-Newton optimization method implemented in the MTP framework. Fundamentally, training the MTP model with atomic configurations entails finding the parameter set $\{\Theta\}$ by solving the minimization problem presented in Equation 3.5. We trained the refined unified training sets as detailed in Tab. S7 (supporting information) and Tab. 3.5. First, MTP potentials with parameter sets ranging from 300 to 1600, corresponding to levels 18 to 26, were used to train set 1 (Tab. S7 and Fig. S1(a)) as part of preliminary works, including optimization, training mode, and testing. The preliminary works are also presented in supplemental information from Fig. S5 to Fig. S8. At this stage, two training modes were employed : vibration mode and structures weighting mode. Specifically, the potential resulting from the vibration mode was used as input for the structures weighting mode training. The training process iterated until a desirable level of accuracy was achieved. As outlined in the supporting information, we assessed the validation error using the resulting potentials. We

employed two independent validation sets, denoted as Validation 1 and Validation 2, which atom distributions are shown in the supplementary Fig. S2 and Fig. S3, respectively. Validation 1 was randomly selected concurrently with the training set from the curated database. On the other hand, Validation 2 consisted of the AIMD cooling and equilibration trajectories at 300 K. These validation sets were utilized as part of preliminary work. For the final implementation, we utilized the level 28. Based on the preliminary work, we opted for the structure weighting mode. The two-step training mode, as applied to large-size configurations in set 1 (Tab. S7 and Fig. S1 a), was deemed unnecessary. Given that our final training set (Tab. 3.5) comprises small cell configurations, we exclusively employed the structure weighting mode. . We utilized the resulting potential to conduct both static calculations and molecular dynamic simulations, with the outcomes presented in the main text.

$$\sum_{i=1}^n \left[w_e (E^{mtp}(x^{(i)}, \Theta) - E^{qm}(x^{(i)}))^2 + \right. \\ \left. w_f \sum_{j=1}^{N_a(x^{(i)})} |F_j^{mtp}(x^{(i)}, \Theta) - F_j^{qm}(x^{(i)})|^2 \right. \\ \left. + w_\sigma |\sigma^{mtp}(x^{(i)}, \Theta) - \sigma^{qm}(x^{(i)})|^2 \right] \rightarrow \min$$

Here, E^{qm} , F^{qm} , σ^{qm} denotes the values of energy, force, and stress computed by the quantum mechanical approach (DFT), while E^{mtp} , F^{mtp} , σ^{mtp} represents the corresponding values obtained from the MTP model. w_e , w_f , w_σ are the relative weights indicating the importance of the energy, the force and stress in optimization procedure.

3.5.6 Static calculations

This section summarizes the mathematical procedure used to determine static properties presented in the article.

For the chemical component with a formula $X_l Y_m Z_n$, we calculate the cohesive energy as follows :

$$E_{coh} = E_{X_l Y_m Z_n} - (lE_X + mE_Y + nE_Z). \quad (3.7)$$

Where $E_{X_l Y_m Z_n}$ represents the energy of the supercell of the compound, while E_X , E_Y , and E_Z correspond to the energies of the isolated atoms. The subscripts l , m , and n indicate the number of X, Y, and Z atoms present in the building block $E_{X_l Y_m Z_n}$ of the material. Due to variations in the number of atoms within the primitive cell of each polymorph compared to the standard structural configuration, we normalize the cohesive energy by dividing it by the number of atoms present in the regular phase.

Points defects formation properties such as vacancy formation energy (E_v^f) was calculated using this equations :

$$E_v^f = E_{N_0-1} - \frac{N_0 - 1}{N_0} * E_{N_0}. \quad (3.8)$$

For interstitial formation energy E_i^f , we used :

$$E_i^f = E_{N_0+1} - \frac{N_0 + 1}{N_0} * E_{N_0}. \quad (3.9)$$

In equations (9) and (10), N_0 and E_{N_0} correspond to the number of atoms and total energy of a perfect supercell.

Particularly, vacancies in SiO₂ polymorphs were estimated considering a neutral state. Thus, the formation energy in SiO₂ polymorphs was calculated using :

$$E_f = E_{vac} - E_{bulk} + \mu_O. \quad (3.10)$$

In this equation, E_{vac} and E_{bulk} represent the energy of the supercell containing the oxygen vacancy and the energy of the bulk supercell, respectively. The chemical potential is defined as half of the energy of a dioxygen molecule ($\mu_O = 1/2 * E_{O_2}$).

The equilibrium bulk modulus which correspond to the curvature of the energy-volume curve at its minimum was derived from the second order elastic constants (Pandit *et al.*, 2023). We calculate elastic stiffness constant C_{ij} using central finite difference formula.

$$C_{ij} = \frac{P_i^{(+\varepsilon_j)} - P_i^{(-\varepsilon_j)}}{2 * \varepsilon_j}. \quad (3.11)$$

where $P_i^{(+\varepsilon_j)}$ is the i^{th} component of the stress tensor when the configuration is strained only by j^{th} component (ε_j) of the strain vector ($\vec{\varepsilon}$). After applying directional or isotropic deformation, the atomic positions undergo relaxation while the overall box size remains fixed. We compute the generalized stacking fault energy ($\gamma(u)$) by incrementally shifting the upper crystal half along the slip direction and assessing energy differences per unit area (A) of the fault plane.

$$\gamma(u) = \frac{E(u) - E_o}{A}. \quad (3.12)$$

where, E_o represents the energy of the perfect crystal, while $E(u)$ denote the energy of the supercell with the fault vector u which is directly proportional to the Burgers vector (b). Surface energy is also calculated using the following expression :

$$\gamma = \frac{E_{slab} - NE_{bulk}}{2A}. \quad (3.13)$$

In this context, A refers to the area of the slab, N represents the number of atoms in the slab, while E_{slab} and E_{bulk} denote the total energy of the slab and the bulk energy per atom, respectively.

3.5.7 Si - SiO₂ Interface construction

Previous reports indicate that defects are commonly observed at the interface Si - SiO₂ due to the imperfect matching of the two materials. To avoid considerable lattice mismatch, we utilize specific techniques. First, we rotate the alpha-quartz structure to achieve a tetragonal configuration. Next, we duplicate both the silicon crystal and alpha-quartz structure, ensuring that the lattice dimensions perpendicular to the interface direction closely match. This approach

enables us to apply a small strain (less than 2 %) to the lattice vectors before forming the interface. Technically, the lattice mismatch α can be defined as the relative difference in lattice parameters between two crystalline materials, often expressed as a percentage or in terms of the absolute difference in lattice constants along specific crystallographic directions :

$$\alpha = \frac{n * L_1 - m * L_2}{n * L_1 + m * L_2}. \quad (3.14)$$

Lattice duplication factors are represented by integers n and m ; L_1 and L_2 denote the lattice parameters of a given direction. In both cases, symmetric and asymmetric interfaces were constructed for both oxygen-terminated and silicon-terminated quartz slabs, incorporating Si (100), Si (110), and Si (111) slab orientations. Our objective is not solely to construct a flawless interface representation of a naturally occurring or real-world interface, but rather to explore the versatility of the potential. We then estimate the interface energy using :

$$\gamma = \frac{E_S - (n_{SiO_2} * E_{SiO_2} + m_{Si} * E_{Si})}{A}. \quad (3.15)$$

Where A represents the area of the interface, n_{SiO_2} and m_{Si} represent the number of formula units of SiO_2 and Si in the interface system. E_S is the energy of the supercell containing the interface. The terms E_{SiO_2} and E_{Si} correspond to the energy of silica and silicon per formula unit, respectively. Due to the impractical size of duplicated models for energy computation via density functional theory (DFT), smaller superlattices were also constructed involving Si (100) interfacing with α -quartz, β -cristobalite, α -cristobalite, β -tridymite, and amorphous silica. This facilitated comparison between results obtained using MTP potentials and those from DFT. Each simulation box contains two distinct interfaces. Initially, these small models were relaxed at 0 K using MTP potentials. Subsequently, MTP-driven molecular dynamics (MD) simulations were performed at temperatures of 300 K, 500 K, 800 K, and 1200 K for 100 ps, and configurations were selected from the trajectories. The energies of these selected configurations, as well as the relaxed configurations, were then computed using DFT-based single-point energy calculations.

3.5.8 Molecular dynamics simulation

Ab initio molecular dynamics (AIMD) simulation was carried out using Quantum Espresso using the parameters as described in the section "*Ab initio* calculations details." The integration timestep was set to 1 fs for Si and 2 fs for SiO₂. The ionic temperature during simulations was controlled using velocity rescaling.

Force-field MD simulations were performed using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) software package (Thompson *et al.*, 2022). While it is impractical to perfectly replicate the MD settings in Quantum Espresso within LAMMPS, we aimed to make them as close as possible. To generate disordered structures, we employed the velocity rescaling thermostat to control the temperature during the simulations. The time integration used the same time steps as in the AIMD simulations. For studying point defects diffusion in Si and self-diffusion in SiO₂, the Nosee-Hoover thermostat (Evans & Holian, 1985) was employed. The latter simulations were performed using a time step of 1 fs, and the damping parameter was set to 100 fs.

Data availability

The Si-O-SiO₂ database and the unified potentials can be found at <https://gitlab.com/Kazongogit/MTPu>.

Code availability

The main codes used for this work are Quantum ESPRESSO (version 6.8), LAMMPS (version 2022), PHONOPY, and MLIP-2 (version 2). They are available at <https://www.quantum-espresso.org>, <https://lammps.sandia.gov>, <https://phonopy.github.io/phonopy> and <https://mlip.skoltech.ru> respectively. Further details can be found in the GitLab repository <https://gitlab.com/Kazongogit/MTPu>. Additionally, Python scripts were written for data manipulation and processing; most are available on the GitLab repository.

3.6 Supplementary Information for : A unified moment tensor potential for silicon, oxygen, and silica

3.6.1 Database preparation

The ab initio reference data plays a crucial role as one of the main components in machine learning inter-atomic potential (MLIP) methods. The success of an MLIP in achieving near DFT-accuracy and transferability relies not only on the model itself but also on the quality and characteristics of the reference data used. The features of the reference data should encompass the quality of the quantum mechanical calculations, the diversification of atomic configurations, and the coverage of the configurational space. To construct a data-driven reactive force field, it is essential to consider various bonding and coordination scenarios, including charge transfer. This is made possible by examining polymorphs of the materials and conducting AIMD simulations at different temperatures and pressures. For instance, the coordination number changes from 4 in cubic diamond silicon to 6 in beta-tin and hexagonal silicon at higher pressure (Lenosky *et al.*, 2000). Some of the silicon polymorphs exhibit varying bond angles in their minimum energy state (Kumagai *et al.*, 2007; Kasper & Richards, 1964). Another example includes silica polymorphs, where silicon and oxygen are typically four-fold and two-fold coordinated, except in the case of seifertite and sishovite structures, in which silicon is six-fold coordinated and oxygen is three-fold coordinated, respectively. In addition, during molecular dynamics simulations, over/under coordination is sometime inevitable. Thus, we have considered diverse configurations, including crystal structures, strained crystal structures, point defects, extended defects, and disordered structures, in which our materials of interest may chemically exist. Indeed, these configurations represent different states or phenomena of the materials under external conditions. Most of the phenomena, such as defect migration mechanisms, defect stability, and phase changes, were considered in the construction of our database. We followed the procedure described in (Zongo *et al.*, 2022) to generate our database. However, in this article, several additional techniques and methods were incorporated to complement the ones described in (Zongo *et al.*, 2022) to create a comprehensive database of Si/SiO₂/O. The construction of the database involved various types of calculations, including single point calculations, geometry

optimizations, and ab initio molecular dynamics simulations (AIMD). The final step of each geometry optimization was considered and included in the database. For the AIMD simulations, we followed a two-step process. In the first step, we performed AIMD simulations with low values of parameters, such as k-points and kinetic energy cutoff (ecutwfc). For instance, we set the kpoints to 1 1 1 and energy criterion for convergence of self-consistent field (SCF) iterations to 10^{-8} eV. In the second step, snapshots were taken at intervals of 100 time steps to minimize correlation. The selected configurations were then recalculated with higher values of k-points and kinetic energy cutoff (ecutwfc). All the calculations were performed using the Quantum Espresso package. The majority of numerical details can be located in the "Ab Initio Calculation Details" section of the main article. We consulted multiple databases, including the Materials Project (Jain *et al.*, 2013), the American Mineralogist Crystal Structure Database (Downs & Hall-Wallace, 2003), and the Automatic Flow for Materials Discovery (AFLOW) (Curtarolo *et al.*, 2012). The majority of our crystal structures, which are labeled with atomic coordinates and lattice parameters, are sourced from experimental X-ray diffraction (XRD) or quantum mechanical calculations available in these databases. Subsequently, we utilized these inputs for geometry relaxation, employing thoroughly converged density functional theory (DFT) parameters. It is important to highlight that the development of the comprehensive database and molecular dynamic simulations was made possible through the use of various modeling packages. Among them were AtomsK (Hirel, 2015), Vesta (Momma *et al.*, 2008), Ovito (Stukowski, 2009), and Xcrysden (Kokalj, 1999).

3.6.2 Silicon

We considered nearly 13 polymorphs of silicon, along with the liquid and amorphous states. While the majority of the process for generating the silicon database can be found in (Zongo *et al.*, 2022), additional calculations were conducted in this study to enhance the completeness of the database. Ab initio molecular dynamics simulations were conducted using a supercell of 63 atoms containing a vacancy at five different temperatures : 400, 800, 1600, 2600, and 3600 K. In addition to the liquid configurations generated in (Zongo *et al.*, 2022), we utilized a

MTP-trained potential to melt the silicon at temperatures of 2000 K and 3000 K. The resulting melted configurations were then used as inputs for the subsequent AIMD equilibration. Although in (Zongo *et al.*, 2022), AIMD liquid configurations were equilibrated near these temperatures, in this study we used different starting inputs. The aim is to achieve a high degree of diversification in the atomic environments. In contrast to the random displacement configurations used as input for the AIMD melting process in (Zongo *et al.*, 2022), we utilized the MTP-melts as the initial input for the AIMD equilibration in this work. Two models were utilized in this study : one with 64 atoms and another with 216 atoms. These models were gradually heated over a period of 30 ps and subsequently equilibrated for 20 ps using the Stillinger-Weber potential. The resulting liquid structure at each temperature was then used as input for the ab initio molecular dynamics (AIMD) simulation. The AIMD simulations employed a time-step of 1 fs for the 64-atom model and 2 fs for the 216-atom model. For example, the result of the MTP potential-based melting at 2000 K was used as the input for the subsequent AIMD equilibration at 2000 K. Compressed liquid was also incorporated in the database by filling a cubic box with 64 atoms in order to achieve a density of 3.20 g/cm^3 . We also included a small model for AIMD liquid silicon using a unit cell of the diamond structure containing 8 atoms. This model was equilibrated at 2500 K and 3500 K. Additional models were included to simulate amorphous silicon. A cubic box was randomly filled with 120 atoms, aiming for a density equal to the experimental density of amorphous silicon (2.28 g/cm^3). The configuration was then briefly equilibrated using the Stillinger-Webber (Stillinger & Weber, 1985) potential before the AIMD run.

3.6.3 Silica

The silica database was constructed by including all its polymorphs and deformations, following the Voigt notation, similar to the approach used for silicon. In addition, disordered structures were included in the SiO_2 database through a two-step process. First, pre-relaxation was performed using the Beest Kramer van Santen (BKS)(Van Beest *et al.*, 1990) potential, followed by full equilibration using DFT calculations. The density was adjusted to 2.20 g/cm^3 in each process to match the experimental density. This was achieved by utilizing various cubic box models

containing 3, 6, 9, 12, 27, 36, 48, 54, 72, 96, 114, and 192 atoms. The silicon and oxygen atoms were randomly distributed within these boxes according to the stoichiometry. The preliminary relaxation was conducted using the BKS potential at various temperatures ranging from 300 K to 6000 K. Each configuration was gradually heated to the target temperature over a period of 80 ps and subsequently equilibrated for 30 ps using a time step of 1 fs. Each result obtained from the preliminary relaxation was used as input for AIMD equilibration at the corresponding BKS relaxation temperature. As an example, the result of the BKS pre-equilibration at 3000 K was utilized as the input for the subsequent AIMD equilibration at the same temperature. The time step for all AIMD equilibrations was set to 2 fs. The maximum number of time steps recorded was 35 ps. Additionally, random displacements were introduced within the selected configurations at temperatures ranging from 300 K to 1000 K.

3.6.4 Oxygen

The oxygen database, which accounts for 2% of the full database, was constructed by considering only the gaseous state. This choice was made because PBE-based DFT modeling has limitations in accurately describing solid oxygen, where both van der Waals and covalent interactions coexist. Thus, for gaseous phases such as dioxygen, ozone, and others, we first optimized the bond length and bond angle. Then, we varied the bond length of the diatomic molecule and computed the corresponding properties. In the case of bent molecules like ozone, we varied the bond angle and performed two types of calculations. In the first set of calculations, we kept the bond length fixed while varying the bond angle. In the second set of calculations, we optimized the bond length at each angle. Additionally, we added oxygen configurations from chemical reactions using the nudge elastic band method (NEB) (Henkelman *et al.*, 2000). Furthermore, we selected other types of structures, such as oxygen clusters and silicon dimers, from the silica melt at higher temperatures using the MTP potential.

For a comprehensive breakdown of the database composition, please consult Tab 3.13 and (Zongo *et al.*, 2022).

3.6.5 Active learning

In addition to the database generation methods described above, we performed additional techniques to enhance and facilitate the training of the unified potential for the Si-SiO₂ and O system. First to improve our potential and diversity of our database, we added interfacial configurations of Si-SiO₂. We built few models of Si-SiO₂ interface and conducted single point calculations of these models. We also equilibrate these configurations shortly using AIMD simulation at 300 K, 500 K and 800 K and selected few configurations from the short trajectories and compute their energies, forces and stress with DFT. We then merged these configurations to ones from single point calculations of the unrelaxed models and trained with low level MTP potential (level 08). The resulting potential were used to conduct active learning. The active learning approach implemented in the MTP framework uses the D-optimality criterion (Podryabinkin & Shapeev, 2017; Novikov *et al.*, 2020b). This criterion allows for the detection of configurations on which the MTP potential extrapolates during MD simulations. We run MD trajectories at 300 K and 800 K for 100 ps and extrapolative configurations were gathered to complement our database need. In principle, during a 100 ps simulation, the MTP potential is retrained on the fly. However, in this case, we are not interested in the retrained potential ; instead, we focus on the extrapolative configurations. As our primary database contain a lot of large models mainly for vacancy, divacancy and interstitial, we did an active learning on these defects using unit cell of 7, 6, 9 atoms built from the diamond silicon unit cell that contain 8 atoms. The same procedure elaborated for interfaces was used here. First, starting with an 8-atom unit cell, we either removed 1 or 2 atoms or added 1 atom. We then conducted SFC calculations and relaxed the resulting configurations. Thus, for each defect, we obtained 2 configurations, which we used to fit our MTP 08 potential and begin the active learning process at 300 K. For interstitials, separate active learning was conducted for each of the interstitial positions, including bond-breaking interstitial, hexagonal interstitial, and tetragonal interstitial.

3.6.6 Training mode

One of the central challenges in developing a machine learning interatomic potential is the proper optimization of the cost function, which encompasses a multitude of parameters. The number of these parameters typically varies based on the data's characteristics, encompassing configuration types, their quantity, and the distribution of configuration sizes. Given the non-uniform distribution of sizes within the dataset, we employ two distinct training modes : the vibration mode and the structure mode. In the initial phase, we train the clean potential using the default vibration mode, as implemented in the MTP code. In this mode, an equal significance is given to every force vector, irrespective of the size distribution among configurations. This pertains to the simulation of thermal properties within the context of a molecular dynamic simulation. In the second step, we train the potential obtained from the first mode using structure modes. All the configurations have equal weight regardless of the size distribution. This seems to resemble a normalization. We have found the two-step training approach to be highly valuable mainly for calculating cohesive energy. From Fig. 3.19, it is evident that the energy error on isolated atoms is never eliminated when using the vibration mode, whereas the error becomes zero when the structure mode is employed.

3.6.7 Amorphous silicon

A 1000-atom cubic simulation box was used. For all the potentials, the configuration was heated to 3000 K and equilibrated for 30 ps. The configuration was then cooled to 1800 K at a cooling rate of 10 K/ps and equilibrated at 1800 K for 30 ps. Following this, the configuration was cooled to 1300 K at a cooling rate of 8.5 K/ps and equilibrated at 1300 K for another 50 ps. Finally, the configuration was cooled from 1300 K to 500 K at a rate of 5 K/ps and equilibrated at 500 K using the NPT ensemble for 120 ps. Structural data were gathered for the last 20 ps.

Tableau 3.6 Examples of fitting parameters for levels 26 and 28 of MTP potentials, consisting of 1657 and 2638 parameters respectively, are provided for a two-component system. The table highlights both the hyperparameters of the MTP model and the free parameters optimized during the training phase. The hyperparameters, such as radial functions count, radial basis size, interaction cutoff radius, and weight, are manually set before the training process commences. Selecting the right hyperparameters is crucial for achieving the desired performance of the model. In addition to the primary hyperparameters, careful attention to the BFGS iteration number becomes pivotal. This parameter precisely dictates the frequency with which the model undergoes updates through the Broyden–Fletcher–Goldfarb–Shanno (BFGS) iterative algorithm. The impact of this parameter on the convergence and efficiency of the optimization process is particularly pronounced when considering the number of model parameters. Choosing appropriate values for this parameter is paramount, as excessively low values (less training time) may result in underfitting, while overly high values (more training time) can lead to overfitting. Striking the right balance is key, as there exists an optimal number for this parameter that ensures the model converges effectively without sacrificing its ability to generalize

Parameters	Values (level 26)	Values (level 28)
radial functions count f_μ	6 ($\mu = 6$)	6 ($\mu = 6$)
radial basis size $Q^{(\alpha)}$	8 ($\alpha = 8$)	8 ($\alpha = 8$)
free parameters	1657	2638
Cutoff radius (Å)	5.7	5.7
Energy weight	1	1
Force weight	0.01	0.01
Stress weight	0.001	0.001
BFGS iteration number	3000	2000
Structures weighting mode	vibrations	structures
Parameters initialization mode	random	random

3.6.8 Amorphous silica

The configuration is heated from 300 K to 4200 K at a heating rate of 65 K/ps. We then scale the velocity to exactly 4200 K and equilibrate for 10 ps. After that, the velocity is scaled again to 3500 K, and the system is equilibrated for 35 ps at 3500 K. From 3500 K, the configuration is cooled to 2000 K at a cooling rate of 25 K/ps and equilibrated at 2000 K for 50 ps. Finally, the configuration is cooled to 300 K at a cooling rate of 10 K/ps and equilibrated with the NPT ensemble for 120 ps at 300 K. The structural properties were measured during the last 20 ps. The characteristics of the obtained structure match those of the experimental data, as we

reported both in the main text and here. The experimental elastic constant of amorphous silica (Guerette & Huang, 2012) are $C_{11} = 77.93$ GPa, $C_{12} = 15.63$ GPa and $C_{44} = 31.15$ GPa. Our MTP predicted elastic constant are $C_{11} = 76.08$ GPa, $C_{12} = 18.90$ GPa and $C_{44} = 38.36$ GPa. To ensure effective utilization of most semi-empirical potentials and achieve complete melting of the crystalline β -cristobalite, a specific heating and cooling protocol was implemented. The procedure for all semi-empirical models is as follows : The configuration is initially heated from 300 K to 5000 K over 60 ps. It is then held at 5000 K for 100 ps before being cooled to 3000 K at a rate of 20 K/ps. Once at 3000 K, the configuration is equilibrated for 50 ps, after which it is cooled to 2000 K at the same rate of 20 K/ps. Following this, another equilibration period of 50 ps at 2000 K is conducted as in MTP case. Finally, the configuration is cooled to 300 K at a rate of 10 K/ps. It is then equilibrated in the NPT ensemble at 300 K for 120 ps, during which structural data is collected for the last 20 ps.

Tableau 3.7 The computational expenses.

The expense is assessed through the execution of a molecular dynamics (MD) simulation, employing a system comprising 64 atoms for silicon and 96 atoms for silica. The simulation with MTP and semi-empirical potentials is specifically configured to use a single CPU core, simplifying the comparison process. The cost is given in seconds per MD step

Silicon		Silica		Silica	
Method	Times	Method	Times	Method	Times
DFT	118	DFT	16.51	SHK1	0.0016
MTP26	0.062	MTP26	0.16	SHK2	0.0016
MTP28	0.138	MTP28	0.25	TS	0.0003
SW	0.0002	BKS	0.0015	VA	0.0005

Tableau 3.13 Compositon of silica database. This table provides details on the composition of the silica database. The silicon database is also presented in (Zongo *et al.*, 2022)

Content	Replication	Atom/cell	Total
α -quartz			
Bulk	1x1x1	9	1

Tableau 3.13 Compositon of silica database. This table provides details on the composition of the silica database. The silicon database is also presented in (Zongo *et al.*, 2022) (continued)

Content	Replication	Atom/cell	Total
Bulk deformations (tensile, compression, shear)	1x1x1	9	124
Oxygen vacancies	2x2x2	71	19
Silicon vacancies	2x2x2	71	18
Oxygen interstitial	2x2x2	72	20
Silicon interstitial	2x2x2	72	27
Dioxygen interstitial	2x2x2	73	23
Surfaces & others (0001), (-1100)	1x1x8, 2x1x8 2x2x8	8, 12, 16	2
Stacking fault	1x3x1	36	46
Random displacement	1x1x1, 2x2x2 3x3x3	9, 72,243	8
<i>β</i> -quartz			
Bulk	1x1x1	9	1
Bulk deformations	1x1x1	9	94
Oxygen vacancies	2x2x2	95	1
Silicon vacancies	2x2x2	95	1
Random displacement	2x2x2,3x3x3	72, 243	4
<i>α</i> -cristobalite			
Bulk	1x1x1	2	1
Bulk deformations	1x1x1	12	71
Oxygen vacancies	2x2x2	95	4
Silicon vacancies	2x2x2	95	4
Amorphous	2x2x2	96	3

Tableau 3.13 Compositon of silica database. This table provides details on the composition of the silica database. The silicon database is also presented in (Zongo *et al.*, 2022) (continued)

Content	Replication	Atom/cell	Total
Random displacement	1x1x1, 2x2x2 3x3x3	12, 96,324	9
β -cristobalite			
Bulk	1x1x1	6, 24	2
Bulk deformations	1x1x1	6, 24	109
Oxygen vacancies	2x2x2	192	2
Silicon vacancies	2x2x2	192	2
MD 300K	2x2x2	48	15
Liquids MD 3600K, 5000K, 8000K	2x2x2	192	16
Liquids MD ab initio 2000K, 3600K, 4600K	2x2x2	48	31
Amorphous	2x2x2	192	88
Random displacement	1x1x1, 2x2x2 3x3x3	24, 192, 648	9
α -tridymite			
Bulk	1x1x1	12	1
Bulk deformations	1x1x1	12	78
Oxygen vacancies	2x2x2	95	1
Silicon vacancies	2x2x2	95	1
Random displacement	1x1x1	12	10
α -cristobalite II			
Bulk	1x1x1	24	1
Bulk deformations	1x1x1	24	8
Liquids MD ab initio 4000K	1x1x1	24	9
Random displacement	1x1x1	24	2

Tableau 3.13 Compositon of silica database. This table provides details on the composition of the silica database. The silicon database is also presented in (Zongo *et al.*, 2022) (continued)

Content	Replication	Atom/cell	Total
α -cristobalite-212121			
Bulk	1x1x1	24	1
Bulk deformations	1x1x1	24	8
Random displacement	1x1x1	24	2
β -cristobalite-I42d			
Bulk	1x1x1	24	1
Bulk deformations	1x1x1	24	8
Random displacement	1x1x1	24	2
β -cristobalite-213			
Bulk	1x1x1	12	1
Bulk deformations	1x1x1	12	8
Random displacement	1x1x1	12	2
Chabazite			
Bulk	1x1x1	108	1
Bulk deformations	1x1x1	108	8
Random displacement	1x1x1	108	2
Coesite-II			
Bulk	1x1x1	24	1
Bulk deformations	1x1x1	24	8
Random displacement	1x1x1	24	2
Octatetracontaoxo-tetraicosasilicate			
Bulk	1x1x1	288	1
Bulk deformations	1x1x1	288	8
Random displacement	1x1x1	288	2

Tableau 3.13 Compositon of silica database. This table provides details on the composition of the silica database. The silicon database is also presented in (Zongo *et al.*, 2022) (continued)

Content	Replication	Atom/cell	Total
Silica zeolite-2			
Bulk	1x1x1	36	1
Bulk deformations	1x1x1	36	8
Random displacement	1x1x1	36	2
Silica zeolite-GUS-1			
Bulk	1x1x1	96	1
Bulk deformations	1x1x1	96	8
Random displacement	1x1x1	96	2
Silica zeolite-ZSM-5			
Bulk	1x1x1	288	1
Bulk deformations	1x1x1	288	8
Random displacement	1x1x1	288	2
Monoclinic low tridymite			
Bulk	1x1x1	72	1
Bulk deformations	1x1x1	72	8
Random displacement	1x1x1	72	2
Orthorhombic high tridymite			
Bulk	1x1x1	72	1
Bulk deformations	1x1x1	72	8
Random displacement	1x1x1	72	2
Superstructure monoclinic tridymite			
Bulk	1x1x1	72	1
Bulk deformations	1x1x1	72	8
Random displacement	1x1x1	72	2

Tableau 3.13 Compositon of silica database. This table provides details on the composition of the silica database. The silicon database is also presented in (Zongo *et al.*, 2022) (continued)

Content	Replication	Atom/cell	Total
Triclinic tridymite			
Bulk	1x1x1	72	1
Bulk deformations	1x1x1	72	8
Random displacement	1x1x1	72	2
α -2D			
Bulk	1x1x1	12	1
Bulk deformations	1x1x1	12	48
Oxygen vacancies	2x2x2	108	4
Silicon vacancies	2x2x2	108	3
Random displacement	2x2x2	108	2
β -2D			
Bulk deformations	1x1x1	12	48
Oxygen vacancies	6x6x6	108	4
Silicon vacancies	6x6x6	108	3
Random displacement	6x6x6	108	2
δ -2D			
Bulk	1x1x1	3	1
Bulk deformations	1x1x1	3	48
Oxygen vacancies	2x2x2	108	4
Silicon vacancies	2x2x2	108	5
Random displacement	2x2x2	108	2
γ -2D			
Bulk	1x1x1	12	1
Bulk deformations	1x1x1	12	48
Oxygen vacancies	2x2x2	108	4

Tableau 3.13 Compositon of silica database. This table provides details on the composition of the silica database. The silicon database is also presented in (Zongo *et al.*, 2022) (continued)

Content	Replication	Atom/cell	Total
Silicon vacancies	2x2x2	108	8
Random displacement	2x2x2	108	2
<i>β</i> -tridymite			
Bulk	1x1x1	12	1
Bulk deformations	1x1x1	12	81
Oxygen vacancies	2x2x2	95	1
Silicon vacancies	2x2x2	95	1
Random displacement	1x1x1	12	10
Keatite			
Bulk	1x1x1	36	1
Bulk deformations	1x1x1	36	80
Oxygen vacancies	2x2x2	287	1
Silicon vacancies	2x2x2	287	1
Random displacement	1x1x, 2x2x2	36, 288	7
Coesite			
Bulk	1x1x1	24	1
Bulk deformations	1x1x1	24	101
Oxygen vacancies	2x2x2	191	2
Silicon vacancies	2x2x2	191	2
Random displacement	1x1x1, 2x2x2	24, 192	7
Moganite			
Bulk	1x1x1	18	1
Bulk deformations	1x1x1	18	101
Oxygen vacancies	2x2x2	143	2
Silicon vacancies	2x2x2	143	1

Tableau 3.13 Compositon of silica database. This table provides details on the composition of the silica database. The silicon database is also presented in (Zongo *et al.*, 2022) (continued)

Content	Replication	Atom/cell	Total
Random displacement	1x1x1, 2x2x2	18, 144	7
Stishovite			
Bulk	1x1x1	6	1
Bulk deformations	1x1x1	6	83
Oxygen vacancies	3x3x3	161	2
Silicon vacancies	3x3x3	161	2
Random displacement	2x2x2, 3x3x3	48, 162	4
Seifertite			
Bulk	1x1x1	12	1
Bulk deformations	1x1x1	12	83
Oxygen vacancies	2x2x2	95	2
Silicon vacancies	2x2x2	95	2
Random displacement	1x1x1, 2x2x2	12, 96	7

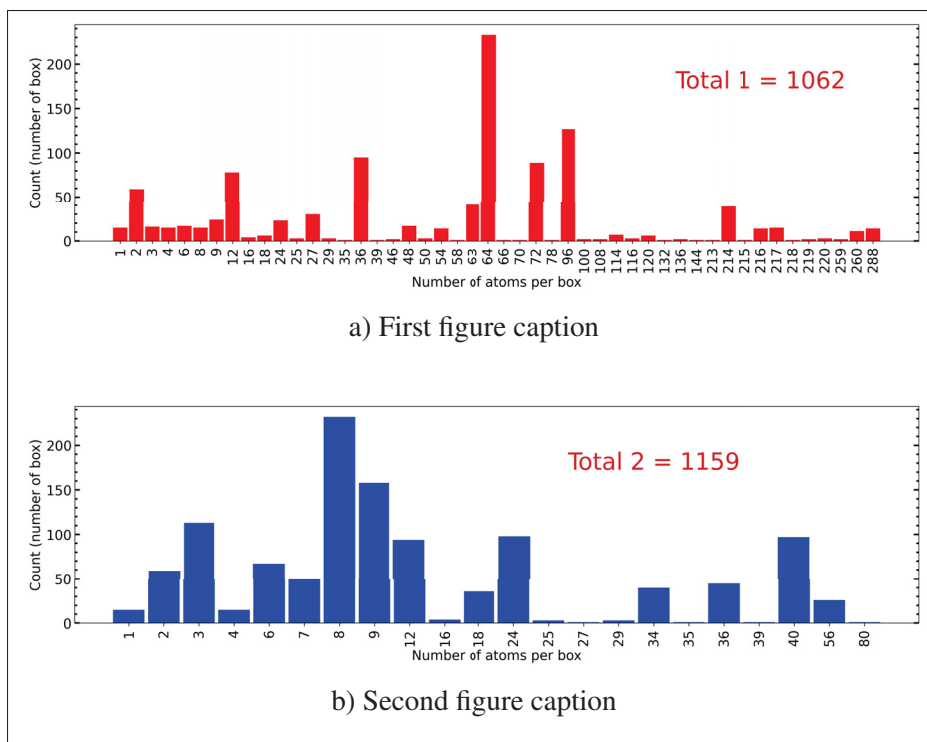


Figure 3.12 Distribution of configuration sizes in the unified training sets ($T_{unified}$) : (a) for preliminary work, testing, and training ; (b) for final implementation. Total 2 = 0.52 x

Total 1 + new DFT calculations . The first training set Fig. 3.12a (a) corresponds to configurations ranging in size from 1 to 288 atoms, without any interfaces. For this set, we implemented a potential of level 26 (pot26.mtp) from the GitLab repository. This set was part of optimizing the database and testing the training mode. Given that we used different types of configurations, the distribution of the configuration size dictates the settings for training, including the training mode and the number of parameters. This allows us to choose a normalization approach, ensuring that all configurations have the same relative importance. Ensuring a diverse distribution of sizes and various configuration types in the dataset is instrumental in enabling the model to acquire knowledge of different atomic environments of the configurational space. This diversity is essential for enhancing the robustness of the interaction potential. However, this introduces a layer of complexity to the decision-making process, particularly when determining the necessary number of parameters for training the model. Given the utilization of various configuration types and the highly distributed configuration sizes, scaling the number of parameters in the model becomes a non-trivial task, as it cannot be directly determined based on the size of the training set. To enhance our work and predict properties more accurately, we used a second training set (Fig. 3.12b (b)) and trained with level 28 (pot28.mtp). Because the configurations are relatively small, we trained directly using the structure mode. The results in the main article correspond to this set, which has a total of 1149 configurations. The first set, with a total of 1062 configurations, was used for preliminary work and testing. Total 2 is made up of 52% of the configurations from Total 1, plus new DFT calculations (active learning). Please refer to Tab. 3.11 for the composition of Total 1 and to Tab. V from the main article for the composition of Total 2

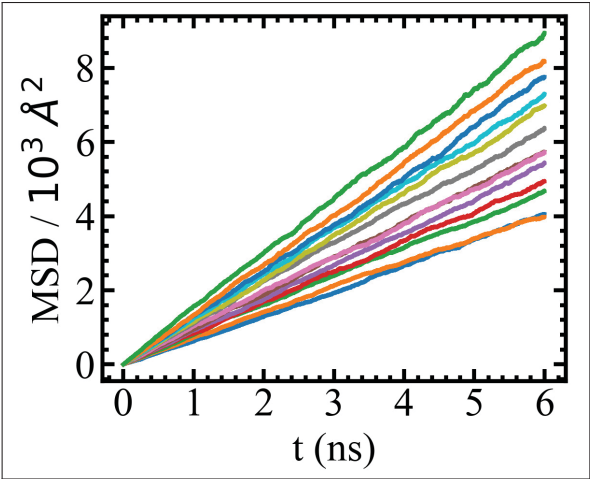


Figure 3.15 Mean square displacement.

The time dependence of mean square displacement in silicon crystal with a vacancy across a temperature range from 1000 K to 1650 K. Based on this dynamical correlation function, we calculate the diffusion coefficient via the Einstein relation and determine the activation energy for diffusion. TThis is crucial as it allows for a comparison with NEB calculations. The MTP results, along with those from SW, are detailed in the main paper

Tableau 3.10 Illustrations of training errors (RMSE) on disordered structures subsets.

