Machine learning and adaptive sampling to predict finite-temperature properties in metallic materials at the atomic scale

Anruo Zhong

université PARIS-SACLAY

Machine Learning and Adaptive Sampling to Predict Finite-Temperature Properties in Metallic Materials at the Atomic Scale

Apprentissage statistique et échantillonnage adaptatif pour prédire les propriétés à température finie des matériaux métalliques à partir de l'échelle atomique

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 564, Physique en Île-de-France (PIF)
Spécialité de doctorat : Physique
Graduate School : Physique. Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Service de Recherche en Corrosion et Comportement des Matériaux (Université Paris-Saclay, CEA)**, sous la direction de **Mihai-Cosmin MARINICA**, directeur de recherche, et la co-direction de **Manuel ATHÈNES**, directeur de recherche.

Thèse soutenue à Paris-Saclay, le 10 octobre 2024, par

Anruo ZHONG

Composition du jury

Membres du jury avec voix délibérative

Marie-Pierre GAIGEOT Professeure des Universités, Université Paris-Saclay	Présidente
David RODNEY Professeur des Universités, Université Claude Bernard Lyon 1	Rapporteur & Examinateur
Anne HÉMERYCK Directrice de Recherche, LAAS-CNRS, Université de Toulouse	Rapporteure & Examinatrice
David WALES Professeur des Universités, University of Cambridge	Examinateur

THESE DE DOCTORAT
NNT : 2024UPASP107

université PARIS-SACLAY | ÉCOLE DOCTORALE Physique en Île-de-France (PIF)

Title: Machine learning and adaptive sampling to predict finite-temperature properties in metallic materials at the atomic scale

Keywords: adaptive sampling, free energy estimation, machine learning, multi-scale modeling, finite-temperature properties, metallic materials

Abstract: The properties and behaviors of materials under extreme conditions are essential for energy systems such as fission and fusion reactors. However, accurately predicting the properties of materials at high temperatures remains challenging. Direct measurements of these properties are constrained by experimental instrument limitations, and atomic-scale simulations based on empirical force fields are often unreliable due to a lack of accuracy. This problem can be addressed using machine learning techniques, which have recently become widely used in materials research. Machine learning force fields achieve the accuracy of *ab initio* calculations; however, their implementation in sampling methods is limited by high computational costs, typically several orders of magnitude greater than those of traditional force fields. To overcome this limitation, this thesis has two objectives: (i) developing machine learning force fields with a better accuracy-efficiency trade-off, and (ii) creating accelerated sampling methods to facilitate the use of computationally expensive machine learning force fields and accurately estimate free energy. For the first objective, we enhance the construction of machine learning force fields by focusing on three key factors: the database, the descriptor of local atomic environments, and the regression model. Within the framework of Gaussian process regression, we propose and optimize descriptors based on Fourier-sampled kernels and novel sparse points selection methods for kernel regression. For the second objective, we develop a fast and robust Bayesian sampling scheme for estimating the fully anharmonic free energy, which is crucial for understanding temperature effects in crystalline solids, utilizing an improved adaptive biasing force method. This method performs a thermodynamic integration from a harmonic reference system, where numerical instabilities associated with zero frequencies are screened off. The proposed sampling method significantly improves convergence speed and overall accuracy. We demonstrate the efficiency of the improved method by calculating the second-order derivatives of the free energy, such as the elastic constants, which are computed several hundred times faster than with standard methods. This approach enables the prediction of the thermodynamic properties of tungsten and Ta-Ti-V-W high-entropy alloys at temperatures that cannot be investigated experimentally, up to their melting point, with *ab initio* accuracy by employing accurate machine learning force fields. An extension of this method allows for the sampling of a specified metastable state without transitions between different energy basins, thereby providing the formation and binding free energies of defective configurations. This development helps to explain the mechanism behind the observation of voids in tungsten, which cannot be explained by existing *ab initio* calculations. The free energy profile of vacancies in the Ta-Ti-V-W system is also computed for the first time. Finally, we validate the application of this free energy sampling method to liquids. The accuracy and numerical efficiency of the proposed computational framework, which combines machine learning force fields and enhanced sampling methods, opens up numerous possibilities for the reliable prediction of finite-temperature material properties.

Titre: Apprentissage statistique et échantillonnage adaptatif pour prédire les propriétés à température finie des matériaux métalliques à partir de l'échelle atomique

Mots clés: échantillonnage adaptatif, estimation de l'énergie libre, apprentissage statistique, modélisation multi-échelle, propriétés à température finie, matériaux métalliques

Résumé: Les propriétés et le comportement des matériaux dans des conditions extrêmes sont essentiels pour les systèmes énergétiques tels que les réacteurs de fission et de fusion. Cependant, prédire avec précision les propriétés des matériaux à haute température reste un défi. Les mesures directes de ces propriétés sont limitées par les instruments expérimentaux, et les simulations à l'échelle atomique basées sur des champs de force empiriques sont souvent peu fiables en raison d'un manque de précision. Ce problème peut être résolu à l'aide de techniques d'apprentissage statistique, qui ont récemment vu leur utilisation exploser en science des matériaux. Les champs de force construits par apprentissage statistique atteignent le degré de précision des calculs *ab initio* ; cependant, leur mise en œuvre dans les méthodes d'échantillonnage est limitée par des coûts de calcul élevés, généralement supérieurs de plusieurs ordres de grandeur à ceux des champs de force traditionnels. Pour surmonter cette limitation, deux objectifs sont poursuivis dans cette thèse : (i) développer des champs de force par apprentissage statistique avec un meilleur compromis précision-efficacité et (ii) créer des méthodes accélérées d'échantillonnage de l'énergie libre afin de faciliter l'utilisation de champs de force d'apprentissage statistique coûteux en termes de calcul. Pour le premier objectif, nous améliorons la construction des champs de force d'apprentissage statistique en nous concentrant sur trois facteurs clés : la base de données, le descripteur de l'environnement atomique local et le modèle de régression. Dans le cadre de la régression par processus gaussien, nous proposons et optimisons des descripteurs basés sur des noyaux échantillonnés par la transformée de Fourier ainsi que de nouvelles méthodes de sélection de points épars pour la régression par noyau. Pour le deuxième objectif, nous développons un schéma d'échantillonnage bayésien rapide et robuste pour estimer l'énergie libre anharmonique, qui est cruciale pour comprendre les effets de la température sur les solides cristallins, à l'aide d'une méthode de force de biais adaptative améliorée. Cette méthode effectue une intégration thermodynamique à partir d'un système de référence harmonique, où les instabilités numériques associées aux fréquences nulles sont éliminées. La méthode d'échantillonnage proposée améliore considérablement la vitesse de convergence et la précision globale. Nous démontrons l'efficacité de la méthode améliorée en calculant les dérivées de second ordre de l'énergie libre, telles que les constantes élastiques, avec une rapidité plusieurs centaines de fois supérieure à celle des méthodes standard. Cette approche permet de prédire les propriétés thermodynamiques du tungstène et des alliages à haute entropie Ta-Ti-V-W à des températures qui ne peuvent être étudiées expérimentalement, jusqu'à leur point de fusion, avec une précision *ab initio* grâce à l'utilisation de champs de force construits par apprentissage statistique. Une extension de cette méthode permet l'échantillonnage d'un état métastable spécifique sans transition entre différents bassins d'énergie, fournissant ainsi l'énergie libre de formation et de liaison d'une configuration défectueuse. Ce développement aide à expliquer le mécanisme derrière l'observation des cavités dans le tungstène, mécanisme qui ne peut pas être expliqué par les calculs *ab initio* existants. Le profil d'énergie libre des lacunes dans le système Ta-Ti-V-W est également calculé pour la première fois. Enfin, nous validons l'application de cette méthode d'échantillonnage de l'énergie libre aux liquides. La précision et l'efficacité numérique du cadre de calcul proposé, qui combine des champs de force d'apprentissage statistique et des méthodes d'échantillonnage améliorées, ouvrent de nombreuses possibilités pour la prédiction fiable des propriétés des matériaux à température finie.