These are the results of the final iteration of the cleaning process applied to disordered silicon and silica. We cleaned disordered silicon and silica separately. The configurations in these subsets were obtained from AIMD simulations at various temperatures, as explained above. For silicon structures, we applied a selection (select-add command), while for silica, we initially utilized a train-remove-train process to exclude off-equilibrium configurations resulting from high-temperature AIMD simulations, followed by the application of the selection

	AIMD Si AL ₁ Total = 272 min 63 max 216 mean 64		AIMD SiO ₂ AL ₂ Total = 325 min 27 max 114 mean 96		T2 = AL ₁ + AL ₂ Total = 597 min 27 max 216 mean 72	
Level	E	F	E	F	E	F
16	5.3	136.0	4.0	178.6	7.2	217.8
18	5.1	133.2	4.4	172.1	5.6	189.3

Tableau 3.11 First refined and optimized training set (set 1). This dataset was derived from an extensive uncured database through curation and optimization as described in the main text, and is detailed from Tab. 3.8 to Fig. 3.12. It was used for preliminary testing and training, as shown in Fig. 3.16 to Fig. 3.19. The distribution of the number of atoms is presented in Fig. 3.12a (a). 52% of the configurations presented in Tab. V of the main text, with the atom distribution shown in Fig. 3.12a (b), are taken from this set

Cleaned and Optimized Subset	Total
Vacancy & Divacancy	51
Interstitial	34
Stacking Fault	40
Others (Dislocation, Embedded Dimer, Slab SiO ₂)	45
Isolated Atom Si and O	2
Deformation Si	111
Deformation SiO ₂	161
Disordered Si (85% Liquid and 15% Low Temperature AIMD)	272
Liquid SiO ₂	325
Oxygen Molecules	21
Unified Training Set	1062

Tableau 3.12 The first unification process that led to training set 1 (see Total 1 in Fig. 3.12a (a) for the distribution of atoms and Tab. 3.11 for the composition) and validation sets. Training error measured on the unified training set 1 and validation error, both expressed as (RMSE), with E representing energy error per atom in meV/atom, F representing force error in meV/Å, and S representing stress error in GPa. Additionally, we performed joint training on defect structures, elastic constants, and molecules, denoted as set T_3 . Subsequently, isolated atoms were introduced to observe their impact on the optimization process. Finally, we augmented the training set by incorporating disordered structures of silicon and silica, resulting in the formation of the first unified training set (Total 1). We utilized the validation set to assess the underfitting/overfitting and generalization capabilities of the unified potentials. The notations R and S represent random mode and structure mode, respectively

	$T_3 = D_1 + T_1$ Total = 463 min 1 max 288 mean 68		$T_4 = T_3 + T_{isolated}$ Total = 465 min 1 max 288 mean 68		$T_f = T_4 + T_2$ Total 1 = 1062 min 1 max 288 mean 67			V_1 Total = 1217 min 1 max 1000 mean 75			V_2 Total = 742 min 1 max 120 mean 74		
Level	E	F	E	F	E	F	S	E	F	S	E	F	S
18 R	11.7	85.7	16.2	100.0	36.7	220.1	2.5	22.8	278.8	1.0	25.3	199.4	0.6
20 R	9.1	79.0	9.1	71.9	17.5	174.9	1.5	17.3	220.2	0.9	19.3	156.4	0.3
22 R	7.7	63.2	7.3	66.6	14.2	153.6	0.9	17.1	216.1	0.6	24.6	146.3	0.3
24 R	-	-	-	-	10.1	135.2	0.7	11.9	187.4	0.5	20.3	135.7	0.2
26 R	-	-	-	-	9.7	123.1	0.6	11.8	202.8	0.5	18.8	123.8	0.2
18 S	-	-	-	-	13.9	226.2	0.9	16.1	283.1	0.7	16.2	204.2	0.4
20 S	-	-	-	-	9.9	178.7	0.7	11.5	224.7	0.6	16.2	157.4	0.3
22 S	-	-	-	-	8.3	157.7	0.5	10.8	213.5	0.5	17.5	149.3	0.3
24 S	-	-	-	-	6.3	137.3	0.4	8.2	191.9	0.4	17.0	139.4	0.2
26 S	-	-	-	-	5.4	126.0	0.3	8.3	208.0	0.4	15.2	126.6	0.2

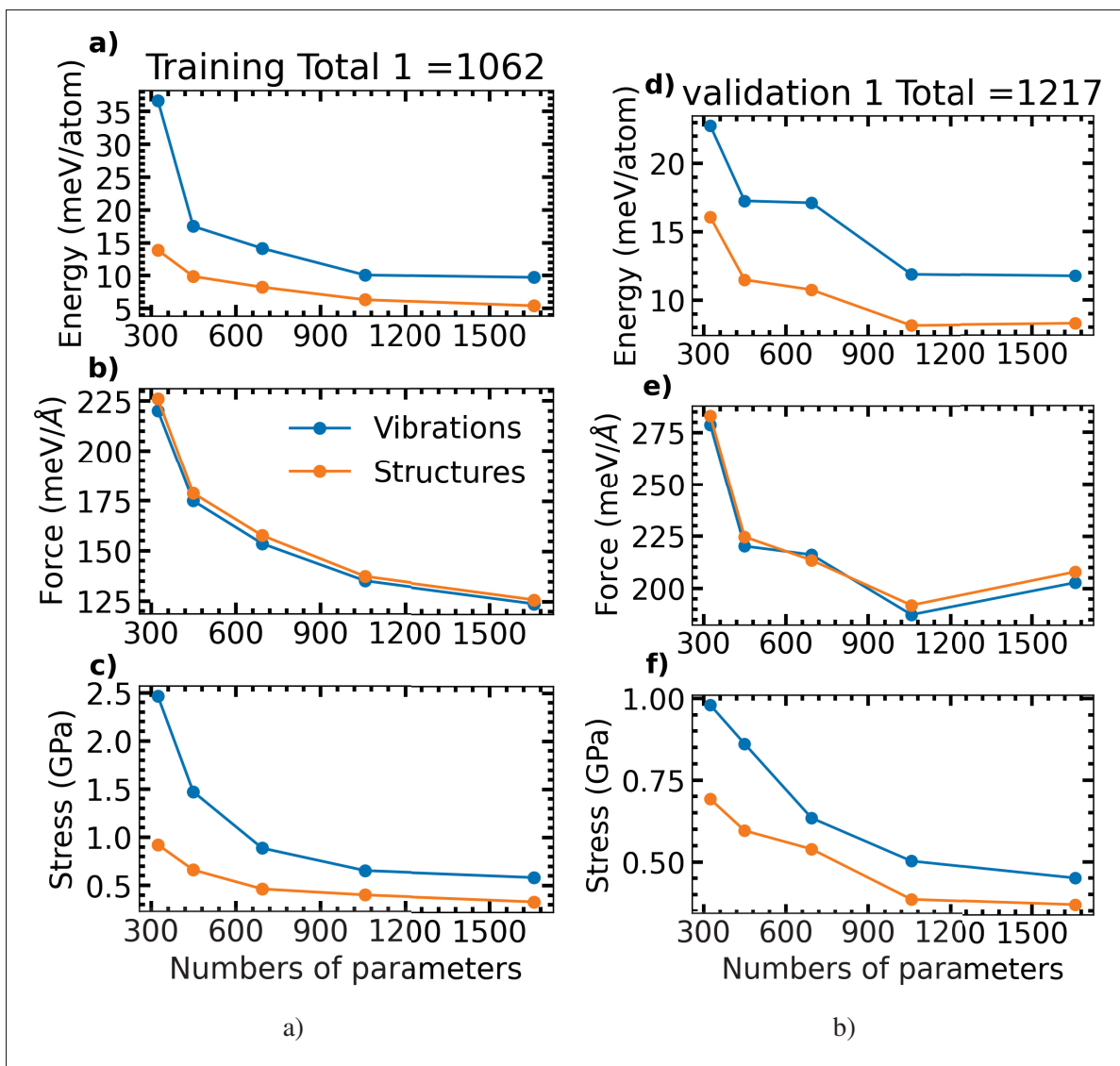


Figure 3.16 Comparison of RMSE in energy, forces, and stress. We assess the RMSE errors with respect to training mode and the number of degrees of freedom. This is conducted separately on the unified training set and validation set 1. It is evident that the training mode reduces the RMSE for both energy and stress, with a noticeable decrease in RMSE as the number of parameters increases. However, the training mode has no influence on the force RMSE, which only decreases with the number of parameters. This figure can assist in selecting the appropriate training mode or deciding which mode to implement for training a diversified training set

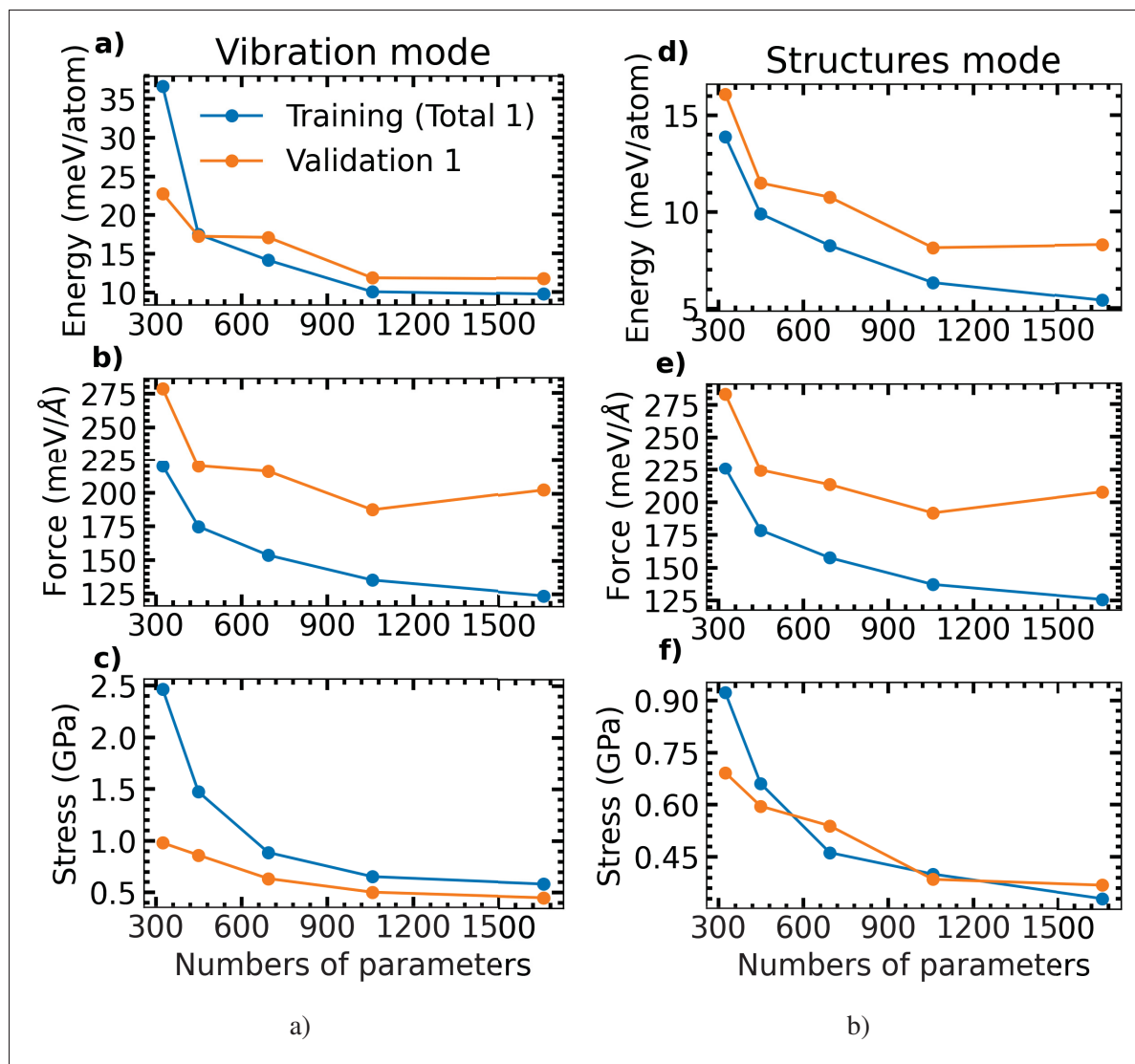


Figure 3.17 Comparison of RMSE in energy, forces, and stress. In this figure, we compare the RMSE error of the training set with that of validation set 1 for each training mode while varying the number of parameters. It's important to note that validation set 1 was randomly selected concurrently with the training set from the database. As observed, neither underfitting nor overfitting is evident, even though validation set 1 is larger than the unified training set

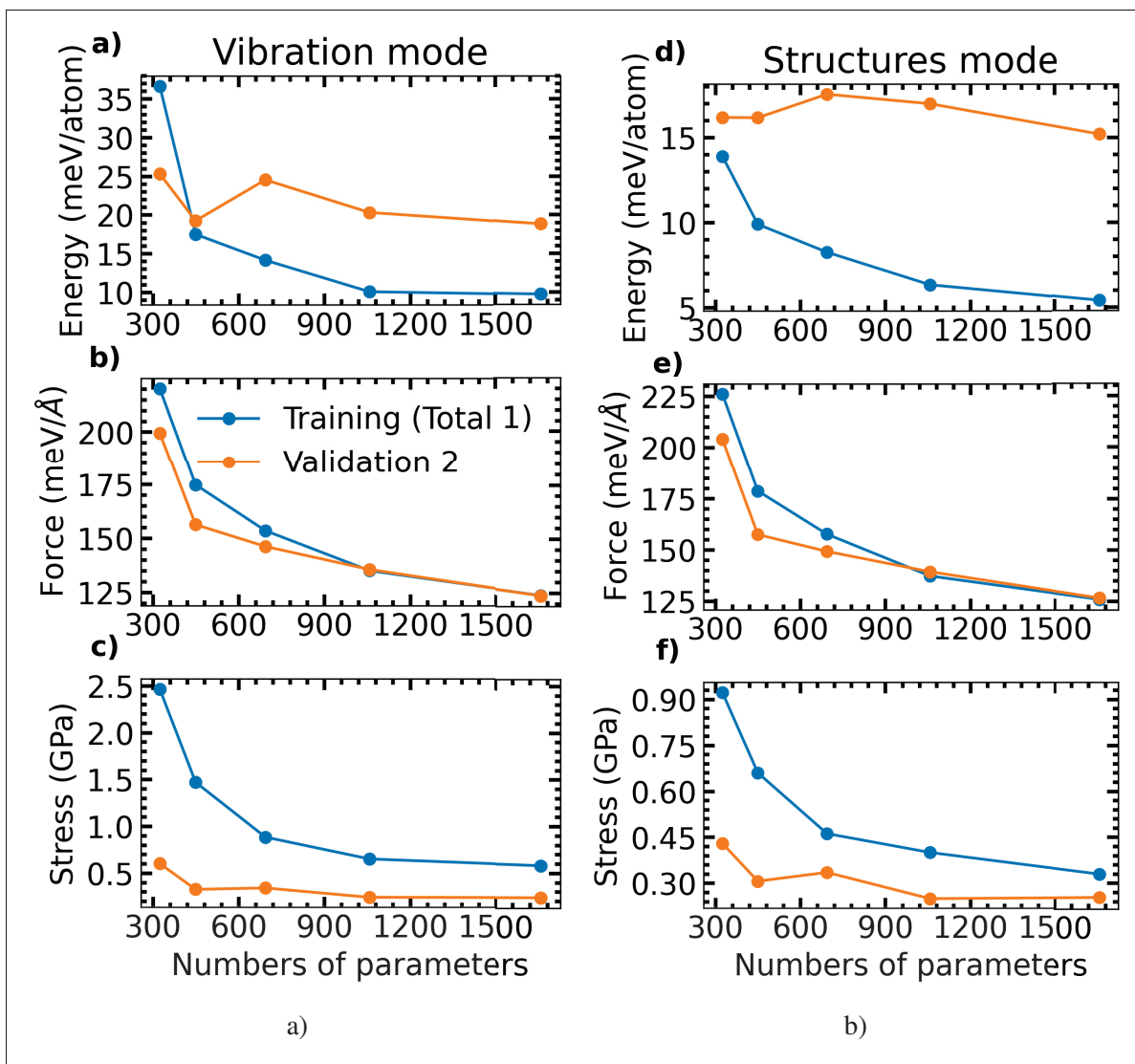


Figure 3.18 Comparison of RMSE in energy, forces, and stress. A second validation set (Validation Set 2) was employed to further explore the investigation of underfitting or overfitting related issues concerning the training mode. With the exception of the energy RMSE for structure mode, all other RMSE values are indicative of a well-trained potential.

Validation Set 2 was not included in the training set; it was selected from AIMD cooling/equilibration trajectories of amorphous silicon and silica. This demonstrates the transferability capability of the unified potential

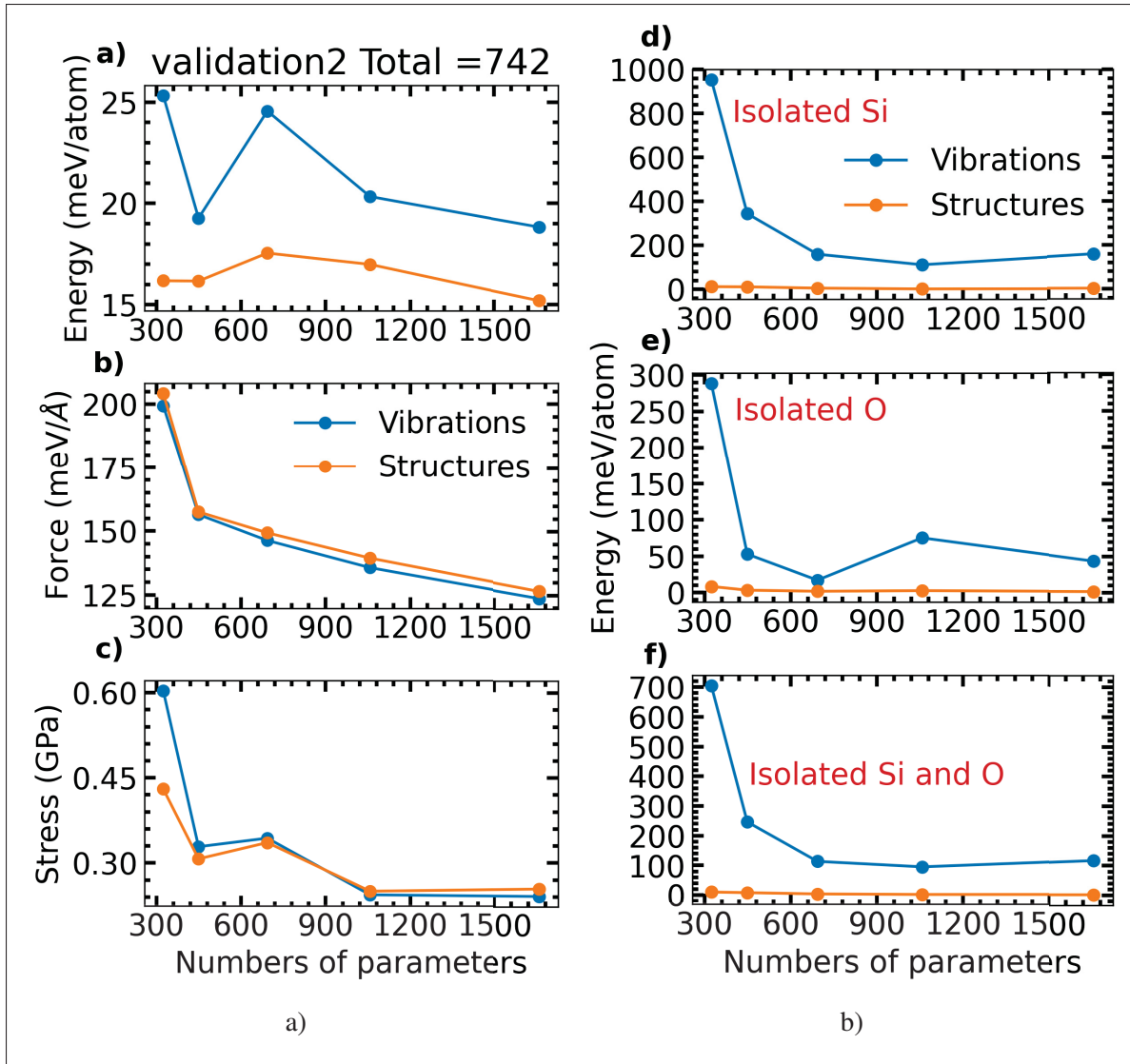


Figure 3.19 Comparison of RMSE in energy, forces and stress. For the second validation set (Validation Set 2), we also assess the influence of training mode while varying the number of parameters. Overall, the structure mode outperforms the vibration mode. The graph on the right corresponds to energy errors of isolated atoms with respect to the training mode. Since we calculate cohesive energy using isolated atoms, this enables us to adopt the correct training mode. As observed, for the structures mode, RMSE never reaches zero, whereas it drops to zero when the structures mode is applied. This indicates that a normalization approach is necessary when the data is highly diversified

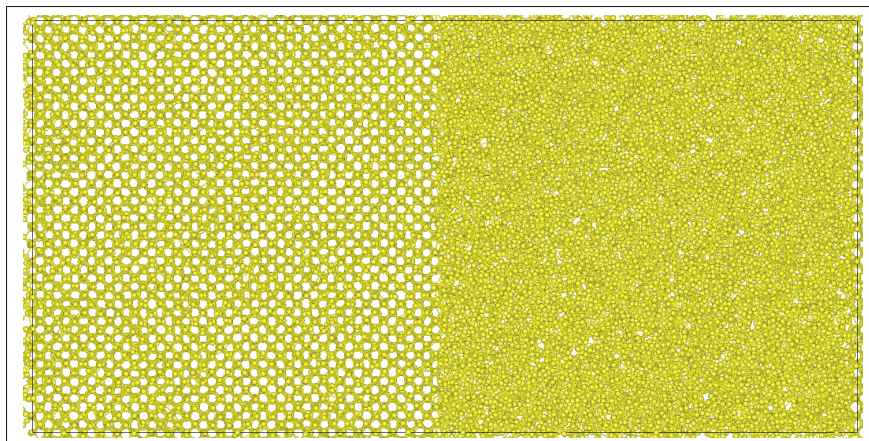


Figure 3.20 Initial state of the simulation block containing both the solid and liquid phases used for the melting point simulation. To predict the silicon crystal's melting behavior, we employed a simulation block consisting of a $12 \times 12 \times 12$ unit cell, housing a total of 13,824 atoms. Within this framework, we considered two distinct boxes : one with solid density and another with liquid density. These boxes were meticulously constructed, each tailored to its respective temperature using specific lattice parameters. Our simulation procedure began with the independent equilibration of the solid and liquid phases. Before bringing the solid box into contact with the liquid phase, we applied a slight strain in the x and y directions. Subsequently, we conducted a comprehensive solid-liquid coexistence simulation, utilizing the NPT ensemble. The outcomes of this simulation were indicative of either solid growth or liquid expansion, contingent on the temperature (T) being either below or above the melting point. A visual representation of the simulation block used for this coexistence simulation can be found in Fig 3.20

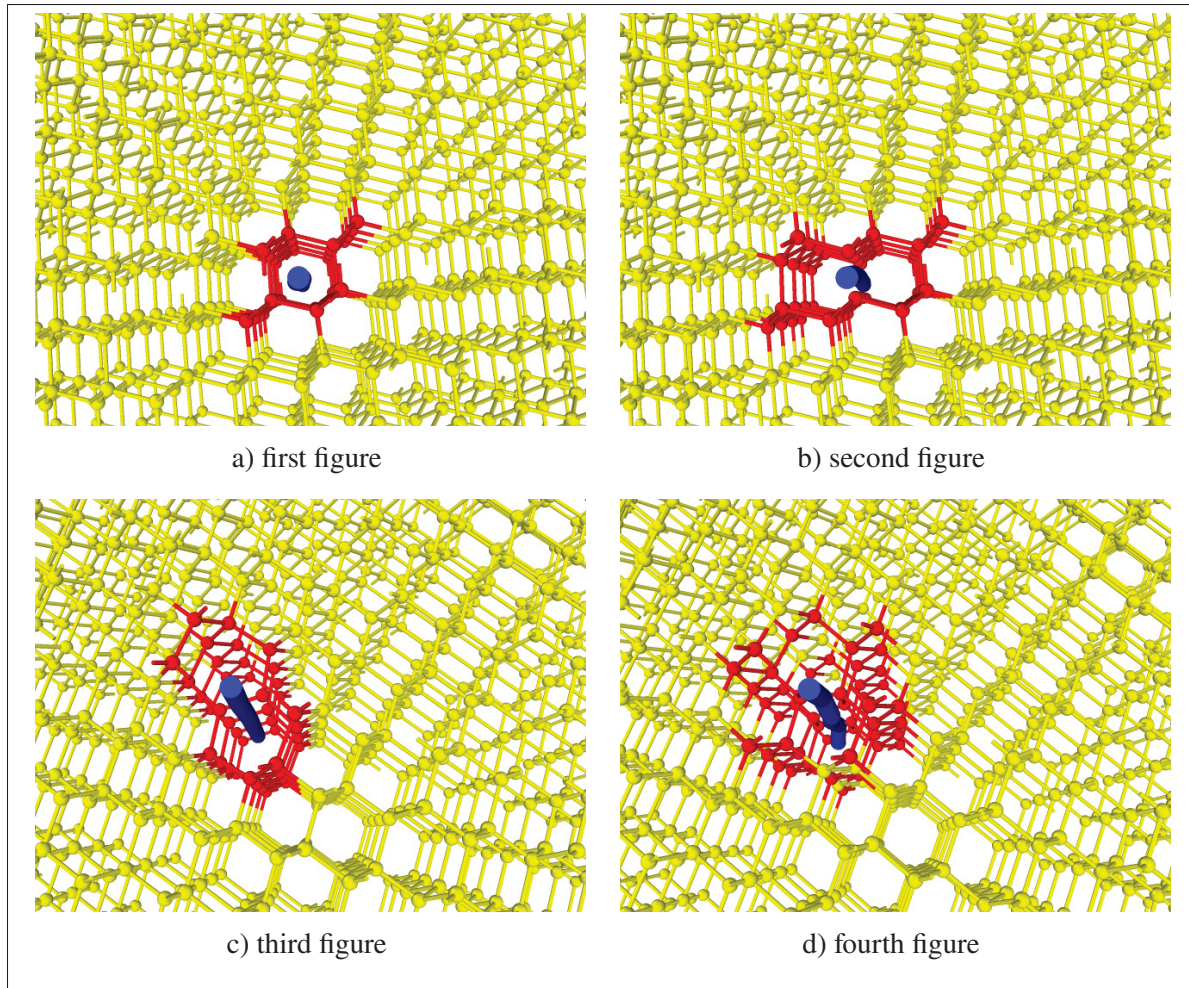


Figure 3.21 Illustrations of screw dislocation core structures obtained using the unified potential. (a) A-core, (b) B-core (which relaxes to A-core), (c) C_1 core extracted after only 5 steps of relaxation, and (d) C_2 core obtained directly from the relaxation of C_1 core. Notably, manual reconstruction, as required in previous studies such as (Huang *et al.*, 2021) and (Guénolé *et al.*, 2010), is not necessary. Atoms in close proximity to the core position are highlighted in red. The blue line represents the dislocation line

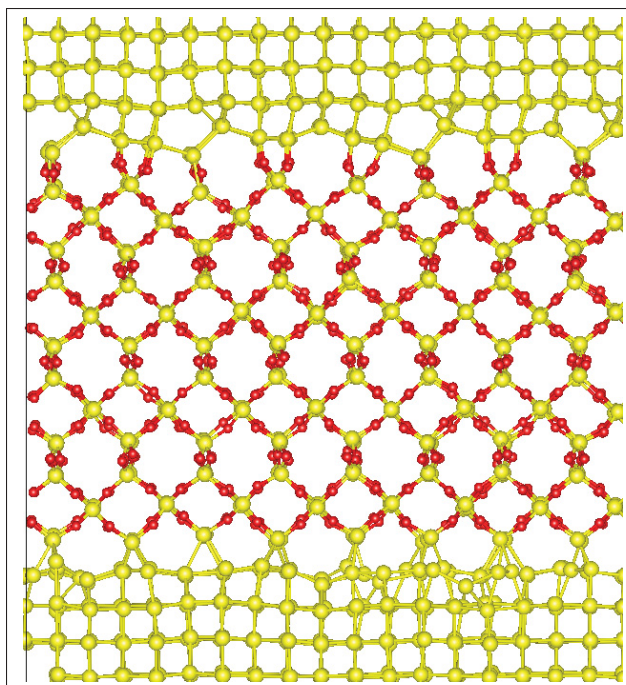


Figure 3.22 Illustrations interface configurations. A larger model of the Si/SiO₂ interface was obtained from energy and force minimization using the MTP potential

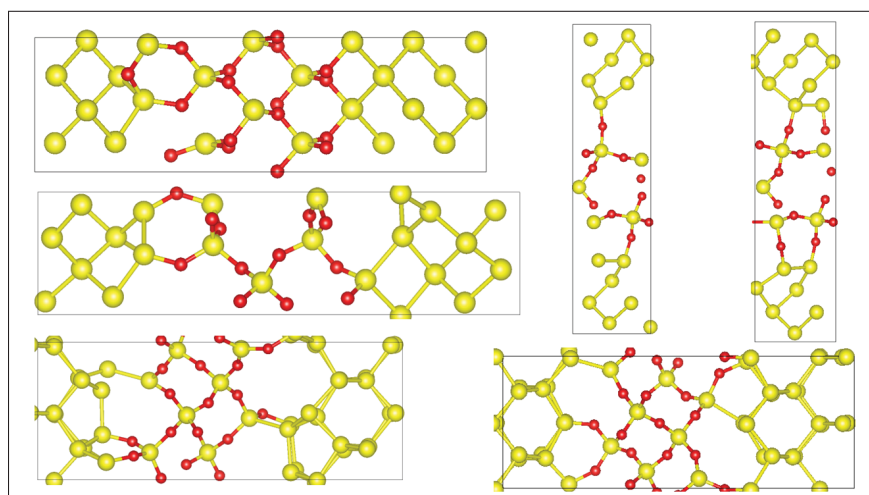


Figure 3.23 Illustrations interface configurations. The small model of Si/SiO₂ interface that was built directly from non-replicated unit cells of silicon crystal, alpha-quartz, and beta-cristobalite. These resulting configurations were obtained from energy and force minimization using the MTP potential

CHAPITRE 4

AMORPHOUS SILICON STRUCTURES GENERATED USING A MOMENT TENSOR POTENTIAL AND THE ACTIVATION RELAXATION TECHNIQUE *NOUVEAU*

Karim Zongo¹, Hao Sun², Claudiane Ouellet-Plamondon¹, Normand Mousseau³, Laurent Karim Béland²

¹ Department of Construction Engineering, École de Technologie Supérieure, Université du Québec

1100 Notre-Dame St W, Montreal (Quebec) H3C 1K3, Canada

² Department of Mechanical and Materials Engineering, Queen's University
99 University Ave, Kingston, Ontario K7L 3N6

³ Department of Physics, University of Montreal
1375 Avenue Thérèse-Lavoie-Roux, Montréal, Quebec H2V 0B3

Paper submitted in Journal of Physical Review B, January 2025.

4.1 Abstract

Preparing realistic atom-scale models of amorphous silicon (a-Si) is a decades-old condensed matter physics challenge. Herein, we combine the Activation Relaxation Technique *nouveau* (ARTn) to a Moment Tensor Potential (MTP) to generate seven a-Si models containing between 216 and 4096 atoms. A thorough analysis of their short-range and medium-range structural properties is performed, alongside assessments of excess energy and mechanical properties. The seven ARTn-MTP models are compared with available experimental data and other high quality a-Si models present in the literature. The seven ARTn-MTP a-Si models are in excellent agreement with available experimental data. Notably, several of our models, including the 216-atom, 512-atom, and 1000-atom a-Si models, exhibit low coordination defects without any traces of crystalline grains. Historically overlooked in previous research, our study underlines the need to assess the validity of the continuous random-network hypothesis for the description of *perfect* amorphous model by characterizing local crystalline environment and to explore the crystallisation process of a-Si through modelling.

4.2 Introduction

Amorphous silicon (a-Si) is used in applications that include photovoltaics, thin-film transistors, battery electrodes, and liquid-crystal displays (Street, 1999; Carlson *et al.*, 1976; Powell, 1989; Cui *et al.*, 2009; Key *et al.*, 2009). Recent advances include its use as coating materials for next-generation gravitational wave detectors (Birney *et al.*, 2018). Additionally, a-Si serves as the canonical disordered solid prototype for the development of modeling methods.

a-Si (Lewis, 2022; Madanchi *et al.*, 2024) is idealized as a continuous random network (CRN) (Zachariasen, 1932), characterized by a local atomic environment akin to crystalline silicon (Wooten *et al.*, 1985; Vink *et al.*, 2001a). It maintains a local tetrahedral crystalline environment without long-range order. Real a-Si, however, is believed to contain defects associated with its disordered nature at a level that depends on its preparation (Roorda *et al.*, 1991; Laaziri *et al.*, 1999a; Laaziri *et al.*, 1999b; Lévesque *et al.*, 2022; Birney *et al.*, 2018). These include local coordination defects, medium-range ring defects and two-level systems—the latter contribute to noise in sensitive gravitational wave detectors and cause decoherence in quantum computers (Lévesque *et al.*, 2022; Birney *et al.*, 2018). Variations in tetrahedral geometry and bond lengths also significantly affect a-Si's performance. Controlling structural anomalies during preparation poses challenges from both an experimental and a theoretical perspective. Indeed, the structure of a-Si is highly influenced by preparation methods (Holmström *et al.*, 2016); for instance, chemical vapor deposition (Kail *et al.*, 2011) and ion implantation (Roorda *et al.*, 1989; Van den Hoven *et al.*, 1992; Laaziri *et al.*, 1999a) can introduce voids and coordination defects. Thermal annealing (Roorda *et al.*, 1989; Fortner & Lannin, 1989; Laaziri *et al.*, 1999a; Haberl *et al.*, 2009) often refines these imperfections, enhancing material quality.

Extensive theoretical research has focused on developing realistic structural models for a-Si (Wooten *et al.*, 1985; Barkema & Mousseau, 1996, 2000; Deringer *et al.*, 2018a; Justo *et al.*, 1998; Ishimaru *et al.*, 1997; Wright *et al.*, 2013). Various methods for simulating materials, including a-Si, can be categorized into dynamic and non-dynamic approaches (Lewis, 2022; Madanchi *et al.*, 2024). Dynamic methods, traditionally based on density functional theory

(DFT) (Sholl & Steckel, 2022) or semiempirical force fields (Torrens, 2012), primarily use melt-quench processes. However, these methods face challenges : experimental a-Si cannot be produced by melt-and-quench, DFT simulations are limited in system size and quenching rate, and semi-empirical force fields are often compromised by limited accuracy and transferability issues.

Non-dynamic methods include the Reverse Monte Carlo (RMC) approach (McGreevy & Pusztai, 1988; McGreevy & Howe, 1992; Gereben *et al.*, 2007; Gurman *et al.*, 1990; McGreevy, 2001). In RMC, the computer-generated a-Si structure is randomly modified in order for its simulated structure factor and radial distribution function $g(r)$ to match experimental data. RMC's main drawback is that it can yield multiple configurations for the same $g(r)$ — in other word the solution is degenerate. To overcome this, hybrid RMC methods incorporate force field interactions, ensuring generated configurations are both consistent with experimental data and physically realistic. Variants like the force-enhanced atomic refinement (Pandey *et al.*, 2015; Igram *et al.*, 2018) and experimentally constrained molecular relaxation (Biswas *et al.*, 2004) utilize semi-empirical and DFT interactions, respectively. Other key non-dynamic methods include the Wooten-Winer-Weaire (WWW) bond-switching algorithm (Wooten *et al.*, 1985), which optimizes atomic configurations by minimizing energy, and the Activation-Relaxation Technique *nouveau* (ARTn) (Barkema & Mousseau, 1996; Malek *et al.*, 2000; Mousseau *et al.*, 2012; Jay *et al.*, 2022), which combines activation and relaxation steps to discover low-energy structures. ARTn can be paired with various force fields and *ab initio* methods, making it suitable for applications that include relaxation of disordered systems and diffusion processes (Mousseau *et al.*, 1999; Lévesque *et al.*, 2022; Ganster *et al.*, 2012; El-Mellouhi *et al.*, 2007; Salles *et al.*, 2017; Joly *et al.*, 2012; Sun *et al.*, 2022, 2023; Trochet *et al.*, 2020; Béland *et al.*, 2011).

One of the key components in both dynamical and non-dynamical methods is the interaction potential. The ongoing challenge of balancing accuracy and computational cost in these methods has led to the development of machine learning interatomic potentials (MLIPs) (Behler & Parrinello, 2007; Deringer *et al.*, 2019; Wang *et al.*, 2024). MLIPs combine descriptors of the atomic environment, regression methods, and DFT data. Many MLIPs

frameworks have been developed, including artificial neural networks (Behler & Parrinello, 2007; Khorshidi & Peterson, 2016), kernel-based methods (Bartok, 2010; Bartók *et al.*, 2018), and linear regression approaches (Thompson *et al.*, 2015; Shapeev, 2016), among others. For a detailed description of MLIPs, including their implementation and testing, the reader is referred to the existing literature (Wang *et al.*, 2024). In the case of a-Si, structures have been modeled using artificial neural networks (Li *et al.*, 2020; Fan *et al.*, 2021; Li *et al.*, 2019) and Gaussian Approximation Potentials (GAP) (Bartok, 2010; Bartók *et al.*, 2018). In initial studies, the GAP potential was applied to a-Si systems containing between 512 and 4,096 atoms (Deringer *et al.*, 2018a). Later, the GAP model was extended to larger systems of up to 100,000 atoms to investigate the atomistic mechanisms underlying various structural transitions in disordered silicon (Deringer *et al.*, 2021). Other MLIPs, such as the Moment Tensor Potential (Shapeev, 2016) (MTP), have also been applied to a-Si (Zongo *et al.*, 2024). In these studies, melt-quench MD simulations were employed to generate a-Si structures.

A natural next step is to couple ARTn (Barkema & Mousseau, 1996; Malek & Mousseau, 2000; Mousseau *et al.*, 2012; Jay *et al.*, 2022) with machine learning potential to generate a-Si models. The objective of this paper is to generate a-Si models using ARTn-ML, which could enhance our understanding of the a-Si model and open new avenues for further research. The manuscript is organized into three main sections. The Methods section outlines the principles of the ARTn method and details our input models, the ARTn-MTP simulation process, and the analytical tools used. The Results and Discussion Section presents findings from our analysis of the configurations generated by ARTn-MTP. Finally, the key insights and discoveries of our study are summarized in the conclusions.

4.3 Methods

4.3.1 Activation Relaxation Technique *nouveau*

ARTn (Barkema & Mousseau, 1996; Malek & Mousseau, 2000) is a computational method used in the study of activated processes, particularly in the context of diffusion and transitions

between states in condensed matter systems. ARTn is designed to simulate rare events that signify substantial changes in the microscopic structure of atomic systems, based on the efficient exploration of potential energy landscapes. ARTn employs an event-driven approach in which the system evolves by transitioning between different states or configurations. The concept is based on the understanding that within the configurational space, there is a distinct set of points representing saddle points and adjacent minima in the potential energy surface of the system. As the name implies, ARTn can be summarized as a two-step process (Barkema & Mousseau, 1996) : activation and relaxation. In the activation step, starting from a local energy minimum, the configuration is pushed toward a nearby first-order saddle point. In the relaxation step, the system is driven over the saddle point and then relaxed to a new energy minimum. The new configuration is accepted based on the Metropolis acceptance probability, which is given by $P_{\text{accept}} = \min\left(1, \exp\left(-\frac{\Delta E}{kT}\right)\right)$, where T is the fictitious temperature, set to 0.25 eV for our simulation, k is the Boltzmann constant, and ΔE is the energy difference between the old and new configurations (Barkema & Mousseau, 1996; N'Tsouaglo, 2015; Barkema & Mousseau, 2000). The search is restricted to paths that include only saddle points. The ARTn method has demonstrated its value in revealing rare events that conventional simulation methods may struggle to capture due to the slow dynamics and relaxation inherent in these systems. Detailed descriptions of the method can be found elsewhere (Barkema & Mousseau, 1996; Malek & Mousseau, 2000; Mousseau *et al.*, 2012; Jay *et al.*, 2020; Gunde *et al.*, 2024).

4.3.2 A moment tensor potential for silicon

In practical terms, ARTn serves as a saddle-point search engine, employing an interaction potential as both a force and energy calculator. The simulation's accuracy is directly influenced by the selection of the force field. Previous studies have relied on semi-empirical force field (Ganster *et al.*, 2012; Barkema & Mousseau, 1996; Mousseau *et al.*, 1998; Lévesque *et al.*, 2022; Eliassen *et al.*, 2019) and ab initio methods (Machado-Charry *et al.*, 2011; Béland *et al.*, 2014; Salles *et al.*, 2017). Here we employ a machine learning-based forcefield. This potential was developed to jointly describe Si, O, and SiO₂ systems (Zongo *et al.*, 2024), and is an improvement

over a previously reported Si MTP (Zongo *et al.*, 2022). This potential is based on the MTP framework (Shapeev, 2016; Novikov *et al.*, 2020b). A detailed description of the development and implementation of this MTP can be found in our previous work (Zongo *et al.*, 2024, 2022).

Tableau 4.1 The models utilized in this study, the preparation methods of the input configurations, the number of atoms in each model, and the preparation methods for the a-Si models

Models name	Input configurations for a-Si preparation	N atoms	a-Si preparation methods
216-R1	Randomly filled	216	ARTn runs using MTP potential
216-R2	Randomly filled	216	ARTn runs using MTP potential
216-FEAR	Randomly filled	216	Force hence atomic refinement (Igram <i>et al.</i> , 2018)
512-R1	Randomly filled	512	ARTn runs using MTP potential
512-R2-CG	Randomly filled	512	ARTn runs using MTP potential
512-MD	Randomly filled	512	MD simulation using MTP potential
512-RMC	Randomly filled	512	Reverse monte carlo simulation (Cliffe <i>et al.</i> , 2017; Deringer <i>et al.</i> , 2018a)
512-INV	Randomly filled	512	Invariant environment refinement technique (Cliffe <i>et al.</i> , 2017; Deringer <i>et al.</i> , 2018a)
512-FEAR	Randomly filled	512	Force hence atomic refinement (Igram <i>et al.</i> , 2018)
512-WWW	Randomly filled	512	Bond switching algorithm (Cliffe <i>et al.</i> , 2017; Deringer <i>et al.</i> , 2018a)
1000-R-CG	Randomly filled	1000	ARTn runs using MTP potential
4096-R-MD	Randomly filled + MD runs	4096	ARTn runs using MTP potential
4096-WWW	Randomly filled	4096	Bond switching algorithm (Igram <i>et al.</i> , 2018)
100k-GAP18	Unknown	100000	MD simulation using GAP potential (Deringer <i>et al.</i> , 2021)
1000-aSi-SW	Amorphous configuration (Barkema & Mousseau, 1996); (Malek & Mousseau, 2000) (Machado-Charry <i>et al.</i> , 2011))	1000	ARTn runs using MTP potential

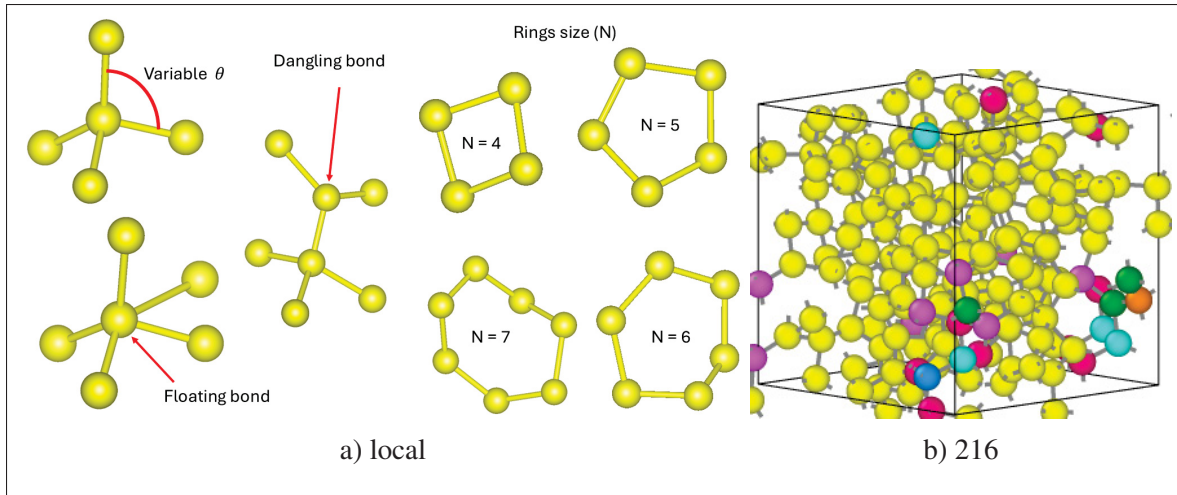


Figure 4.1 Key structural elements characterizing a-Si : (a) bond angle (θ), coordination defects such as dangling bonds (3-fold bonded atoms) and floating bonds (5-fold bonded atoms), and ring structures, including 4-, 5-, 6-, and 7-membered rings. (b) A 100% 4-fold fully coordinated amorphous structure comprising 216 atoms, generated by ARTn-MTP simulation, with 14.40 atoms in a crystalline environment. The atoms shown in yellow within the box are in an amorphous environment, while those colored differently are in a crystalline environment, specifically cubic or hexagonal diamond

4.3.3 Protocol for input models preparation and simulation details

We prepared models containing 216, 512, 1000, and 4096 atoms. For the model with 216 atoms, we used two different simulation boxes corresponding to densities of 2.20 g/cm³ (216-R1) and 2.28 g/cm³ (216-R2), respectively. These two boxes were randomly filled with an overlap distance (initial atomic separation) of 2.3 Å for 2.20 g/cm³ and 2 Å for 2.28 g/cm³. For the 512-atom model, we employed two simulation boxes randomly filled with atoms, using the same densities as those applied to the 216-atom model. The first 512-atom model had a density of 2.20 g/cm³ (512-R1) with an initial overlap of 2.3 Å. The second 512-atom model (512-R2) along with the 1000-atom model (1000-R) were randomly filled with a density of 2.28 g/cm³ and an initial overlap of 2 Å. The search for events during ARTn-MTP simulations was initiated using configurations that were already at local minima. Herein, the ARTn simulations begins by applying random displacements to the chosen central atom and its neighbors, based on a user-defined local cut-off radius. Once the displacements are introduced, the algorithm searches

for convergence to a saddle point, guided by the lowest curvature. After locating the saddle point, the minimization process begins to find the new minimum. The minimization is carried out using either the Steepest Descent (SD) or the Fast Inertial Relaxation Engine (FIRE) algorithm. In all ARTn-MTP simulations, we used a fictitious temperature of 0.25 eV in the Metropolis accept-reject criterion. Each simulation started with event searches initiated with a local cut-off radius of 3 Å, which the algorithm used to identify the region around the selected central atom which will be deformed. This cut-off radius was progressively increased to a maximum cut-off 5.5 Å in order to expand the local zone of initial deformation when the ARTn moves no longer led to energy decrease. The coordination defects level was monitored during the simulations. For coordination defect levels above 5 %, the full list of atoms was used as potential central atoms for ARTn moves. For defect levels at or below 5 %, only atoms with a non-four-fold coordination were allowed to serve as the central atom in ARTn moves.

4.3.4 Structural analysis

We employ a set of criteria to assess the quality of our amorphous structures. These criteria are derived from considerations of short-range structures, coordination defects, and intermediate-range order. Understanding these different structural aspects is crucial in the study of materials properties, as they influence the physical, chemical, and mechanical behavior of materials. We conduct coordination analysis using OVITO (Stukowski, 2009) with a bond-length cutoff of 2.85 Å. The analysis of the ring size distribution, providing insights into intermediate-range order, is performed using the R.I.N.G.S code (Le Roux *et al.*, 2010) based on the King criteria. We employed the ISAACS code (Le Roux *et al.*, 2010) for the calculation of the structure factor while the radial distribution function and angular distribution function were determined using LAMMPS (Thompson *et al.*, 2022) through pseudo simulations.

We also measure the amount of crystallinity in the models. Here, crystallinity refers to environments that are cubic or hexagonal diamond to second neighbour, all atoms at the center of this environment or within the first- and second-neighbour shell of a perfectly atom are considered

as having crystallinity. We use method developed by Ref. (Maras *et al.*, 2016) as implemented in OVITO (Stukowski, 2009).

4.4 Results and discussion

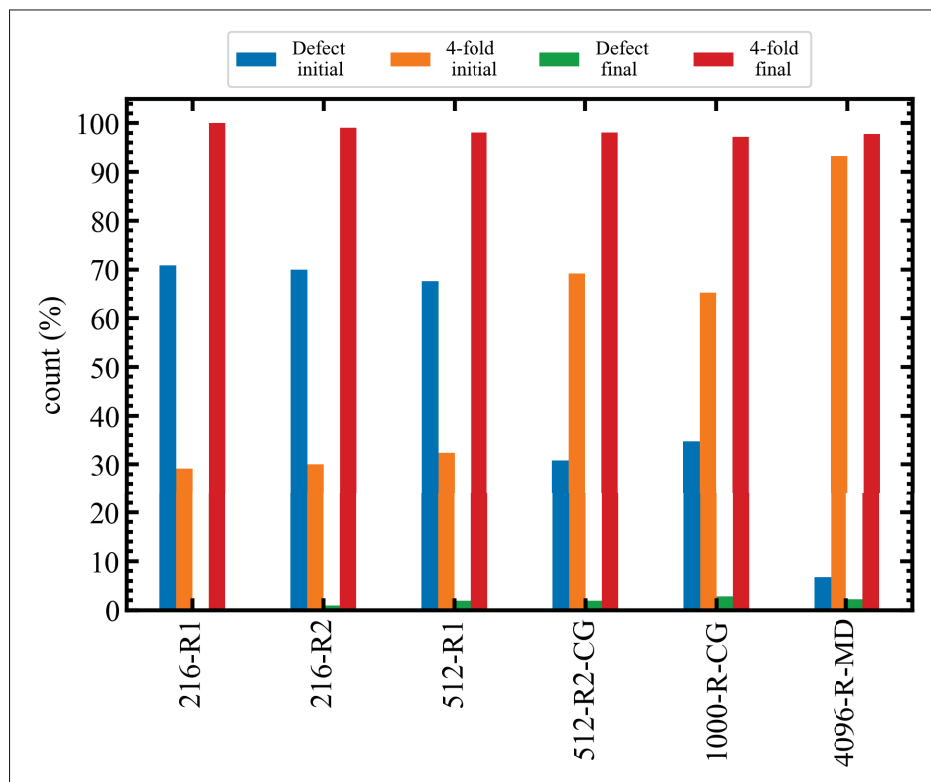


Figure 4.2 Coordination numbers of input configurations and ARTn-MTP optimized configurations (final configurations).

All non-4-fold coordinated atoms, including under- and over-coordinated atoms, are considered defective. The models analyzed include 216, 512, 1000 and 4096 atoms

Tableau 4.2 Structural characteristics of ARTn-MTP generated amorphous models. This includes density, the positions of the first (r_1) and second (r_2) peaks of the RDF, average coordination number (n), average bond angle (with variability indicated in parentheses), the percentages of 3-fold, 4-fold, and 5-fold coordinated atoms in each mode, and crystallinity (C). A comparison is provided with previously published models. Bold fonts are used to highlight models where fewer than 3 % of atoms were mis-coordinated

Model	density	r_1	r_2	n	$\langle\theta\rangle$ (σ)	3-fold (%)	5-fold (%)	4-fold (%)	C (%)
216-R1	2.20	2.36	3.77	3.990	109.20 (9.95)	0.93	0.00	99.07	0.00
216-R2	2.28	2.36	3.70	4.000	109.19 (10.02)	0.00	0.00	100.00	14.40
216-FEAR (Igram <i>et al.</i> , 2018)	2.33	2.36	3.81	4.028	108.52 (15.59)	1.39	4.17	94.44	
512-R1	2.20	2.36	3.90	4.008	109.11 (11.39)	0.59	1.37	98.04	0.00
512-R2	2.28	2.36	3.90	4.015	108.99 (11.76)	0.19	1.76	98.04	0.00
512-MD	2.28	2.37	3.83	3.996	109.05 (10.95)	1.37	0.98	97.65	3.30
512-RMC (Cliffe <i>et al.</i> , 2017; Deringer <i>et al.</i> , 2018a)	2.30	2.34	3.66	4.039	108.57 (14.16)	1.56	5.47	92.96	3.30
512-INV (Cliffe <i>et al.</i> , 2017; Deringer <i>et al.</i> , 2018a)	2.31	2.34	3.86	4.054	108.56 (14.47)	1.95	7.42	90.62	0.00
512-WWW (Cliffe <i>et al.</i> , 2017; Deringer <i>et al.</i> , 2018a)	2.27	2.34	3.94	4.008	109.05 (10.36)	0.19	0.98	98.82	10.00
512-FEAR (Igram <i>et al.</i> , 2018)	2.33	2.35	3.82	4.008		1.17	2.73	95.90	
512-GAP18 (Deringer <i>et al.</i> , 2018a)	2.27	2.36	3.76	4.004	109.20 (9.60)	0.58	0.98	98.44	6.4
1000-R	2.28	2.37	3.76	4.020	108.94 (12.31)	0.6	2.6	97.20	0.00
4096-R-MD	2.28	2.37	3.83	4.012	109.14 (10.65)	0.51	1.44	98.05	6.40
4096-WWW (Igram <i>et al.</i> , 2018)	2.33	2.36	3.78	4.004		0.05	0.49	99.46	
100k-GAP18 (Deringer <i>et al.</i> , 2021)	2.27	2.36	3.83	4.003	109.17 (9.99)	0.69	0.95	98.35	6.5

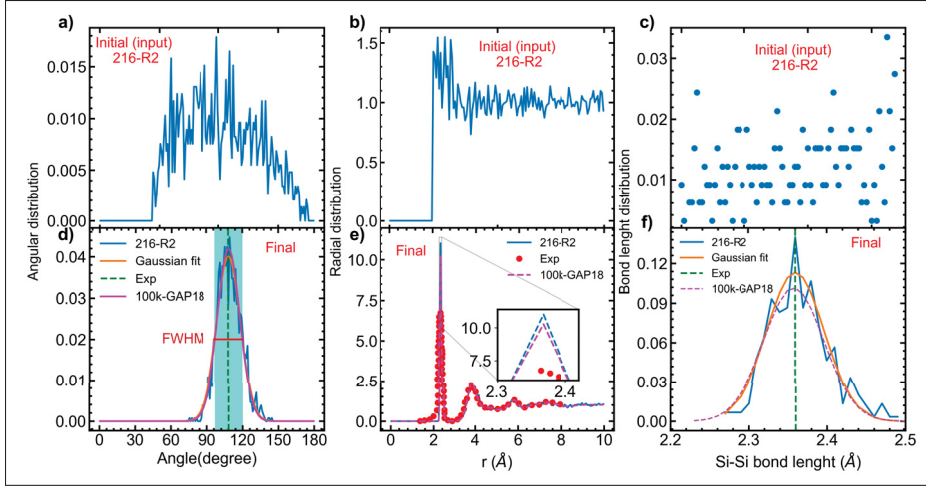


Figure 4.3 Comparison of the local structural properties of input configurations (top panels : a, b, and c) with the final configuration (100% coordinated) (bottom panels : d, e, and f) generated using the ARTn-MTP coupling scheme. The structural properties analyzed include angular distribution, radial distribution, and bond length distribution for a system comprising 216 atoms. Experimental data and data from the 100k-GAP18 model (Deringer *et al.*, 2021) are included to assess the quality of our 100% 4-fold coordinated 216-atom a-Si model

Models of size 216 atoms, 512 atoms, 1000 atoms and 4096 atoms were generated as detailed in table 4.1. We analyze these configurations with respect to short-range order, medium-range order, electronic properties as well as mechanical properties. Fig. 4.2 displays the defects in both the input and final configurations. The input configurations exhibit a higher prevalence of coordination defects than the final configurations : approximately 70% for the 216-R1, 216-R2, and 512-R1 models, compared to around 30% for the 512-R2 and 1000-R models, as illustrated by the blue bars. The green bars represent the total number of defects present in the final configurations.

In all ARTn-MTP a-Si models, the final percentage of defective atoms remained below 2%, with the exception of the 1000-R and 4096-R-MD models, which exhibit a final defect percentage of approximately 3%. The initial 4-fold coordination number (shown by the orange bars in Fig. 4.2) is approximately 30% for the 216-R1, 216-R2, and 512-R1 models, and around 68% for the

512-R2 and 1000-R models. In contrast, the final configurations display markedly higher 4-fold coordination numbers (depicted by the red bars in Fig. 4.2 and Table 4.2), with values of 99.07%, 100%, 98.04%, 98.04%, 97.20%, and 97.80% for the 216-R1, 216-R2, 512-R1, 512-R2, 1000-R and 4096-R-MD models, respectively. No atoms with fewer than 3 neighbors or more than 5 neighbors were observed.

To further contextualize the number of coordination defects in the ARTn-MTP a-Si models, we present a comparison in Fig. 4.5, which includes data from the literature as well. First, all the models presented show no atoms with fewer than 3 neighbors or more than 5 neighbors. Secondly, all of the ARTn-MTP generated models exhibit fewer floating and dangling bonds compared to all other methods, except the WWW model and 100k-GAP18 model. Notably, the 216-R2 model is absent from Fig. 4.5 because it contains 100% four-fold coordination. To our knowledge, no other a-Si models exhibit 100% four-fold coordination, except for some WWW models generated with a Keating-type potential (Pandey *et al.*, 2016b; Barkema & Mousseau, 2000)

The last column of Table 4.2 presents the fraction of atoms in a crystalline environment (cubic or hexagonal diamond). As shown, most of the generated a-Si models, particularly those from MD simulations, contain some proportion of crystallinity. Although WWW is known to produce large a-Si models without coordination defects, the WWW a-Si models still exhibit traces of crystallinity. We identified four models—216-R1, 512-R1, 512-R2, and 1000-R—that are free from crystallinity. Furthermore, as indicated in Table 4.3, both the 1000-aSi-SW initial and 1000-aSi-SW final configurations also contain no crystallinity. It is important to note that all four of our a-Si models without crystallinity, including the 1000-aSi-SW initial model, were generated starting from a random state without any prior MD prerun, meaning their boxes were initially filled randomly. From Table 4.2, we note three potential trends. (i) this fraction does not seem to depend much on the system size. (ii) Starting ARTn from a random state seems to lead to a less crystalline state than starting from a melt-and-quench state. (iii) Fewer coordination defects seem to favor a higher crystallinity content. Now that, with reliable ML potentials, we can generate similar quality models using different pathways (non-dynamic melt-and-quench

methods), we can start examining the link between local crystallinity, defects and the ideal continuous-random networks. This can lead to understanding why, for example, amorphous silicon cannot be produced experimentally from the liquid phase and crystallizes well below melting as well as whether the CRN model corresponds really to ideal a-Si. The ML potential (Zongo *et al.*, 2024, 2022), which leads to high-quality models, makes further study on this question now possible.

4.4.1 Short-range order structural analysis of the models

Fig. 4.3 illustrates the structural properties of the 216-R2 models, including the angular distribution function, radial distribution function, and bond length distribution. Panels a), b), and c) display these properties for the randomly packed input configuration (216-R2), while panels d), e), and f) show the properties of the final configurations (216-R2) produced by ARTn-MTP. The final 216-R2 configuration is 100% fully coordinated, with no coordination defects present. As shown in Fig. 4.3 a), b), and c), the data for the input configurations are randomly distributed and widespread. In contrast, the data for the final configurations, depicted in Fig. 4.3 d), e), and f), follow well-defined distributions. The shape of the radial distribution function matches the experimental data, though the peak is somewhat overestimated. This indicates that the Si-Si pair distances in our model are widely distributed around the experimental pair separation distance. The bond angle distribution function is centered around the tetrahedral value of 109.5° with a spread of 9.96° . A Gaussian fit to the data reveals a full width at half-maximum (FWHM) of 23.45° . The half width at half-maximum (HWHM) is 11.73° , which is similar to the experimental HWHM value of 11° (Fortner & Lannin, 1989). For the bond length, the data for the input configuration (4.3 c) are widespread. In contrast, for the final configuration (4.3 f), the bond lengths follow a Gaussian distribution centered around the experimental value of $2.35 \text{ \AA}$. The mean bond length is $2.36 \text{ \AA}$, with a spread of $0.03 \text{ \AA}$. The angular distribution function, radial distribution function, and bond length distribution of the final model (216-R2) were compared with those of the 100k-GAP18 configuration, which contains 100,000 atoms. As shown in Fig. 4.3 (d), (e), and (f), these properties align well with those of the larger 100k-GAP18 model. This

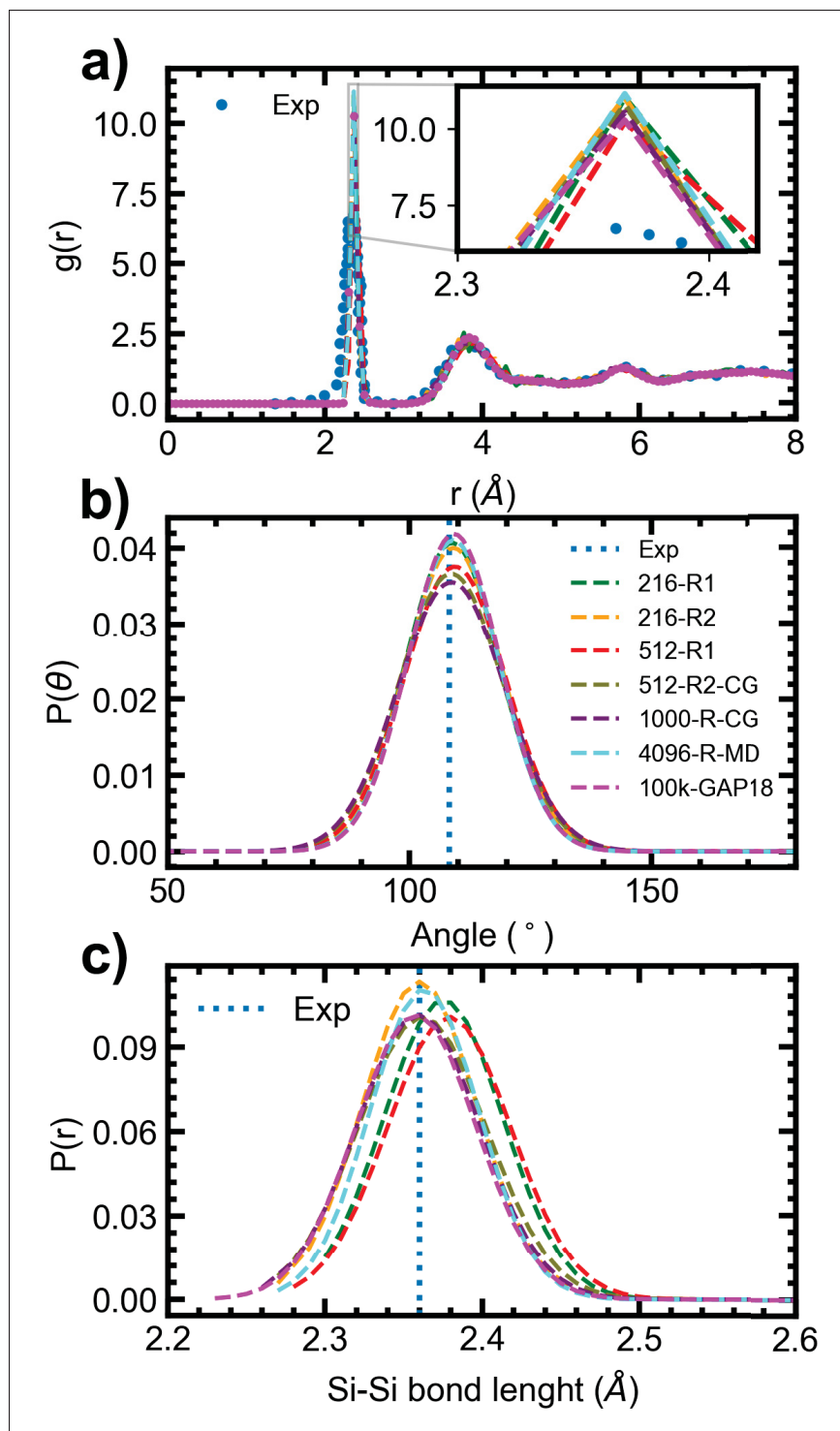


Figure 4.4 Local structural properties of ARTn-MTP generated a-Si models : (a) radial distribution, (b) angular distribution, and (c) bond length distributions. Also shown are the properties obtained from experimental data and the 100k-GAP18 atom simulations

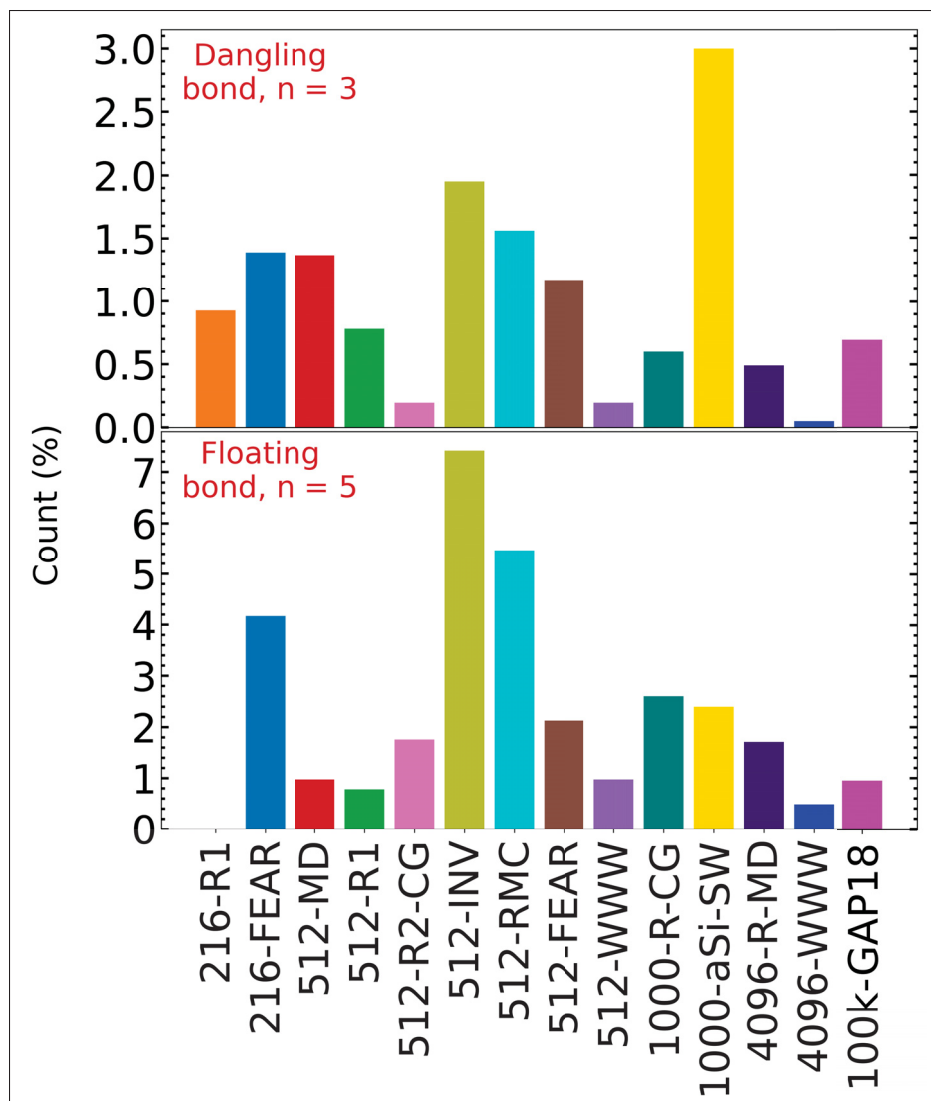


Figure 4.5 Proportion of coordination defects in various a-Si models generated by different methods. 216-R1 has no floating bond and 216-R2 is 100 % fully 4-fold coordinated. No atoms with fewer than 3 neighbors or more than 5 neighbors have been recorded in any model

demonstrates the high quality of our fully coordinated a-Si model. Beyond the analysis of the fully coordinated configuration, we also explore the short-range order structure in our other models. The detailed results are presented in Fig. 4.4 and Tab. 4.2. Once again, the structural properties—including the radial distribution function (Fig. 4.4 a), bond angle distribution function (Fig. 4.4 b), and bond length distribution (Fig. 4.4 c)—closely match experimental data

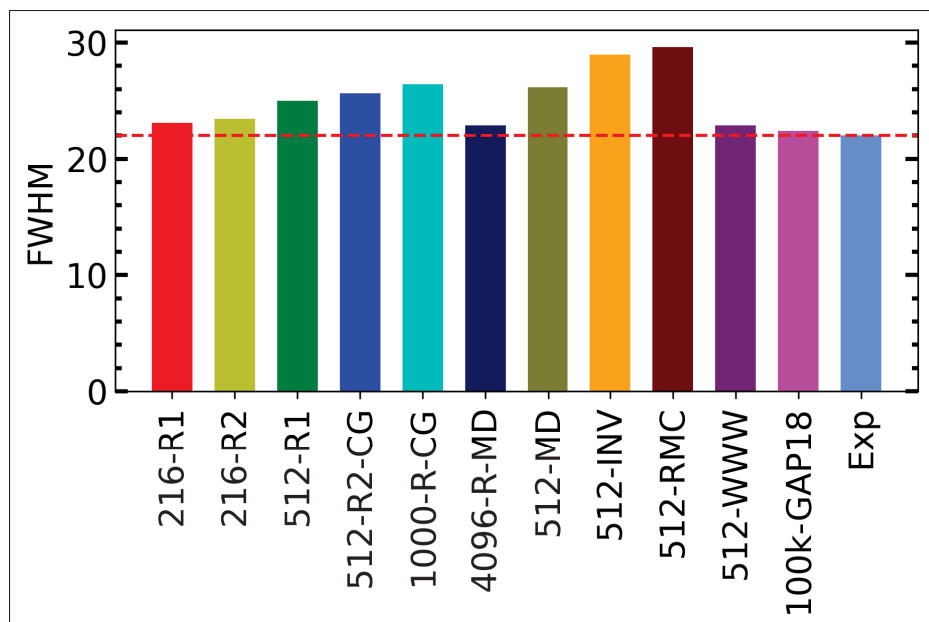


Figure 4.6 FWHM of the bond-angle distribution. The FWHM values are computed by fitting a Gaussian distribution to the bond angle distributions. The red line denotes the experimental value of FWHM (Fortner & Lannin, 1989)

and those of 100k-GAP18 models. The data for all the models presented in Fig. 4.4 show similar behavior, with the exception of the Si-Si bond length distribution. Notably, the bond lengths for two of our models—216-R1 (in green) and 512-R1 (in red)—are shifted. These two models have a density of 2.20 g/cm^3 , which is lower than that of the other models, which is 2.28 g/cm^3 . Since amorphous silicon features local tetrahedral bonding, we quantify the deviation from ideal tetrahedral geometry by calculating the FWHM for all models included in this study, as presented in Fig. 4.6. The FWHM values for all our models closely align with the experimental data, as well as with the WWW model and the 100k-GAP18 models.

4.4.2 Medium-range order structural analysis of the models

We further analyze the ARTn-MTP a-Si models at intermediate length scales using the structure factor and ring statistics. Fig. 4.3 illustrates the static structure factor of our fully coordinated 216-R2 a-Si model, alongside experimental data and the 100k-GAP18 model. Additionally, the

structure factors for the 512-R2, 1000-R, and 4096-R-MD models are included to provide a comprehensive comparison with both the experimental $S(q)$ and the $S(q)$ of the 100k-GAP18 models. As shown, the structure factors of all our models align well with the experimental data, with the exception of the first peak at $S(q)$ near $q = 2 \text{ \AA}^{-1}$, which is underestimated by the 216-R2, 512-R2, and 1000-R models. This is consistent with other results from the literature (Igram *et al.*, 2018; Deringer *et al.*, 2018a) and is attributed to the system size rather than a specific method. Thus, the 4096-R-MD model, which consists of 4096 atoms, and the 100k-GAP18 model, comprising 100,000 atoms, both exhibit a strong match with the first peak of the experimental $S(q)$. The first peak, known as the first sharp diffraction peak, is typically underestimated in systems with fewer than a thousand atoms (Igram *et al.*, 2018; Deringer *et al.*, 2018a). The height of this peak serves as a measure of structural ordering.

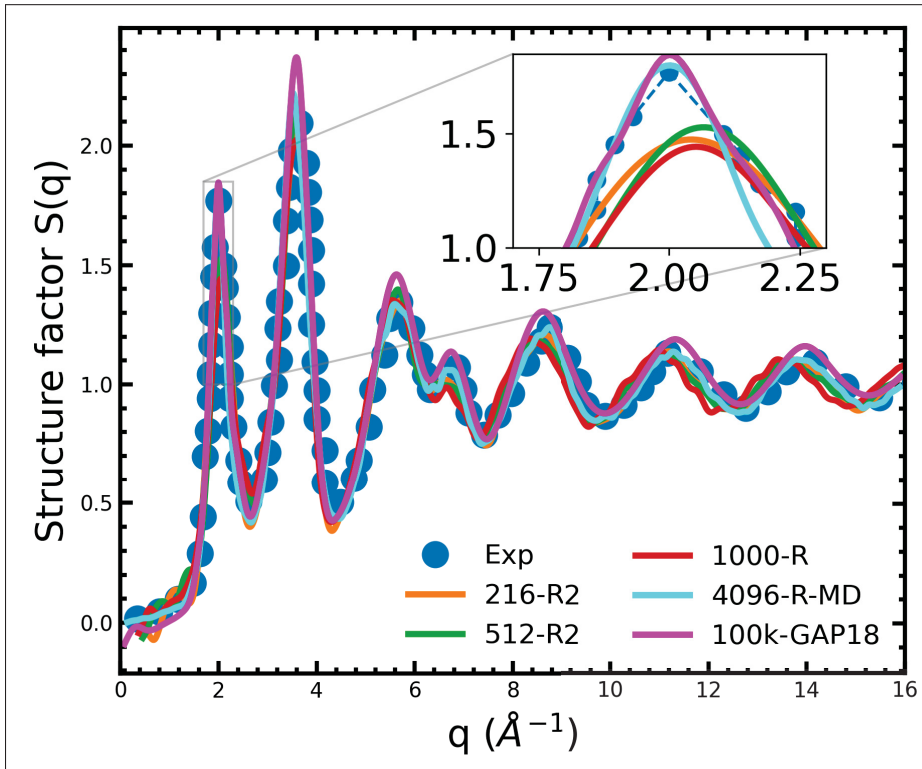


Figure 4.7 Structure factor for ARTn-MTP generated models compared with that obtained from experimental data and 100k-GAP18 atom simulations

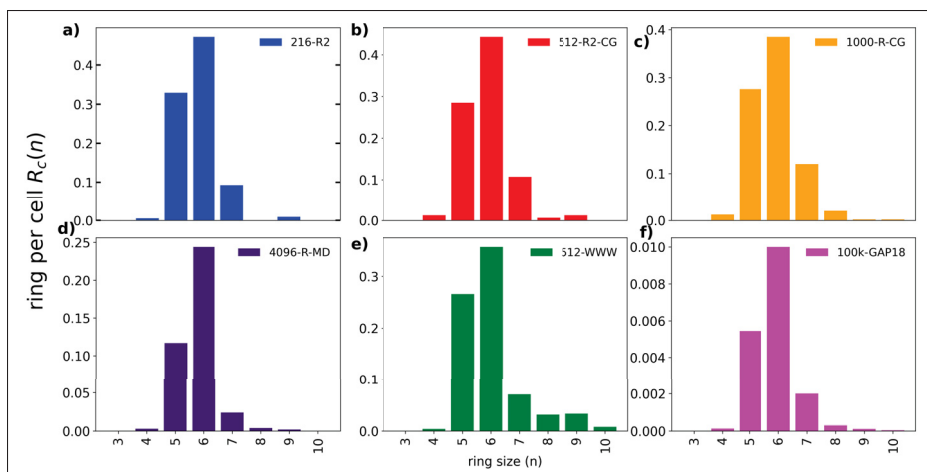


Figure 4.8 Ring statistics for ARTn-MTP generated models as well as those derived from the 512-WWW model and 100k-GAP18 atom models

Ring statistics are illustrated in Fig. 4.8. Crystalline silicon, characterized by its diamond-type structure, primarily features 6-membered rings, while amorphous silicon tends to favor rings of sizes 5, 6, and 7. Rings with fewer than 5 members or more than 7 are typically regarded as ring defects. Notably, no 3-membered rings are observed in any of the models. Additionally, very few 4-membered rings, which are known to introduce structural stress, were recorded across all models. As shown in Fig. 4.8 (a-f), each model predominantly contains 5, 6, and 7-membered rings, consistent with existing literature. These ring sizes are energetically favorable for a-Si. Furthermore, all models exhibit a few larger ring defects with sizes greater than 7, as detailed in Fig. 4.8.

4.4.3 Refinement of ART-SW optimized model : 1000-aSi-SW

As we first test the capability of ARTn coupled with MLIP to model amorphous systems, it is essential to explore several scenarios. A key consideration is the starting input configuration (see Table 4.1). Starting from a completely random state presents a challenge, particularly for larger systems. Ideally, the process should begin from a local minimum, which can be obtained through methods such as MD runs, CG relaxation, or ARTn coupled with a semi-empirical model. In

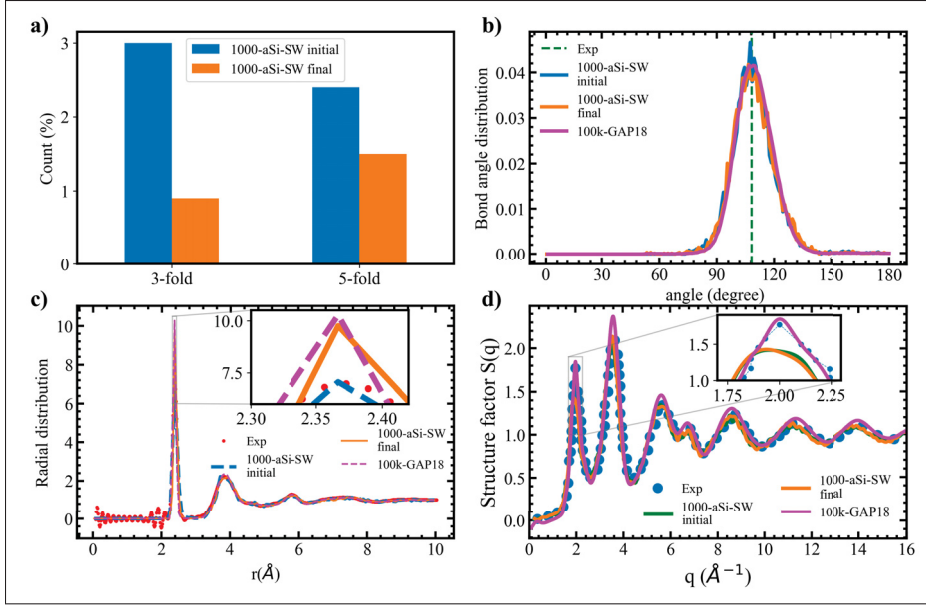


Figure 4.9 Local structural properties of ARTn-MTP refined configurations (1000-aSi-SW final) compared to ART-SW optimized configurations (1000-aSi-SW initial). a) Initial and final counts of 3-fold and 5-fold defects, b) bond angle distribution function, c) radial distribution function, and d) structure factor. For figures (b), (c), and (d), experimental data and results from the 100k-GAP-18 atom model are included to assess the quality of both the initial and final configurations of the 1000-aSi-SW model. The ART-SW optimized configurations served as the input for the ARTn-MTP refined configurations

this section, we present preliminary results that demonstrate the capability of ARTn-MTP to push the limits of relaxation for a configuration already optimized by ARTn coupled with the Stillinger-Weber (SW) potential (Stillinger & Weber, 1985).