Acknowledgements

Cela a été une véritable chance de compléter ma thèse au sein de la Section de Recherche de Métallurgie Physique (SRMP) du CEA Saclay. Je souhaite d'abord exprimer ma profonde gratitude envers mes encadrants, Mihai-Cosmin Marinica et Manuel Athènes. Un grand merci pour votre soutien, votre expertise et la confiance que vous m'avez accordée. Je suis tellement chanceuse d'avoir pu travailler à vos côtés lors de ces trois années. Ce travail de recherche n'aurait pas pu aboutir sans vous, car il s'agit avant tout d'un travail d'équipe. Cosmin : merci de m'avoir guidée pour explorer le monde du machine learning et ses applications aux matériaux. Je suis heureuse d'avoir pu travailler sur un sujet aussi passionnant avec toi ! Tu es toujours au courant des derniers avancements dans ce domaine, et tu partages constamment des idées très intéressantes. Merci pour ta pédagogie détaillée, à la fois théorique et pratique, grâce à laquelle j'ai tant appris, ta curiosité sans limites, ton rire et sourire chaleureux, ton optimisme et tes encouragements lorsque je me suis sentie perdue. Manuel : merci de m'avoir introduite à la physique statistique, de m'avoir donné le temps de réfléchir et de découvrir l'univers fascinant des équations et des méthodes numériques, un domaine que tu connais comme ta poche. Chaque discussion avec toi m'a énormément apporté, notamment tes réflexions et propositions face aux problèmes. Merci aussi pour ton immense patience devant mes nombreuses questions, ta rigueur à laquelle aucun détail n'échappe, ton enthousiasme et ton envie de toujours faire mieux. Encore une fois, je vous suis reconnaissante, Cosmin et Manuel, pour tout ce que vous avez fait pour ma thèse, ainsi que pour ma formation en tant que chercheuse.

Je souhaite ensuite remercier Anne Hémeryck et David Rodney d'avoir accepté d'être les rapporteurs de ma thèse. Je remercie également la présidente du jury, Marie-Pierre Gaigeot, et l'examinateur, David Wales, pour leur participation au jury de ma soutenance. Mes remerciements s'adressent aussi à Jean-Bernard Maillet. Merci à vous tous d'avoir accepté d'évaluer ce travail de thèse. Vos remarques constructives, vos conseils avisés et nos discussions éclairantes m'ont permis de mieux apprécier les possibilités et les limites de mes travaux, et m'ont beaucoup inspirée pour aller plus loin.

Je tiens aussi à adresser mes remerciements à mes collègues de l'équipe de recherche et les collaborateurs. Vos aides inestimables ont été indispensables pour le bon avancement de mes travaux de thèse, et je vous suis reconnaissante pour le temps que vous y avez consacré. Clovis : merci d'être là et de m'aider sans hésitation dans les moments les plus "chauds" pendant ma thèse—au début et également à la fin. Sasha : j'ai été ravie de pouvoir travailler avec toi, merci pour ta gentillesse, ta créativité et ta disponibilité tout au long du chemin. Alexandre : cela a été un grand plaisir que nous ayons pu apprendre, discuter et découvrir le monde de la recherche ensemble. Arnaud : merci pour ton aide sur l'utilisation de PAFI. Jacopo: thank you for your excellent idea about improving sampling efficiency in free energy estimation. Jan Wróbel: thank you for giving me the opportunity to work on HEAs, what a fascinating topic! Kazuto Arakawa: thank you for helping me learn about the experimental methods

related to my research. Tom Swinburne: thank you for providing interesting insights on probabilistic sampling, AI, and free energy computations.

Je tiens aussi à remercier tous mes collègues à la SRMP, qui maintiennent une atmosphère dynamique, tant sur le plan de la recherche que dans la vie quotidienne. Grâce à vous, je garde d'excellents souvenirs de ces trois dernières années : les journées annuelles et les réunions de service, les discussions professionnelles et amicales dans la salle café, pendant les repas de midi et même après le boulot, les courses de relais pendant les fêtes de la musique... J'exprime tout d'abord mes remerciements particuliers à Jean-Luc Béchade, notre chef de section, pour avoir fait de ton mieux afin que les thèses se déroulent dans les meilleures conditions ; et à Rosabelle Berger, notre secrétaire, pour les efforts que tu as fournis pour faciliter les tâches administratives. Je tiens aussi à remercier Jean-Paul Crocombette d'avoir fait partie de mon comité de suivi avec Jean-Luc, et de m'avoir donné de précieux conseils pour me former en tant que chercheuse. Merci également à Jean-Paul, Thomas Jourdan, Maylise Nastar et Estelle Meslin d'avoir participé aux répétitions de ma soutenance en proposant des discussions enrichissantes concernant ma thèse. Un grand merci à tous mes collègues doctorants, post-doctorants et stagiaires de la "Nowhere generation" ! Thank you to Kajal, my office mate, for the girls' talks about everything. Merci à Maxime et Arthur, les doctorants de la même promotion que moi, avec qui on s'est débrouillés dans tous les "messes" administratifs. Merci à Kangming, Liangzhao et Kan, les anciens qui ont suivi le même parcours que moi et qui m'ont montré d'excellents exemples dans la recherche. Merci à Daphné, Pamela et Xixi pour votre gentillesse et bonne humeur depuis toujours. D'ailleurs, je n'oublie pas de remercier les autres camarades qui m'ont accueillie chaleureusement à mon arrivée et m'ont beaucoup aidée à m'intégrer à la SRMP : Baptiste, Charbel, Emile, Guilhem, Océane, Orane, Quentin, Savneet, Yunho et Zhengxuan. Je remercie également les collègues que j'ai rencontrés au cours de ces trois années, pour les excellents moments passés ensemble : Adrien, Antoine (merci pour le badminton !), Camilo, Gabrielle, Hilal, Jasurbek, Jean-Baptiste, Lautaro, Lisa Lefort, Lisa Vessier-Alary, Mattia, Pablo, Rohit, Roshal, Tarek, Thomas Bilyk, Thomas Leveau, Thomas Mainguy et Zexin. J'ai eu la chance d'appartenir à une unité aussi sympathique, où l'entraide est très présente. Je remercie encore une fois l'ensemble des non-permanents et des permanents de la SRMP.

I am also grateful to Kan Lai and Yangfan Hu, my bachelor's and master's thesis supervisors, who opened the door to scientific research for me and helped me discover its appeal. Je souhaite également exprimer ma gratitude à Bertrand Mercier, qui m'a fait découvrir les opportunités de recherche en France, sans qui je n'aurais pas eu la chance de réaliser une thèse au CEA.

A heartfelt thank you to my dear friends Jiayi, Liyi, Wenjia, and Chuanyong, who explored Paris, France, and Europe with me! It has always been a great pleasure to spend time with you. Thank you for giving me an amazing out-of-PhD life. Many thanks also to my friends who are or were pursuing their PhDs around the world: Shasha, Jiewei, Zhibo, Jian, Zhaoxuan, and Guiling, with whom I shared the highs and lows of research. Thanks as well to my friends who are not doing a PhD: Anqi, Liqi, Weikang, Hanyu, Huaiqin, Jinlin, and Ziqi, for always being there to encourage me, despite the long distances between us. Un merci particulier à Romain, du fond de mon cœur, pour tout le soutien que tu m'as apporté pendant ma thèse. Ton écoute, tes conseils et ta capacité à m'encourager ont été inestimables pour que je puisse aller aussi loin. Je me sens tellement chanceuse de t'avoir à mes côtés.

Finally, this manuscript is dedicated to my parents, Liqin and Zhengguang, who have supported me throughout these three years, just as they always have. I am endlessly grateful to my parents for nurturing my curiosity, standing by my dreams, and providing me with unconditional love.

List of abbreviations

ABF	adaptive biasing force
ABP	adaptive biasing potential
ACE	Atomic Cluster Expansion
AI	artificial intelligence
AM	Armiento-Mattson functional
BABF	Bayesian adaptive biasing force
bcc	body-centered cubic
DFT	density functional theory
DL	deep learning
eABF	extended adaptive biasing force
EA	Einstein approximation
EAM	embedded atom method
FEP	free energy perturbation
GAP	Gaussian Approximation Potential
GP	Gaussian process
GNN	Graph Neural Network
HA	harmonic approximation
HA-SVD	SVD-filtered harmonic approximation
HEA	high-entropy alloy
h-LML	linear machine learning model with hybrid descriptors
KNML	Kernel Noise Machine Learning
LAE	local atomic environment
LDA	local density approximation

LML	Linear Machine Learning
MC	Monte Carlo
MD	molecular dynamics
MEAM	modified EAM
MEP	minimum energy path
MFEP	minimum free energy path
MiLaDy	Machine Learning Dynamics
ML	machine learning
MTP	Moment Tensor Potential
NEB	Nudge Elastic Band
NN	neural network
PAFI	Projected Average Force Integrator
PBE	Perdew-Burke-Ernzerhof functional
PCA	principal component analysis
PFC	plasma-facing component
QNML	Quadratic Noise Machine Learning
qSNAP	quadratic Spectral Neighbor Analysis Potential
RKHS	reproducing kernel Hilbert space
RMSE	root mean square error
SIA	self-interstitial atom
SOAP	Smooth Overlap of Atomic Positions
SVD	singular value decomposition
TEM	transmission electron microscopy
TI	thermodynamic integration
TU-TILD	two-stage upsampled thermodynamic integration using Langevin dynamics
UF	Uhlenbeck-Ford model
UP-TILD	upsampled thermodynamic integration using Langevin dynamics
ZBL	Ziegler-Biersack-Littmark potential

Contents

1 Introduction **1**
- 1.1 Interatomic potential: traditional methods and limitations . . . 2
- 1.2 Machine learning approaches: the renaissance of empirical potential . . . 4
- 1.3 Advanced sampling schemes and estimation of free energy . . . 7
- 1.4 Machine-learnt free energy profile: introduction to the thesis . . . 9

2 Machine learning force fields **11**
- 2.1 Basics of machine learning force fields . . . 12
  - 2.1.1 Database for machine learning force fields . . . 15
    - 2.1.1.1 Database from DFT and construction of target vector . . . 16
    - 2.1.1.2 Active learning and database updating . . . 18
  - 2.1.2 Representation of local atomic environment . . . 19
    - 2.1.2.1 Local atomic descriptor: definition and examples . . . 20
    - 2.1.2.2 Construction of design matrix . . . 24
  - 2.1.3 Regression model . . . 27
    - 2.1.3.1 Linear model . . . 27
    - 2.1.3.2 Nonlinear model . . . 28
    - 2.1.3.3 Loss function and error evaluation . . . 30
- 2.2 Kernel regression . . . 36
  - 2.2.1 Kernel model: definition and formulation . . . 38
  - 2.2.2 Selection of sparse points . . . 40
  - 2.2.3 Examples of kernel functions . . . 46
    - 2.2.3.1 Squared-exponential, polynomial and Mahalanobis kernels . . . 47
    - 2.2.3.2 2- and 3-body kernels . . . 48
  - 2.2.4 Fourier-sampled kernels . . . 51
    - 2.2.4.1 Invariant n-body kernels and Fourier sampling . . . 51
    - 2.2.4.2 Fourier-sampled n-body descriptors . . . 55
    - 2.2.4.3 High-order kernels: incomplete but very fast kernels . . . 60
- 2.3 Short-range correction: Ziegler-Biersack-Littmark potential . . . 61
- 2.4 Conclusion of the chapter . . . 64

3 Enhanced sampling methods for free energy landscape **65**
- 3.1 Free energy estimation: general perspectives . . . 66
  - 3.1.1 Preliminaries in statistical physics . . . 67
    - 3.1.1.1 Canonical ensemble . . . 67
    - 3.1.1.2 Sampling in the canonical ensemble . . . 68
    - 3.1.1.3 Harmonic approximation and its limitations . . . 69

3.1.2 Including anharmonicity in free energy . . . 70
3.1.2.1 Thermodynamic integration-based calculations . . . 71
3.1.2.2 Free energy perturbation and adaptive sampling . . . 74
3.2 Accelerated Bayesian adaptive biasing force method . . . 77
3.2.1 Bayesian adaptive biasing force: theoretical framework and implementation . . . 77
3.2.2 Optimizations of BABF method . . . 81
3.2.2.1 Reference system: choices and numerical instabilities filter . . . 81
3.2.2.2 Acceleration of convergence: reweighting sampling . . . 86
3.2.3 Validation: comparison with MD simulations and TI . . . 87
3.3 Bound BABF method: sampling a metastable state . . . 91
3.3.1 Motivation: sampling free energy landscape of defects . . . 91
3.3.2 Constrained configuration space exploration without bias . . . 93
3.3.3 Parallelization based on shared bias . . . 97
3.4 Fluid-phase free energy computations . . . 99
3.5 Conclusion of the chapter . . . 101

4 Applications **105**
4.1 Applications to bcc tungsten . . . 106
4.1.1 High-temperature thermodynamic properties . . . 107
4.1.1.1 Performance of EAM potentials . . . 109
4.1.1.2 Prediction of elastic properties based on existing ML force fields . . . 110
4.1.1.3 Improvement of ML force fields . . . 112
4.1.1.4 Prediction model of elastic properties . . . 113
4.1.2 Free energy landscape of vacancies . . . 114
4.1.2.1 Experimental observations of void formation . . . 116
4.1.2.2 ML force fields for vacancy study . . . 117
4.1.2.3 Binding free energy of di-vacancies . . . 120
4.2 Applications to bcc Ta-Ti-V-W high-entropy alloys . . . 123
4.2.1 ML force fields for Ta-Ti-V-W HEAs . . . 124
4.2.2 Finite-temperature elastic properties . . . 126
4.2.3 Finite-temperature formation free energy of vacancy . . . 127
4.3 Correlation between harmonic and anharmonic contributions . . . 129
4.4 Conclusion of the chapter . . . 133

5 Conclusion and outlook **135**

Résumé étendu en français **139**

Chapter 1

Introduction

Multi-scale modeling stands as a pivotal paradigm in materials science, offering a comprehensive framework to understand and predict the behavior of materials across varying length and time scales. By integrating knowledge from atomistic, mesoscopic, and continuum levels, multi-scale modeling enables researchers to elucidate complex material phenomena, ranging from atomic rearrangements and phase transformations to macroscopic mechanical responses and material failure. Multi-scale modeling seeks to bridge the gap between fundamental atomic interactions and macroscopic material properties, recognizing that phenomena occurring at different length and time scales are inherently interconnected. As presented in Fig. 1.1, computational modeling and simulations grant access to length and time scales unachievable by experiments, thereby enabling a deeper understanding of macroscopic material properties and behaviors, and guiding the design of novel materials.

The development of multi-scale modeling techniques has been propelled by advances in computational power and simulation algorithms. These techniques encompass a diverse array of methodologies, including *ab initio* calculations of electronic structures, atomic-scale molecular dynamics (MD), mesoscopic Monte Carlo (MC) simulations and dislocation dynamics, and finite element analysis of macroscopic objects, each tailored to address problems at specific length and time scales. In general, the application of methods at smaller scales is limited by the computational cost associated with their high accuracy. The objective of this work is to overcome this limitation, enabling atomic-scale simulations to achieve *ab initio* accuracy and providing properties that emerge in the macroscopic assessment of materials. To this aim, we mainly focus on interatomic potentials, which serve as the "engine" of atomic-scale simulations. We demonstrate the improved design and applications of reinforced interatomic potentials based on machine learning (ML) techniques.

In this chapter, we briefly introduce the background of this work. The concept of interatomic potentials and the limitations of their traditional design are presented in Section 1.1, allowing readers to understand why improvements are indispensable. In Section 1.2, we show how artificial intelligence (AI) benefits the physics and materials science communities, particularly in the design of interatomic potentials. With better tools available, it is essential to find effective ways to use them, which calls for improved simulation methods in this context. Thus,

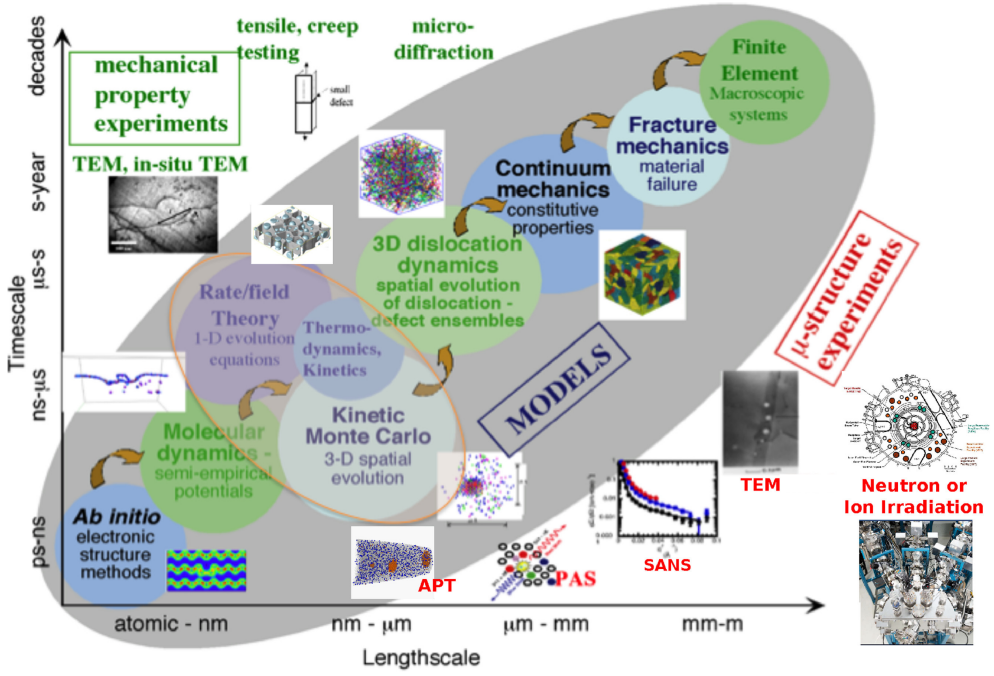

Figure 1.1: Illustration of an integrated experimental and computational approach for multi-scale investigation of materials [1]. The multi-scale modeling methods are depicted within the gray region.

in Section 1.3, we present the scheme of free energy sampling techniques, where ML-based potentials can be employed to predict material properties. Finally, in Section 1.4, we outline the general framework of this thesis, which includes novel ML-based interatomic potentials and enhanced free energy sampling approaches, for the accurate prediction of finite-temperature properties of metallic materials.