A previously published a-Si model (Barkema & Mousseau, 1996; Malek & Mousseau, 2000) using ARTn with the SW potential (Stillinger & Weber, 1985), modified by Vink (Vink *et al.*, 2001b), served as input for an ARTn-MTP process. The 1000-atom ARTn-SW a-Si configuration (1000-aSi-SW initial) was refined through ARTn-MTP, resulting in the (1000-aSi-SW final) configuration. The goal was to assess the ARTn-MTP process's ability to refine configurations optimized with a semi-empirical potential. The energy dropped by -23.25

meV/atom, and structural properties, shown in Fig. 4.9 and summarized in Tab. 4.3, reveal significant improvements. The number of dangling and floating bonds decreased, reducing total defects by 3%. Bond angles (Fig. 4.9 b) are closer to the ideal 109.5° , with a narrower spread in the final model. Bond lengths were concentrated around $2.38 \text{ \AA}$, aligning with experimental values, and the radial distribution function in Fig. 4.9 c confirmed this. The 1000-aSi-SW final and 100k-GAP18 models showed better agreement with experimental data than the initial configuration. The structure factor, in Fig. 4.9 d, also matched experimental $S(q)$. The excess energy (ΔE) decreased from 0.2021 eV/atom (initial) to 0.1789 eV/atom (final), falling within the experimental range of $0.135 - 0.205 \text{ eV/atom}$ (Limbu *et al.*, 2018). The final configuration, with lower excess energy, is more stable and realistic. The bulk modulus increased from 65.8573 GPa (initial) to 68.8690 GPa (final), indicating enhanced stiffness. These results demonstrate that the ARTn-MTP coupling scheme effectively relaxes a semi-empirically optimized amorphous configuration.

Tableau 4.3 Structural and elastic properties for 1000-aSi-SW initial and 1000-aSi-SW final configurations : coordination defects, coordination number, average bond length, average bond angle, bulk modulus and energy difference per atom between amorphous state and crystalline Si. Standard deviations are provided in parentheses. The ART-SW optimized configurations (1000-aSi-SW initial) served as the input for the ARTn-MTP refined configurations (1000-aSi-SW final)

Properties	1000-aSi-SW initial	1000-aSi-SW final
Density (g/cm^3)	2.20	2.24
Fourfold Si (%)	94.60	97.60
Threefold Si (%)	3.00	0.90
Fivefold Si (%)	2.40	1.50
Average bond length (σ)	2.38 (0.05)	2.38 (0.03)
Average bond angle (σ)	109.11 (10.91)	109.12 (11.26)
Average coordination number	3.99	4.01
ΔE (eV / atom)	0.2021	0.1789
Bulk modulus (GPa)	65.8573	68.8690
Crystallinity (%)	0.00	0.00

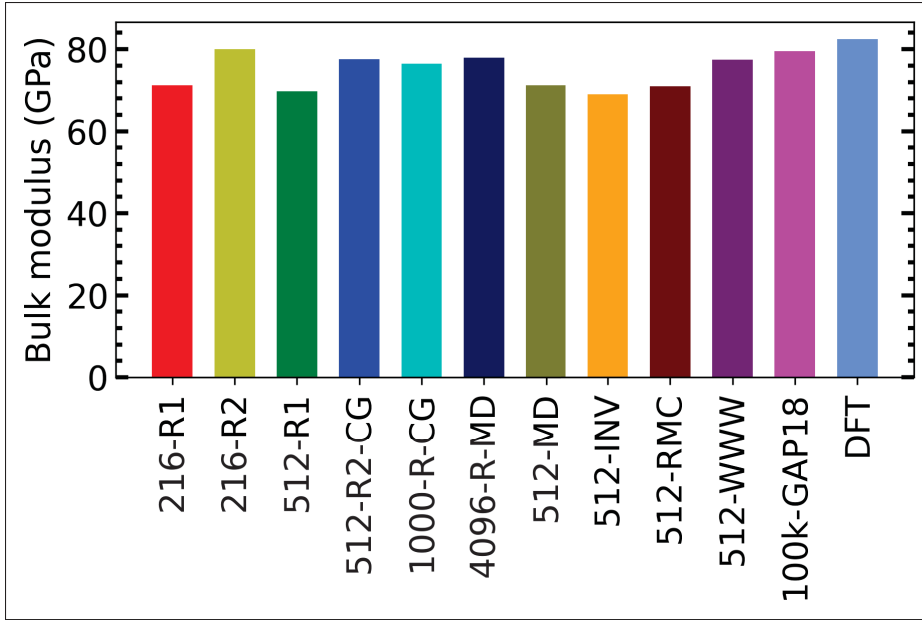


Figure 4.10 Bulk modulus for ARTn-MTP generated models as well as models from the literature

4.4.4 Stiffness

We calculated the bulk modulus B of our models and of the models found in the literature. The results are presented in Fig. 4.10. We include the bulk modulus value obtained from density functional theory (DFT) simulations reported in Ref. (Durandurdu & Drabold, 2001), which is 82.5 GPa. For our 216-R2 configuration with 100% 4-fold coordination, we obtained a bulk modulus of 80.01 GPa. The 512-R2 model yields a bulk modulus of 77.53 GPa, closely matching the 512-WWW model's value of 77.41 GPa. In contrast, other models such as 512-MD, 512-INV, and 512-RMC exhibit lower bulk modulus values of 71.18 GPa, 69.00 GPa, and 70.90 GPa, respectively. The 1000-R model demonstrates a bulk modulus of 76.46 GPa. Our largest models, the 4096-R-MD and 100k-GAP18, yield bulk modulus values of 77.93 GPa and 79.56 GPa, respectively. All of these values were obtained using our unified machine learning potential developed for the Si, O, and SiO₂ systems. Other potentials may yield different results. To our knowledge, there are no clear experimental benchmark values for the bulk modulus of amorphous silicon.

4.4.5 Potential energy

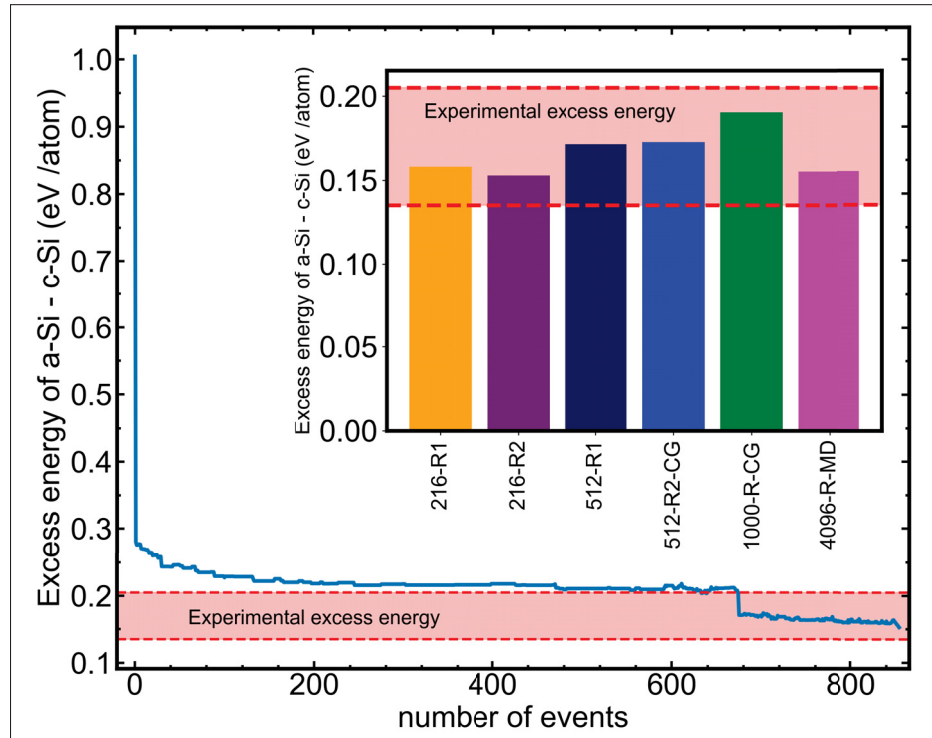


Figure 4.11 Excess energy between ART-MTP configurations (a-Si) and crystalline silicon (c-Si) as a function of the number of accepted events. A total of 4 events per atom was required to achieve 100% 4-fold coordination in 216-R2 a-Si model

Fig. 4.11 shows the relationship between energy decrease and the number of events for the model 216-R2. As can be seen, the energy decreases as the number of events increase. A detailed analysis of the events and key parameters is provided in the Nature of the Relaxation section. The experimental excess energy interval is indicated by the light red shaded region in the figure for reference. The inset of Fig. 4.11 shows the excess energy per atom for each of our models. These values are also presented in table 4.4. All models fall within the experimental excess energy interval (Limbu *et al.*, 2018), as detailed in the inset. Notably, the 216-R2 model, which initially contained over 70% coordination defects (see Fig. 4.2), was efficiently relaxed to achieve 100% 4-fold coordination and the lowest excess energy of 0.17 eV/atom with just 4 events per atom. To our knowledge, with the exception of a few Keating-potential-based WWW models (Wooten

Tableau 4.4 Initial (ρ_i) and final (ρ_f) density for the ARTn-MTP generated a-Si model.

Each model was randomly filled and the simulation box sizes was adjusted to match the initial density, which was maintained from the beginning to the end of the ARTn simulation.

The final density was obtained by relaxing each of the generated models to 0.5 K as described in main test. The values in parentheses were obtained by relaxing each model at 300 K and 1 atm. The excess energy in eV per atom (ΔE) relative to crystalline silicon is presented alongside the experimental excess energy

Models	ρ_i (g/cm ³)	ΔE (eV / atom)	ρ_f (g/cm ³)
216-R1	2.23	0.16	2.24 (2.23)
216-R2	2.28	0.15	2.25 (2.24)
512-R1	2.23	0.18	2.23 (2.22)
512-R2	2.28	0.18	2.25 (2.24)
1000-R	2.28	0.19	2.25 (2.24)
4096-R-MD	2.28	0.16	2.26 (2.25)
Exp	2.28 (Custer <i>et al.</i> , 1994)	0.135 – 0.205 (Limbu <i>et al.</i> , 2018)	-

et al., 1985; Pandey *et al.*, 2016b; Barkema & Mousseau, 2000), no other method, including DFT (Pedersen *et al.*, 2017), has yielded a model of a-Si with no coordination defects.

4.4.6 Density

We computed the density of our models at 0 K. The final density was obtained by running a 50 ps molecular dynamics simulation, cooling the system from 300 K to 0.5 K, and then holding it at 0.5 K for an additional 50 ps, while barostatting the system at 0 Pa. In separate simulations, we also equilibrated each of the aforementioned models at 300 K and 1 atm for 100 ps. The results are presented in table 4.4. The ART-MTP a-Si density values at 0 K are 0.71% to 1.94% larger than the experimental density measured at room temperature. ART-MTP a-Si density at 300 K and 1 atm is characterized by density values 1.18% to 2.43% larger than experimental densities. The MTP potential is trained using a generalized gradient approximation exchange-correlation functional, which is known to overestimate the density of crystalline silicon by 2 % (Jain *et al.*, 2015). Retraining the MTP on a higher-accuracy exchange-correlation functional, such as the restored regularized strongly constrained and appropriately normed meta-generalized gradient

approximation (Furness *et al.*, 2020) is likely to solve this small discrepancy with experimental values.

4.4.7 Nature of the relaxation

The primary goal is not only to generate high-quality amorphous structures but also to understand the underlying physics of the relaxation process in amorphous systems over time. Thus, to understand the relaxation process in our amorphous system, we analyze simulation data, focusing on key parameters such as barrier distribution, atomic displacements, active atoms, and energy decrease, considering only accepted events as detailed in Fig. 4.12. These parameters can help provide insights into the nature and mechanism of the relaxation process. The data were collected under normal deformation, excluding shear deformation during the simulation. We first examine the correlation between atomic displacement and energy barriers. A clear correlation is observed: larger displacements are typically associated with higher energy barriers and more active atoms, as shown in Fig. 4.12 (a). To further validate this, we calculate the correlation between the square of the atomic displacement and the energy barriers. According to Hooke's law, we expect a quadratic relationship between energy and displacement. When we plot energy against the square of the displacement, we expect a linear relationship, which our results confirm, yielding a strong correlation. Next, we find no significant correlation between the energy barriers and the energy gain, nor between energy gain and the number of active atoms, which is consistent with previous studies (Kallel *et al.*, 2010). Furthermore, no correlation exists between the number of atoms displaced up to the saddle point and the final state. As illustrated in Fig. 4.12(d), a correlation exists between energy barriers and the number of active atoms, with higher energy barriers typically involving a greater number of active atoms. On average, the number of active atoms varies from 10 to 30, with a mean of around 20 atoms that moves by more than 0.1 Å per event. This suggests that each event is localized, typically involving a cutoff radius of less than 7 Å. The microstructural changes primarily occur within a few neighboring cells, indicating a localized relaxation process. The relationship between the aforementioned parameters and the

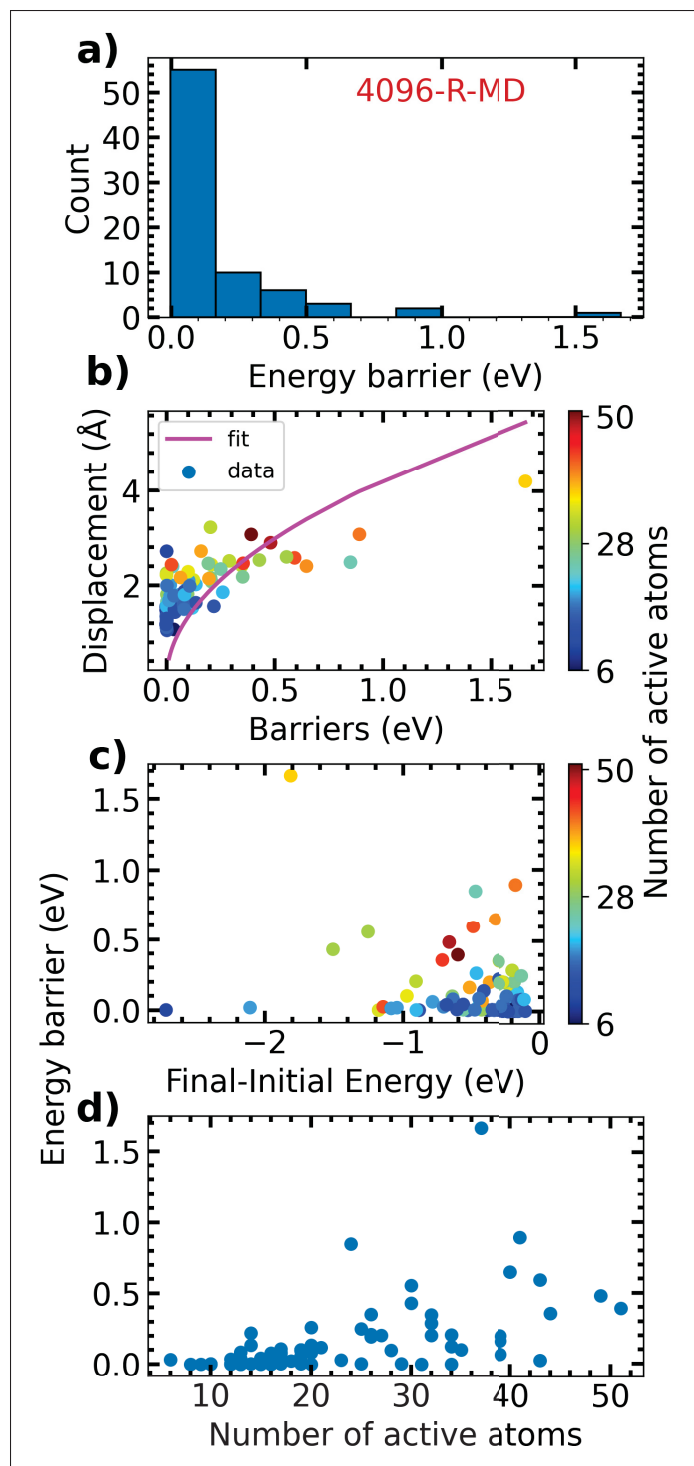


Figure 4.12 Analysis of accepted Events in ARTn-MTP optimization : a) Barrier distribution, b) square root of squared atomic displacement versus energy barrier, c) energy barrier as a function of energy decrease, and d) energy barrier versus active atoms during events

number of randomly displaced atoms at the start of ARTn searches (which depends on the cutoff and the central atom) was not investigated in this work.

4.4.8 Computational cost and nonlinear optimization process

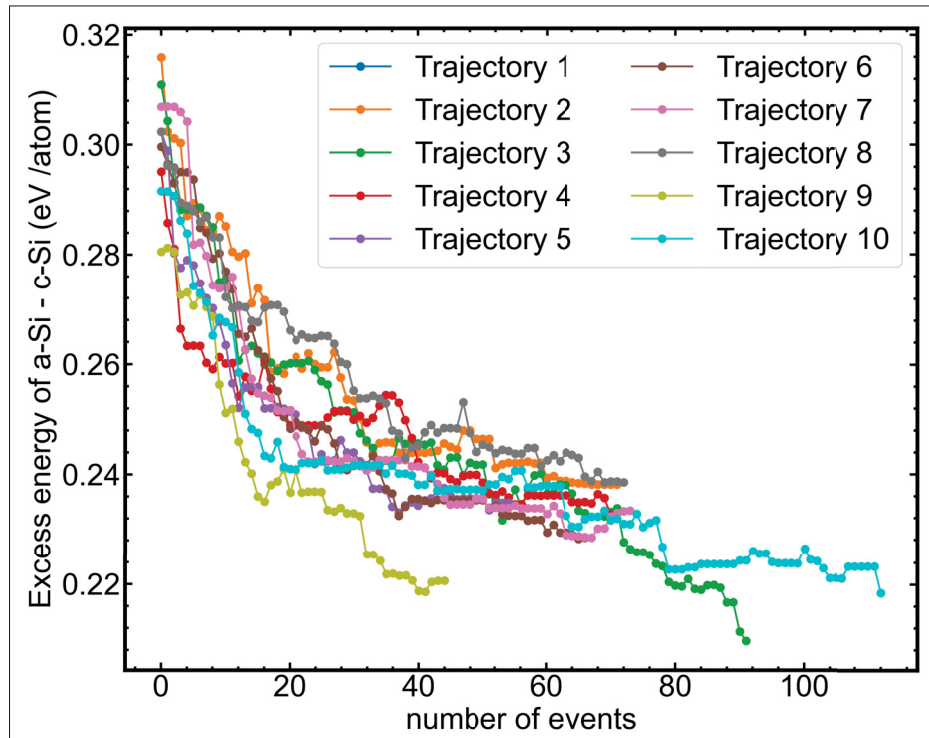


Figure 4.13 Nonlinear optimization effects characterized by excess energy per atom and number of events computed from independent runs with identical input Configuration

One of the primary challenges in computational materials modeling is finding the right balance between accuracy and computational cost. In this work, we present the first-ever coupling between ARTn and machine learning potentials, and we document the computational cost involved in preparing our amorphous system.

For the system of 216 atoms, the optimization of configurations required 0.0615 core-years and involved 55.20 millions force evaluations. Notably, one of the 216 atoms was optimized to achieve 100% 4-fold coordination. For the larger systems, the computational costs were as follows : the 512-atom system took 0.22 core-years, and the 1000-atom system required 0.34

core-years. Finally, for the 4096-atom system, we recorded 0.09 core-years to refine the MD configuration, with 37.25 million force evaluations. As mentioned in previous sections, this configuration achieved 98.05% 4-fold coordination and an excess energy of 0.15 eV per atom, which is closer to the experimentally observed low excess energy value.

The nonlinear optimization process within the ARTn framework was explored using a randomly generated 216-atom input configuration, with independent runs conducted using the same number of steps. Data were collected from these independent trajectories. Fig. 4.13 illustrates the relationship between excess energy levels and the number of accepted events for each trajectory. The results reveal varying numbers of accepted events and relaxation levels across the trajectories. Among the independent runs, the lowest number of accepted events was 45, while the highest was 113. Despite this difference, both systems achieved nearly identical excess energy levels. The mean number of accepted events was approximately 75, with most systems exhibiting higher excess energy than the system with 92 accepted events, which showed the lowest excess energy. These findings are indicative of a nonlinear optimization process.

4.5 Conclusions

In this paper, we investigated a-Si using ARTn coupled with a MTP. Seven models of varying sizes were constructed. They displayed excellent agreement with both experimental data and data obtained using a state-of-the-art 100,000-atom MD-based model leveraging a different MLIP. The proposed ARTn-MTP optimizer can further refine and relax configurations that have already been previously optimized using ARTn with a semi-empirical model. Additionally, the relaxation level of a-Si generated through standard molecular dynamics simulations via the melt-quench process can also be improved using the ARTn-MTP coupling scheme. We have generated high-quality amorphous structures of various sizes, including 216, 512, and 1000 atoms, without any traces of local crystallinity. We also found that starting the ARTn simulation with a randomly filled configuration results in an optimized structure with less crystallinity compared to initiating the simulation with an MD prerelaxed configuration. The question of local crystallinity, its link with defects and the ideal amorphous structure has not been addressed previously, due to the

lack of good quality models generated through different pathways (non-dynamic, such as with ARTn, and melt-and-quench, for example). The results presented here raise numerous related research questions that could further enhance our understanding of a-Si and its fabrication. Future research could focus on increasing system sizes while minimizing coordination and ring defects, for example, by targeting over-coordinated, under-coordinated, and high-energy atoms during event searches. Further studies will investigate the origins of crystallinity, explore its formation during the preparation of amorphous silicon, and examine methods for producing large-scale amorphous a-Si free of crystallinity.

It is now clear that ARTn coupled with MLIPs is a very effective approach to model a-Si. However, if the primary goal is to replicate experimental conditions as accurately as possible, further innovation is necessary. Experimentally, a-Si is produced mainly by chemical vapor deposition and ion implantation; representative elementary volumes range from scales of tens of nanometers to micrometers, and processing times on the order of seconds to hours. Accessing such time- and length-scales would enable both applied studies—e.g. device-scale simulations—and fundamental studies—e.g. assessing the degree of hyperuniformity of a-Si. Although modeling large systems is straightforward thanks to parallel computing, reaching long timescales remains an elusive challenge. While there are examples of ARTn-based techniques reaching such timescales in irradiated c-Si and a-Si (Béland *et al.*, 2013; Joly *et al.*, 2013; N'Tsouaglo *et al.*, 2015; Béland *et al.*, 2015), further development is necessary in order for these methods to routinely achieve such outcome.

4.6 Acknowledgements

We thank the Digital Research Alliance of Canada for generous allocation of compute resources. Financial support was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Association canadienne-française pour l'avancement des sciences (ACFAS), and the Canada Research Chair on Sustainable Multifunctional Construction Materials.

Author contributions

L.K.B. initiated and coordinated the research project. K.Z. performed all ARTn-MTP and molecular dynamics simulations, as well as the analyses presented in this work. All authors—K.Z., H.S., N.M., L.K.B., and C.O.P.—contributed to the writing of the paper. C.O.P., N.M., and L.K.B. supervised K.Z., while C.O.P. and L.K.B. provided the necessary computing resources.

4.7 Competing interests

The authors declare no conflict of interest.

CHAPITRE 5

THE COST OF MODELLING SILICON-BASED MATERIALS WITH MACHINE LEARNING POTENTIAL

Karim Zongo¹, Laurent Karim Bland², Claudiane Ouellet-Plamondon¹

¹ Department of Construction Engineering, École de Technologie Supérieure,
1100 Notre-Dame St W, Montreal (Quebec) H3C 1K3, Canada

² Department of Mechanical and Materials Engineering, Queen's University
99 University Ave, Kingston, Ontario K7L 3N6

Paper published in Journal of Springer Nature Switzerland, p. 275-282, March 2025

Abstract

The emergence of machine learning potential has revolutionized material modelling, yet its development involves thousands of quantum mechanical calculations. These calculations, along with potential training and molecular dynamic simulations, require substantial computational resources. In this study, we aim to elucidate the resource consumption involved in developing machine learning interatomic potential for silicon-based materials modelling. The consumption data presented here are extracted from the Digital Research Alliance of Canada account, automatically saved over four calendar years by their internal computation system. This model is the start of a model aiming to represent construction materials.

Keywords : Silicon, Moment tensor potential, Parallel computing.

5.1 Introduction

Multiscale modelling serves as a cornerstone in civil engineering, enabling a comprehensive understanding of materials at the atomic level and aiding in risk mitigation strategies. Silicon and silica (SiO₂) play integral roles in various civil engineering materials such as aggregates, cement, and concrete (Karlina *et al.*, 2023). Their interaction with water is crucial to the long-term durability of infrastructure. Despite the valuable insights offered by first principles calculations based on density functional theory, their computational demands pose significant challenges in

terms of cost and scalability (Kohn & Sham, 1965). In contrast, machine learning (ML)-based material modelling leverages extensive databases derived from quantum mechanical calculations (Behler & Parrinello, 2007). ML force fields, developed through optimization procedures, encode geometric and ab initio properties without explicitly representing electronic structures (Zongo *et al.*, 2022, 2024). However, the computational demands of such methodologies necessitate access to high-performance computer clusters, surpassing the capabilities of standard computing resources. In practice, the development of the ML potential involves three main steps : database generation, model training, and molecular dynamics simulations (Bartók *et al.*, 2018; Zuo *et al.*, 2020) (or coupling with other modelling engines such as ARTn). Every step involves the consumption of computational resources. The database generations involved thousands of calculations. Estimating computation consumption at each step becomes very complicated due to the extensive computational time required for database generations, training, including database optimization. This article delves into an analysis of resource consumption during the development of a machine learning potential database, utilizing the Digital Alliance cluster in Canada (Alliance Canada, 2023). The cluster, distributed across Canada, enables the use of parallel computing techniques. The computation was conducted across five distinct clusters : Cedar, Graham, Beluga, Narval, and Niagara. Data collection spanned four calendar years, from 2020 to 2023, within the Alliance account. Every time a job is initiated on the Digital Research Alliance of Canada cluster, the consumption is automatically logged by the compute job monitoring system and transmitted to the user's account. Notably, at the onset of data collection, Narval had not yet been integrated into the Alliance cluster system. These data correspond to consumption recorded by Digital Research Alliance of Canada for each client. By comprehensively analyzing resource consumption across various modelling techniques, this article aims to optimize computational workflows and enhance efficiency in civil engineering material modelling endeavours. Through the integration of advanced computational techniques and high-performance computing resources, researchers can advance our understanding of material properties, leading to safer and more resilient infrastructure designs.

5.2 Computational methods

This section details respectively our ML model (the moment tensor potential), the database preparation with density functional theory and the main computing procedure.

5.2.1 Moment tensor potential

The moment tensor potential (MTP) formalism is based on the partitioning property of potential energy surface (PES) within a certain medium mainly non-polarized or non-charged medium in which the PES can be approximated as the contribution of individual atomic energies given by their local environments (Shapeev, 2016). The MTP model linearly expands the potential energy V_{local} of an atom into a complete set of basis functions.

$$V_{local} = \sum_{\beta} c_{\beta} B_{\beta} \quad (5.1)$$

The basis functions are determined from the scalar tensorial contraction of the local atomic environment descriptors give as follow :

$$M_{\mu, \nu}(r_{ij}, \tau_i, \tau_j) = \sum_j f_{\mu}(|r_{ij}|, \tau_i, \tau_j) \underbrace{r_{ij} \otimes \cdots \otimes r_{ij}}_{\nu \text{ times}} \quad (5.2)$$

$M_{\mu\nu}$ is a tensor of rank ν enclosing the radial distribution ($f_{\mu}(|r_{ij}|, \tau_i, \tau_j)$) and the angular information ($\underbrace{r_{ij} \otimes \cdots \otimes r_{ij}}_{\nu \text{ times}}$) of the local environment of the central atom . In practice, the contraction of $M_{\mu\nu}$ to a scalar is formulated by $n \times n$ symmetric matrix β which off-diagonal elements indicates the number of dimensions to be contracted between tensors M_{lm} and M_{pq} . The indexes $u \wedge v$ enables the bounding of basis size for practical implementation and allow the estimation of computational complexity and efficiency. This mathematical description of atomic interactions is implemented in MTP code (Novikov *et al.*, 2020a).

5.2.2 Generation of a database on parallel computing

The fitting of the MTP model mainly relies on a database from the first-principal calculation. Thus, in this work database consisting of Si, SiO₂, O and H₂O structures was generated using quantum mechanical calculations in the framework of density functional theory. The procedure for creating the database is illustrated in Figure 1 below. Several modelling softwares were used for manipulating, conversion, and visualization of atomic structures. Packages like XCrySDen (Kokalj, 1999), VESTA (Momma & Izumi, 2008), AtomsK (Hirel, 2015), and OVITO (Stukowski, 2009) were employed for these purposes. These software tools are set up on local computers and do not account for resource consumption. On the other hand, Quantum ESPRESSO (QE) (Giannozzi *et al.*, 2009) is utilized specifically for electronic structure calculations. The QE package, along with LAMMPS (Thompson *et al.*, 2022) and MTP code (Novikov *et al.*, 2020a), are solely responsible for the resource consumption presented in this work. Firstly, the QE package is employed to generate the database. Subsequently, the MTP code is utilized to train the model, followed by the utilization of the LAMMPS code to perform molecular dynamics simulations. All these codes are highly parallelized, maximizing computational efficiency. We prepare the input files for jobs on Digital Research Alliance of Canada, previously called Compute Canada, from our local computer, as depicted in Figure 2. Additionally, we detail the job monitoring and performance analysis in Figure 2. This performance analysis enables us to adjust the parallelization scheme for optimal calculations.

5.3 Results

This section presents the database generated using the QE package, along with the training of machine learning models and the resulting consumption. We generated approximately 7000 configurations comprising Si, SiO₂, and O structures. This number corresponds to approximately 10,000 single-point DFT calculations, as the selections were made from ab initio trajectories and variable cell relaxations. The table below presents the summary of the database. Another significant contributor to computational resource consumption is the training of the machine learning (ML) model. The training process begins with optimizing the database. Figure 4

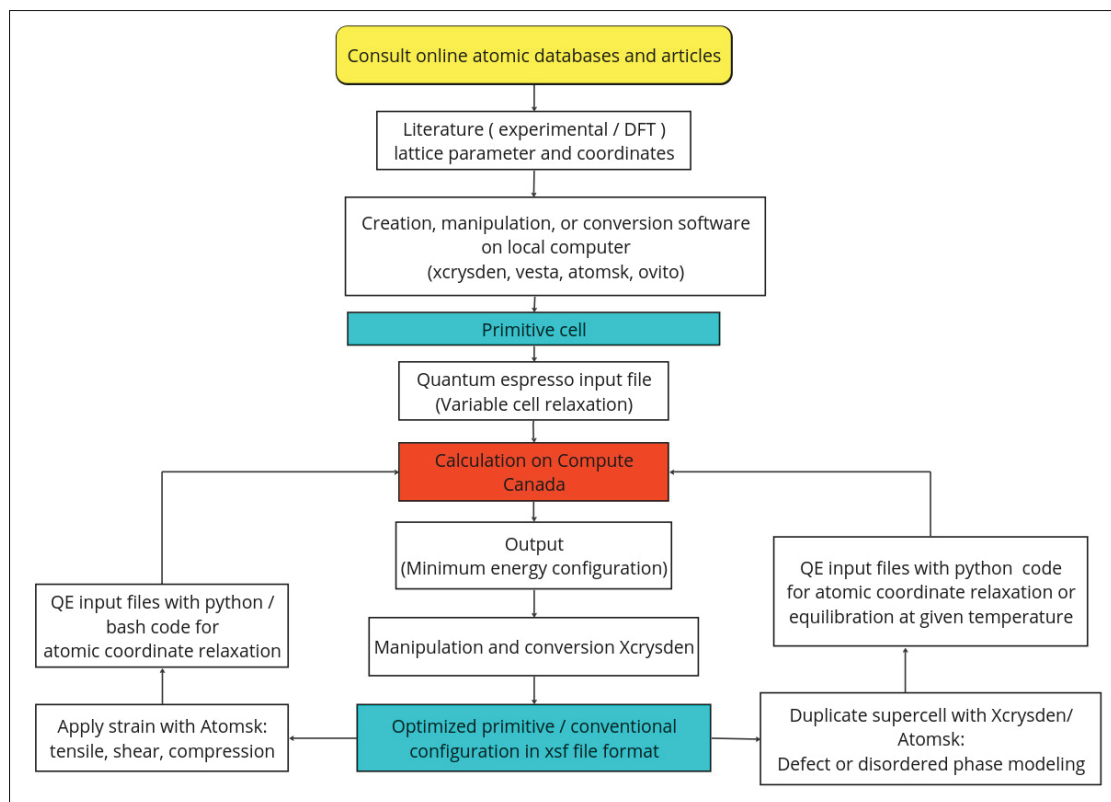


Figure 5.1 Creating a database of Si and SiO₂ Configurations : A Blueprint for database creation by combining two or more modelling software such as Xcrysden, VESTA, Atomsk, OVITO, and Quantum ESPRESSO

Tableau 5.1 Total number of configurations in the database and fraction utilized for the implementation of the unified Si-SiO₂-O potential (Zongo *et al.*, 2024)

Compounds	Number configurations	Fraction used for the implementation of the potential (%)
Si	3900	13
SiO ₂	3500	14
O	100	20
H ₂ O	2500	-

provides an illustration of the cost function during the training of the MTP model using level 26. Since the parameters are initially randomly initialized, finding the optimal model and data

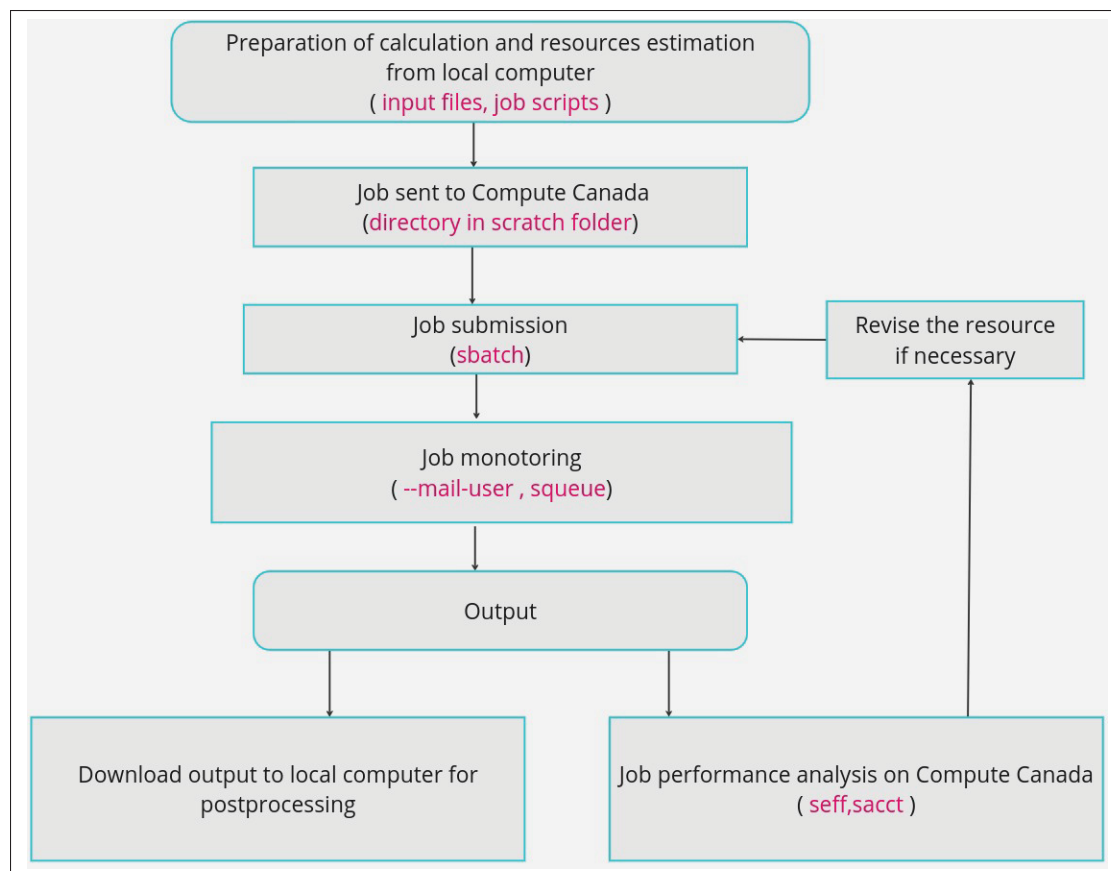


Figure 5.2 Migration of computation files and management, monitoring of computations on the Digital Research Alliance of Canada cluster

set for atomistic modelling necessitates several rounds of training. Consequently, the training and database optimization are iteratively repeated until a satisfactory model is obtained, which can then be utilized for molecular dynamics simulations. The costs associated with database generation, training, and molecular dynamics are depicted in Figure 4. The left figure illustrates the consumption for each year. Notably, the year 2021 exhibits the highest consumption. During this period, quantum mechanics calculations were exclusively performed, along with preliminary testing and training of the MTP model. In 2022 and 2023, consumption levels are nearly identical. During these years, classical molecular dynamics simulations were conducted concurrently with the generation of water databases. Notably, in 2022, the database underwent refinement, and the MTP model was trained. In 2020, we used fewer resources as our focus was primarily on

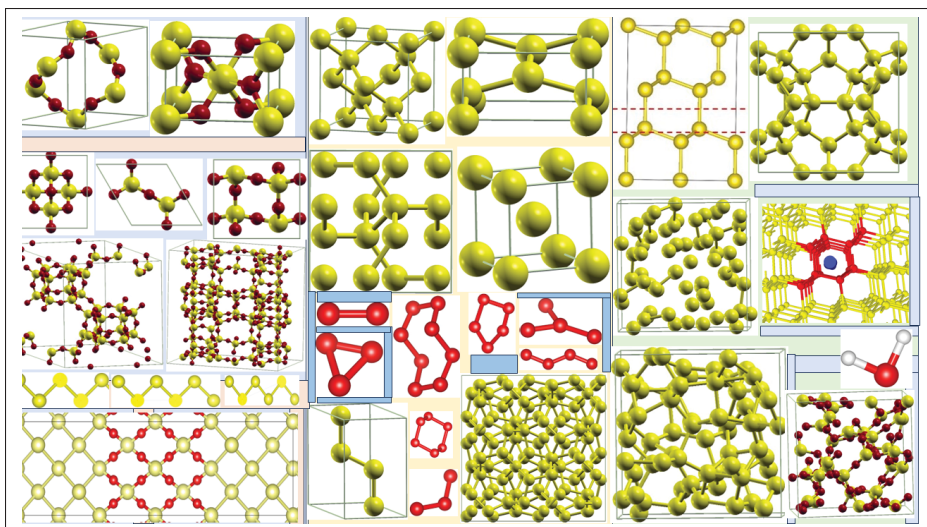


Figure 5.3 Atomic structures in the database : Liquid, amorphous, defects of silicon and silica, oxygen molecules, and water molecules. The configurations are labeled with energies, forces, and stress values

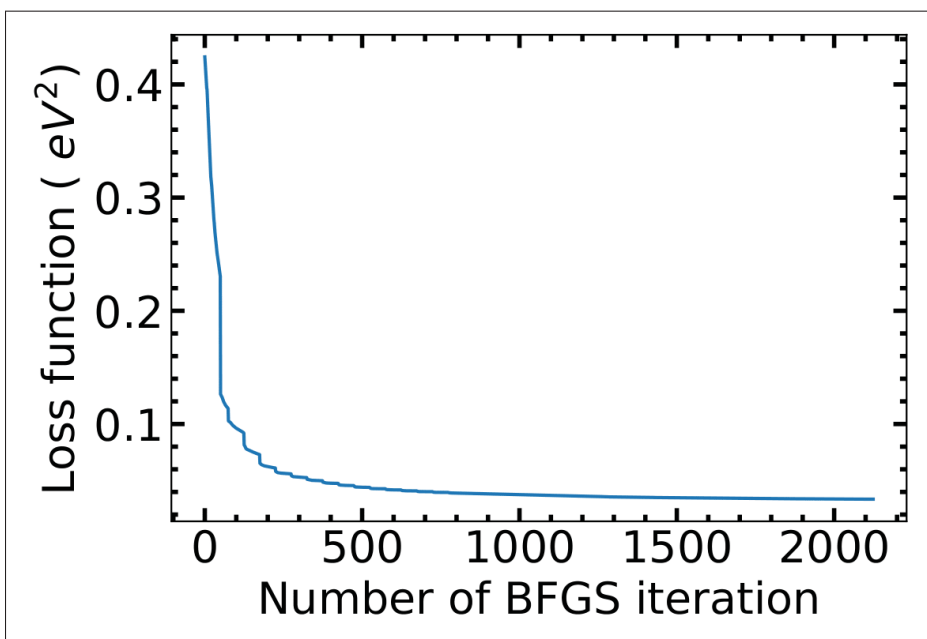


Figure 5.4 Evolution of the cost function during training on Digital Research Alliance of Canada Cluster

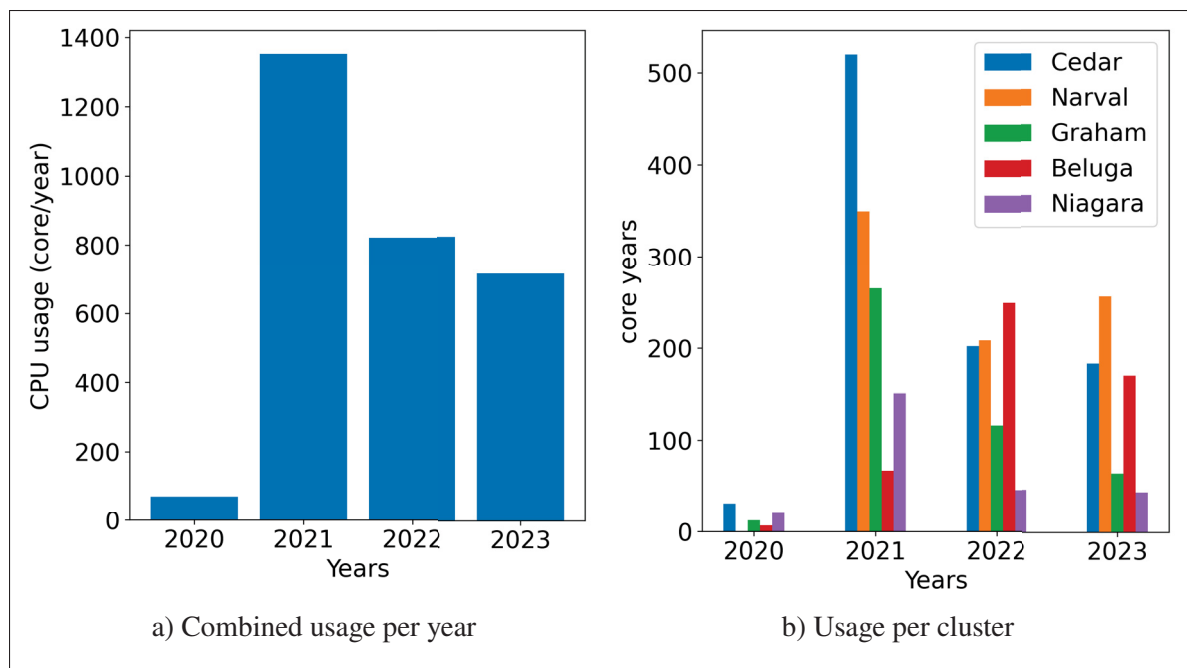


Figure 5.5 Resource consumption for our modelling projects over four calendar years

theoretical research activities. In terms of cluster usage, Cedar computers were predominantly employed for computations. Following closely, the Narval cluster also saw significant utilization. This can be attributed to the high number of computer nodes available on both Cedar and Narval, resulting in minimal waiting times. Additionally, the node configurations are noteworthy. For example, the nodes on Narval boast 64 and 48 cores each. Cedar boasts nodes with 24, 32, and 48 cores each, offering a range of computational power. Additionally, Narval surpasses Cedar in speed and efficiency when using our modelling packages. Other clusters, such as Graham and Beluga, also played significant roles in supporting our research endeavours.

5.4 Conclusion

In conclusion, our study provides valuable insights into resource consumption during the development of machine learning potential for material modelling. The overall consumption can be attributed to the extensive calculations performed to explore a wide range of material structures. It's worth noting that, due to a lack of experience and a lack of adherence to

proper methodologies at the outset of our research, some of the calculations may have been unnecessary. However, the resulting configurations can still be valuable for future research endeavours. This highlights the importance of experience and adherence to established practices in minimizing resource consumption in computational research. By shedding light on the computational demands involved in this process, researchers can better understand and manage resource utilization, ultimately leading to more efficient and sustainable modelling practices. In summary, by implementing strategies such as refining algorithms, optimizing code, and effectively leveraging high-performance computing resources, researchers can significantly enhance the efficiency and sustainability of their modelling endeavours.

Acknowledgement

The authors thank the Digital Research Alliance of Canada, Canada Research Chair Program (award CRC-2019-00074), ACFAS mobility scholarship, Nuclear Waste Management Organization of Canada (NWMO), and NSERC to support the research.

CHAPITRE 6

BASE DE DONNÉES DE L'EAU

Ce chapitre présente la création d'une base de données complète pour l'eau, conçue pour la modélisation par apprentissage automatique en utilisant le modèle du potentiel par tenseur de moment (MTP). La base de données couvre une large gamme de phases de l'eau, y compris l'eau liquide à différentes températures, les phases solides et les états gazeux. De plus, elle intègre d'autres formes de molécules d'eau, enrichissant ainsi la diversité de l'ensemble de données. Les calculs ont été réalisés à l'aide de simulations de dynamique moléculaire *ab initio* standard et de techniques de dynamique moléculaire Car-Parrinello (Car *et al.*, 1985b).

6.1 Introduction

L'eau est la substance la plus essentielle pour tous les êtres vivants. De la fonction des organismes à la croissance des plantes, l'eau est le composant le plus crucial de notre planète. Elle joue un rôle vital dans diverses industries, que ce soit en tant qu'ingrédient dans la fabrication de produits ou en tant que liquide de refroidissement et conservateur pour les produits réfrigérés ou congelés. De plus, l'eau est un élément fondamental dans des matériaux largement utilisés, tels que le ciment et le béton (Ley-Hernandez *et al.*, 2018; Hover, 2011; Bauchy *et al.*, 2014; Pellenq *et al.*, 2009). L'eau joue également un rôle crucial dans les systèmes d'argile (Brigatti *et al.*, 2006; Madsen *et al.*, 1989; Tambach *et al.*, 2004), comme la bentonite, qui est susceptible d'être utilisée dans les systèmes de stockage souterrain des déchets nucléaires et d'autres applications environnementales (Galamboš *et al.*, 2010; Ma *et al.*, 2019). De plus, l'eau est présente dans les zéolites, qui sont précieuses pour capturer les polluants environnementaux (Esposito *et al.*, 2015; Wang *et al.*, 2010). L'eau est au cœur des matériaux cimentaires (béton, ciment, ...) et des matériaux innovants conçus pour réduire l'empreinte carbone, tels que les polymères inorganiques activés par alcalins (Škvára, 2007; Davidovits, 2017; Ruan *et al.*, 2024). L'eau est essentielle à la fois pour la vie et pour l'industrie, mettant en évidence son rôle indispensable dans notre monde. La structure de l'eau peut être envisagée comme un réseau de molécules d'eau interconnectées par des liaisons hydrogène (Chen *et al.*, 2017; Villard *et al.*, 2024). Différents

types d'interactions contribuent à ses propriétés uniques, y compris les liaisons covalentes intramoléculaires, les liaisons hydrogène intermoléculaires et les forces de van der Waals (Chen *et al.*, 2017; Zheng *et al.*, 2018). Chaque molécule d'eau individuelle est formée par des liaisons covalentes fortes, tandis que les liaisons hydrogène jouent un rôle crucial pour maintenir les molécules d'eau ensemble dans la phase condensée. Cette organisation conduit à la formation de structures tétraédriques ou quasi-tétraédriques (Zheng *et al.*, 2018; Chen *et al.*, 2017).

Néanmoins, l'eau a fait l'objet de nombreuses recherches en raison de la complexité de sa structure. Les études et observations précédentes ont révélé plusieurs anomalies associées à l'eau liquide, notamment son anomalie de densité, sa capacité calorifique élevée, ses points d'ébullition et de fusion élevés, sa tension superficielle importante, sa constante diélectrique élevée et sa viscosité accrue (Debenedetti *et al.*, 2003; Speedy *et al.*, 1976; Mishima *et al.*, 1998; Dasgupta *et al.*, 2021; Gartner III *et al.*, 2020; Malenkov, 2006; Villard *et al.*, 2024). Par exemple, l'eau gèle en une phase solide qui est moins dense à température ambiante et sous pressions modérées, présente une densité maximale à une température spécifique et montre une viscosité réduite sous pression (Debenedetti & Stanley, 2003; Speedy & Angell, 1976; Mishima & Stanley, 1998; Nilsson *et al.*, 2011). De plus, des augmentations significatives de la compressibilité, de la capacité calorifique et du coefficient d'expansion thermique ont été observées à mesure que l'eau liquide se refroidit (Debenedetti & Stanley, 2003; Speedy & Angell, 1976; Mishima & Stanley, 1998; Nilsson & Pettersson, 2011). La stabilité des polymorphes de glace et la description précise du diagramme de phases de l'eau demeurent des questions centrales nécessitant davantage d'investigations (Salzmann, 2019; Della Pia *et al.*, 2022; Zhang *et al.*, 2021; Reinhardt *et al.*, 2021). De plus, une contraction thermique à basse température a également été signalée pour glace hexagonale (Ih) (Salim *et al.*, 2016). Afin de mieux comprendre ces anomalies, de nombreuses études expérimentales ont été menées, se concentrant principalement sur les liaisons hydrogène et la dynamique du réarrangement du réseau de liaisons hydrogène. Diverses techniques, telles que la diffraction des rayons X et des neutrons, la spectroscopie d'absorption des rayons X (XAS) et la spectroscopie infrarouge, entre autres, ont été utilisées pour sonder les propriétés microscopiques des structures de l'eau liquide et de la

glace (Head-Gordon *et al.*, 2002; Nilsson & Pettersson, 2011; Fecko *et al.*, 2003; Nibbering *et al.*, 2004; Winter *et al.*, 2004; Nilsson & Pettersson, 2011). Cependant, une compréhension complète de l'eau basée uniquement sur des données expérimentales reste difficile à atteindre (Zheng *et al.*, 2018). Plusieurs défis ont entravé la caractérisation complète de l'eau, notamment en ce qui concerne ses propriétés dynamiques. Par exemple, les fluctuations thermiques perturbent le réseau de liaisons hydrogène dans l'état liquide, ces fluctuations se produisant sur des échelles de temps de picosecondes à nanosecondes, ce qui représente un défi majeur pour les investigations expérimentales (Zheng *et al.*, 2018; Chen *et al.*, 2017). De plus, il a été noté que de nombreuses mesures expérimentales s'appuient sur des modèles théoriques pour leur interprétation (Chen *et al.*, 2017; Villard *et al.*, 2024). Compte tenu de la structure complexe de l'eau, les simulations numériques ont de plus en plus servi de substitut aux expériences, permettant aux chercheurs d'explorer des régions qui ne sont pas facilement accessibles par les méthodes expérimentales établies (Gillan *et al.*, 2016; Schwegler *et al.*, 2004; Chen *et al.*, 2017; Della Pia *et al.*, 2022; Salzmann, 2019; Della Pia *et al.*, 2022; Zhang *et al.*, 2021; Reinhardt & Cheng, 2021).

La modélisation théorique et la simulation de l'eau ont été rendues possibles grâce au développement continu des outils informatiques. En conséquence, divers modèles d'eau—allant des modèles rigides et flexibles aux modèles polarisables et non polarisables—peuvent être simulés. Les simulations traditionnelles de l'eau ont utilisé des potentiels semi-empiriques, tels que ceux de la famille TIPnP (Abascal *et al.*, 2005; Vega *et al.*, 2011), ainsi que le modèle SPC/E (Berendsen *et al.*, 1987), en plus de diverses approximations de la théorie de la fonctionnelle de la densité (DFT) (Hohenberg *et al.*, 1964; Sholl & Steckel, 2011; Tuckerman, 2002; Payne *et al.*, 1992; Car & Parrinello, 1985b). Les modèles TIPnP (où n représente le nombre de sites d'interaction) sont des potentiels moléculaires utilisés en dynamique moléculaire pour simuler les interactions entre les molécules d'eau. Ces modèles, tels que TIP3P, TIP4P et TIP5P, diffèrent par le nombre de sites d'interaction qu'ils introduisent pour représenter la molécule d'eau, permettant ainsi une modélisation plus précise de ses propriétés thermodynamiques et structurales. Le modèle SPC/E (Simple Point Charge/Extended) est une variante du modèle SPC (Simple Point Charge). Il utilise trois sites d'interaction : un atome d'oxygène et deux atomes d'hydrogène,

chacun portant une charge partielle. Ce modèle est conçu pour améliorer la représentation des propriétés structurales et dynamiques de l'eau, notamment en reproduisant plus fidèlement la densité et la diffusion de l'eau liquide. Les modèles semi-empiriques et la DFT ont tous deux leurs propres limitations. Les potentiels semi-empiriques, bien que computationnellement efficaces, sont souvent trop simplistes pour capturer avec précision la complexité inhérente de la structure de l'eau sous ses différentes formes. D'autre part, bien que la DFT offre une plus grande précision, elle est limitée aux petits systèmes et aux courtes échelles de temps, ce qui la rend inadaptée pour saisir les propriétés dynamiques de l'eau à travers ses différentes phases. De plus, les résultats de la DFT sont fortement sensibles au choix de la fonctionnelle d'échange-corrélation (XC) (Perdew & Schmidt, 2001; Villard *et al.*, 2024). Par exemple, les fonctionnelles PBE et celles qui y sont associées ont du mal à décrire avec précision les systèmes d'eau (Gillan *et al.*, 2016; Villard *et al.*, 2024; Chen *et al.*, 2017). La fonctionnelle récemment développée SCAN (Strongly Constrained and Appropriately Normed) (Sun *et al.*, 2015) a montré une grande précision dans la modélisation des structures de l'eau (Chen *et al.*, 2017; Zheng *et al.*, 2018). SCAN est une fonctionnelle non empirique de type approximation de gradient généralisé (meta-GGA) spécifiquement conçue pour satisfaire 17 contraintes physiques exactes et pour récupérer avec précision plusieurs interactions non liées (Sun *et al.*, 2015). Malgré la grande précision de SCAN, des limitations spatio-temporelles demeurent un défi.

Les avancées récentes incluent l'application de potentiels d'apprentissage automatique pour étudier les propriétés uniques de l'eau. Des modèles d'apprentissage automatique notables, basés sur des potentiels de réseaux de neurones, ont été mis en œuvre avec succès pour les études sur l'eau (Schran *et al.*, 2019, 2021; Zhai *et al.*, 2024; Wang *et al.*, 2018). Bien que ces potentiels de réseaux de neurones aient permis de révéler des propriétés clés de l'eau, il est encore nécessaire d'explorer d'autres formes de potentiels d'apprentissage automatique (Wang *et al.*, 2024). Cet article présente une base de données pour la modélisation de l'eau, utilisant le potentiel par tenseur de moment (MTP). La base de données a deux objectifs principaux : d'abord, elle permettra la paramétrisation du potentiel MTP pour l'étude des structures de l'eau et des phénomènes associés ; ensuite, elle servira de bases pour la paramétrisation d'un potentiel

pour le gel de silice, facilitant l'étude des interactions de l'eau avec les matériaux en silice. Cela constitue une étape essentielle pour la modélisation des matériaux aluminosilicatés et de l'argile en utilisant des approches d'apprentissage automatique.

6.2 Méthodes

6.2.1 Génération de la base de données

La base de données a été construite en utilisant plusieurs techniques pour échantillonner l'espace de configuration, y compris les simulations par apprentissage actif (Podryabinkin & Shapeev, 2017). Pour construire notre ensemble actif, nous avons placé une seule molécule d'eau dans une boîte de simulation cubique, avec une densité correspondant à celle de l'eau liquide à température ambiante. La longueur de la liaison O-H dans la molécule a été modifiée pour générer quatre structures initiales, pour lesquelles l'énergie, les forces et les contraintes ont été calculées à l'aide de la DFT. Ces quatre structures ont ensuite été utilisées pour entraîner le potentiel MTP de niveau 08 (Novikov *et al.*, 2020b). Le potentiel résultant a ensuite été appliqué dans le processus d'apprentissage actif. L'approche par apprentissage actif mise en œuvre dans le cadre du MTP utilise le critère D-optimalité (Podryabinkin & Shapeev, 2017; Novikov *et al.*, 2020b), qui détecte les configurations où le potentiel MTP extrapole pendant les simulations MD. Nous avons exécuté des trajectoires MD à 300 K pendant 100 ps, en collectant des configurations extrapolées. Bien que le potentiel MTP soit réentraîné en temps réel pendant la simulation de 100 ps, notre objectif était de nous concentrer sur les configurations extrapolées, plutôt que sur le potentiel réentraîné. La simulation a été arrêtée une fois que le grade d'extrapolation maximal autorisé a été dépassé. À ce stade, les configurations extrapolées ont été ajoutées à l'ensemble actif initial, le potentiel a été réentraîné, et le processus a été répété. L'objectif était d'équilibrer la molécule d'eau. De plus, nous avons effectué un apprentissage actif en utilisant les ensembles NPT (Nombre de particules, Pression et Température constant) et NVT (Nombre de particules, Volume et Température constant) à 500 K et 700 K.

Pour étendre davantage notre base de données, nous avons effectué des simulations à la fois de dynamique moléculaire Born-Oppenheimer (BOMD) (Born *et al.*, 1985) et de dynamique moléculaire Car-Parrinello (CP-MD) (Car & Parrinello, 1985b). Dans les simulations BOMD, nous avons utilisé un pas de temps de 1 fs pour équilibrer divers modèles, y compris l'eau pure, le peroxyde d'hydrogène, l'acide orthosilicique contenant des molécules d'eau et un modèle de silice amorphe avec de l'eau, le tout à environ 300 K. Les tailles des systèmes variaient de 3 à 192 atomes. Pour les simulations de dynamique moléculaire Car-Parrinello, nous nous sommes concentrés sur l'équilibrage de l'eau liquide à 300 K. Une boîte de simulation contenant 64 molécules d'eau a été utilisée, avec une densité correspondant à celle de l'eau à température ambiante. Nous avons exploré deux ensembles, NVT et NPT, pendant les simulations. Le pas de temps était défini sur 2 unités atomiques (environ 0,0484 fs), avec la masse de l'hydrogène définie comme étant de 2 unités de masse atomique. Le temps total de simulation était de 13,62 picosecondes.

En plus des runs dynamiques précédemment mentionnés, nous avons effectué des calculs à un point pour diverses molécules, y compris l'eau, l'hydrogène, le peroxyde d'hydrogène, l'acide orthosilicique, l'acide pyrosilicique, des structures de glace et de l'hydrogène solide.

Toutes les simulations ont utilisé des pseudopotentiels Vanderbilt normés optimisés (ONCV), et les interactions d'échange-corrélation ont été traitées avec la fonctionnelle meta-GGA SCAN (Strongly Constrained and Appropriately Normed) (Sun *et al.*, 2015). Le choix des k-points variait en fonction du type de molécule et de la taille des boîtes de simulation. De plus, nous avons consulté plusieurs bases de données, y compris le Materials Project (Jain *et al.*, 2013), et avons utilisé divers outils pour la gestion et la manipulation des fichiers, tels qu'Atomsk (Hirel, 2015), Vesta (Momma & Izumi, 2008), Ovito (Stukowski, 2009) et Xcrysden (Kokalj, 1999).

6.3 Résultats

Cette section présente un résumé de la base de données créée pour l'eau et les composés associés. Le tableau 6.1 ci-dessous présente les entrées de notre base de données, incluant des détails

tels que le type de calculs, le nombre d'atomes dans chaque configuration, et le nombre total d'entrées pour chaque type de calcul. Un total de 2 500 configurations atomiques ont été générées pour ce projet.

Tableau 6.1 Base de données

Eau pure		
Type	Nombre d'atome	Total
Molécule H_2O	3	4
MD 240 K, 260 K, 300 K, 375 K	3, 12, 18, 32, 48, 64, 192	1716
Apprentissage actif 300 K, 500 K, 700 K	3	404
Hydrogène		
Molécule H_2	2	1
H_2 solid	2, 4, 8	8
Peroxyde d'hydrogène		
Molécule H_2O_2	4	1
MD 300K	32	44
Acide orthosilicique		
Molécule $\text{Si}(\text{OH})_4$	9	1
MD 300K	54, 81	102
Acide pyrosilicique		
Molécule $\text{H}_6\text{Si}_2\text{O}_7$	15	1
Défaut hydrogarnet (Silice amorphe)		
MD 300K	120, 54, 57	62

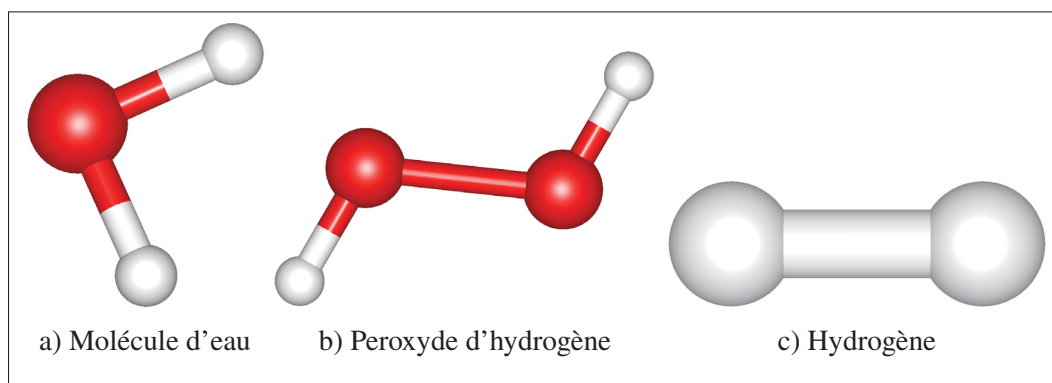


Figure 6.1 Modèles moléculaires de l'eau, du peroxyde d'hydrogène et du dihydrogène

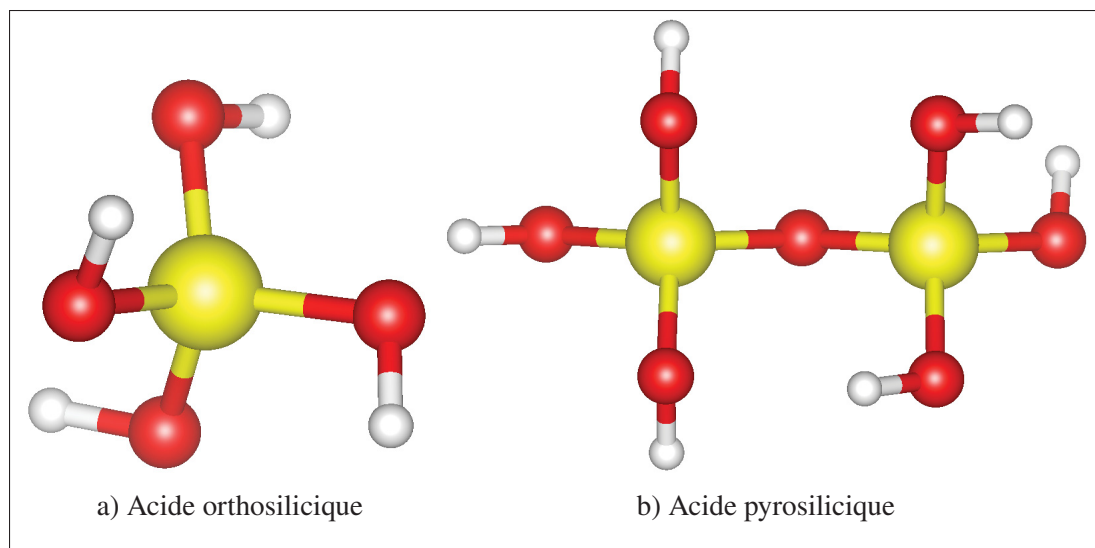


Figure 6.2 Structures moléculaires de l'acide orthosiliciques et de l'acide pyrosiliciques

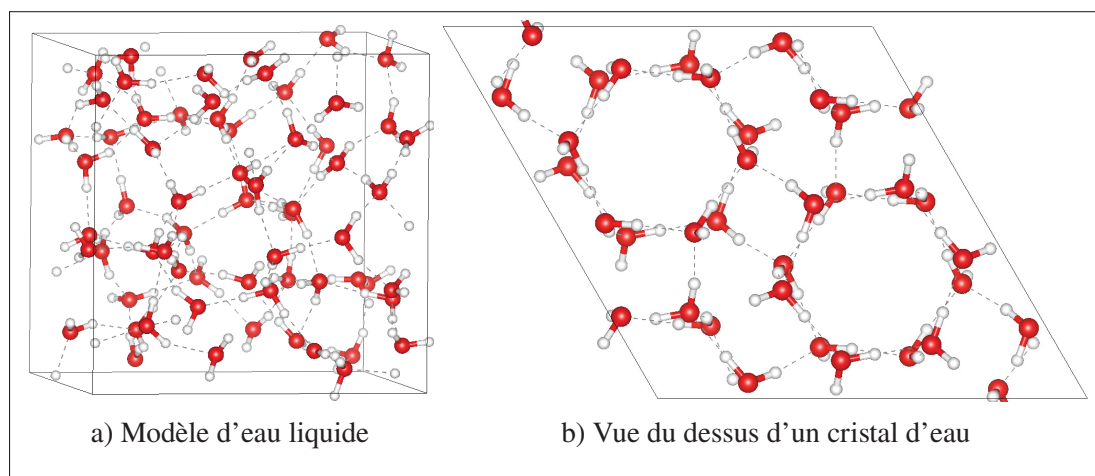


Figure 6.3 Eau liquide dans une boîte de simulation cubique et vue de dessus d'une structure de glace hexagonale

6.4 Conclusion

Dans ce travail, une base de données conçue pour les simulations de l'eau a été générée à l'aide de calculs basés sur la mécanique quantiques. La fonctionnelle SCAN (Strongly Constrained and Appropriately Normed) a été utilisée avec succès pour ces calculs. Les données sont destinées à entraîner un potentiel d'apprentissage automatique basé sur le modèle de Potentiel par Tenseur

de Moment (MTP). Les travaux futurs consisteront à entraîner le potentiel MTP sur cette base de données, suivi de simulations de dynamique moléculaire. De plus, les efforts futurs permettront d'étendre la base de données sur l'eau pour inclure davantage de structures pour les simulations de gel de silice. Une partie du jeu de données sur le silicium et des molécules d'oxygène issues d'études précédentes (Zongo *et al.*, 2024) pourra être recalculée à l'aide de la fonctionnelle SCAN afin d'enrichir la base de données présentée ici, améliorant ainsi l'entraînement du potentiel pour le gel de silice.

CONCLUSION

L'objectif principal de cette thèse est de modéliser les interactions interatomiques au sein du silicium et de la silice grâce à des approches d'apprentissage automatique. Le silicium et la silice représentent des éléments fondamentaux dans la composition de matériaux complexes tels que les argiles, les matériaux cimentaires, les polymères inorganiques activés par des agents alcalins, ainsi que les zéolithes. Cette recherche vise à approfondir notre compréhension des mécanismes à l'œuvre dans ces matériaux, en tirant parti des capacités prédictives de l'apprentissage automatique pour améliorer la modélisation de leurs propriétés. Dans cette section, nous résumerons les travaux réalisés au cours de cette thèse, en mettant en lumière les contributions de chaque chapitre. Nous proposerons également des recommandations et des pistes pour des travaux futurs, afin de poursuivre l'exploration des sujets abordés et d'approfondir notre compréhension des interactions interatomiques. Cette synthèse vise à fournir une vision claire des résultats obtenus et à tracer des voies prometteuses pour les recherches à venir.

Cette thèse se compose de cinq chapitres, dont quatre mettent en lumière nos contributions directes et indirectes, en adéquation avec nos objectifs de recherche. Dans le deuxième chapitre, une base de données a été conçue, générée et testée avec succès, constituant ainsi l'une des trois composantes essentielles au développement d'un modèle d'interaction basé sur l'apprentissage automatique. Il est important de rappeler que le processus de développement d'un potentiel par apprentissage automatique repose principalement sur trois éléments clés : une base de données, un descripteur d'environnement atomique, et une méthode de régression appropriée. Les descripteurs, sous forme de tenseurs covariants, représentent les environnements locaux des atomes dans la base de données de manière mathématique, adaptée à l'apprentissage automatique. Cela permet de transformer les configurations spatiales en ensembles de données exploitables. La régression linéaire constitue la dernière composante essentielle dans le développement du modèle par tenseur de moment. Dans le troisième chapitre, nous avons d'abord élargi la base de données développée dans le chapitre 2 pour y intégrer des données sur la silice et l'oxygène. Ensuite,

le potentiel unifié par tenseur de moment a été paramétrisé conjointement pour le silicium, la silice et l'oxygène. Cette approche a permis de tester avec succès la capacité du descripteur du modèle par tenseur de moment à prendre en compte le transfert de charge ainsi que trois entités disjointes (Si, O, SiO₂) dans l'espace configurationnel. Le potentiel unifié implémenté a été évalué avec succès sur des cas n'ayant pas été inclus dans l'ensemble de données de paramétrisation. Les résultats des tests, réalisés à l'aide de calculs statiques et de simulations de dynamique moléculaire, montrent une bonne concordance avec les données expérimentales. De plus, ils rivalisent avec les modèles semi-empiriques pour certaines propriétés et les surpassent dans la prédiction d'autres caractéristiques. Les bases de données générées dans les chapitres 2 et 3 peuvent enrichir d'autres projets de recherche, tout en fournissant à la communauté un potentiel unifié pour le système (Si, O, SiO₂). De plus, suite à la publication des travaux sur ce potentiel unifié, la communauté bénéficie d'une expertise et de techniques en modélisation conjointe, qui pourraient être appliquées à d'autres matériaux importants. Pour évaluer plus en profondeur la qualité de notre potentiel unifié, le quatrième chapitre examine le couplage de ce potentiel avec la technique d'activation et de relaxation (Activation Relaxation Techniques - ART) afin de modéliser le silicium amorphe. Cette approche a permis de générer des configurations de silicium amorphe de haute qualité, dont les propriétés et les caractéristiques structurales sont comparables aux résultats expérimentaux, tout en surpassant celles des modèles de silicium amorphe issus des méthodes traditionnelles. Ce couplage innovant souligne l'efficacité et la pertinence de notre potentiel unifié dans la simulation de matériaux complexes. Ce test sur le silicium amorphe apportera également une contribution précieuse à la communauté scientifique en explorant la possibilité de coupler les potentiels d'apprentissage automatique avec des méthodes telles que ART ou FEAR (Force Enhanced Atomic Refinement). Cette synergie pourrait révolutionner notre capacité à développer des modèles réalistes des matériaux amorphes et à approfondir notre compréhension de leurs propriétés. Le succès de la description conjointe du système disjoint (Si, SiO₂, O) par le modèle unifié a ouvert la voie à la conception et à la génération d'une base

de données pour l'eau, présentée dans le cinquième chapitre. Cette nouvelle base de données constitue une extension significative de celle dédiée au système (Si, SiO₂, O), en intégrant l'hydrogène. Cette avancée représente une étape cruciale dans notre exploration des matériaux complexes et de leurs propriétés. Enfin, dans le cinquième chapitre, nous présentons un résumé de la consommation des ressources computationnelles de l'Alliance Canada. Cette analyse permet d'évaluer l'efficacité des ressources utilisées et d'identifier les domaines d'amélioration potentiels, contribuant ainsi à une meilleure gestion des infrastructures computationnelles dans le cadre de nos recherches. En conclusion, un composé et ses constituants élémentaires peuvent être décrits conjointement par un modèle d'interaction basé sur l'apprentissage automatique. De manière globale, il ressort que les coordonnées de réaction chimiques sont suffisantes pour capturer les transferts de charges, et qu'un choix judicieux du nombre de paramètres du modèle peut atténuer les problèmes liés à l'espace configurationnel disjoint. Cependant, cette approche peut engendrer une complexité computationnelle accrue, nécessitant un équilibre entre précision et efficacité dans le développement des modèles.

RECOMMANDATIONS

Cette section propose des suggestions pour les futures orientations de recherche, fondées sur les résultats et les limites identifiées dans cette thèse.

Compréhension théorique des phénomènes à modéliser et construction de modèles réalistes

En plus des techniques de conception de bases de données développées dans cette thèse, celles-ci peuvent être améliorées pour des projets futurs. Par exemple, une revue de la littérature bien planifiée pourrait être réalisée pour mieux comprendre la théorie physico-chimique sous-jacente à chaque phénomène à modéliser ou à inclure dans la base de données. Cela inclut, sans s'y limiter, les défauts, la diffusion de défauts, ainsi que la formation de phases désordonnées, liquides ou amorphes. Il est également impératif de construire un modèle réaliste du phénomène afin qu'il puisse être correctement simulé ou calculé à l'aide de la théorie de la fonctionnelle de la densité (DFT). Rappelons que l'un des critères essentiels pour garantir la qualité des données en mécanique quantique est de construire un modèle physico-chimique qui reflète le phénomène réel. Chaque configuration ou entrée dans la base de doit être refléter un phénomène réel. Par exemple, pour inclure une configuration de surface orientée (111) dans la base de données, il est nécessaire de construire un modèle qui reflète cette surface selon l'orientation spécifiée dans la réalité.

Amélioration de la méthode d'échantillonnage de la base de données

Il est essentiel de couvrir au maximum l'espace de configuration, ce qui s'avère difficile à réaliser avec les méthodes standard d'échantillonnage de bases de données. Il serait judicieux d'explorer d'autres méthodes alternatives, telles que l'apprentissage actif. En se concentrant sur les données les plus informatives, l'apprentissage actif (referred to in English as active learning) permet d'atteindre de meilleures performances du modèle avec moins d'instances étiquetées, par

rapport aux méthodes d'apprentissage traditionnelles. Cette approche pourrait considérablement améliorer l'efficacité et la pertinence des modèles développés.

Modélisation à petite échelle

La taille des modèles à inclure dans la base de données peut être optimisée. L'objectif est de développer un modèle d'interaction généraliste, ce qui nécessite une couverture maximale de l'espace configurationnel et donc un grand nombre de configurations. Toutefois, il est également crucial de limiter la taille des modèles afin de minimiser le coût des simulations par la théorie de la fonctionnelle de la densité (DFT). L'optimisation de la taille des modèles facilitera l'entraînement et permettra d'améliorer l'efficacité globale. Par exemple, nous pouvons choisir des modèles de boîte contenant entre 2 et 12 atomes pour des matériaux de structures simple comme le silicium ou la silice. Cependant, cela dépend fortement du niveau de complexité de la structure. Il existe une limite à la simplification possible. Par exemple, si l'on considère la montmorillonite, le minéral principal de la bentonite, l'unité cellulaire contient 192 atomes, ce qui réduit considérablement les possibilités de modélisation avec des boîtes de simulation de petite taille. La construction de modèles représentatifs devrait donc se faire avec des modèles de taille moyenne ou grande. Ainsi, la modélisation à petite échelle pourrait constituer une solution viable pour développer un modèle généraliste moins coûteux et plus performant.

Entraînement du modèle et hyperparamètres

Une des tâches les plus difficiles dans le développement de modèles d'interaction est l'entraînement du modèle. Cela nécessite un ensemble de données bien organisé et optimisé, des rayons de coupure R_c appropriés, ainsi qu'un nombre d'itérations adéquat. Il est également impératif de choisir un nombre optimal de cœurs pour un processeur (*CPU : Central Processing Unit*) pour garantir l'efficacité de l'entraînement, étant donné que les codes ne sont pas infiniment

parallélisables. Une planification minutieuse de ces paramètres contribuera à améliorer les routines d'entraînement, ainsi que les performances et la précision du modèle.

Formalisme mathématique du modèle d'interaction

Une autre préoccupation importante est de pouvoir décrire des espaces configurationnels disjoints tout en tenant compte simultanément des transferts de charge. En plus des techniques recommandées précédemment pour améliorer la description des environnements atomiques variés, il peut être pertinent de revoir le formalisme mathématique du modèle. On pourrait se demander si une modification de ce formalisme permettrait de décrire plus aisément les espaces configurationnels disjoints ainsi que les transferts de charge de manière simultanée. Cette réflexion pourrait ouvrir de nouvelles perspectives pour la modélisation des interactions complexes.