1.1 Interatomic potential: traditional methods and limitations

In classical mechanics, the evolution of a system is determined by the distribution of local minima and saddle points on the potential energy surface. Configurations located in the minima drive the system's thermodynamics, while the saddle points and the pathways connecting the local minima govern the system's kinetic evolution. In atomic-scale simulations, the potential energy surface is determined using an interatomic potential (also known as a force field), where the energy of interaction between atoms is expressed as a function of atomic coordinates, with electronic degrees of freedom only implicitly accounted for. The connection between local atomic energy and local atomic environments was established early in the field of atomistic materials science. For metals, the tight-binding approximation [2, 3] formalized the basis of this relationship. With the advent of semi-empirical potentials [3, 4, 5], the second moment tight-binding model was replaced by ad-hoc local functions that are fitted to bulk properties, defect formation energies, migration energies, and other characteristics. Not limited to metals, the functional form of the local energy as a function of local coordination forms the

basis of empirical many-body force fields. These functions have simple analytic forms, such as the number of first and/or second neighbors, radial functions [6, 4, 7, 8, 9, 10, 11, 12, 13, 14], or somewhat more complex functions that account for angular information [15]. Regardless of their analytic form, all these functions have the same utility and they provide fingerprints of atomic environments. For example, in metals where the functional form of the density of states is relatively simple, Embedded Atom Method (EAM) potentials [16, 6, 4, 17, 18, 19] are valid and widely applied. EAM provides an accurate description of the complex many-body interactions by incorporating the effects of the local atomic environment. In this way, the total potential energy of a system is expressed as a sum of local contributions from individual atoms, where the energy of each atom is a function of its local electron density, which is influenced by its neighboring atoms. For materials subjected to irradiation, EAM potentials have yielded many important results. A bibliographic review covering the past decade of atomistic studies on metals crucial for current (e.g., fission) or future (e.g., fusion) power generation reveals over 1500 studies utilizing EAM potentials for iron (Fe) [8, 13, 20] and tungsten (W) [21, 22]. Given this success, one might question why there is a need to explore alternative methods for parameterizing interatomic interactions in complex systems. There are at least two major reasons.

Firstly, the fitting process of a traditional potential, such as EAM, is often limited and rigid. As computational power continues to increase, *ab initio* approaches frequently reveal new, previously hidden aspects of defects, necessitating the integration of these findings into empirical models. Consequently, the task of developing empirical potentials becomes increasingly complex and labor-intensive. Improving certain properties often comes at the expense of others, suggesting that the underlying formalism should be changed to enhance the flexibility of the functional form. However, such changes should not result in overfitting or compromising the fundamental physics underlying the potential. Over the past three decades of effort, various empirical potentials, including Modified EAM (MEAM) [23], Bond-Order Potentials (BOP) [24], Reactive Force Field (ReaxFF) [25], and Charge-Optimized Many-Body (COMB) potentials [26], have been developed. Although these potentials generally offer greater accuracy than EAM in comparison with *ab initio* calculations, they also encounter issues related to transferability. Transferability refers to the ability of a potential to accurately describe a wide range of different systems and conditions beyond those used in its initial parameterization. A potential with high transferability can reliably predict properties across various materials (e.g., alloys with the same base elements but different compositions or atomic percentages) and environments, whereas one with low transferability may perform well only within a limited set of conditions. In this context, exploring new fitting solutions and support functions inspired by the AI and ML communities offers promising avenues to overcome the limitations of classical potentials.

Secondly, physical models constrain the scalability of numerical methods [27]. A detailed analysis of existing numerical methods in materials science reveals a gap between less accurate empirical methods, which scale as N_s^2 or lower (N_s being the number of atoms in system s), and more accurate electronic structure calculations, which scale as N_s^3 or higher. Although electronic structure methods, such as tight-binding or hybrid quantum mechanics/molecular mechanics (QM/MM) approach, attempt to bridge the gap between *ab initio* $\geq N_s^3$ methods and empirical $\leq N_s^2$ methods, they often fall short and retain unfavorable scaling. The rigorous parameterization and increasing accuracy of AI/ML methods provide a possibility to bridge

this gap.

Building upon these two considerations, the subsequent section will elucidate the application of AI/ML methodologies in materials science, especially focusing on the development of interatomic potentials. The present-day ML potentials [28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38] propose a direct multivariate regression in the descriptor space, between the local atomic environments and the atomic energy.

1.2 Machine learning approaches: the renaissance of empirical potential

Nowadays, artificial intelligence and machine learning influence many fields of physics and materials science. However, AI/ML methods cannot fully replace traditional approaches in these disciplines. In the fields of physics and materials science, dynamical systems are endowed with a well-defined Hamiltonian and structure, and their evolution within their phase space is governed by the Schrödinger equation for states (wave functions) or the Liouville-von Neumann equation for the density operator.

This space is too vast and complex to be adequately described solely by the inherent statistical correlations within the data points. To our knowledge, no statistical methods alone, generically called ML and its subclass deep learning (DL), can provide a valuable alternative to the laws of physics. To provide reliable results in the field of physics, ML and DL should be trained on robust, coherent data provided by well-established methods from the physics community. Statistical methods trained on physical datasets can be immensely helpful when traditional methods are limited or their direct application is hindered by factors such as high computational costs or insufficient computer memory. The synergy of ML approaches with traditional methods opens up numerous research opportunities in materials science. In particular, it enhances materials modeling by providing access to crucial physical properties, such as accurate energy within molecular dynamics trajectories [39, 40, 41], free energy sampling with *ab initio* accuracy [42, 43, 44], formation and migration energy of large defects such as straight dislocations, kink pairs, loops, and large 3D clusters [45, 46], as well as the investigation of continuum mechanics [47]. Similar approaches could be applied to large systems of biological and chemical molecules. Statistical methods help overcome the limitations of traditional methods from the community of physics and bridge the gap between different length and time scales, thereby enabling further progress and developments in these fields.

The first attempt to couple AI and high-dimensional problems in atomic-scale materials science was proposed by Behler and Parrinello in 2007 [28]. In contrast to traditional force fields, where the performance and limitations of the potential are mainly defined by the physical formalism, the performance and accuracy of ML potentials are determined by three equally important components: the database containing density functional theory (DFT) calculations (i.e., *ab initio* calculations), the representation of atomic environment in descriptor space, and the regression algorithm used for fitting. This feature is indicated in Fig. 1.2.

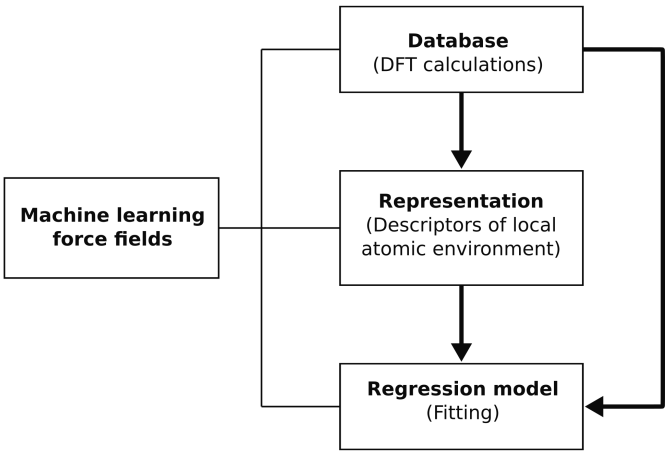

Figure 1.2: Three important components that determine the performance and accuracy of machine learning potentials.

Like all ML algorithms, ML potentials require an extensive training database, as its content significantly impacts the accuracy and transferability of the potential. The design of the database, including the selection of relevant information and the choice of pertinent instances [48, 49], referred to as sparsification, is a crucial step in developing an effective ML potential.

Atomic descriptors provide a specific numerical representation of atomic structures from the database and offer an invariant description with respect to the symmetries of the Hamiltonian of the system (e.g., permutation of identical atoms, rotation, translation). Thus, rather than using an $\mathbb{R}^{3N_s}$-dimensional description of all local atomic environments in a system containing N_s atoms, one employs a descriptor space in $\mathbb{R}^D$, where D represents the dimensionality of the descriptor space. The value of D typically ranges from a few tens to a few thousand [50, 51, 52]. Since each atomic environment is described individually, this space has a fixed dimensionality that remains constant regardless of the system size. The descriptors encode the local geometry of neighboring atoms using various methods, basically including distances and/or angles between atoms [28, 27, 50], spectral analysis of local atomic environments [27, 50] and tensorial description of atomic coordinates [53, 54]. Besides, Mallat *et al.* [55, 56] proposed innovative descriptors based on the scaling wavelets transformation. Following more specialised approaches, Bruneval *et al.* [57] used physical observables such as Mulliken charges in order to define quantum mechanical-informed descriptors and Yeo *et al.* [58] adopted partial histograms of electronic density of states. Moreover, a systematic basis that preserves the symmetry of the potential energy function with respect to rotations and permutations can also be developed by expressing the total energy as a sum of atomic body-ordered terms in permutation-invariant polynomials [59, 60]. The similarity distance descriptors characterize the distances between pairs of atomic environments. Examples include the Smooth Overlap of Atomic Positions (SOAP) [50] and the graph version [61, 62], which is defined through a functional representation of atomic positions. Similar to SOAP but employing a different framework, the Atomic Cluster Expansion (ACE) can be used to construct a complete basis of invariant polynomials by combining radial and spherical harmonic functions [36, 37]. The framework of deep learning neural networks (NNs) itself can also be employed to construct

descriptors [63, 39, 64, 65, 66, 67, 68, 69]. Recently, the use of Graph Neural Networks (GNNs) for embedding local environments has gained significant popularity, especially for surrogate models [70, 71, 72, 73, 74].

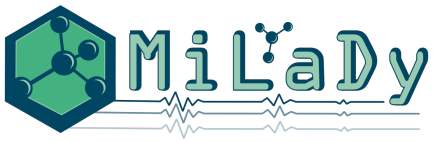

Figure 1.3: Logo of `MiLaDy` package. The full introduction and utilities of the code can be found at online documentation.

The fitting of ML potentials is conducted in descriptor space, where the statistical ML procedure used for the fit defines the performance and limitations of the potential. The relationship between atomic energies and the components of the descriptors can be either linear [75, 34, 35, 76, 77, 78, 38, 79] or non-linear. Non-linear relationships are most commonly modeled using dense NNs [28, 31, 32, 80, 81], invariant/equivariant GNNs [82, 83, 84, 85, 86, 82, 87, 88, 89, 90] or kernel methods [91, 92, 93, 94, 95, 96, 40, 97]. Using a linear kernel corresponds to performing linear regression, whereas a polynomial kernel is equivalent to linear regression with a basis set composed of the outer products of the elements of the feature vectors [98]. Some kernel models are formalized within the ever-growing field of statistical on-the-fly learning methods [96, 97, 42], while others are built in the form of potentials, such as Gaussian Approximation Potentials (GAP) [27, 33], which is a widely used variant of kernel-based potentials. Generally speaking, highly non-linear methods are suitable for interpolating multivariate functions, but they often exhibit poor performance in the extrapolation regime [99, 100]. This inconvenience can be partially mitigated through well-chosen regularization, continuous augmentation of the database, or by employing on-the-fly active learning techniques [97, 42, 101] to constantly expand the boundaries of the interpolation regime. However, modeling complex energy landscapes, such as those encountered in the study of irradiation damage or phase transformations, often necessitates extremely large fitting databases to maintain accuracy. In this context, sparsification approaches can reduce the database size and enhance regression and evaluation efficiency [102] without impairing extrapolation performance.

The general workflow for constructing a ML force field based on a given database involves the following steps: a user-specified descriptor is computed for the local atomic environments consisting of the atomic coordinates in the database. Then the descriptors in the training dataset are fitted to the DFT values using a user-specified regression model to determine the model parameters. The model is validated on a testing dataset to prevent overfitting, and further validated on its prediction of basic physical parameters to ensure the preservation of physical nature. Once the model is validated in terms of both statistical accuracy and physical validity, the ML potential can be employed in simulations.

Constructed from a database of DFT calculations, ML force fields enable atomic-scale simulations to achieve *ab initio*-level accuracy. However, it is important to emphasize that ML potentials are typically orders of magnitude slower than empirical potentials, although they remain orders of magnitude faster than DFT calculations [103].

`MiLaDy` (Machine Learning Dynamics) is a package designed by M.-C. Marinica and A. M. Goryaeva at SRMP, CEA Saclay, for the construction and application of machine learning force fields (Fig. 1.3). All algorithm developments for ML potentials in this thesis were implemented within this framework, and all newly built ML potentials were created using this toolkit.

1.3 Advanced sampling schemes and estimation of free energy

The thermodynamic properties of materials, including heat capacity, thermo-elasticity, and phase stability, are critical benchmarks in materials design, as they dictate the performance and practical applications of a given material. Consequently, this knowledge is indispensable for comprehending the behavior of materials at finite temperatures. Experimental measurements of the thermal quantities are often time-consuming, expensive or even unfeasible under extreme conditions, e.g., at high temperatures and/or pressures. Atomic-scale simulations are therefore widely used to predict the thermodynamic quantities of practical interest and/or extrapolate them beyond experimental conditions.

Thermodynamic properties are well characterized by the free energy and its derivatives. In crystalline solids, an accurate representation of the free energy includes three contributions: (i) the contribution accounting for chemical disorder and including configurational entropy, which can be computed using MC simulations or approximated by assuming that the lattice is rigid, (ii) the contribution of the electronic excitations, and (iii) the harmonic and anharmonic contributions of the lattice vibrations, i.e., the interactions of phonons with themselves and with other modes of excitation. In this thesis, we thus focus on the vibrational contribution to free energy. Its harmonic part can be straightforwardly obtained by computing the phonon spectrum resorting to the harmonic or quasi-harmonic approximations [104, 105, 106]. These two approximations, however, become inaccurate at elevated temperatures, where phonon softening and broadening need to be considered [107].

The anharmonic free energy contribution, which is crucial for deriving the finite-temperature properties, can be directly evaluated by thermodynamic integration (TI) from a suitable reference system [108, 109, 105]. In TI, the first derivative of free energy is first estimated using a sampling algorithm and then integrated. TI-based calculations of thermodynamics properties including the impact of anharmonicity from accurate electronic structure calculations, i.e., *ab initio* calculations, were initiated in 2001 [110, 111]. Such brute-force integration from electronic structure calculations is often computationally prohibitive in practice, as it necessitates sampling too many configurations along the integration path [42]. Therefore, several improvements have been proposed to make TI-based methods more feasible and amenable, particularly for electronic structure calculations. Notably, upsampled thermodynamic integration using Langevin dynamics (UP-TILD) method was developed [112], which enhances the performance of TI by combining the DFT calculations using "reduced" DFT parameters (energy cutoff of kinetic energy and k-points sampling of the Brillouin zone) with an almost configuration-independent offset with respect to the fully converged energy, where only a small number of configurations are required to evaluate the fully converged term. Based on UP-TILD, an improved version, referred to as two-stage upsampled thermodynamic integra-

tion using Langevin dynamics (TU-TILD), was developed to further accelerate the convergence of calculation [113], wherein TI is split into two stages, first from the harmonic to an intermediate potential, and then from the intermediate potential to the full DFT Hamiltonian. The TU-TILD scheme has recently been applied with moment tensor potentials [114], a class of ML potentials that have demonstrated good accuracy and efficiency [42, 43]. However, these approaches assume that the points sampled from the reference distribution faithfully represent the target distribution associated with the full Hamiltonian. This assumption can be true for high quality reference or intermediate potentials, such as the moment tensor potentials within the ML framework, but is not very reliable when using simple reference systems such as harmonic or quasi-harmonic Hamiltonians [42]. Generally speaking, when ML potentials are chosen as intermediate potential, the free energy from the reference should be computed with great accuracy. Nevertheless, this can be a crude task because, at least for crystalline materials, the ML force fields are from a few tens times slower up to four orders of magnitude slower in terms of CPU times than, for instance, traditional EAM potentials.