Augmentation du niveau de complexité du système modéliser

Rappelons que, dans cette thèse, le silicium, l'oxygène et la silice ont servi de prototypes pour comprendre le fonctionnement du modèle par tenseur de moment. Ce modèle d'apprentissage automatique est basé sur des descripteurs d'environnement atomique et sur une régression linéaire, sans considération de la structure électronique. Comme nous pouvons le constater, les résultats détaillés dans les chapitres précédents démontrent la capacité du potentiel par tenseur de moment à décrire un système disjoint tout en intégrant les transferts de charge. Ce système inclut la silice, le silicium et l'oxygène. Plus précisément, le modèle a d'abord été testé sur du silicium pur, avant d'incorporer de l'oxygène pour modéliser conjointement la silice, le silicium et l'oxygène. Pour ce faire, nous avons généré une base de données comprenant le silicium, l'oxygène, la silice et l'hétérosystème formé par le silicium et la silice. À présent, nous nous interrogeons sur la manière dont le modèle peut prendre en compte une entité supplémentaire. Pour cela, nous allons ajouter l'hydrogène au système (Si, O, SiO₂), afin de former un ensemble

(Si, O, SiO₂, H) qui permettra de modéliser le gel de silicice. Ce gel constitue un point de départ prometteur pour la modélisation des aluminosilicates, puis de la montmorillonite, minéral principal de la roche bentonite qui est prévue d'être utilisée dans le système de stockage souterrain des déchets nucléaire. Au niveau du gel, de nombreuses molécules et structures doivent être prises en compte pour concevoir la base de données, en fonction de la structure spécifique du gel. Nous aborderons principalement les questions suivantes :

1. **Formation du gel** : La première étape consiste à imiter la formation de défauts hydrogarnet par le retrait aléatoire d'atomes de silicium. Ensuite, les oxygènes non saturés sont complétés par des atomes d'hydrogène, formant ainsi des groupes silanol. Dans certains cas, la formation du gel peut débuter avec des monomères silicatés isolés de Si(OH)₄.
2. **Hydratation et déshydratation** : Nous examinerons les interactions entre le gel sec et l'eau.
3. **Diffusion** : Une attention particulière sera portée aux processus de diffusion, en considérant la circulation des molécules d'eau à travers les pores du gel.
4. **Dissolution de la silice** : Enfin, nous explorerons la possibilité de dissolution de la silice, qui se produit par la diffusion des molécules SiO₄ dans l'eau.
5. **Plusieurs réactions chimiques à considérer** : La formation de la molécule d'eau par la réaction chimique $2\text{H}_2 + \text{O}_2 \rightarrow 2\text{H}_2\text{O}$, ainsi que l'interaction hydrogène-hydrogène, doivent être prises en compte pour assurer la qualité globale du potentiel d'interaction.

En tenant compte des points 1 à 5, l'ajout de l'hydrogène au système (Si, O, SiO₂) doit se faire en intégrant divers scénarios de données. Voici les étapes proposées :

- **Optimisation de la molécule d'hydrogène** : Commencer par optimiser la structure de la molécule d'hydrogène, suivie d'une variation de la longueur de liaison H-H.
- **Optimisation de la molécule d'eau** : Procéder ensuite à l'optimisation de la molécule d'eau, en variant l'angle et la longueur de la liaison O-H.

- **Formation de la molécule d'eau** : Placer un atome d'oxygène et deux atomes d'hydrogène dans une boîte, puis équilibrer le système pour permettre la formation de la molécule d'eau en fonction de la température.
- **Modèles de configurations d'eau** : Équilibrer des modèles contenant de l'eau sous ses phases liquide et solide à différentes températures.
- **Modèles de silice amorphe contenant des groupes silanol** : Construire et équilibrer de petits modèles de silice amorphe contenant des atomes d'hydrogène, supprimer quelques atomes de silicium et ajouter de l'hydrogène pour saturer les oxygènes vacants. Il est également nécessaire d'introduire des molécules d'eau dans les configurations de silice amorphe contenant des groupes silanol.
- **Ajout d'autres types de molécules** : Il est également important d'intégrer d'autres molécules, telles que le peroxyde d'hydrogène (H_2O_2), le silane (SiH_4), l'acide orthosilicique ($\text{Si}(\text{OH})_4$) et l'acide pyrosilicique ($\text{H}_6\text{Si}_2\text{O}_7$). Les deux dernières molécules correspondent respectivement aux monomères silicatés (Q_0) et aux dimères silicatés (Q_1), où Q_n est défini comme le nombre d'oxygènes de pont associés à un atome de silicium.
- **Modèle de Q_n avec de l'eau** : On peut équilibrer des modèles comprenant soit des monomères silicatés en présence d'eau, soit des dimères silicatés contenant de l'eau.

DÉPÔT DES DONNÉES ATOMIQUES

Les données atomiques générées dans le cadre des travaux de cette thèse sont stockées en ligne. Nous avons choisi **Borealis**, un site de dépôt de données de recherche en ligne. Borealis est un dépôt Dataverse canadien soutenu par les bibliothèques universitaires et les établissements de recherche à travers tout le Canada. Les données sont séparées en deux ensembles dans Borealis. Le premier ensemble de données porte sur le silicium, la silice et l'oxygène, tandis que le deuxième ensemble concerne l'eau.

Base de données de configurations atomiques sur le silicium et la silice

Cet ensemble contient les données relatives au silicium, à la silice et à l'oxygène. Il est accessible via le DOI suivant :

<https://doi.org/10.5683/SP3/RQZCVS>

Base de données de configurations atomiques sur l'eau

Cet ensemble héberge les données concernant l'eau. Il est accessible via le DOI suivant :

<https://doi.org/10.5683/SP3/FDCKFY>

Pour accéder aux données, il suffit de suivre les DOI appropriés mentionnés ci-dessus.

ANNEXE I

ARTICLE DE CONFÉRENCE

1. La consommation des ressources de calcul au sein de l'Alliance de recherche numérique du Canada

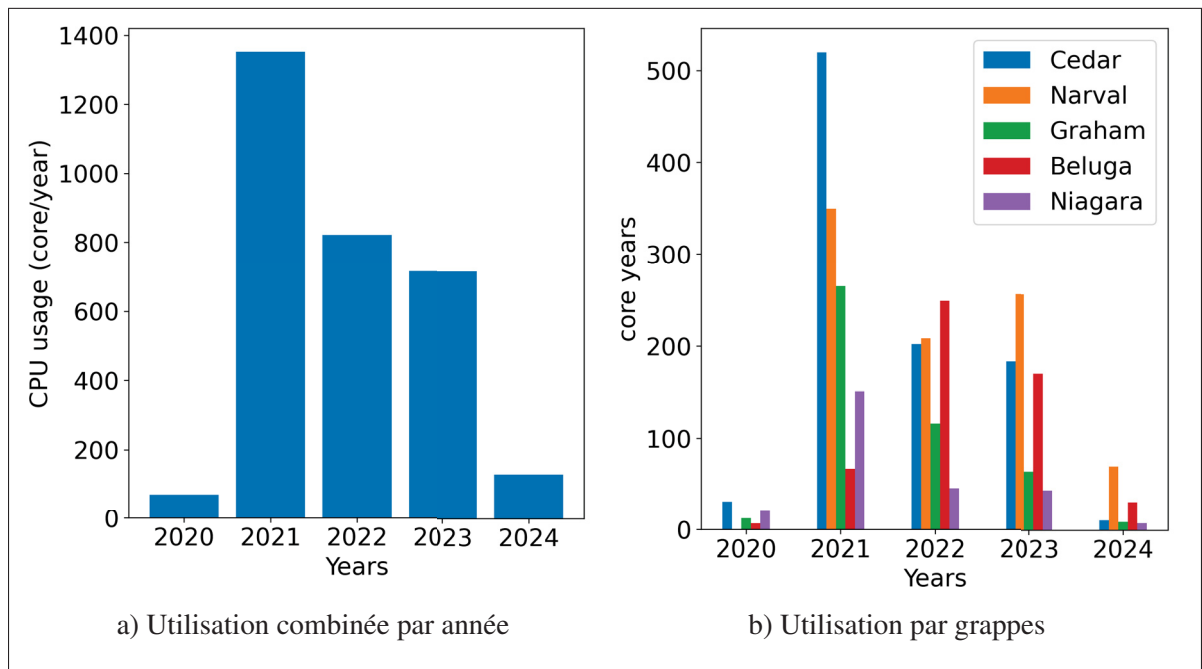


Figure-A I-1 Consommation des ressources pour nos projets de modélisation sur cinq années civiles

2. Conférence CNS/CNA en ligne

Cette section présente l'article rédigé dans le cadre de ma participation à la conférence de la Société nucléaire canadienne (CNS). Cet article a été préparé dans le cadre d'une étude préliminaire portant sur l'évaluation du potentiel par tenseur de moment appliqué à une base de données en silicium. Les figures I-1 à I-5 reproduisent les cinq pages de l'article présenté lors de cette conférence. L'article est disponible à l'adresse suivante : <https://www.xcdsystem.com/cns/proceedings/prof34444.html>

40th Annual Conference of the Canadian Nuclear Society and 45th Annual CNS/CNA Student Conference
Virtual Conference, June 6 – June 9, 2021

Improving atom-scale models of clay minerals using machine learning

K. Zongo¹, L.K. Béland², C. Ouellet-Plamondon³

^{1,3}Ecole de Technologie Supérieure, Montréal, Quebec, Canada
karim.zongo.2@ens.etsmtl.ca

²Queen's University, Kingston, Ontario, Canada

A Ph.D. Level Submission

Abstract. Bentonite clay is a geomaterial with numerous interesting physicochemical properties. Notably, it shows promise as a buffer material for underground long-term storage of radioactive waste. The material is predominantly composed of smectite montmorillonite, as well as considerable quantities of other minerals such as quartz, illite, feldspars, and others. Understanding this material at the atomic scale can help understand how minute changes in the environment or the potential release of radioisotopes would affect a spent fuel long-term repository. I will present a model to describe the interatomic interactions within the main components of bentonite clay via machine learning. Specifically, I will introduce the moment tensor potential (MTP), how it will be adapted to describe clay minerals, and present the first set of results, involving an interatomic model of silicon and silica, that constitute the backbone of clay minerals.

Keywords: Bentonite, clays, silicon, machine learning, force fields, moment tensor potential

1. Introduction

Bentonite clay will likely be used as a buffer material in long-term repositories of radioactive waste. Atomistic simulations of bentonite can help evaluate its performance when exposed to hostile underground conditions or radioactive isotopes potentially released from storage canisters. Deep geological repositories are designed to last for a very long time in the order of tens of thousands of years. This is well beyond the duration of any conceivable experiments. Nanoscale computer simulations can provide qualitative and quantitative insights into the mechanisms which could be of importance over these very long timescales that could be missed by macroscopic and microscopic scale experiments. These simulations can also help interpret experiments, including transport and sorption measurements, mechanical testing, and imaging-based on advanced microscopy and diffraction techniques. Current computer-based experiments fall into two categories: Quantum mechanical calculations and force field-based approach. For a given atomic or molecular system both calculate the total energy based on

Figure-A I-2 Article présenté à la Conférence CNS/CNA 2021

different approximations. Quantum mechanical approximations such as the density functional theory (DFT) offer great accuracy and transferability, but they are computationally expensive and limited to small systems typically containing hundreds of atoms where the time scale for phenomenon or event is approximately hundreds of picoseconds[1, 2]. Semi-empirical force field such as CLAYFF can model systems of size beyond the limits of *ab initio* approaches, but they lack flexibility. Their scalability is at the cost of accuracy[1]. In other words, one method has spatiotemporal limitations, while the other suffers from transferability and accuracy limitations. To address these concerns, we propose a machine learning interatomic potential that will improve upon the existing atom-scale model of the clay systems. We employ a strategy based on the so-called moment tensor potential (MTP) model[3]. The MTP is a machine learning model that has shown great promise in materials such as metallic alloys[4], semiconductors[1], and water[5]. The model is not extended yet to complex, multi-component and multiphase systems such as clays. Thus, we adopt the development of the moment tensor potential for the clay system in a stepwise manner. This will allow us to test and validate the model at a simple level and to ensure the full understanding as well as handling of moment tensor descriptors mainly at a complex level. Consequently, as a starting point, this paper is devoted to silicon, one of the key elemental constituents of the clay system.

2. Computational methods

This section details respectively the moment tensor potential and the database preparation with density functional theory.

2.1. Moment tensor potential

The moment tensor potential (MTP) was first introduced in 2016 by Alexander Shapeev[3]. Like other MLIAP, the formalism is based on the partitioning property of potential energy surface (PES) within a certain medium mainly non-polarized or non-charged medium in which the PES can be approximated as the contribution of individual atomic energies given by their local environments. The MTP model linearly expands the potential energy $V_i(r)$ of an atom i into a complete set of basis functions $B_\beta(r)$.

$$V_i(r) = \sum_{\beta} c_{\beta} B_{\beta}(r) \quad (1)$$

The basis functions $B_{\beta}(r)$ are determined from the scalar tensorial contraction of the local atomic environment descriptors given as follow:

$$M_{\mu,\theta}(r) = \sum_j f_{\mu}(r_{ij}) \underbrace{r_{ij} \otimes \dots \otimes r_{ij}}_{v \text{ times}} \quad (2)$$

Figure-A I-3 Article présenté à la Conférence CNS/CNA 2021

$M_{\mu,\vartheta}(r)$ is a tensor of rank ϑ enclosing the radial distribution ($f_{\mu}(r_{ij})$) and the angular information ($r_{ij} \otimes \dots \otimes r_{ij}$) of the local environment of the central atom i .

In practice, the contraction of $M_{\mu,\vartheta}(r)$ to a scalar is formulated by $n \times n$ symmetric matrix β which off-diagonal elements indicates the number of dimensions to be contracted between tensors M_{lm} and M_{pq} . The indexes μ and ϑ enables the bounding of basis size for practical implementation and also allow the estimation of computational complexity and efficiency.

2.2. Database

Similarly, to the other machines learning interatomic potential, the fitting of the MTP model mainly relies on a database from the first-principles calculation[3, 6]. Thus, in this work database consisting of 1279 silicon structures was generated using quantum mechanical calculations in the framework of density functional theory. All the calculations were conducted using the Quantum Espresso package[7] and the projector augmented wave pseudopotential (PAW)[8] with the generalized gradient approximation (GGA) of Perdew, Burke, and Ernzerhof(PBE). Convergence criteria for geometry optimization including atomic relaxation were set to 10^{-5} eV and 10^{-3} eV / Å for energy and forces, respectively. Before calculation on each crystallographic phase, we performed a convergence test for both kinetic cut-off energy and k-point. Thus, a value of 884 eV was set as the kinetic cut-off energy for the expansion of the wave function into plane waves. Monkhorst-Pack grid[9] was employed for Brillouin zone sampling and various k-point were used for this database generation.

Most of the possible configurations and crystallographic phases in which the silicon may chemically exist were computed. We built most of these configurations based on experimental parameters and atomic coordinates found in the literature. We begin by computing the ground state for each crystallographic phase and then apply strain in 6 different modes as in[10]. Some configurations such as liquids, amorphs, and surfaces are taken from [1],[11], [12-14] respectively and computed with our own DFT code and parameters.

3. Results

This section summarizes the results obtained with preliminary testing of the MTP model on silicon. We investigated basic properties of diamond phase and generalized stacking fault one of the planar defects in crystalline material crucial for macroscopic behavior such as plasticity, creep, and fracture. Here we used the slab method[15] in two directions $[\bar{1}10]$ and $[11\bar{2}]$ representing respectively full dislocation and partial dislocation directions in the (111) glide plane of silicon. We consider a supercell consisting of 36 atoms. Unrelaxed gamma surfaces for both displacement directions are presented in figure 1 and figure 2. First, our DFT results are similar to that of the literature [16]. Second, the MTP model shows great performance in reproducing the DFT results.

	DFT	MTP	Relative error
a	5.4690	5.4699	-0.016%
C_{11}	153.4478	149.6989	2.44%
C_{12}	56.7133	58.1083	-2.46%
C_{44}	76.8221	71.5311	6.89%
B	88.8406	88.63	0.24%
γ_{us}	1.71977	1.6421	4.52%
γ_{in}	2.23438	2.07109	9.49%

Table 1: Properties comparison between DFT and MTP

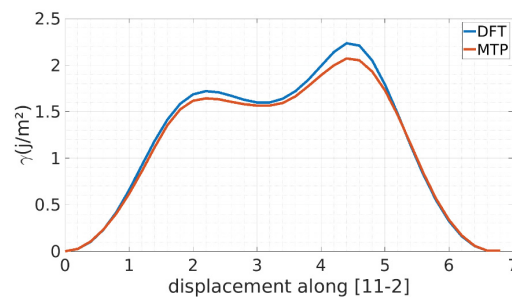


Figure 1: Generalized stacking fault curve

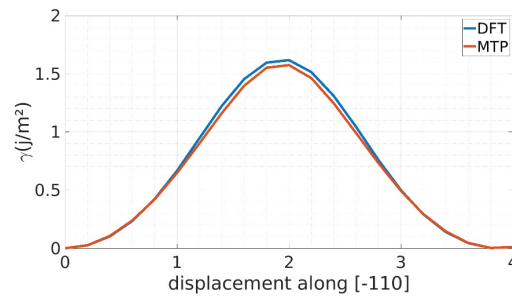


Figure 2: Stacking fault energy curve

4. Conclusion

We have successfully implemented an MTP-based model of Si. It shows good performance in reproducing DFT energies, forces, stresses as well as related properties such as lattice constant, Bulk modulus, elastic constant and stacking fault energies.

5. References

1. Sellin, P. and O.X. Leupin, *The use of clay as an engineered barrier in radioactive-waste management—a review*. Clays and Clay Minerals, 2013. **61**(6): p. 477-498.
2. Ma, Z., et al., *Review of application of molecular dynamic simulations in geological high-level radioactive waste disposal*. Applied Clay Science, 2019. **168**: p. 436-449.
3. Zuo, Y., et al., *Performance and cost assessment of machine learning interatomic potentials*. The Journal of Physical Chemistry A, 2020. **124**(4): p. 731-745.
4. Shapeev, A.V., *Moment tensor potentials: A class of systematically improvable interatomic potentials*. Multiscale Modeling & Simulation, 2016. **14**(3): p. 1153-1173.
5. Gubaev, K., et al., *Accelerating high-throughput searches for new alloys with active learning of interatomic potentials*. Computational Materials Science, 2019. **156**: p. 148-156.
6. Lomaka, A. and T. Tamm, *Linearization of moment tensor potentials for multicomponent systems with a preliminary assessment for short-range interaction energy in water dimer and trimer*. The Journal of chemical physics, 2020. **152**(16): p. 164115.
7. Novikov, I.S., et al., *The MLIP package: Moment Tensor Potentials with MPI and Active Learning*. Machine Learning: Science and Technology, 2020.
8. Giannozzi, P., et al., *QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials*. Journal of physics: Condensed matter, 2009. **21**(39): p. 395502.
9. Blöchl, P.E., *Projector augmented-wave method*. Physical review B, 1994. **50**(24): p. 17953.
10. Monkhorst, H.J. and J.D. Pack, *Special points for Brillouin-zone integrations*. Physical review B, 1976. **13**(12): p. 5188.
11. Tritsaris, G.A., et al., *Diffusion of lithium in bulk amorphous silicon: a theoretical study*. The Journal of Physical Chemistry C, 2012. **116**(42): p. 22212-22216.
12. Tran, R., et al., *Surface energies of elemental crystals*. Scientific data, 2016. **3**(1): p. 1-13.
13. Bartók, A.P., et al., *Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons*. Physical review letters, 2010. **104**(13): p. 136403.

Figure-A I-6 Article présenté à la Conférence CNS/CNA 2021

BIBLIOGRAPHIE

- Abascal, J. L. & Vega, C. (2005). A general purpose model for the condensed phases of water : TIP4P/2005. *The Journal of chemical physics*, 123(23).
- Abdolhosseini Qomi, M., Krakowiak, K., Bauchy, M., Stewart, K., Shahsavari, R., Jagannathan, D., Brommer, D. B., Baronnet, A., Buehler, M. J., Yip, S. et al. (2014). Combinatorial molecular optimization of cement hydrates. *Nature communications*, 5(1), 4960.
- Adams, G. B., O’Keeffe, M., Demkov, A. A., Sankey, O. F. & Huang, Y.-M. (1994). Wide-band-gap Si in open fourfold-coordinated clathrate structures. *Physical Review B*, 49(12), 8048.
- Aggarwal, P., Singh, R. P. & Aggarwal, Y. (2015). Use of nano-silica in cement based materials—A review. *Cogent Engineering*, 2(1), 1078018.
- Ahmed, A. (2019). *Étude des propriétés physico-chimiques de nouveaux supports argileux modifiés par des liquides ioniques dans la perspective d’applications au traitement de déchets toxiques*. (Thèse de doctorat, Université de Picardie Jules Verne).
- Alfè, D. & Gillan, M. (2003). Exchange-correlation energy and the phase diagram of Si. *Physical Review B*, 68(20), 205212.
- Allen, A. E., Dusson, G., Ortner, C. & Csányi, G. (2021). Atomic permutationally invariant polynomials for fitting molecular force fields. *Machine Learning : Science and Technology*, 2(2), 025017.
- Allen, M. P. et al. (2004). Introduction to molecular dynamics simulation. *Computational soft matter : from synthetic polymers to proteins*, 23(1), 1–28.
- Alliance Canada. [Accessed : YYYY-MM-DD]. (2023). Technical Documentation. Repéré à https://docs.alliancecan.ca/wiki/Technical_documentation.
- Anderson, B., Hy, T. S. & Kondor, R. (2019). Cormorant : Covariant molecular neural networks. *Advances in neural information processing systems*, 32.
- Annad, M., Lefkir, A., Mammar-Kouadri, M. & Bettahar, I. (2021). Development of a local scour prediction model clustered by soil class. *Water Practice & Technology*, 16(4), 1159–1172.
- Arai, N., Takeda, S. & Kohyama, M. (1997). Self-interstitial clustering in crystalline silicon. *Physical review letters*, 78(22), 4265.

- Arns, R. G. (1998). The other transistor : early history of the metal-oxide semiconductor field-effect transistor. *Engineering Science & Education Journal*, 7(5), 233–240.
- Artioli, G., Lamberti, C. & Marra, G. (2000). Neutron powder diffraction study of orthorhombic and monoclinic defective silicalite. *Acta Crystallographica Section B : Structural Science*, 56(1), 2–10.
- Artrith, N. & Urban, A. (2016). An implementation of artificial neural-network potentials for atomistic materials simulations : Performance for TiO₂. *Computational Materials Science*, 114, 135–150.
- Artrith, N., Morawietz, T. & Behler, J. (2011). High-dimensional neural-network potentials for multicomponent systems : Applications to zinc oxide. *Physical Review B*, 83(15), 153101.
- Ashby, M. F. (1992). Physical modelling of materials problems. *Materials Science and Technology*, 8(2), 102–111.
- Babaei, H., Guo, R., Hashemi, A. & Lee, S. (2019). Machine-learning-based interatomic potential for phonon transport in perfect crystalline Si and crystalline Si with vacancies. *Physical Review Materials*, 3(7), 074603.
- Balamane, H., Halicioglu, T. & Tiller, W. (1992). Comparative study of silicon empirical interatomic potentials. *Physical Review B*, 46(4), 2250.
- Balyakin, I., Rempel, S., Ryltsev, R. & Rempel, A. (2020). Deep machine learning interatomic potential for liquid silica. *Physical Review E*, 102(5), 052125.
- Barkema, G. T. & Mousseau, N. (2000). High-quality continuous random networks. *Physical Review B*, 62(8), 4985.
- Barkema, G. & Mousseau, N. (1996). Event-based relaxation of continuous disordered systems. *Physical review letters*, 77(21), 4358.
- Barth, T. (1932). Cristobalite structures ; II, Low-cristobalite. *American Journal of Science*, 24(140), 97–110.
- Bartok, A. P. (2010). *The Gaussian Approximation Potential : an interatomic potential derived from first principles quantum mechanics*. Springer Science & Business Media.
- Bartók, A. P., Kondor, R. & Csányi, G. (2013). On representing chemical environments. *Physical Review B—Condensed Matter and Materials Physics*, 87(18), 184115.

- Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. (2018). Machine learning a general-purpose interatomic potential for silicon. *Physical Review X*, 8(4), 041048.
- Bartók, A. P., Kondor, R. & Csányi, G. (2013). On representing chemical environments. *Physical Review B*, 87(18), 184115.
- Baskes, M. (1987). Application of the embedded-atom method to covalent materials : a semiempirical potential for silicon. *Physical review letters*, 59(23), 2666.
- Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csányi, G. (2022). MACE : Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35, 11423–11436.
- Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E. & Kozinsky, B. (2022). E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1), 2453.
- Bauchy, M., Abdolhosseini Qomi, M. J., Bichara, C., Ulm, F.-J. & Pellenq, R. J.-M. (2014). Nanoscale structure of cement : viewpoint of rigidity theory. *The Journal of Physical Chemistry C*, 118(23), 12485–12493.
- Bauza, D. (2001). Thermal oxidation of silicon and Si–SiO₂ interface morphology, structure, and localized states. *Handbook of surfaces and interfaces of materials*, 1, 115–216.
- Becquart, C. & Perez, M. (2016). Dynamique moléculaire appliquée aux matériaux 2010. *Visité le*, 26(11).
- Behler, J. (2021). Four generations of high-dimensional neural network potentials. *Chemical Reviews*, 121(16), 10037–10072.
- Behler, J. (2011). Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7).
- Behler, J. (2016). Perspective : Machine learning potentials for atomistic simulations. *The Journal of chemical physics*, 145(17), 170901.
- Behler, J. (2014). Representing potential energy surfaces by high-dimensional neural network potentials. *Journal of Physics : Condensed Matter*, 26(18), 183001.
- Behler, J. & Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14), 146401.

- Béland, L. K. & Mousseau, N. (2013). Long-time relaxation of ion-bombarded silicon studied with the kinetic activation-relaxation technique : Microscopic description of slow aging in a disordered system. *Physical Review B*, 88(21), 214201.
- Béland, L. K., Brommer, P., El-Mellouhi, F., Joly, J.-F. & Mousseau, N. (2011). Kinetic activation-relaxation technique. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 84(4), 046704.
- Béland, L. K., Anahory, Y., Smeets, D., Guihard, M., Brommer, P., Joly, J.-F., Pothier, J.-C., Lewis, L. J., Mousseau, N. & Schiettekatte, F. (2013). Replenish and relax : Explaining logarithmic annealing in ion-implanted c-Si. *Physical review letters*, 111(10), 105502.
- Béland, L. K., Machado-Charry, E., Pochet, P. & Mousseau, N. (2014). Strain effects and intermixing at the Si surface : Importance of long-range elastic corrections in first-principles calculations. *Physical Review B*, 90(15), 155302.
- Béland, L. K., Osetsky, Y. N., Stoller, R. E. & Xu, H. (2015). Slow relaxation of cascade-induced defects in Fe. *Physical Review B*, 91(5), 054108.
- Bereau, T., DiStasio, R. A., Tkatchenko, A. & Von Lilienfeld, O. A. (2018). Non-covalent interactions across organic and biological subsets of chemical space : Physics-based potentials parametrized from machine learning. *The Journal of chemical physics*, 148(24).
- Berendsen, H. J., Postma, J. v., Van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8), 3684–3690.
- Berendsen, H. J., Grigera, J.-R. & Straatsma, T. P. (1987). The missing term in effective pair potentials. *Journal of Physical Chemistry*, 91(24), 6269–6271.
- Berghout, A., Tunega, D. & Zaoui, A. (2010). Density functional theory (DFT) study of the hydration steps of Na⁺/Mg²⁺/Ca²⁺/Sr²⁺/Ba²⁺-exchanged montmorillonites. *Clays and Clay Minerals*, 58(2), 174-187.
- Billeter, S. R., Curioni, A., Fischer, D. & Andreoni, W. (2006). Ab initio derived augmented Tersoff potential for silicon oxynitride compounds and their interfaces with silicon. *Physical Review B—Condensed Matter and Materials Physics*, 73(15), 155329.
- Birney, R., Steinlechner, J., Tornasi, Z., MacFoy, S., Vine, D., Bell, A., Gibson, D., Hough, J., Rowan, S., Sortais, P. et al. (2018). Amorphous silicon with extremely low absorption : Beating thermal noise in gravitational astronomy. *Physical review letters*, 121(19), 191101.

- Biswas, P., Atta-Fynn, R., Drabold, D. et al. (2004). The inclusion of experimental information in first principles modelling of materials. *Journal of Physics : Condensed Matter*, 16(44), S5173.
- Biswas, R. & Hamann, D. (1985). Interatomic potentials for silicon structural energies. *Physical review letters*, 55(19), 2001.
- Blank, T. B., Brown, S. D., Calhoun, A. W. & Doren, D. J. (1995). Neural network models of potential energy surfaces. *The Journal of chemical physics*, 103(10), 4129–4137.
- Blöchl, P. E. (1994). Projector augmented-wave method. *Physical review B*, 50(24), 17953.
- Bofill, J. M. (2003). Remarks on the updated Hessian matrix methods. *International journal of quantum chemistry*, 94(6), 324–332.
- Borisenko, K. B., Haberl, B., Liu, A. C., Chen, Y., Li, G., Williams, J. S., Bradby, J. E., Cockayne, D. J. & Treacy, M. M. (2012). Medium-range order in amorphous silicon investigated by constrained structural relaxation of two-body and four-body electron diffraction data. *Acta Materialia*, 60(1), 359–375.
- Born, M. & Heisenberg, W. (1985). Zur quantentheorie der molekeln. *Original Scientific Papers Wissenschaftliche Originalarbeiten*, 216–246.
- Botu, V., Batra, R., Chapman, J. & Ramprasad, R. (2017). Machine learning force fields : construction, validation, and outlook. *The Journal of Physical Chemistry C*, 121(1), 511-522.
- Brandes, E. A. & Brook, G. (2013). *Smithells metals reference book*. Elsevier.
- Brigatti, M. F., Galan, E. & Theng, B. (2006). Structures and mineralogy of clay minerals. *Developments in clay science*, 1, 19–86.
- Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1), 76–90.
- Buehler, M. J., Van Duin, A. C. & Goddard III, W. A. (2006). Multiparadigm Modeling of Dynamical Crack Propagation in Silicon Using a Reactive Force Field. *Physical review letters*, 96(9), 095505.
- Bussi, G., Donadio, D. & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1).

- Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. (2018a). Machine learning for molecular and materials science. *Nature*, 559(7715), 547-555.
- Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. (2018b). Machine learning for molecular and materials science. *Nature*, 559(7715), 547-555.
- Cai, J. (1999). New Simple Analytical Many-Body Potential for Covalent Materials. *physica status solidi (b)*, 212(1), 9–18.
- Cannon, J. F. (1974). Behavior of the elements at high pressures. *Journal of Physical and Chemical Reference Data*, 3(3), 781–824.
- Canterakis, N. (1999). 3D Zernike Moments and Zernike Affine Invariants for 3D Image Analysis and Recognition. Repéré à <https://api.semanticscholar.org/CorpusID:8307002>.
- Car, R. & Parrinello, M. (1985a). Unified approach for molecular dynamics and density-functional theory. *Physical review letters*, 55(22), 2471.
- Car, R. & Parrinello, M. (1985b). Unified approach for molecular dynamics and density-functional theory. *Physical review letters*, 55(22), 2471.
- Carlson, D. E. & Wronski, C. R. (1976). Amorphous silicon solar cell. *Applied Physics Letters*, 28(11), 671–673.
- Caro, M. A. (2019). Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials. *Physical Review B*, 100(2), 024112.
- Carré, A., Ispas, S., Horbach, J. & Kob, W. (2016). Developing empirical potentials from ab initio simulations : The case of amorphous silica. *Computational Materials Science*, 124, 323–334.
- Cauchy, A. et al. (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847), 536–538.
- Cellai, D., Carpenter, M., Kirkpatrick, R., Salje, E. & Zhang, M. (1995). Thermally induced phase transitions in tridymite : an infrared spectroscopy study. *Physics and Chemistry of Minerals*, 22, 50–60.
- Chen, C. & Ong, S. P. (2022). A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11), 718–728.

- Chen, M., Ko, H.-Y., Remsing, R. C., Calegari Andrade, M. F., Santra, B., Sun, Z., Selloni, A., Car, R., Klein, M. L., Perdew, J. P. et al. (2017). Ab initio theory and modeling of water. *Proceedings of the National Academy of Sciences*, 114(41), 10846–10851.
- Cheng, Y., Zhu, L., Wang, G., Zhou, J., Elliott, S. R. & Sun, Z. (2020). Vacancy formation energy and its connection with bonding environment in solid : A high-throughput calculation and machine learning study. *Computational Materials Science*, 183, 109803.
- Christensen, A. S., Bratholm, L. A., Faber, F. A. & Anatole von Lilienfeld, O. (2020). FCHL revisited : Faster and more accurate quantum machine learning. *The Journal of chemical physics*, 152(4).
- Christidis, G. E. (2011). The concept of layer charge of smectites and its implications for important smectite-water properties.
- Cliffe, M. J., Bartók, A. P., Kerber, R. N., Grey, C. P., Csányi, G. & Goodwin, A. L. (2017). Structural simplicity as a restraint on the structure of amorphous silicon. *Physical Review B*, 95(22), 224108.
- Clouet, E. (2020). Ab initio models of dislocations. *Handbook of Materials Modeling : Methods : Theory and Modeling*, 1503–1524.
- Coombs, P., De Natale, J., Hood, P., McElfresh, D., Wortman, R. & Shackelford, J. (1985). The nature of the Si-O-Si bond angle distribution in vitreous silica. *Philosophical Magazine B*, 51(4), L39–L42.
- Coulombe, A. (2023). *Étude des comportements de sorption et de la capacité des argiles sensibles de l'Est du Canada à adsorber et séquestrer les métaux lourds et hydrocarbures aromatiques diméthylés*. (Thèse de doctorat, Université du Québec à Rimouski).
- Crain, J., Clark, S., Ackland, G., Payne, M., Milman, V., Hatton, P. & Reid, B. (1994). Theoretical study of high-density phases of covalent semiconductors. I. Ab initio treatment. *Physical Review B*, 49(8), 5329.
- Cui, L.-F., Ruffo, R., Chan, C. K., Peng, H. & Cui, Y. (2009). Crystalline-amorphous core-shell silicon nanowires for high capacity and high current battery electrodes. *Nano letters*, 9(1), 491–495.
- Curtarolo, S., Setyawan, W., Hart, G. L., Jahnatek, M., Chepulskii, R. V., Taylor, R. H., Wang, S., Xue, J., Yang, K., Levy, O. et al. (2012). AFLOW : An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58, 218–226.

- Custer, J. S., Thompson, M. O., Jacobson, D., Poate, J., Roorda, S., Sinke, W. & Spaepen, F. (1994). Density of amorphous Si. *Applied physics letters*, 64(4), 437–439.
- Cvitkovich, L., Waldhör, D., El-Sayed, A.-M., Jech, M., Wilhelmer, C. & Grasser, T. (2022). Computational Modeling of Initial Stage Si (100) Surface Oxidation. *arXiv preprint arXiv :2204.13954*.
- Cvitkovich, L., Waldhör, D., El-Sayed, A.-M., Jech, M., Wilhelmer, C. & Grasser, T. (2023). Dynamic modeling of Si (100) thermal oxidation : Oxidation mechanisms and realistic amorphous interface generation. *Applied Surface Science*, 610, 155378.
- Cygan, R. T., Kubicki, J. & Bleam, W. (1998). Molecular models of metal sorption on clay minerals. *Molecular modeling of clays and mineral surfaces*, 12.
- Cygan, R. T., Liang, J.-J. & Kalinichev, A. G. (2004). Molecular models of hydroxide, oxyhydroxide, and clay phases and the development of a general force field. *The Journal of Physical Chemistry B*, 108(4), 1255-1266.
- Cygan, R. T., Greathouse, J. A., Heinz, H. & Kalinichev, A. G. (2009). Molecular models and simulations of layered materials. *Journal of Materials Chemistry*, 19(17), 2470-2481.
- Da Silva, J., Pinatti, D., Anderson, C. & Rudee, M. (1975). A refinement of the structure of vitreous silica. *The Philosophical Magazine : A Journal of Theoretical Experimental and Applied Physics*, 31(3), 713–717.
- Das Sarma, S., Deng, D.-L. & Duan, L.-M. (2019). Machine learning meets quantum physics. *Physics Today*, 72(3), 48–54.
- Dasgupta, S., Lambros, E., Perdew, J. P. & Paesani, F. (2021). Elevating density functional theory to chemical accuracy for water simulations through a density-corrected many-body formalism. *Nature communications*, 12(1), 6359.
- Davidovits, J. (2017). Geopolymers : Ceramic-like inorganic polymers. *J. Ceram. Sci. Technol*, 8(3), 335–350.
- Daw, M. S. & Baskes, M. I. (1983). Semiempirical, quantum mechanical calculation of hydrogen embrittlement in metals. *Physical review letters*, 50(17), 1285.
- Daw, M. S. & Baskes, M. I. (1984). Embedded-atom method : Derivation and application to impurities, surfaces, and other defects in metals. *Physical Review B*, 29(12), 6443.

- De Jong, M., Chen, W., Angsten, T., Jain, A., Notestine, R., Gamst, A., Sluiter, M., Ande, C. K., Van Der Zwaag, S. & Plata, J. J. (2015). Charting the complete elastic properties of inorganic crystalline compounds. *Scientific data*, 2(1), 1-13.
- Debenedetti, P. G. & Stanley, H. E. (2003). Supercooled and glassy water. *Physics Today*, 56(6), 40–46.
- Della Pia, F., Zen, A., Alfè, D. & Michaelides, A. (2022). DMC-ICE13 : Ambient and high pressure polymorphs of ice from diffusion Monte Carlo and density functional theory. *The Journal of Chemical Physics*, 157(13).
- Deng, B., Zhong, P., Jun, K., Riebesell, J., Han, K., Bartel, C. J. & Ceder, G. (2023). CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9), 1031–1041.
- Denteneer, P. & Van Haeringen, W. (1987). Stacking-fault energies in semiconductors from first-principles calculations. *Journal of Physics C : Solid State Physics*, 20(32), L883.
- Dera, P., Prewitt, C. T., Boctor, N. Z. & Hemley, R. J. (2002). Characterization of a high-pressure phase of silica from the Martian meteorite Shergotty. *American Mineralogist*, 87(7), 1018–1023.
- Deringer, V. L., Bernstein, N., Bartók, A. P., Cliffe, M. J., Kerber, R. N., Marbella, L. E., Grey, C. P., Elliott, S. R. & Csányi, G. (2018a). Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics. *The journal of physical chemistry letters*, 9(11), 2879–2885.
- Deringer, V. L., Bernstein, N., Bartók, A. P., Cliffe, M. J., Kerber, R. N., Marbella, L. E., Grey, C. P., Elliott, S. R. & Csányi, G. (2018b). Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics. *The journal of physical chemistry letters*, 9(11), 2879–2885.
- Deringer, V. L., Caro, M. A. & Csányi, G. (2019). Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46), 1902765.
- Deringer, V. L., Caro, M. A. & Csányi, G. (2020). A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nature communications*, 11(1), 5461.
- Deringer, V. L., Bernstein, N., Csányi, G., Ben Mahmoud, C., Ceriotti, M., Wilson, M., Drabold, D. A. & Elliott, S. R. (2021). Origins of structural and electronic transitions in disordered silicon. *Nature*, 589(7840), 59–64.