In addition to the integration of the first derivative, the free energy difference between two distinct Hamiltonian states can be computed directly by implementing the free energy perturbation (FEP) method. This entails the sampling of one of the two states and the estimation of a partition function ratio [115]. In practice, the degree of accuracy of the FEP method is controlled by the extent of overlap between the reference and target distributions [116]. High energy barriers limit the ergodic sampling of phase space by inhibiting transitions between connected basins of attraction and preventing the system from exploring the important regions of phase space that contribute most to the free energy difference. To enhance numerical ergodicity, substantial effort has been directed at artificially favouring the rare excursions to important regions. The central idea of the so-called importance-sampling techniques [117] is to sample from another distribution, one that exhibits good overlapping properties with both the target and reference distribution. The first importance-sampling technique used in free energy calculations has been dubbed umbrella sampling [118, 119], precisely because the sampling distribution specified by the biasing potential should cover simultaneously the region of configuration space relevant to both the target and reference systems. Establishing such a biasing potential that provides good overlapping properties is a complex task, and significant progress has been made in this area. In particular, the adaptive dynamics employing an on-the-fly adjustment of biasing potential reduce significantly the intervention from the user. In a MC run, the adaptive process can be realized by the Wang-Landau algorithm [120, 121], that is, modifying the MC acceptance probability every time a new configuration is visited. For MD simulations, metadynamics was proposed, wherein the system evolution is biased by a time-dependent potential constructed as the sum of Gaussian functions [122, 123, 124, 125]. Inspired by metadynamics, temperature accelerated molecular dynamics (TAMD) [126, 127] was subsequently introduced. This method demonstrates that sampling can be accelerated by applying an artificially high temperature to the collective variables. Recently, Swinburne *et al.* introduced an analytical mean-force model able to directly compute the anharmonic free energy of a general bond lattice within meV/atom of accuracy [128]. Although computationally inexpensive, this model is restricted to perfect crystals and difficult to extend in presence of imperfections.

Despite the advancements, accurate and general determination of thermodynamics properties remains challenging, since those quantities of interest rely on the numerical convergence of

first and second derivatives of free energy, which requires considerable precision and stability of the algorithm. Within the framework of adaptive algorithms, the adaptive biasing force (ABF) method [129, 130, 131, 132, 133] combines the computationally expensive TI with the importance sampling strategy, giving access to the numerically exact free energy profile. In the ABF technique, the on-the-fly estimate of the free energy gradient from TI is used as the biasing force, which corresponds to the gradient of the biasing potential. It has been rigorously proved that the free energy is obtained in the long time limit, and the convergence is exponentially fast in time [134]. In recent decades, various improvements of ABF have been proposed, including extended ABF (eABF) method [135, 136, 137, 138] which utilizes fictitious variables that are harmonically coupled to the transition coordinates instead of the original coordinates, as well as metadynamics-ABF hybrid technique [139] that is augmented with machine learning techniques, such as Gaussian process regression [140] and deep learning [141, 142]. The latter is often applied in biological research of organic molecules, such as protein folding. Moreover, a Bayes-formula estimator has been previously used in the eABF context in combination with MD [143] or MC [144] simulations, which allows to systematically reduce the statistical variance when the converged biasing force is frozen [145], compared with other standard estimators. However, the Bayesian formulation of TI has never been profiled and used to compute the anharmonic free energy of a crystalline solid. We present a rationale for why this formulation offers numerous advantages, including rapid convergence and robustness, which are crucial for an adaptive sampling algorithm aimed at estimating free energy.

1.4 Machine-learnt free energy profile: introduction to the thesis

This thesis introduces a comprehensive computational framework that integrates ML force fields with advanced sampling methods to accurately estimate free energy and predict the finite-temperature properties of materials, achieving a level of precision comparable to that of DFT calculations. The DFT-level accuracy of this framework is ensured by ML potentials. However, as mentioned in Section 1.2, their use in atomic-scale simulations is limited due to high computational costs, being slower by a factor of several tens to four orders of magnitude compared to empirical potentials. To make the investigation of free energy profiles from computationally expensive ML potentials feasible at elevated temperatures, this thesis focuses on two main objectives: (i) developing ML potentials with an improved accuracy-efficiency trade-off, and (ii) creating faster and more robust sampling methods for free energy computations.

This thesis is divided into three main chapters:

- Chapter 2 addresses the first objective, detailing the three essential components of ML force fields (Fig. 1.2) in Section 2. We propose optimizations for constructing ML potentials based on these components. The main developments of this work, focusing on kernel regression, are presented in Section 2.2. Additionally, a correction strategy for short interatomic distances is introduced in Section 2.3.
- Chapter 3 focuses on the second objective, wherein Section 1.3 presents the fundamental

concepts and important existing approaches within the framework of free energy calculations. Building on this, we introduce a Bayesian adaptive sampling algorithm for the fast evaluation of anharmonic free energy in Section 3.2. An extension of this method for sampling a specified metastable state is presented in Section 3.3, and Section 3.4 briefly demonstrates how to sample the free energy profile of a liquid system.

- Chapter 4 covers applications of the current ML-based free energy sampling framework. Using this toolkit, we investigate the thermodynamic properties and stability of vacancies in tungsten (Section 4.1) and Ta-Ti-V-W high-entropy alloys (Section 4.2) over a wide temperature range up to their melting points. The results for tungsten show great consistency with experimental observations. Moreover, in Section 4.3, we identify and validate a direct shortcut from harmonic approximation to full anharmonic free energy for tungsten and iron, benefiting from the proposed method.

Briefly, the first two main chapters elaborate on the methodological innovations of this thesis, while the final chapter demonstrates the practical utility of these methods by addressing real-world problems in materials science.

Chapter 2

Machine learning force fields

Contents

2.1 Basics of machine learning force fields 12
2.1.1 Database for machine learning force fields 15
2.1.1.1 Database from DFT and construction of target vector . . . 16
2.1.1.2 Active learning and database updating 18
2.1.2 Representation of local atomic environment 19
2.1.2.1 Local atomic descriptor: definition and examples 20
2.1.2.2 Construction of design matrix 24
2.1.3 Regression model 27
2.1.3.1 Linear model 27
2.1.3.2 Nonlinear model 28
2.1.3.3 Loss function and error evaluation 30
2.2 Kernel regression 36
2.2.1 Kernel model: definition and formulation 38
2.2.2 Selection of sparse points 40
2.2.3 Examples of kernel functions 46
2.2.3.1 Squared-exponential, polynomial and Mahalanobis kernels 47
2.2.3.2 2- and 3-body kernels 48
2.2.4 Fourier-sampled kernels 51
2.2.4.1 Invariant n-body kernels and Fourier sampling 51
2.2.4.2 Fourier-sampled n-body descriptors 55
2.2.4.3 High-order kernels: incomplete but very fast kernels 60
2.3 Short-range correction: Ziegler-Biersack-Littmark potential 61
2.4 Conclusion of the chapter 64

Atomic-scale modeling of materials has made remarkable progress, yet it remains fundamentally constrained by the high computational cost of explicit electronic structure methods, such as DFT calculations. This chapter demonstrates how machine learning is currently enabling a new level of realism in materials modeling. By "learning" electronic structure data, ML-based force fields provide access to atomistic simulations that achieve DFT-level accuracy but are orders of magnitude faster. Firstly, in Section 2.1, we provide a comprehensive perspective of machine learning interatomic potentials by detailing the construction process and highlighting possible optimization strategies from the three key components: the database (Subsection 2.1.1), the representation of local atomic environments (LAEs) (Subsection 2.1.2), and the regression model (Subsection 2.1.3). Following this, we focus on a specific model type, the kernel model based on Gaussian process (GP) regression, discussed in Section 2.2. This section includes the definition of the model in Subsection 2.2.1 and several essential examples in Subsection 2.2.3. Additionally, we propose optimizations for the kernel regression structure, specifically the sparse points selection method, in Subsection 2.2.2, and introduce a novel LAE representation based on the Fourier transform-sampled n-body kernels in Subsection 2.2.4. This chapter concludes by presenting a correction strategy designed to handle very short interatomic distances (Section 2.3).

2.1 Basics of machine learning force fields

The foundation of any empirical potential concept states a correlation between the LAE and the local atomic energy. Here, we define the local energy as the energy attributable to atoms located within the neighborhood or LAE of a central atom, up to a cutoff distance R_{cut} as illustrated in Fig. 2.1. The relationship between total energy and local atomic energy was established in the early days of atomistic materials science. For metals, the tight-binding approximation [2, 146, 3] formalized the basis of this relationship. According to this formalism, the total energy E_s of a system s containing N_s atoms can be expressed as the sum of the local energies $\epsilon_{s,a}$ of each a^{th} atom:

$$E_s = \sum_{a=1}^{N_s} \epsilon_{s,a}. \tag{2.1}$$

It should be noted that the above form of the total energy is a crude approximation for systems where electronic correlations are important, or in cases where charge screening is ineffective (e.g., insulators), and the charge interaction between ions cannot be neglected. Such systems require more sophisticated formalisms [147], such as explicitly accounting for long-range interactions beyond R_{cut}, which is beyond the scope of the present study. Consequently, the total energy, forces, and stress are related to their representations in the descriptor space through various functions, ranging from linear to highly non-linear, as well as all combinations in between. The advantage of this reduction is that it allows the projection of any system into a unified descriptor space. The case of simple proportionality between the local energies and the descriptor representations of the LAEs is illustrated in Fig. 2.2(a). Similar to the case of energy (Eq. 2.1), atomic forces exhibit a linear relationship with the force descriptors and are characterized by the same regression parameters, as shown in Fig. 2.2(b). Using this

type of regression, we can achieve unprecedented accuracy in energy ($< 10\,\mathrm{meV/atom}$), force ($< 8\,\mathrm{meV/Å}$) and stress ($< 20\,\mathrm{meV/Å^3}$) compared to DFT targets (Fig. 2.2(c)) [148, 49].