- Di Mare, M., Inumberable, N., Brisebois, P. P. & Ouellet-Plamondon, C. M. (2022). Combined experimental and computational prediction of the piezoresistivity of alkali-activated inorganic polymers. *The Journal of Physical Chemistry C*, 126(35), 14995–15000.
- Díaz-Cabañas, M.-J. & Barrett, P. A. (1998). Synthesis and structure of pure SiO₂ chabazite : the SiO₂ polymorph with the lowest framework density. *Chemical Communications*, (17), 1881–1882.
- Ding, J.-T. & Li, Z. (2002). Effects of metakaolin and silica fume on properties of concrete. *Materials Journal*, 99(4), 393–398.
- Dodson, B. W. (1987). Development of a many-body Tersoff-type potential for silicon. *Physical Review B*, 35(6), 2795.
- Dorner, F., Sukurma, Z., Dellago, C. & Kresse, G. (2018). Melting Si : beyond density functional theory. *Physical Review Letters*, 121(19), 195701.
- Dove, M. T. (2008). An introduction to atomistic simulation methods. *Seminarios de la SEM*, 4, 7–37.
- Downs, R. T. & Hall-Wallace, M. (2003). The American Mineralogist crystal structure database. *American Mineralogist*, 88(1), 247–250.
- Downs, R. & Palmer, D. (1994). The pressure behavior of α cristobalite. *American Mineralogist*, 79(1-2), 9–14.
- Dragoni, D., Daff, T. D., Csányi, G. & Marzari, N. (2018). Achieving DFT accuracy with a machine-learning interatomic potential : Thermomechanics and defects in bcc ferromagnetic iron. *Physical Review Materials*, 2(1), 013808.
- Drautz, R. (2019). Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1), 014104.
- Du, Y. A., Barr, S. A., Hazzard, K. R., Lenosky, T. J., Hennig, R. G. & Wilkins, J. W. (2005). Fast diffusion mechanism of silicon tri-interstitial defects. *Physical Review B*, 72(24), 241306.
- Dufresne, A., Arayro, J., Zhou, T., Ioannidou, K., Ulm, F.-J., Pellenq, R. & Béland, L. K. (2018). Atomistic and mesoscale simulation of sodium and potassium adsorption in cement paste. *The Journal of Chemical Physics*, 149(7), 074705.
- Durandurdu, M. & Drabold, D. (2001). Ab initio simulation of first-order amorphous-to-amorphous phase transition of silicon. *Physical Review B*, 64(1), 014101.

- Ebina, T., Iwasaki, T., Onodera, Y. & Chatterjee, A. (1999). A comparative study of DFT and XPS with reference to the adsorption of caesium ions in smectites. *Computational Materials Science*, 14(1-4), 254-260.
- Ebrahimi, D. (2013). *Multiscale modeling of clay-water systems*. (Thesis, Massachusetts Institute of Technology).
- Ebrahimi, D., Pellenq, R. J.-M. & Whittle, A. J. (2016). Mesoscale simulation of clay aggregate formation and mechanical properties. *Granular Matter*, 18(3), 49.
- El-Mellouhi, F. & Mousseau, N. (2007). Ab initio characterization of arsenic vacancy diffusion pathways in GaAs with SIEST-A-RT. *Applied Physics A*, 86, 309–312.
- El-Mellouhi, F., Mousseau, N. & Ordejón, P. (2004). Sampling the diffusion paths of a neutral vacancy in silicon with quantum mechanical calculations. *Physical Review B*, 70(20), 205202.
- Eliassen, S. N., Friis, J., Ringdalen, I. G., Mousseau, N., Trochet, M. & Li, Y. (2019). Atomistic approach to simulate kink migration and kink-pair formation in silicon : The kinetic activation-relaxation technique. *Physical Review B*, 100(15), 155305.
- Erhard, L. C., Rohrer, J., Albe, K. & Deringer, V. L. (2022). A machine-learned interatomic potential for silica and its relation to empirical models. *npj Computational Materials*, 8(1), 90.
- Escamilla-Roa, E., Huertas, F. J., Hernández-Laguna, A. & Sainz-Díaz, C. I. (2017). A DFT study of the adsorption of glycine in the interlayer space of montmorillonite. *Physical Chemistry Chemical Physics*, 19(23), 14961-14971.
- Esposito, S., Marocco, A., Dell'Agli, G., De Gennaro, B. & Pansini, M. (2015). Relationships between the water content of zeolites and their cation population. *Microporous and Mesoporous Materials*, 202, 36–43.
- Evans, D. J. & Holian, B. L. (1985). The nose–hoover thermostat. *The Journal of chemical physics*, 83(8), 4069–4074.
- Evrard, G. & Pusztai, L. (2005). Reverse Monte Carlo modelling of the structure of disordered materials with RMC++ : a new implementation of the algorithm in C++. *Journal of Physics : Condensed Matter*, 17(5), S1.
- Faber, F., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. (2015). Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry*, 115(16), 1094–1101.

- Fan, Z., Zeng, Z., Zhang, C., Wang, Y., Song, K., Dong, H., Chen, Y. & Ala-Nissila, T. (2021). Neuroevolution machine learning potentials : Combining high accuracy and low cost in atomistic simulations and application to heat transport. *Physical Review B*, 104(10), 104309.
- Farnan, I., Grandinetti, P., Baltisberger, J., Stebbins, J., Werner, U., Eastman, M. & Pines, A. (1992). Quantification of the disorder in network-modified silicate glasses. *Nature*, 358(6381), 31–35.
- Fecko, C., Eaves, J., Loparo, J., Tokmakoff, A. & Geissler, P. (2003). Ultrafast hydrogen-bond dynamics in the infrared spectroscopy of water. *Science*, 301(5640), 1698–1702.
- Feng, X.-L. (2003). Molecular dynamics simulation of thermal conductivity of nanoscale thin silicon films. *Microscale thermophysical engineering*, 7(2), 153–161.
- Fermeglia, M. (2008). Multiscale molecular modeling : a tool for the design of nano structured materials. *Computer Aided Chemical Engineering*, 25. doi : 10.1016/S1570-7946(08)80126-5.
- Finzi, M., Stanton, S., Izmailov, P. & Wilson, A. G. (2020). Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. *International Conference on Machine Learning*, pp. 3165–3176.
- Fischer, C. F. (1977). Hartree–Fock method for atoms. A numerical approach.
- Fischer, D., Curioni, A., Billeter, S. & Andreoni, W. (2006). The structure of the SiO₂/ Si (100) interface from a restraint-free search using computer simulations. *Applied physics letters*, 88(1).
- Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, 13(3), 317–322.
- Flores, Y. C., Cordeiro, G. C., Toledo Filho, R. D. & Tavares, L. M. (2017). Performance of Portland cement pastes containing nano-silica and different types of silica. *Construction and Building Materials*, 146, 524–530.
- Fortner, J. & Lannin, J. (1989). Radial distribution functions of amorphous silicon. *Physical Review B*, 39(8), 5527.
- France-Lanord, A., Soukiassian, P., Glattli, C. & Wimmer, E. (2016). Ab initio parameterization of a charge optimized many-body forcefield for Si–SiO₂ : Validation and thermal transport in nanostructures. *The Journal of Chemical Physics*, 144(10).

- Freysoldt, C., Grabowski, B., Hickel, T., Neugebauer, J., Kresse, G., Janotti, A. & Van de Walle, C. G. (2014). First-principles calculations for point defects in solids. *Reviews of Modern Physics*, 86(1), 253.
- Friederich, P., Häse, F., Proppe, J. & Aspuru-Guzik, A. (2021). Machine-learned potentials for next-generation matter simulations. *Nature Materials*, 20(6), 750–761.
- Fuchs, F., Worrall, D., Fischer, V. & Welling, M. (2020). Se (3)-transformers : 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33, 1970–1981.
- Furness, J. W., Kaplan, A. D., Ning, J., Perdew, J. P. & Sun, J. (2020). Accurate and numerically efficient r2SCAN meta-generalized gradient approximation. *The journal of physical chemistry letters*, 11(19), 8208–8215.
- Gadzhiev, O. B., Ignatov, S. K., Kulikov, M. Y., Feigin, A. M., Razuvaev, A. G., Sennikov, P. G. & Schrems, O. (2013). Structure, Energy, and Vibrational Frequencies of Oxygen Allotropes O_n (n = 6) in the Covalently Bound and van der Waals Forms : Ab Initio Study at the CCSD (T) Level. *Journal of chemical theory and computation*, 9(1), 247–262.
- Galamboš, M., Kufčáková, J., Roskopfová, O. & Rajec, P. (2010). Adsorption of cesium and strontium on natrified bentonites. *Journal of Radioanalytical and Nuclear Chemistry*, 283(3), 803–813.
- Galamboš, M., Kufčáková, J., Roskopfová, O. & Rajec, P. (2010). Adsorption of cesium and strontium on natrified bentonites. *Journal of Radioanalytical and Nuclear Chemistry*, 283(3), 803–813.
- Ganster, P., Trégliá, G., Lancon, F. & Pochet, P. (2010a). Molecular dynamics simulation of silicon oxidization. *Thin Solid Films*, 518(9), 2422–2426.
- Ganster, P., Trégliá, G. & Saúl, A. (2010b). Atomistic modeling of strain and diffusion at the Si/SiO₂ interface. *Physical Review B—Condensed Matter and Materials Physics*, 81(4), 045315.
- Ganster, P., Béland, L. K. & Mousseau, N. (2012). First stages of silicon oxidation with the activation relaxation technique. *Physical Review B*, 86(7), 075408.
- Gartner III, T. E., Zhang, L., Piaggi, P. M., Car, R., Panagiotopoulos, A. Z. & Debenedetti, P. G. (2020). Signatures of a liquid–liquid transition in an ab initio deep neural network model for water. *Proceedings of the National Academy of Sciences*, 117(42), 26040–26046.

- Gastegger, M., Schwiedrzik, L., Bittermann, M., Berzsenyi, F. & Marquetand, P. (2018). wACSF—Weighted atom-centered symmetry functions as descriptors in machine learning potentials. *The Journal of chemical physics*, 148(24).
- Gasteiger, J., Giri, S., Margraf, J. T. & Günnemann, S. (2020a). Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv :2011.14115*.
- Gasteiger, J., Groß, J. & Günnemann, S. (2020b). Directional message passing for molecular graphs. *arXiv preprint arXiv :2003.03123*.
- Gasteiger, J., Becker, F. & Günnemann, S. (2021). Gemnet : Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34, 6790–6802.
- Gereben, O., Jóvári, P., Temleitner, L. & Pusztai, L. (2007). A new version of the RMC++ Reverse Monte Carlo programme, aimed at investigating the structure of covalent glasses. *Journal of Optoelectronics and Advanced Materials*, 9(10), 3021.
- Geysermans, P., Gorse, D. & Pontikis, V. (2000). Molecular dynamics study of the solid–liquid interface. *The Journal of Chemical Physics*, 113(15), 6382–6389.
- Ghayaza, M. (2012). *Réactivité argiles-polluants métalliques : simulation des barrières argileuses des sites de stockage des déchets*. (Thèse de doctorat, Université d’Orléans).
- Ghoufi, A., Malfreyt, P. & Tildesley, D. J. (2016). Computer modelling of the surface tension of the gas–liquid and liquid–liquid interface. *Chemical Society Reviews*, 45(5), 1387–1409.
- Giannozzi, P., Baroni, S., Bonini, N., Calandra, M., Car, R., Cavazzoni, C., Ceresoli, D., Chiarotti, G. L., Cococcioni, M. & Dabo, I. (2009). QUANTUM ESPRESSO : a modular and open-source software project for quantum simulations of materials. *Journal of physics : Condensed matter*, 21(39), 395502.
- Gibbs, R. E. (1926). The polymorphism of silicon dioxide and the structure of tridymite. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 113(764), 351–368.
- Gillan, M. J., Alfe, D. & Michaelides, A. (2016). Perspective : How good is DFT for water? *The Journal of chemical physics*, 144(13).
- Godwin, J., Schaarschmidt, M., Gaunt, A., Sanchez-Gonzalez, A., Rubanova, Y., Veličković, P., Kirkpatrick, J. & Battaglia, P. (2021). Simple gnn regularisation for 3d molecular property prediction & beyond. *arXiv preprint arXiv :2106.07971*.

- Goedecker, S. (1999). Linear scaling electronic structure methods. *Reviews of Modern Physics*, 71(4), 1085.
- Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109), 23–26.
- González, M. (2011). Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique*, 12, 169-200.
- Gottardi, G. & Galli, E. (2012). *Natural zeolites*. Springer Science & Business Media.
- Goyal, A., Li, Y., Chernatynskiy, A., Jayashankar, J. S., Kautzky, M. C., Sinnott, S. B. & Phillpot, S. R. (2021). The influence of alloying on the stacking fault energy of gold from density functional theory calculations. *Computational Materials Science*, 188, 110236.
- Greathouse, J. A., Hart, D. B., Bowers, G. M., Kirkpatrick, R. J. & Cygan, R. T. (2015). Molecular simulation of structure and diffusion at smectite–water interfaces : Using expanded clay interlayers as model nanopores. *The Journal of Physical Chemistry C*, 119(30), 17126-17136.
- Grimley, D. I., Wright, A. C. & Sinclair, R. N. (1990). Neutron scattering from vitreous silica IV. Time-of-flight diffraction. *Journal of Non-Crystalline Solids*, 119(1), 49–64.
- Grocholski, B., Shim, S.-H. & Prakapenka, V. (2013). Stability, metastability, and elastic properties of a dense silica polymorph, seifertite. *Journal of Geophysical Research : Solid Earth*, 118(9), 4745–4757.
- Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. (2018). Machine learning of molecular properties : Locality and active learning. *The Journal of chemical physics*, 148(24), 241727.
- Gubaev, K., Podryabinkin, E. V., Hart, G. L. & Shapeev, A. V. (2019). Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. *Computational Materials Science*, 156, 148-156.
- Guénolé, J., Godet, J. & Pizzagalli, L. (2010). Determination of activation parameters for the core transformation of the screw dislocation in silicon. *Modelling and Simulation in Materials Science and Engineering*, 18(6), 065001.
- Guerette, M. & Huang, L. (2012). A simple and convenient set-up for high-temperature Brillouin light scattering. *Journal of Physics D : Applied Physics*, 45(27), 275302.

- Gunde, M., Jay, A., Poberžnik, M., Salles, N., Richard, N., Landa, G., Mousseau, N., Martin-Samos, L. & Hemeryck, A. (2024). Exploring Potential Energy Surfaces to Reach Saddle Points above Convex Regions. *The Journal of Chemical Physics*, 160(23), 232501. doi : 10.1063/5.0210097.
- Gurman, S. & McGreevy, R. (1990). Reverse Monte Carlo simulation for the analysis of EXAFS data. *Journal of Physics : Condensed Matter*, 2(48), 9463.
- Haberl, B., Liu, A. C. Y., Bradby, J. E., Ruffell, S., Williams, J. S. & Munroe, P. (2009). Structural characterization of pressure-induced amorphous silicon. *Physical Review B—Condensed Matter and Materials Physics*, 79(15), 155209.
- Haghighatlari, M., Li, J., Guan, X., Zhang, O., Das, A., Stein, C. J., Heidar-Zadeh, F., Liu, M., Head-Gordon, M., Bertels, L. et al. (2022). NewtonNet : a Newtonian message passing network for deep learning of interatomic potentials and forces. *Digital Discovery*, 1(3), 333–343.
- Hansson, T., Oostenbrink, C. & van Gunsteren, W. (2002). Molecular dynamics simulations. *Current opinion in structural biology*, 12(2), 190–196.
- Hautier, G., Fischer, C., Ehrlicher, V., Jain, A. & Ceder, G. (2011). Data mined ionic substitutions for the discovery of new compounds. *Inorganic chemistry*, 50(2), 656-663.
- Hayoun, M. (2002). La méthode de monte carlo metropolis. *École «Simulation Numérique en Matière Condensée» (Paris, 29-31 mai 2002)*.
- Head-Gordon, T. & Hura, G. (2002). Water structure from scattering experiments and simulation. *Chemical reviews*, 102(8), 2651–2670.
- Henkelman, G., Uberuaga, B. P. & Jónsson, H. (2000). A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics*, 113(22), 9901–9904.
- Hicks, D., Mehl, M. J., Gossett, E., Toher, C., Levy, O., Hanson, R. M., Hart, G. & Curtarolo, S. (2019). The AFLOW library of crystallographic prototypes : part 2. *Computational Materials Science*, 161, S1–S1011.
- Himpfel, F., McFeely, F., Taleb-Ibrahimi, A., Yarmoff, J. & Hollinger, G. (1988). Microscopic structure of the SiO₂/Si interface. *Physical review B*, 38(9), 6084.
- Hine, N. D., Haynes, P. D., Mostofi, A. A., Skylaris, C.-K. & Payne, M. C. (2009). Linear-scaling density-functional theory with tens of thousands of atoms : Expanding the scope and scale of calculations with ONETEP. *Computer Physics Communications*, 180(7), 1041–1053.

- Hirel, P. (2015). AtomsK : A tool for manipulating and converting atomic data files. *Computer Physics Communications*, 197, 212–219.
- Hohenberg, P. & Kohn, W. (1964). Density functional theory (DFT). *Phys. Rev*, 136(1964), B864.
- Holmström, E., Haberl, B., Pakarinen, O. H., Nordlund, K., Djurabekova, F., Arenal, R., Williams, J. S., Bradby, J. E., Petersen, T. C. & Liu, A. (2016). Dependence of short and intermediate-range order on preparation in experimental and modeled pure a-Si. *Journal of Non-Crystalline Solids*, 438, 26–36.
- Horn, R. A. & Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.
- Hover, K. C. (2011). The influence of water on the performance of concrete. *Construction and Building Materials*, 25(7), 3003–3013.
- Hu, J. Z., Merkle, L. D., Menoni, C. S. & Spain, I. L. (1986). Crystal data for high-pressure phases of silicon. *Physical Review B*, 34(7), 4679.
- Hu, L., Su, R., Huang, B. & Liu, F. (2019). An Accurate and Transferable Machine-Learning Interatomic Potential for Silicon. *arXiv preprint arXiv :1901.01638*.
- Huang, X., Hu, Y.-J. & An, Q. (2021). Locking of Screw Dislocations in Silicon due to Core Structure Transformation. *The Journal of Physical Chemistry C*, 125(44), 24710–24718.
- Hünenberger, P. H. (2005). Thermostat algorithms for molecular dynamics simulations. *Advanced computer simulation : Approaches for soft matter sciences I*, 105–149.
- Huo, H. & Rupp, M. (2022). Unified representation of molecules and crystals for machine learning. *Machine Learning : Science and Technology*, 3(4), 045017.
- Hwang, G. S. & Goddard III, W. A. (2002). Diffusion and dissociation of neutral divacancies in crystalline silicon. *Physical Review B*, 65(23), 233205.
- Igram, D., Bhattarai, B., Biswas, P. & Drabold, D. A. (2018). Large and realistic models of amorphous silicon. *Journal of Non-Crystalline Solids*, 492, 27–32.
- Imbalzano, G., Anelli, A., Giofré, D., Klees, S., Behler, J. & Ceriotti, M. (2018). Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *The Journal of chemical physics*, 148(24).
- Indiamart. [Accessed : 2024-09-05]. (2024). Natural Silica (Silicon Dioxide). Repéré à <https://www.indiamart.com/proddetail/natural-silica-silicon-dioxide-7861071473.htm>.

- Ioannidou, K., Carrier, B., Vandamme, M. & Pellenq, R. (2017). The potential of mean force concept for bridging (length and time) scales in the modeling of complex porous materials. *EPJ Web of Conferences*, 140, 01009.
- Ishimaru, M., Munetoh, S. & Motooka, T. (1997). Generation of amorphous silicon structures by rapid quenching : A molecular-dynamics study. *Physical Review B*, 56(23), 15133.
- Jaccodine, R. (1963). Surface energy of germanium and silicon. *Journal of the electrochemical society*, 110(6), 524.
- Jain, A. & McGaughey, A. J. (2015). Effect of exchange–correlation on first-principles-driven lattice thermal conductivity predictions of crystalline silicon. *Computational Materials Science*, 110, 115–120.
- Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G. et al. (2013). Commentary : The Materials Project : A materials genome approach to accelerating materials innovation. *APL materials*, 1(1).
- Jay, A., Huet, C., Salles, N., Gunde, M., Martin-Samos, L., Richard, N., Landa, G., Goiffon, V., de Gironcoli, S., Hemeryck, A. & Mousseau, N. (2020). Finding Reaction Pathways and Transition States : R-ARTn and d-ARTn as an Efficient and Versatile Alternative to String Approaches. *Journal of Chemical Theory and Computation*, 16(10), 6726–6734. doi : 10.1021/acs.jctc.0c00541.
- Jay, A., Gunde, M., Salles, N., Poberžnik, M., Martin-Samos, L., Richard, N., de Gironcoli, S., Mousseau, N. & Hémercyck, A. (2022). Activation–Relaxation Technique : An efficient way to find minima and saddle points of potential energy surfaces. *Computational Materials Science*, 209, 111363.
- Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4), 620.
- Jeanmairet, G. (2014). *Une théorie de la fonctionnelle de la densité moléculaire pour la solvation dans l'eau*. (Thesis).
- Joly, J.-F., Béland, L. K., Brommer, P., El-Mellouhi, F. & Mousseau, N. (2012). Optimization of the kinetic activation-relaxation technique, an off-lattice and self-learning kinetic monte-carlo method. *Journal of Physics : Conference Series*, 341(1), 012007.
- Joly, J.-F., Béland, L. K., Brommer, P. & Mousseau, N. (2013). Contribution of vacancies to relaxation in amorphous materials : A kinetic activation-relaxation technique study. *Physical Review B—Condensed Matter and Materials Physics*, 87(14), 144204.

- Juan, Y.-M. & Kaxiras, E. (1996). Generalized stacking fault energy surfaces and dislocation properties of silicon : a first-principles theoretical study. *Philosophical Magazine A*, 74(6), 1367-1384.
- Justo, J. F., Bazant, M. Z., Kaxiras, E., Bulatov, V. V. & Yip, S. (1998). Interatomic potential for silicon defects and disordered phases. *Physical Review B*, 58(5), 2539.
- Kail, F., Farjas, J., Roura, P., Secouard, C., Nos, O., Bertomeu, J. & Cabarrocas, P. R. i. (2011). The configurational energy gap between amorphous and crystalline silicon. *physica status solidi (RRL)–Rapid Research Letters*, 5(10-11), 361–363.
- Kallel, H., Mousseau, N. & Schiettekatte, F. (2010). Evolution of the potential-energy surface of amorphous silicon. *Physical Review Letters*, 105(4), 045503.
- Kandy, A. K. A., Rossi, K., Raulin-Foissac, A., Laurens, G. & Lam, J. (2023). Comparing transferability in neural network approaches and linear models for machine-learning interaction potentials. *Physical Review B*, 107(17), 174106.
- Karakasidis, T. E. & Charitidis, C. (2007). Multiscale modeling in nanomaterials science. *Materials Science and Engineering : C*, 27(5-8), 1082-1089.
- Karlina, A. I., Karlina, Y. I. & Gladkikh, V. A. (2023). Analysis of Experience in the Use of Micro-and Nanoadditives from Silicon Production Waste in Concrete Technologies. *Minerals*, 13(12), 1525.
- Kasper, J. & Richards, S. (1964). The crystal structures of new forms of silicon and germanium. *Acta Crystallographica*, 17(6), 752–755.
- Ke, Q., Gong, X., Liao, S., Duan, C. & Li, L. (2022). Effects of thermostats/barostats on physical properties of liquids by molecular dynamics simulations. *Journal of Molecular Liquids*, 365, 120116.
- Keen, D. A. (2001). A comparison of various commonly used correlation functions for describing total scattering. *Journal of Applied Crystallography*, 34(2), 172–177.
- Kellouai, W. (2022). *Adsorption and Transport in Zeolites : Molecular Simulation and Neutron Scattering Experiments*. (Thèse de doctorat, Université Grenoble Alpes [2020-....]).
- Kermode, J. R., Cereda, S., Tangney, P. & De Vita, A. (2010). A first principles based polarizable O (N) interatomic force field for bulk silica. *The Journal of chemical physics*, 133(9).

- Key, B., Bhattacharyya, R., Morcrette, M., Seznec, V., Tarascon, J.-M. & Grey, C. P. (2009). Real-time NMR investigations of structural changes in silicon electrodes for lithium-ion batteries. *Journal of the American Chemical Society*, 131(26), 9239–9249.
- Kholmurodov, K., Abasheva, M. & Yasuoka, K. (2010). Molecular dynamics simulations of valinomycin interactions with potassium and sodium ions in water solvent. *Advances in Bioscience and Biotechnology*, 1(3), 145-240.
- Khorshidi, A. & Peterson, A. A. (2016). Amp : A modular approach to machine learning in atomistic simulations. *Computer Physics Communications*, 207, 310-324.
- Khouchaf, L., Boulahya, K., Das, P. P., Nicolopoulos, S., Kis, V. K. & Lábár, J. L. (2020). Study of the microstructure of amorphous silica nanostructures using high-resolution electron microscopy, electron energy loss spectroscopy, X-ray powder diffraction, and electron pair distribution function. *Materials*, 13(19), 4393.
- Kim, D. Y., Stefanoski, S., Kurakevych, O. O. & Strobel, T. A. (2015). Synthesis of an open-framework allotrope of silicon. *Nature Materials*, 14(2), 169–173.
- Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. (2021). A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nature communications*, 12(1), 398.
- Kobayashi, K., Nagai, Y., Itakura, M. & Shiga, M. (2021). Self-learning hybrid Monte Carlo method for isothermal–isobaric ensemble : Application to liquid silica. *The Journal of Chemical Physics*, 155(3).
- Kobayashi, K., Okumura, M., Nakamura, H., Itakura, M., Machida, M. & Cooper, M. W. (2022). Machine learning molecular dynamics simulations toward exploration of high-temperature properties of nuclear fuel materials : Case study of thorium dioxide. *Scientific reports*, 12(1), 9808.
- Kohn, W. & Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A), A1133.
- Kokalj, A. (1999). XCrySDen—a new program for displaying crystalline structures and electron densities. *Journal of Molecular Graphics and Modelling*, 17(3-4), 176–179.
- Kumagai, T., Izumi, S., Hara, S. & Sakai, S. (2007). Development of bond-order potentials that can reproduce the elastic constants and melting point of silicon for classical molecular dynamics simulation. *Computational materials science*, 39(2), 457–464.

- Laaziri, K., Kycia, S., Roorda, S., Chicoine, M., Robertson, J., Wang, J. & Moss, S. (1999a). High resolution radial distribution function of pure amorphous silicon. *Physical review letters*, 82(17), 3460.
- Laaziri, K., Kycia, S., Roorda, S., Chicoine, M., Robertson, J., Wang, J. & Moss, S. (1999b). High-energy x-ray diffraction study of pure amorphous silicon. *Physical Review B*, 60(19), 13520.
- Lager, G. A., Jorgensen, J. & Rotella, F. (1982). Crystal structure and thermal expansion of α -quartz SiO₂ at low temperatures. *Journal of Applied Physics*, 53(10), 6751–6756.
- Lambros, E., Hu, J. & Paesani, F. (2021). Assessing the accuracy of the scan functional for water through a many-body analysis of the adiabatic connection formula. *Journal of Chemical Theory and Computation*, 17(6), 3739–3749.
- Landau, D. & Binder, K. (2021). *A guide to Monte Carlo simulations in statistical physics*. Cambridge university press.
- Larentzos, J. P., Greathouse, J. A. & Cygan, R. T. (2007). An ab initio and classical molecular dynamics investigation of the structural and vibrational properties of talc and pyrophyllite. *The Journal of Physical Chemistry C*, 111(34), 12752–12759.
- Laskowski, L., Laskowska, M., Vila, N., Schabikowski, M. & Walcarius, A. (2019). Mesoporous Silica-Based Materials for Electronics-Oriented Applications. *Molecules*, 24(13). Repéré à <https://www.mdpi.com/1420-3049/24/13/2395>.
- Lawrimore, W., Paliwal, B., Chandler, M., Johnson, K. & Horstemeyer, M. (2016). Hierarchical multiscale modeling of Polyvinyl Alcohol/Montmorillonite nanocomposites. *Polymer*, 99, 386–398.
- Le Roux, S. & Jund, P. (2010). Ring statistics analysis of topological networks : New approach and application to amorphous GeS₂ and SiO₂ systems. *Computational Materials Science*, 49(1), 70–83.
- Le Roux, S. & Petkov, V. (2010). ISAACS–interactive structure analysis of amorphous and crystalline systems. *Journal of Applied Crystallography*, 43(1), 181–185.
- Leardini, L., Quartieri, S., Vezzalini, G., Martucci, A. & Dmitriev, V. (2013). Elastic behavior and high pressure-induced phase transition in chabazite : New data from a natural sample from Nova Scotia. *Microporous and Mesoporous Materials*, 170, 52–61.
- Lee, B.-J. (2007). A modified embedded atom method interatomic potential for silicon. *Calphad*, 31(1), 95–104.

- Lee, E., Lee, K.-R., Baskes, M. & Lee, B.-J. (2016). A modified embedded-atom method interatomic potential for ionic systems : 2 nnmeam+ qeq. *Physical Review B*, 93(14), 144110.
- Léger, J.-M., Haines, J. & Chateau, C. (2001). The high-pressure behaviour of the “moganite” polymorph of SiO₂. *European Journal of Mineralogy*, 13(2), 351–359.
- Lenosky, T. J., Sadigh, B., Alonso, E., Bulatov, V. V., de la Rubia, T. D., Kim, J., Voter, A. F. & Kress, J. D. (2000). Highly optimized empirical potential model of silicon. *Modelling and Simulation in Materials Science and Engineering*, 8(6), 825.
- LeSar, R. (2013). *Introduction to Computational Materials Science : Fundamentals to Applications*. Cambridge University Press.
- Lévesque, C., Roorda, S., Schiettekatte, F. & Mousseau, N. (2022). Internal mechanical dissipation mechanisms in amorphous silicon. *Physical Review Materials*, 6(12), 123604.
- Levien, L. & Prewitt, C. T. (1981). High-pressure crystal structure and compressibility of coesite. *American Mineralogist*, 66(3-4), 324–333.
- Levien, L., Prewitt, C. T. & Weidner, D. J. (1980). Structure and elastic properties of quartz at pressure. *American Mineralogist*, 65(9-10), 920–930.
- Lewis, L. J. (2022). Fifty years of amorphous silicon models : the end of the story ? *Journal of Non-Crystalline Solids*, 580, 121383.
- Ley-Hernandez, A. M., Lapeyre, J., Cook, R., Kumar, A. & Feys, D. (2018). Elucidating the effect of water-to-cement ratio on the hydration mechanisms of cement. *ACS omega*, 3(5), 5092–5105.
- Li, C., Wang, C., Han, J., Yan, L., Deng, B. & Liu, X. (2018). A comprehensive study of the high-pressure–temperature phase diagram of silicon. *Journal of materials science*, 53(10), 7475–7485.
- Li, D., Bancroft, G., Kasrai, M., Fleet, M., Secco, R., Feng, X., Tan, K. & Yang, B. (1994). X-ray absorption spectroscopy of silicon dioxide (SiO₂) polymorphs : The structural characterization of opal. *American Mineralogist*, 79(7-8), 622–632.
- Li, H., Kang, T., Zhang, B., Zhang, J. & Ren, J. (2016). Influence of interlayer cations on structural properties of montmorillonites : A dispersion-corrected density functional theory study. *Computational Materials Science*, 117, 33–39.

- Li, R., Lee, E. & Luo, T. (2020). A unified deep neural network potential capable of predicting thermal conductivity of silicon in different phases. *Materials Today Physics*, 12, 100181.
- Li, W., Lu, S., Hu, Q.-M., Kwon, S. K., Johansson, B. & Vitos, L. (2014). Generalized stacking fault energies of alloys. *Journal of Physics : Condensed Matter*, 26(26), 265005.
- Li, W. & Ando, Y. (2018). Comparison of different machine learning models for the prediction of forces in copper and silicon dioxide. *Physical Chemistry Chemical Physics*, 20(47), 30006–30020.
- Li, W. & Ando, Y. (2019). Dependence of a cooling rate on structural and vibrational properties of amorphous silicon : A neural network potential-based molecular dynamics study. *The Journal of Chemical Physics*, 151(11).
- Li, X.-P., Chen, G., Allen, P. & Broughton, J. (1988). Energy and vibrational spectrum of the Si (111)(7×7) surface from empirical potentials. *Physical Review B*, 38(5), 3331.
- Liang, T., Devine, B., Phillpot, S. R. & Sinnott, S. B. (2012). Variable charge reactive potential for hydrocarbons to simulate organic-copper interactions. *The Journal of Physical Chemistry A*, 116(30), 7976–7991.
- Liang, Z., Evans, W. & Keblinski, P. (2013). Equilibrium and nonequilibrium molecular dynamics simulations of thermal conductance at solid-gas interfaces. *Physical Review E*, 87(2), 022119.
- Lide, D. R. (2004). *CRC handbook of chemistry and physics*. CRC press.
- Limbu, D. K., Atta-Fynn, R., Drabold, D. A., Elliott, S. R. & Biswas, P. (2018). Information-driven inverse approach to disordered solids : Applications to amorphous silicon. *Physical Review Materials*, 2(11), 115602.
- Liu, H., Fu, Z., Li, Y., Sabri, N. F. A. & Bauchy, M. (2019). Balance between accuracy and simplicity in empirical forcefields for glass modeling : insights from machine learning. *Journal of Non-Crystalline Solids*, 515, 133–142.
- Liu, Y., Wang, L., Liu, M., Zhang, X., Oztekin, B. & Ji, S. (2021a). Spherical message passing for 3d graph networks. *arXiv preprint arXiv :2102.05013*.
- Liu, Y., Zhao, T., Ju, W. & Shi, S. (2017). Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3), 159-177.

- Liu, Z., Lin, L., Jia, Q., Cheng, Z., Jiang, Y., Guo, Y. & Ma, J. (2021b). Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning. *Journal of chemical information and modeling*, 61(3), 1066–1082.
- Lomaka, A. & Tamm, T. (2020). Linearization of moment tensor potentials for multicomponent systems with a preliminary assessment for short-range interaction energy in water dimer and trimer. *The Journal of Chemical Physics*, 152(16), 164115.
- López-Zorrilla, J., Aretxabaleta, X. M., Yeu, I. W., Etxebarria, I., Manzano, H. & Artrith, N. (2023). $\text{\ae}$ net-PyTorch : a GPU-supported implementation for machine learning atomic potentials training. *The Journal of Chemical Physics*, 158(16).
- Lustig, R. (1994). Statistical thermodynamics in the classical molecular dynamics ensemble. I. Fundamentals. *The Journal of chemical physics*, 100(4), 3048–3059.
- Ma, Z., Gamage, R. P., Rathnaweera, T. & Kong, L. (2019). Review of application of molecular dynamic simulations in geological high-level radioactive waste disposal. *Applied Clay Science*, 168, 436–449.
- Machado-Charry, E., Béland, L. K., Caliste, D., Genovese, L., Deutsch, T., Mousseau, N. & Pochet, P. (2011). Optimized energy landscape exploration using the ab initio based activation-relaxation technique. *The Journal of Chemical Physics*, 135(3).
- Madanchi, A., Azek, E., Zongo, K., Béland, L. K., Mousseau, N. & Simine, L. (2024). Is the Future of Materials Amorphous ? Challenges and Opportunities in Simulations of Amorphous Materials. *arXiv preprint arXiv :2410.05035*.
- Madsen, F. T. & Müller-Vonmoos, M. (1989). The swelling behaviour of clays. *Applied Clay Science*, 4(2), 143–156.
- Maiti, A., Mahan, G. & Pantelides, S. (1997). Dynamical simulations of nonequilibrium processes—Heat flow and the Kapitza resistance across grain boundaries. *Solid State Communications*, 102(7), 517–521.
- Makov, G. & Payne, M. C. (1995). Periodic boundary conditions in ab initio calculations. *Physical Review B*, 51(7), 4014.
- Malek, R. & Mousseau, N. (2000). Dynamics of Lennard-Jones clusters : A characterization of the activation-relaxation technique. *Physical Review E*, 62(6), 7723.
- Malenkov, G. (2006). Structure and dynamics of liquid water. *Journal of Structural Chemistry*, 47, S1–S31.

- Mao, Q., Feng, M., Jiang, X. Z., Ren, Y., Luo, K. H. & van Duin, A. C. (2023). Classical and reactive molecular dynamics : Principles and applications in combustion and energy systems. *Progress in Energy and Combustion Science*, 97, 101084.
- Maras, E., Trushin, O., Stukowski, A., Ala-Nissila, T. & Jónsson, H. (2016). Global Transition Path Search for Dislocation Formation in Ge on Si(001). *Computer Physics Communications*, 205, 13–21. doi : 10.1016/j.cpc.2016.04.001.
- Mazloom, M., Ramezani pour, A. & Brooks, J. (2004). Effect of silica fume on mechanical properties of high-strength concrete. *Cement and Concrete Composites*, 26(4), 347–357.
- McGreevy, R. & Howe, M. (1992). RMC : modeling disordered structures. *Annual Review of Materials Science*, 22(1), 217–242.
- McGreevy, R. & Pusztai, L. (1988). Reverse Monte Carlo simulation : a new technique for the determination of disordered structures. *Molecular Simulation*, 1(6), 359–367.
- McGreevy, R. L. (2001). Reverse monte carlo modelling. *Journal of Physics : Condensed Matter*, 13(46), R877.
- McMahon, M., Nemes, R., Wright, N. & Allan, D. (1994). Pressure dependence of the Imma phase of silicon. *Physical Review B*, 50(2), 739.
- McSkimin, H., Andreatch Jr, P. & Thurston, R. (1965). Elastic moduli of quartz versus hydrostatic pressure at 25 and- 195.8 C. *Journal of Applied Physics*, 36(5), 1624–1632.
- Mehl, M. J., Hicks, D., Toher, C., Levy, O., Hanson, R. M., Hart, G. & Curtarolo, S. (2017). The AFLOW library of crystallographic prototypes : part 1. *Computational Materials Science*, 136, S1-S828.
- Mehta, A. & Ashish, D. K. (2020). Silica fume and waste glass in cement concrete production : A review. *Journal of Building Engineering*, 29, 100888.
- Mei, Q., Benmore, C. & Weber, J. (2007). Structure of liquid SiO₂ : a measurement by high-energy X-ray diffraction. *Physical Review Letters*, 98(5), 057802.
- Meidanshahi, R. V., Bowden, S. & Goodnick, S. M. (2019). Electronic structure and localized states in amorphous Si and hydrogenated amorphous Si. *Physical Chemistry Chemical Physics*, 21(24), 13248–13257.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087–1092.