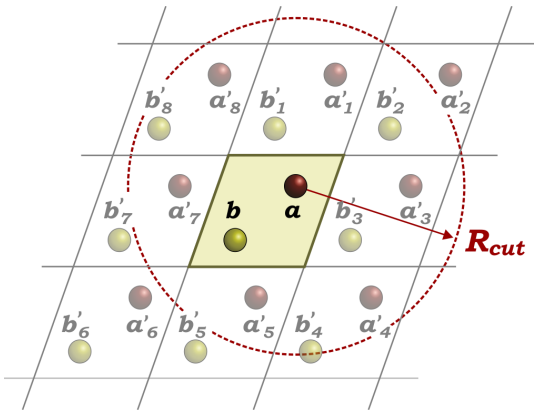

Figure 2.1: Cutoff distance in atomic systems. The derivatives of the descriptors are nontrivial to evaluate. However, we consider that any observable and their descriptor is zero outside the cutoff distance. In this case, the derivatives are easier to handle using a list of neighbors of atom a within the radius R_{cut}.

The primary distinction of the proposed approach from conventional potentials lies in its unambiguous control over the accuracy of target properties, without causing a degradation of other already adjusted properties. Furthermore, by combining descriptors with varying levels of accuracy, one can manage the trade-off between computational cost and the precision of the potential [38]. The selection of hybrid descriptors, as well as the mixing proportions of each component, should be guided by the specific physical problem and the desired accuracy in the fitting process. The user-defined control over the balance between cost and accuracy in ML potentials enables a hierarchical approach to materials property calculations. Here, we present an example of such hierarchical calculations by computing the phonon spectrum of body-centered cubic (bcc) tungsten (W) (Fig. 2.3). The accuracy of the phonons improves incrementally with the fidelity of ML potentials, corresponding to the increased accuracy of the associated descriptors. This methodology allows the generation of a set of ML potentials that achieve comparable accuracy levels for certain material properties (e.g., lattice parameter a_0 and elastic constants), while the predictive accuracy of phonons and related finite temperature properties continues to improve with higher numerical costs associated with the potential. Figure 2.3 demonstrates how enhancing the completeness of descriptors progressively enhances the capability of ML potentials to predict phonon dispersion along high symmetry directions in the first Brillouin zone. The pure bispectrum SO(4) descriptor with $j_{max} = 4.0$ determines the accuracy limit for the calculations, where minor deviations between ML predictions and experimental data primarily arise from systematic errors in the DFT calculations contributing to the W database. Among the tested ML potentials, hybrid descriptors provide reasonably accurate phonon predictions several times faster than the pure bispectrum descriptors.

The proposed hierarchical approach opens up many research avenues in the field of molecular dynamics calculations, particularly those utilizing thermodynamic adaptive sampling methods such as adapting biasing force [129, 149] or adapting biasing potential dynamics [150]. The less accurate yet faster hybrid descriptors can be employed to comprehensively explore the region of interest across phase space. By focusing on specific regions, one can gradually increase

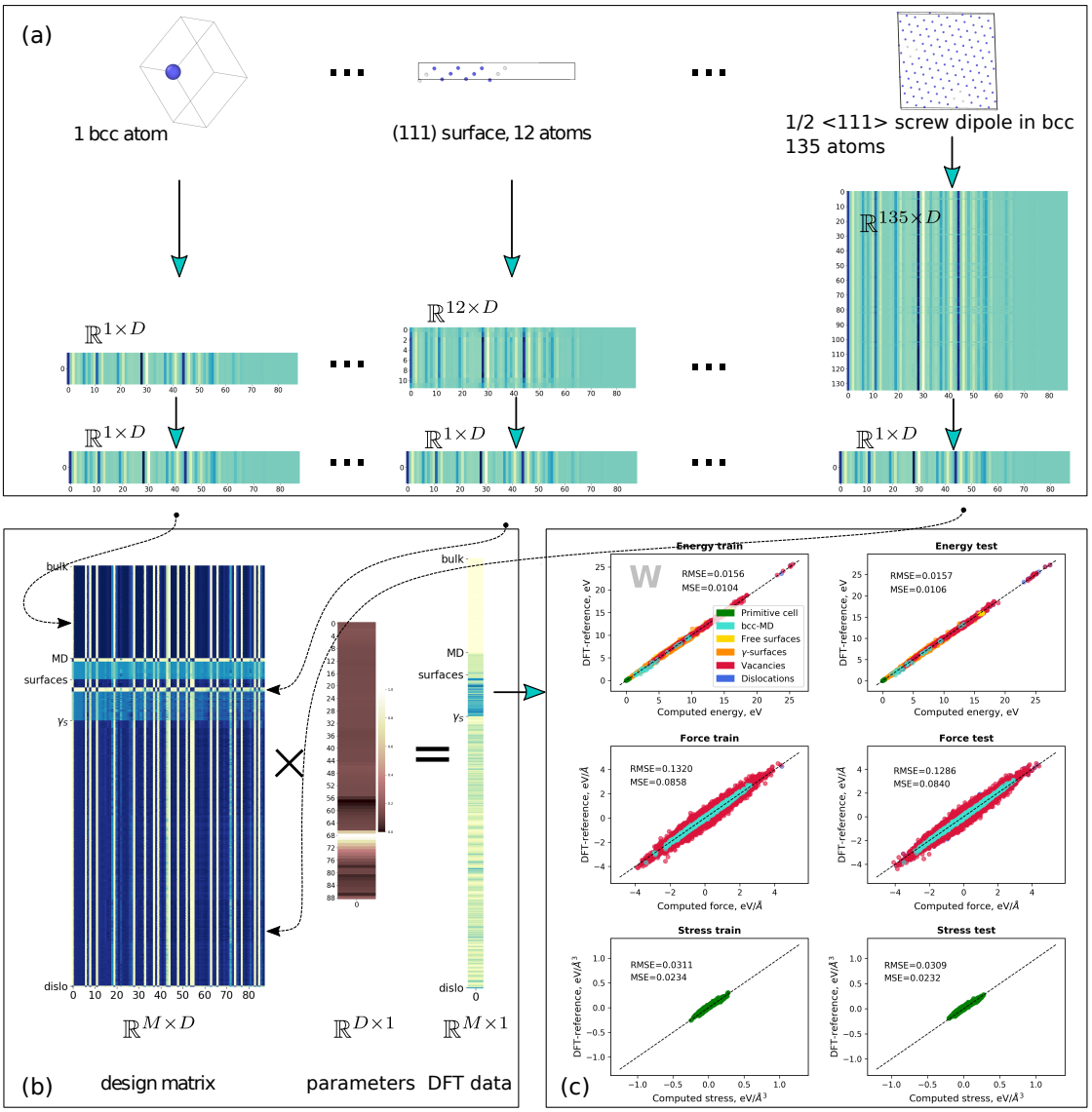

Figure 2.2: Framework of linear ML potential: (a) The projection of atomic systems into a descriptor space of dimension D. (b) After constructing the design matrix, linear regression is performed to obtain the ML potential parametrization, i.e., the regression parameters. (c) Error analysis based on a test dataset, between the ML-predicted values and the DFT target values of energy, force and stress.

sampling accuracy using more accurate but slower ML potentials. This strategy significantly improves the quality of thermodynamic sampling while keeping computational costs within acceptable limits.