- Miehe, G., Graetsch, H. et al. (1992). Crystal structure of moganite : a new structure type for silica. *European Journal of Mineralogy*, 4(4), 693–706.
- Mignon, P., Ugliengo, P., Sodupe, M. & Hernandez, E. R. (2010). Ab initio molecular dynamics study of the hydration of Li⁺, Na⁺ and K⁺ in a montmorillonite model. Influence of isomorphic substitution. *Physical Chemistry Chemical Physics*, 12(3), 688-697.
- Mishima, O. & Stanley, H. E. (1998). The relationship between liquid, supercooled and glassy water. *Nature*, 396(6709), 329–335.
- Mishin, Y. (2021). Machine-learning interatomic potentials for materials science. *Acta Materialia*, 214, 116980.
- Mishra, R. K., Mohamed, A. K., Geissbühler, D., Manzano, H., Jamil, T., Shahsavari, R., Kalinichev, A. G., Galmarini, S., Tao, L., Heinz, H. et al. (2017). cemff : A force field database for cementitious materials including validations, applications and opportunities. *Cement and Concrete Research*, 102, 68–89.
- Mohapatra, S., Ranjan, S., Dasgupta, N., Kumar, R. & Thomas, S. (2018). *Nanocarriers for Drug Delivery : Nanoscience and Nanotechnology in Drug Delivery*. Elsevier.
- Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.
- Momma, K. & Izumi, F. (2008). VESTA : a three-dimensional visualization system for electronic and structural analysis. *Journal of Applied Crystallography*, 41(3), 653–658.
- Monkhorst, H. J. & Pack, J. D. (1976). Special points for Brillouin-zone integrations. *Physical review B*, 13(12), 5188.
- Morris, J. R., Wang, C., Ho, K. & Chan, C. T. (1994). Melting line of aluminum from simulations of coexisting phases. *Physical Review B*, 49(5), 3109.
- Mortazavi, B., Podryabinkin, E. V., Roche, S., Rabczuk, T., Zhuang, X. & Shapeev, A. V. (2020). Machine-learning interatomic potentials enable first-principles multiscale modeling of lattice thermal conductivity in graphene/borophene heterostructures. *Materials Horizons*, 7(9), 2359-2367.
- Mousseau, N. & Barkema, G. T. (1999). Exploring high-dimensional energy landscapes. *Computing in Science & Engineering*, 1(2), 74–80.

- Mousseau, N. & Barkema, G. (1998). Traveling through potential energy landscapes of disordered materials : The activation-relaxation technique. *Physical Review E*, 57(2), 2419.
- Mousseau, N., Béland, L. K., Brommer, P., Joly, J.-F., El-Mellouhi, F., Machado-Charry, E., Marinica, M.-C. & Pochet, P. (2012). The activation-relaxation technique : Art nouveau and kinetic art. *Journal of Atomic and Molecular Physics*, 2012.
- Mozzi, R. & Warren, n. B. (1969). The structure of vitreous silica. *Journal of Applied Crystallography*, 2(4), 164–172.
- Mueller, T., Kusne, A. G. & Ramprasad, R. (2016). Machine learning in materials science : Recent progress and emerging applications. *Reviews in Computational Chemistry*, 29, 186-273.
- Mueller, T., Hernandez, A. & Wang, C. (2020). Machine learning for interatomic potential models. *The Journal of Chemical Physics*, 152(5), 050902.
- Munetoh, S., Motooka, T., Moriguchi, K. & Shintani, A. (2007). Interatomic potential for Si–O systems using Tersoff parameterization. *Computational Materials Science*, 39(2), 334–339.
- Murray, H. H. (2006). Bentonite applications. *Developments in Clay Science*, 2, 111–130.
- Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C. J., Kornbluth, M. & Kozinsky, B. (2023). Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1), 579.
- Musil, F., Grisafi, A., Bartók, A. P., Ortner, C., Csányi, G. & Ceriotti, M. (2021). Physics-inspired structural representations for molecules and materials. *Chemical Reviews*, 121(16), 9759–9815.
- Nebgen, B., Lubbers, N., Smith, J. S., Sifain, A. E., Lokhov, A., Isayev, O., Roitberg, A. E., Barros, K. & Tretiak, S. (2018). Transferable dynamic molecular charge assignment using deep neural networks. *Journal of chemical theory and computation*, 14(9), 4687–4698.
- Neville, A. M. & Brooks, J. J. (1987). *Concrete technology*. Longman Scientific & Technical England.
- Neyt, J.-C., Wender, A., Lachet, V. & Malfreyt, P. (2011). Prediction of the temperature dependence of the surface tension of SO₂, N₂, O₂, and Ar by Monte Carlo molecular simulations. *The Journal of Physical Chemistry B*, 115(30), 9421-9430.

- Nguyen, T. S. (2018). Thermo-Hydro-Mechanical-Chemical processes in geological disposal of radioactive waste—An example of regulatory research. *Advances in Geo-Energy Research*, 2(2).
- Nibbering, E. T. & Elsaesser, T. (2004). Ultrafast vibrational dynamics of hydrogen bonds in the condensed phase. *Chemical Reviews*, 104(4), 1887–1914.
- Nilsson, A. & Pettersson, L. G. (2011). Perspective on the structure of liquid water. *Chemical Physics*, 389(1-3), 1–34.
- Nodehi, M. & Taghvaei, V. M. (2022). Alkali-activated materials and geopolymer : a review of common precursors and activators addressing circular economy. *Circular Economy and Sustainability*, 2(1), 165–196.
- Novikov, I. S. & Shapeev, A. V. (2019). Improving accuracy of interatomic potentials : more physics or more data ? A case study of silica. *Materials Today Communications*, 18, 74–80.
- Novikov, I. S., Suleimanov, Y. V. & Shapeev, A. V. (2018). Automated calculation of thermal rate coefficients using ring polymer molecular dynamics and machine-learning interatomic potentials with active learning. *Physical Chemistry Chemical Physics*, 20(46), 29503–29512.
- Novikov, I. S., Gubaev, K., Podryabinkin, E. & Shapeev, A. V. (2020a). The MLIP package : Moment Tensor Potentials with MPI and Active Learning. *Machine Learning : Science and Technology*.
- Novikov, I. S., Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. (2020b). The MLIP package : moment tensor potentials with MPI and active learning. *Machine Learning : Science and Technology*, 2(2), 025002.
- Novoselov, I., Yanilkin, A., Shapeev, A. & Podryabinkin, E. (2019). Moment tensor potentials as a promising tool to study diffusion processes. *Computational Materials Science*, 164, 46–56.
- Novotni, M. & Klein, R. (2004). Shape retrieval using 3D Zernike descriptors. *Computer-Aided Design*, 36(11), 1047–1062.
- N'Tsouaglo, K. G. (2015). Caractérisation de la surface d'énergie potentielle des matériaux complexes et son application sur la cinétique du SiO₂/Si.
- Nuclear Waste Management Organization. [Accessed : 2024-09-15]. (2002). Official Website. Repéré à <https://www.nwmo.ca/fr>.

- Nuclear Waste Management Organization. (2020). *NWMO Technical Report* (Rapport n°TR201907). Repéré à <https://www.nwmo.ca/media/site/reports/2020/07/16/16/03/nwmo-tr201907.ashx?la=en>.
- N'Tsouaglo, G. K., Béland, L. K., Joly, J.-F., Brommer, P., Mousseau, N. & Pochet, P. (2015). Probing potential energy surface exploration strategies for complex systems. *Journal of Chemical Theory and Computation*, 11(4), 1970–1977.
- Ohbitsu, M., Yokoi, T., Noda, Y., Kamiyama, E., Ushiro, T., Nagakura, H., Sueoka, K. & Matsunaga, K. (2022). Atomic structures and stability of finite-size extended interstitial defects in silicon : Large-scale molecular simulations with a neural-network potential. *Scripta Materialia*, 214, 114650.
- Onat, B., Ortner, C. & Kermode, J. R. (2020). Sensitivity and dimensionality of atomic environment representations used for machine learning interatomic potentials. *The Journal of Chemical Physics*, 153(14), 144106.
- Opletal, G., Petersen, T., O'Malley, B., Snook, I., McCulloch, D. G., Marks, N. A. & Yarovsky, I. (2002). Hybrid approach for generating realistic amorphous carbon structure using metropolis and reverse Monte Carlo. *Molecular Simulation*, 28(10-11), 927–938.
- Opletal, G., Petersen, T. C., O'Malley, B., Snook, I., McCulloch, D. & Yarovsky, I. (2008). HRMC : Hybrid Reverse Monte Carlo method with silicon and carbon potentials. *Computer Physics Communications*, 178(10), 777–787.
- Ortega-Castro, J., Hernández-Haro, N., Hernández-Laguna, A. & Sainz-Díaz, C. (2008). DFT calculation of crystallographic properties of dioctahedral 2 : 1 phyllosilicates. *Clay Minerals*, 43(3), 351-361.
- Pabst, W. & Gregorová, E. (2013). Elastic properties of silica polymorphs—a review. *Ceramics-Silikaty*, 57(3), 167–184.
- Pandey, A., Biswas, P. & Drabold, D. (2015). Force-enhanced atomic refinement : Structural modeling with interatomic forces in a reverse Monte Carlo approach applied to amorphous Si and SiO₂. *Physical Review B*, 92(15), 155205.
- Pandey, A., Biswas, P., Bhattarai, B. & Drabold, D. (2016a). Realistic inversion of diffraction data for an amorphous solid : The case of amorphous silicon. *Physical Review B*, 94(23), 235208.
- Pandey, A., Biswas, P. & Drabold, D. A. (2016b). Inversion of diffraction data for amorphous materials. *Scientific Reports*, 6(1), 33731.

- Pandit, A. & Bongiorno, A. (2023). A first-principles method to calculate fourth-order elastic constants of solid materials. *Computer Physics Communications*, 288, 108751.
- Papin, R. (1956). L'utilisation des argiles dans l'industrie du pétrole. *Bulletin du Groupe français des Argiles*, 7(2), 1–14.
- Paquet, E. & Viktor, H. L. (2018). Computational methods for Ab initio molecular dynamics. *Advances in Chemistry*, 2018, 14.
- Park, C. W., Kornbluth, M., Vandermause, J., Wolverton, C., Kozinsky, B. & Mailoa, J. P. (2021). Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture. *npj Computational Materials*, 7(1), 73.
- Pasquarello, A., Hybertsen, M. S. & Car, R. (1998). Interface structure between silicon and its oxide by first-principles molecular dynamics. *Nature*, 396(6706), 58–60.
- Pastewka, L., Pou, P., Pérez, R., Gumbsch, P. & Moseler, M. (2008). Describing bond-breaking processes by reactive potentials : Importance of an environment-dependent interaction range. *Physical Review B—Condensed Matter and Materials Physics*, 78(16), 161402.
- Payne, M. C., Teter, M. P., Allan, D. C., Arias, T. & Joannopoulos, a. J. (1992). Iterative minimization techniques for ab initio total-energy calculations : molecular dynamics and conjugate gradients. *Reviews of modern physics*, 64(4), 1045.
- Pearson, E., Takai, T., Halicioglu, T. & Tiller, W. A. (1984). Computer modeling of Si and SiC surfaces and surface processes relevant to crystal growth from the vapor. *Journal of Crystal Growth*, 70(1-2), 33–40.
- Pedersen, A., Pizzagalli, L. & Jónsson, H. (2017). Optimal atomic structure of amorphous silicon obtained from density functional theory calculations. *New Journal of Physics*, 19(6), 063018.
- Pellenq, R. J.-M., Kushima, A., Shahsavari, R., Van Vliet, K. J., Buehler, M. J., Yip, S. & Ulm, F.-J. (2009). A realistic molecular model of cement hydrates. *Proceedings of the National Academy of Sciences*, 106(38), 16102–16107.
- Peng, C., Min, F., Liu, L. & Chen, J. (2016). A periodic DFT study of adsorption of water on sodium-montmorillonite (001) basal and (010) edge surface. *Applied Surface Science*, 387, 308–316.
- Peng, J., Jury, E. C., Dönnies, P. & Ciurtin, C. (2021). Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases : applications and challenges. *Frontiers in pharmacology*, 12, 720694.

- Perdew, J. P. & Schmidt, K. (2001). Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conference Proceedings*, 577(1), 1–20.
- Perdew, J. P., Burke, K. & Ernzerhof, M. (1996). Generalized gradient approximation made simple. *Physical review letters*, 77(18), 3865.
- Persson, K. [An optional note]. (2014, 11). Materials Data on SiO₂ (SG :15) by Materials Project.
- Phillpot, S. R., Antony, A. C., Shi, L., Fullarton, M. L., Liang, T., Sinnott, S. B., Zhang, Y. & Biner, S. B. (2018). Charge Optimized Many Body (COMB) potentials for simulation of nuclear fuel and clad. *Computational Materials Science*, 148, 231–241.
- Pisana, S., Lazzeri, M., Casiraghi, C., Novoselov, K. S., Geim, A. K., Ferrari, A. C. & Mauri, F. (2007). Breakdown of the adiabatic Born–Oppenheimer approximation in graphene. *Nature materials*, 6(3), 198–201.
- Pizzagalli, L. (2004). Classical Atomistic Simulations in Materials Sciences : An Introduction. *MATERIALS SCIENCE (MEDŽIAGOTYRA)*, 10(2), 125.
- Plévert, J., Kubota, Y., Honda, T., Okubo, T. & Sugi, Y. (2000). GUS-1 : a mordenite-like molecular sieve with the 12-ring channel of ZSM-12. Electronic supplementary information (ESI) available : details of the synthesis of the SDA, crystal data and fractional atomic coordinates for GUS-1. See <http://www.rsc.org/suppdata/cc/b0/b005225f>. *Chemical Communications*, (23), 2363–2364.
- Podryabinkin, E. V. & Shapeev, A. V. (2017). Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, 140, 171–180.
- Podryabinkin, E. V., Tikhonov, E. V., Shapeev, A. V. & Oganov, A. R. (2019). Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Physical Review B*, 99(6), 064114.
- Powell, M. J. (1989). The physics of amorphous-silicon thin-film transistors. *IEEE transactions on Electron Devices*, 36(12), 2753–2763.
- Provis, J. L. (2018). Alkali-activated materials. *Cement and Concrete Research*, 114, 40–48. doi : <https://doi.org/10.1016/j.cemconres.2017.02.009>.
- Pun, G. P. & Mishin, Y. (2017). Optimized interatomic potential for silicon and its application to thermal stability of silicene. *Physical Review B*, 95(22), 224103.

- Qiao, Z., Christensen, A. S., Welborn, M., Manby, F. R., Anandkumar, A. & Miller III, T. F. (2022). Informing geometric deep learning with electronic interactions to accelerate quantum chemistry. *Proceedings of the National Academy of Sciences*, 119(31), e2205221119.
- Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. (2017). Machine learning in materials informatics : recent applications and prospects. *npj Computational Materials*, 3(1), 1-13.
- Rappe, A. K. & Goddard III, W. A. (1991). Charge equilibration for molecular dynamics simulations. *The Journal of Physical Chemistry*, 95(8), 3358–3363.
- Refson, K., Park, S.-H. & Sposito, G. (2003). Ab initio computational crystallography of 2 : 1 clay minerals : 1. Pyrophyllite-1Tc. *The Journal of Physical Chemistry B*, 107(48), 13376-13383.
- Reinhardt, A. & Cheng, B. (2021). Quantum-mechanical exploration of the phase diagram of water. *Nature communications*, 12(1), 588.
- Reungoat, J. (2007). *Etude d'un procédé hybride couplant adsorption sur zéolithes et oxydation par l'ozone. Application au traitement d'effluents aqueux industriels*. (Thèse de doctorat, Institut National des Sciences Appliquées de Toulouse).
- Roorda, S., Doorn, S., Sinke, W., Scholte, P. & Van Loenen, E. (1989). Calorimetric evidence for structural relaxation in amorphous silicon. *Physical review letters*, 62(16), 1880.
- Roorda, S., Sinke, W., Poate, J., Jacobson, D., Dierker, S., Dennis, B., Eaglesham, D., Spaepen, F. & Fuoss, P. (1991). Structural relaxation and defect annihilation in pure amorphous silicon. *Physical review B*, 44(8), 3702.
- Ross, N. L., Shu, J. & Hazen, R. M. (1990). High-pressure crystal chemistry of stishovite. *American Mineralogist*, 75(7-8), 739–747.
- Rowe, P., Deringer, V. L., Gasparotto, P., Csányi, G. & Michaelides, A. (2020). An accurate and transferable machine learning potential for carbon. *The Journal of Chemical Physics*, 153(3), 034702.
- Ruan, S., Gao, R., Tu, W., Yan, D. & Zhang, M. (2024). Alkali-activated materials with organics : A critical review. *Composites Part B : Engineering*, 111712.
- Rupp, M., Tkatchenko, A., Müller, K.-R. & Von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5), 058301.

- Salim, M. A., Willow, S. Y. & Hirata, S. (2016). Ice Ih anomalies : Thermal contraction, anomalous volume isotope effect, and pressure-induced amorphization. *The Journal of Chemical Physics*, 144(20).
- Salles, N., Richard, N., Mousseau, N. & Hémerlyck, A. (2017). Strain-driven diffusion process during silicon oxidation investigated by coupling density functional theory and activation relaxation technique. *The Journal of Chemical Physics*, 147(5), 054701.
- Salzmann, C. G. (2019). Advances in the experimental exploration of water's phase diagram. *The Journal of chemical physics*, 150(6).
- Satorras, V. G., Hoogeboom, E. & Welling, M. (2021). E (n) equivariant graph neural networks. *International conference on machine learning*, pp. 9323–9332.
- Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. (2019). Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1), 1-36.
- Schommers, W. (2019). Chapter 2 - Theoretical and Computational Methods. Dans Schommers, W. (Éd.), *Basic Physics of Nanoscience* (pp. 91-202). Elsevier. doi : <https://doi.org/10.1016/B978-0-12-813718-5.00002-8>.
- Schran, C., Behler, J. & Marx, D. (2019). Automated fitting of neural network potentials at coupled cluster accuracy : Protonated water clusters as testing ground. *Journal of chemical theory and computation*, 16(1), 88–99.
- Schran, C., Briec, F. & Marx, D. (2021). Transferability of machine learning potentials : Protonated water neural network potential applied to the protonated water hexamer. *The Journal of Chemical Physics*, 154(5).
- Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A. & Müller, K.-R. (2017a). Schnet : A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30.
- Schütt, K., Unke, O. & Gastegger, M. (2021). Equivariant message passing for the prediction of tensorial properties and molecular spectra. *International Conference on Machine Learning*, pp. 9377–9388.
- Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. (2017b). Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8(1), 13890.

- Schütt, K. T., Chmiela, S., Von Lilienfeld, O. A., Tkatchenko, A., Tsuda, K. & Müller, K.-R. (2020). Machine learning meets quantum physics. *Lecture Notes in Physics*.
- Schwegler, E., Grossman, J. C., Gygi, F. & Galli, G. (2004). Towards an assessment of the accuracy of density functional theory for first principles simulations of water. II. *The Journal of chemical physics*, 121(11), 5400–5409.
- Senac, C. (2018). *Solvants et sites de liaison hydrophobes*. (Thesis).
- Seppälä, A., Puhakka, E. & Olin, M. (2016). Effect of layer charge on the crystalline swelling of Na⁺, K⁺ and Ca²⁺ montmorillonites : DFT and molecular dynamics studies. *Clay Minerals*, 51(2), 197–211.
- Shan, T.-R., Devine, B. D., Hawkins, J. M., Asthagiri, A., Phillpot, S. R., Sinnott, S. B. et al. (2010). Second-generation charge-optimized many-body potential for Si/SiO₂ and amorphous silica. *Physical Review B*, 82(23), 235302.
- Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation*, 24(111), 647–656.
- Shapeev, A. V. (2016). Moment tensor potentials : A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation*, 14(3), 1153–1173.
- Shelby, J. E. (2020). *Introduction to glass science and technology*. Royal society of chemistry.
- Shi, C., Jiménez, A. F. & Palomo, A. (2011). New cements for the 21st century : The pursuit of an alternative to Portland cement. *Cement and concrete research*, 41(7), 750–763.
- Sholl, D. & Steckel, J. A. (2011). *Density functional theory : a practical introduction*. John Wiley & Sons.
- Sholl, D. S. & Steckel, J. A. (2022). *Density functional theory : a practical introduction*. John Wiley & Sons.
- Shropshire, J., Keat, P. P. & Vaughan, P. A. (1959). The crystal structure of keatite, a new form of silica. *Zeitschrift für Kristallographie-Crystalline Materials*, 112(1-6), 409–413.
- Siddique, R. (2011). Utilization of silica fume in concrete : Review of hardened properties. *Resources, conservation and recycling*, 55(11), 923–932.
- Singh, N. B. (2022). Clays and clay minerals in the construction industry. *Minerals*, 12(3), 301.
- Škvára, F. (2007). Alkali activated materials or geopolymers. *Ceramics-Silikáty*, 51(3), 173–177.

- Slater, J. C. (1963). The electronic structure of atoms—the Hartree-Fock method and correlation. *Reviews of Modern Physics*, 35(3), 484.
- Smith, J. S., Isayev, O. & Roitberg, A. E. (2017). ANI-1 : an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical science*, 8(4), 3192–3203.
- Smith, J. S., Nebgen, B., Mathew, N., Chen, J., Lubbers, N., Burakovsky, L., Tretiak, S., Nam, H. A., Germann, T., Fensin, S. et al. (2021). Automated discovery of a robust interatomic potential for aluminum. *Nature communications*, 12(1), 1257.
- Smyth, J. R., Swope, R. J. & Pawley, A. R. (1995). H in rutile-type compounds : II. Crystal chemistry of Al substitution in H-bearing stishovite. *American Mineralogist*, 80(5-6), 454–456.
- Šolc, R., Gerzabek, M. H., Lischka, H. & Tunega, D. (2011). Wettability of kaolinite (001) surfaces—molecular dynamic study. *Geoderma*, 169, 47–54.
- Soper, A. (1996). Empirical potential Monte Carlo simulation of fluid structure. *Chemical Physics*, 202(2), 295-306. doi : [https://doi.org/10.1016/0301-0104\(95\)00357-6](https://doi.org/10.1016/0301-0104(95)00357-6).
- Speedy, R. & Angell, C. (1976). Isothermal compressibility of supercooled water and evidence for a thermodynamic singularity at- 45 C. *The Journal of Chemical Physics*, 65(3), 851–858.
- Stillinger, F. H. & Weber, T. A. (1985). Computer simulation of local order in condensed phases of silicon. *Physical review B*, 31(8), 5262.
- Street, R. (1999). *Technology and applications of amorphous silicon*. Springer Science & Business Media.
- Streetman, B. G., Banerjee, S. et al. (2000). *Solid state electronic devices*. Prentice hall New Jersey.
- Stukowski, A. (2009). Visualization and analysis of atomistic simulation data with OVITO—the Open Visualization Tool. *Modelling and Simulation in Materials Science and Engineering*, 18(1), 015012.
- Sun, H. & Béland, L. K. (2022). Statistical distribution of spontaneous recombination radii of Frenkel pairs in FCC and BCC metals. *Acta Materialia*, 229, 117814.
- Sun, H. & Béland, L. K. (2023). Calculation of the recombination radii between various point defects and defect clusters in nickel via kinetic Activation Relaxation Technique. *Journal of Nuclear Materials*, 586, 154663.

- Sun, J., Ruzsinszky, A. & Perdew, J. P. (2015). Strongly constrained and appropriately normed semilocal density functional. *Physical review letters*, 115(3), 036402.
- Sun, W. & Ceder, G. (2013). Efficient creation and convergence of surface slabs. *Surface Science*, 617, 53–59.
- Sundararaman, S., Huang, L., Ispas, S. & Kob, W. (2018). New optimization scheme to obtain interaction potentials for oxide glasses. *The Journal of Chemical Physics*, 148(19).
- Sutton, R. S. & Barto, A. G. (2018). *Reinforcement learning : An introduction*. MIT press.
- Sze, S. M., Li, Y. & Ng, K. K. (2021). *Physics of semiconductor devices*. John wiley & sons.
- Szlachta, W. J., Bartók, A. P. & Csányi, G. (2014). Accuracy and transferability of Gaussian approximation potential models for tungsten. *Physical Review B*, 90(10), 104108.
- Takahashi, N., Yamasaki, T. & Kaneta, C. (2014). Molecular dynamics simulations on the oxidation of Si (100)/SiO₂ interface : Emissions and incorporations of Si-related species into the SiO₂ and substrate. *physica status solidi (b)*, 251(11), 2169–2178.
- Takamoto, S., Izumi, S. & Li, J. (2022). TeaNet : Universal neural network interatomic potential inspired by iterative electronic relaxations. *Computational Materials Science*, 207, 111280.
- Tambach, T. J., Hensen, E. J. & Smit, B. (2004). Molecular simulations of swelling clay minerals. *The Journal of Physical Chemistry B*, 108(23), 7586–7596.
- Tangney, P. & Scandolo, S. (2002). An ab initio parametrized interatomic force field for silica. *The Journal of chemical physics*, 117(19), 8898–8904.
- Tavares, S. R., Vaiss, V. S., Antunes, F. P. N., Fonseca, C. G., Nangoi, I. M., Moraes, P. I. R., Soares, C. V., Haddad, J. F., Lima, L. L. & Silva, B. N. (2018). DFT calculations for structural prediction and applications of intercalated lamellar compounds. *Dalton Transactions*, 47(9), 2852-2866.
- Tersoff, J. (1986). New empirical model for the structural properties of silicon. *Physical review letters*, 56(6), 632.
- Tersoff, J. (1988a). Empirical interatomic potential for silicon with improved elastic properties. *Physical Review B*, 38(14), 9902.
- Tersoff, J. (1988b). Empirical interatomic potential for silicon with improved elastic properties. *Physical Review B*, 38(14), 9902.

- Tholke, P. & De Fabritiis, G. (2022). Torchmd-net : equivariant transformers for neural network based molecular potentials. *arXiv preprint arXiv :2202.02541*.
- Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K. & Riley, P. (2018). Tensor field networks : Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv :1802.08219*.
- Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. J. (2015). Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *Journal of Computational Physics*, 285, 316–330.
- Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., in't Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D. et al. (2022). LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271, 108171.
- Tian, F.-Y., Chen, N.-X., Delczeg, L. & Vitos, L. (2012). Interlayer potentials for fcc (1 1 1) planes of Pd–Ag random alloys. *Computational materials science*, 63, 20–27.
- Tian, Z., Li, Y., Zheng, J. & Wang, S. (2019). A state-of-the-art on self-sensing concrete : Materials, fabrication and properties. *Composites Part B : Engineering*, 177, 107437.
- Torrens, I. (2012). *Interatomic potentials*. Elsevier.
- Toutanji, H. A. & El-Korchi, T. (1995). The influence of silica fume on the compressive strength of cement paste and mortar. *cement and Concrete Research*, 25(7), 1591–1602.
- Tozzini, V. (2010). Multiscale modeling of proteins. *Accounts of chemical research*, 43(2), 220-230.
- Tran, R., Xu, Z., Radhakrishnan, B., Winston, D., Sun, W., Persson, K. A. & Ong, S. P. (2016). Surface energies of elemental crystals. *Scientific data*, 3(1), 1-13.
- Trochet, M., Mousseau, N., Béland, L. K. & Henkelman, G. (2020). Off-lattice kinetic Monte Carlo methods. *Handbook of materials modeling : Methods : Theory and modeling*, 715–743.
- Trolier-McKinstry, S. & Newnham, R. E. (2017). *Materials Engineering : Bonding, Structure, and Structure-Property Relationships*. Cambridge University Press.
- Tsuneyuki, S., Tsukada, M., Aoki, H. & Matsui, Y. (1988). First-principles interatomic potential of silica applied to molecular dynamics. *Physical Review Letters*, 61(7), 869.

- Tucker, M., Keen, D., Dove, M. & Trachenko, K. (2005). Refinement of the Si–O–Si bond angle distribution in vitreous silica. *Journal of Physics : Condensed Matter*, 17(5), S67.
- Tuckerman, M. E. (2002). Ab initio molecular dynamics : basic concepts, current trends and novel applications. *Journal of Physics : Condensed Matter*, 14(50), R1297.
- Ulian, G., Moro, D. & Valdrè, G. (2021). DFT Simulation of the Water Molecule Interaction with the (001) Surface of Montmorillonite. *Minerals*, 11(5), 501.
- Unke, O. T. & Meuwly, M. (2019). PhysNet : A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of chemical theory and computation*, 15(6), 3678–3693.
- Van Beest, B., Kramer, G. J. & Van Santen, R. (1990). Force fields for silicas and aluminophosphates based on ab initio calculations. *Physical review letters*, 64(16), 1955.
- Van den Hoven, G., Liang, Z., Niesen, L. & Custer, J. (1992). Evidence for vacancies in amorphous silicon. *Physical review letters*, 68(25), 3714.
- Van Deventer, J., Provis, J., Duxson, P. & Lukey, G. (2007). Reaction mechanisms in the geopolymeric conversion of inorganic waste to useful products. *Journal of hazardous materials*, 139(3), 506–513.
- Van Duin, A. C., Dasgupta, S., Lorant, F. & Goddard, W. A. (2001). ReaxFF : a reactive force field for hydrocarbons. *The Journal of Physical Chemistry A*, 105(41), 9396–9409.
- Van Duin, A. C., Strachan, A., Stewman, S., Zhang, Q., Xu, X. & Goddard, W. A. (2003). ReaxFFSiO reactive force field for silicon and silicon oxide systems. *The Journal of Physical Chemistry A*, 107(19), 3803–3811.
- Vashishta, P., Kalia, R. K., Rino, J. P. & Ebbsjö, I. (1990). Interaction potential for SiO₂ : A molecular-dynamics study of structural correlations. *Physical Review B*, 41(17), 12197.
- Vega, C. & Abascal, J. L. (2011). Simulating water with rigid non-polarizable models : a general perspective. *Physical Chemistry Chemical Physics*, 13(44), 19663–19688.
- Verlet, L. (1967). Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical review*, 159(1), 98.
- Villard, J., Bircher, M. P. & Rothlisberger, U. (2024). Structure and dynamics of liquid water from ab initio simulations : adding Minnesota density functionals to Jacob's ladder. *Chemical Science*, 15(12), 4434–4451.

- Villars, P. (1985). Pearson's handbook : crystallographic data for intermetallic phases. (*No Title*).
- Vink, R. L. C. (2002). Computer simulations of amorphous semiconductors. *Ph. D. Thesis*.
- Vink, R., Barkema, G., Stijnman, M. & Bisseling, R. (2001a). Device-size atomistic models of amorphous silicon. *Physical Review B*, 64(24), 245214.
- Vink, R., Barkema, G., Van der Weg, W. & Mousseau, N. (2001b). Fitting the Stillinger–Weber potential to amorphous silicon. *Journal of non-crystalline solids*, 282(2-3), 248–255.
- Volz, S. G. & Chen, G. (1999). Molecular dynamics simulation of thermal conductivity of silicon nanowires. *Applied Physics Letters*, 75(14), 2056–2058.
- Vukceovich, M. (1972). A new interpretation of the anomalous properties of vitreous silica. *Journal of Non-Crystalline Solids*, 11(1), 25–63.
- Wang, G., Sun, Y., Zhou, J. & Sun, Z. (2023). PotentialMind : graph convolutional machine learning potential for Sb–Te binary compounds of multiple stoichiometries. *The Journal of Physical Chemistry C*, 127(51), 24724–24733.
- Wang, G., Wang, C., Zhang, X., Li, Z., Zhou, J. & Sun, Z. (2024). Machine learning interatomic potential : Bridge the gap between small-scale models and realistic device-scale simulations. *Iscience*, 27(5).
- Wang, H., Zhang, L., Han, J. & Weinan, E. (2018). DeePMD-kit : A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications*, 228, 178-184.
- Wang, H. & Yang, W. (2018). Force field for water based on neural network. *The journal of physical chemistry letters*, 9(12), 3232–3240.
- Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. (2004). Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9), 1157–1174.
- Wang, S. & Peng, Y. (2010). Natural zeolites as effective adsorbents in water and wastewater treatment. *Chemical engineering journal*, 156(1), 11–24.
- Wang, Z., Wang, C., Zhao, S., Du, S., Xu, Y., Gu, B.-L. & Duan, W. (2021). Symmetry-adapted graph neural networks for constructing molecular dynamics force fields. *Science China Physics, Mechanics & Astronomy*, 64(11), 117211.

- Wang, Z., Wang, C., Zhao, S., Xu, Y., Hao, S., Hsieh, C. Y., Gu, B.-L. & Duan, W. (2022). Heterogeneous relational message passing networks for molecular dynamics simulations. *npj Computational Materials*, 8(1), 53.
- Watanabe, T. W. T., Fujiwara, H. F. H., Noguchi, H. N. H., Hoshino, T. H. T. & Ohdomari, I. O. I. (1999). Novel interatomic potential energy function for Si, O mixed systems. *Japanese journal of applied physics*, 38(4A), L366.
- Wei, J., Chu, X., Sun, X., Xu, K., Deng, H., Chen, J., Wei, Z. & Lei, M. (2019). Machine learning in materials science. *InfoMat*, 1(3), 338-358.
- Weinan, E. & Li, D. (2008). The Andersen thermostat in molecular dynamics. *Communications on pure and applied mathematics*, 61(1), 96–136.
- Wikipedia contributors. [Accessed : 2024-09-05]. (2023). Silicium. Repéré à <https://fr.wikipedia.org/wiki/Silicium>.
- Wikipedia contributors. [Consulté le 15 septembre 2024]. (2024). Condition périodique aux limites. Repéré à https://fr.wikipedia.org/wiki/Condition_p%C3%A9riodique_aux_limites.
- Wikipédia. [Consulté le 15 septembre 2024]. (2024). Algorithme du gradient. Repéré à https://fr.wikipedia.org/wiki/Algorithme_du_gradient.
- Winetrout, J. J., Kanhaiya, K., Sachdeva, G., Pandey, R., Damirchi, B., van Duin, A., Odegard, G. & Heinz, H. (2021). Implementing reactivity in molecular dynamics simulations with the interface force field (iff-r) and other harmonic force fields. *arXiv preprint arXiv :2107.14418*.
- Winter, B., Weber, R., Widdra, W., Dittmar, M., Faubel, M. & Hertel, I. (2004). Full valence band photoemission from liquid water using EUV synchrotron radiation. *The Journal of Physical Chemistry A*, 108(14), 2625–2632.
- Wooten, F., Winer, K. & Weaire, D. (1985). Computer generation of structural models of amorphous Si and Ge. *Physical review letters*, 54(13), 1392.
- Worrell, E., Price, L., Martin, N., Hendriks, C. & Meida, L. O. (2001). Carbon dioxide emissions from the global cement industry. *Annual review of energy and the environment*, 26(1), 303–329.
- Wright, A. & Thorpe, M. (2013). Eighty years of random networks. *physica status solidi (b)*, 250(5), 931–936.

- Wright, A. & Lehmann, M. (1981). The structure of quartz at 25 and 590 C determined by neutron diffraction. *Journal of solid state chemistry*, 36(3), 371–380.
- Wu, Z. & Cohen, R. E. (2006). More accurate generalized gradient approximation for solids. *Physical Review B*, 73(23), 235116.
- Xie, T. & Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14), 145301.
- Xu, P., Guidez, E. B., Bertoni, C. & Gordon, M. S. (2018). Perspective : Ab initio force field methods derived from quantum mechanics. *The Journal of Chemical Physics*, 148(9), 090901.
- Yang, N., Zhang, G. & Li, B. (2008). Ultralow thermal conductivity of isotope-doped silicon nanowires. *Nano letters*, 8(1), 276–280.
- Yao, K., Herr, J. E., Toth, D. W., Mckintyre, R. & Parkhill, J. (2018). The TensorMol-0.1 model chemistry : a neural network augmented with long-range physics. *Chemical science*, 9(8), 2261–2269.
- Yasukawa, A. (1996). Using an extended Tersoff interatomic potential to analyze the static-fatigue strength of SiO₂ under atmospheric influence. *JSME international journal. Ser. A, Mechanics and material engineering*, 39(3), 313–320.
- Yu, J., Sinnott, S. B. & Phillpot, S. R. (2007). Charge optimized many-body potential for the Si/ SiO₂ system. *Physical Review B—Condensed Matter and Materials Physics*, 75(8), 085311.
- Yusslee, E. & Beskhyroun, S. (2022). The potential of one-part alkali-activated materials (AAMs) as a concrete patch mortar. *Scientific Reports*, 12(1), 15902.
- Zachariasen, W. H. (1932). The atomic arrangement in glass. *Journal of the American Chemical Society*, 54(10), 3841–3851.
- Zaverkin, V. & Kästner, J. (2020). Gaussian moments as physically inspired molecular descriptors for accurate and scalable machine learning potentials. *Journal of Chemical Theory and Computation*, 16(8), 5410–5421.
- Zhai, Y., Rashmi, R., Palos, E. & Paesani, F. (2024). Many-body interactions and deep neural network potentials for water. *The Journal of Chemical Physics*, 160(14).

- Zhang, L., Han, J., Wang, H., Car, R. & E, W. (2018). Deep potential molecular dynamics : a scalable model with the accuracy of quantum mechanics. *Physical review letters*, 120(14), 143001.
- Zhang, L., Wang, H., Car, R. & E, W. (2021). Phase diagram of a deep potential water model. *Physical review letters*, 126(23), 236001.
- Zhao, H., Tang, Z., Li, G. & Aluru, N. (2006). Quasiharmonic models for the calculation of thermodynamic properties of crystalline silicon under strain. *Journal of Applied Physics*, 99(6).
- Zhao, S., Hahn, E., Kad, B., Remington, B. A., Wehrenberg, C., Bringa, E. M. & Meyers, M. A. (2016). Amorphization and nanocrystallization of silicon under shock compression. *Acta Materialia*, 103, 519-533.
- Zheng, H., Li, X.-G., Tran, R., Chen, C., Horton, M., Winston, D., Persson, K. A. & Ong, S. P. (2020). Grain boundary properties of elemental metals. *Acta Materialia*, 186, 40-49.
- Zheng, L., Chen, M., Sun, Z., Ko, H.-Y., Santra, B., Dhuvad, P. & Wu, X. (2018). Structural, electronic, and dynamical properties of liquid water by ab initio molecular dynamics based on SCAN functional within the canonical ensemble. *The Journal of Chemical Physics*, 148(16).
- Zhou, T., Ioannidou, K., Ulm, F.-J., Bazant, M. Z. & Pellenq, R.-M. (2019). Multiscale poromechanics of wet cement paste. *Proceedings of the National Academy of Sciences*, 116(22), 10652–10657.
- Zhu, H., Whittle, A. J., Pellenq, R. J.-M. & Ioannidou, K. (2019). Mesoscale simulation of aggregation of imogolite nanotubes from potential of mean force interactions. *Molecular Physics*, 117(22), 3445-3455.
- Ziesche, P., Kurth, S. & Perdew, J. P. (1998). Density functionals from LDA to GGA. *Computational materials science*, 11(2), 122-127.
- Zongo, K., Béland, L. & Ouellet-Plamondon, C. (2021). Improving atom-scale models of clay minerals using machine learning. *Canadian Nuclear Society*. Repéré à <https://www.xcdsystem.com/cns/proceedings/prof34444.html>.
- Zongo, K., Béland, L. & Ouellet-Plamondon, C. (2022). First-principles database for fitting a machine-learning silicon interatomic force field. *MRS Advances*, 7(2-3), 39–47.
- Zongo, K., Sun, H., Ouellet-Plamondon, C. & Béland, L. K. (2024). A unified moment tensor potential for silicon, oxygen, and silica. *npj Computational Materials*, 10(1), 218.

Zuo, Y., Chen, C., Li, X., Deng, Z., Chen, Y., Behler, J., Csanyi, G., Shapeev, A., Thompson, A. & Wood, M. (2020). Performance and Cost Assessment of Machine Learning Interatomic Potentials. *Journal of Physical Chemistry A*, 124(4), 731-745.