The present section details how to build a ML interatomic potential step by step within the hierarchical framework of efficiency/accuracy trade-off. As mentioned in Subsection 1.2, three key factors contribute to the design of a machine learning potential: the database containing DFT calculations, the descriptors of LAEs and the regression model. In accordance with these factors, we present the method of database construction in Subsection 2.1.1, the approach to

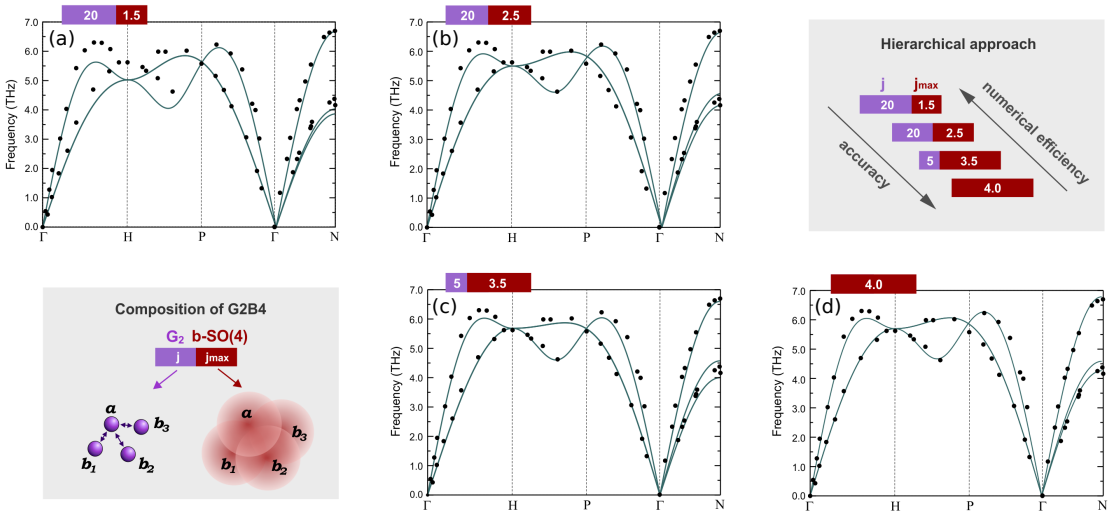

Figure 2.3: Hierarchical approach for the phonons calculations enabled by the hybrid descriptors [38]. (a-d) Phonon dispersion curves in W, plotted along high symmetry directions of the first Brillouin zone, computed using four different hybrid descriptors. Experimental data [151, 152] is depicted with black points; the LML calculations are shown with solid green lines. The inserts with gray background schematically illustrate the structure of the hybrid descriptors (bottom left corner) and their hierarchical performance (top right corner).

build the design matrix based on LAE descriptors in Subsection 2.1.2, and the procedure for determining the parameters that relate the descriptors (design matrix) to the target DFT data (target vector) in Subsection 2.1.3.

2.1.1 Database for machine learning force fields

A database plays a crucial role in the development, training, and evaluation of machine learning models. It provides the necessary data samples for training various models, as well as separate datasets for testing the trained models to evaluate their performance. For machine learning force fields in materials science, it is essential that the generation of the database is consistent and representative of the properties under investigation. In this section, we first outline the methodology employed to create the database, followed by the approach used to construct the target vector in Subsection 2.1.1.1. Subsequently, we discuss in Subsection 2.1.1.2 the maintenance and updating of the database utilizing the concept of active learning. For our first application, we focus on the case of W, but this procedure is general and can be applied to other crystalline materials.

2.1.1.1 Database from DFT and construction of target vector

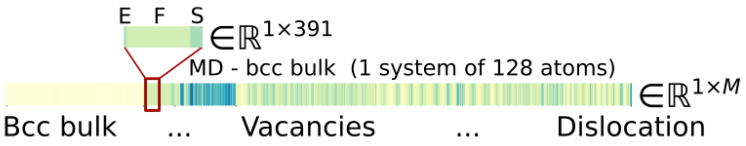

Figure 2.4: The transpose of the target vector ($\mathbf{y}^\top \in \mathbb{R}^{1\times M}$), filled by the DFT values of energy (E), force components (F) and stress components (S) system by system.

To study the behavior of materials under irradiation, the designed DFT databases for W include atomic environments relevant to the physics of various types of defects. We take into account configurations of both perfect and distorted bcc structures, as well as various types of defects, including self-interstitial atoms (SIAs), vacancies, free surfaces, γ-surfaces, dislocations and the liquid state. For most of the atomic systems, we compute the system's energy and the forces on each atom. Additionally, for some systems, the virial stress is also considered. Therefore, a system containing N_s atoms provides at most $7{+}3N_s$ observables to fit: 1 energy (M_E), $3N_a$ force components (M_F) and 6 independent stress components (M_S). The full content of an example database and the total number of ($M_E{+}M_F{+}M_S$) observables are detailed in Table 2.1. To build the target vector $\mathbf{y} \in \mathbb{R}^{M\times 1}$ with $M = M_E + M_F + M_S$, the observables are arranged in the order shown in Fig. 2.4. The DFT values are recorded system by system: for each system, we sequentially list 1 energy value, $3N_s$ force component values, and 6 stress component values (if considered), and then repeat the process for the next system.

The database is calculated with VASP 6.2.0 [153] package using projector augmented wave (PAW) pseudopotentials that account for 14 valence electrons [Xe4f^{14}]5s^{2}5p^{6}6s^{1}5d^5 for W. The exchange-correlation energy is evaluated using the Perdew-Burke-Ernzerhof (PBE) parameterization of the Generalized Gradient Approximation (GGA). The plane wave energy cutoff is set to 500 eV and the Hermite-Gaussian broadening-width for Brillouin zone integration is 0.1 eV. The k-point grid of the Brillouin zone is chosen such that each configuration in the database has a similar density of k-points and corresponds to that of the cubic unit cells of W at $a_0 = 3.1854$ Å with Monkhorst-Pack (MP) 20×20×20 grid.

The database contains information from three types of DFT calculations: (i) structural optimization at 0 K, (ii) minimum energy pathways at 0 K, and (iii) finite temperature MD calculations. The 0 K minimization is performed using conjugate gradients until the maximum magnitude of the atomic forces becomes lower than 0.01 eV/Å. The minimum energy pathway (MEP) calculations are performed using the climbing image version of Nudge Elastic Band (NEB) method [154, 155, 156] with 7-9 images and the same criterion on the maximum force as mentioned above. The MD-DFT simulations sample finite-temperature trajectories of bcc (perfect bulk or perfect bulk with a few defects) and the liquid state using MiLaDy-Lammps package [157], and the collected configurations are recomputed by VASP 6.2.0 [153]. High temperature MD-DFT simulations are used to sample the Boltzmann distribution of positions at a given temperature. The MD-DFT calculations are performed in the NVT ensemble. The time step of MD integration is set to 1.0 fs. The shape of the simulation boxes used for MD is cubic and is fixed to $4a_0{\times}4a_0{\times}4a_0$ with the value of a_0 set to the lattice parameter at 0 K from the DFT calculations. The chosen temperatures are 300 K, 1000 K and 3000 K.

Table 2.1: Details of the different classes of DFT database used for the development of ML potentials for the W system. A short description of the calculations is given in the "Content" column. The number of atoms in the cell used for DFT calculations is presented in the "Atoms per cell" column. The "Properties to fit" column corresponds to the fitted properties for training/testing (energy E, force F, stress S). The "$M_E + M_F + M_S$" column corresponds to the sum of a number of energies (M_E), forces (M_F), and stresses (M_S) in the train/test data that are used to fit/test the potential for each considered class. Finally, the total number of configurations for each considered class that have been used for the training and testing purposes are given in the "Configurations train/test" column.

DB class	Content	Atoms per cell	Properties to fit	$M_E + M_F + M_S$ train/test	Configurations train/test
1	Bcc bulk	128	E	1/0	1/0
2	Elasticity (bcc)	2	ES	525/0	75/0
3	Elasticity (fcc)	4	E	33/0	33/0
4	Deformations	1	ES	10500/3500	1500/500
5	SIA $I_1 - I_4$	129-132	E	12/0	12/0
6	Vacancies $V_1 - V_4$	124-128	E	4/0	4/0
7	Vacancy clusters $V_8 - V_{16}$	240-248	EF	17/0	12636/0
8	Free surfaces	12	E	45/15	45/15
9	γ-surfaces	12	E	100/22	100/22
10	MD - bcc bulk (300 K, 1000 K, 3000 K)	128	EFS	17595/5865	45/15
11	MD - liquid (1000 K, 3000 K)	100	EF	9632/2408	32/8
12	MD - vacancies $V_1 - V_2$ (1000 K)	126-127	EFS	11980/3477	31/9
13	MD - vacancies $V_2 - V_3$ (300 K)	125-126	EF	6330/1886	16/5
14	MD - vacancies V_5 (300 K)	123	EF	4440/1110	12/3
15	MD - (110) γ-surface with vacancy V_1 (300 K)	47	EF	14200/ 4260	100/30
16	MD - $\frac{1}{2}\langle 111\rangle$ screw dislocation (300 K, 1000K)	135	EF	40600/0	100/0
17	Migration of mono-vacancy V_1	53	EF	800/0	5/0
18	Migration of $\frac{1}{2}\langle 111\rangle$ screw dislocation	135	EF	1624/0	4/0
Total				131057/22543	2132/607

2.1.1.2 Active learning and database updating

Active learning is a specialized subset of ML wherein the learning algorithm can interactively query a user to perform specific tasks on data points. This concept can be applied to the development of ML potentials. The main idea is to enlarge the training database for the new potential by adding the information lacking in the previous potential. In the following paragraphs, we will give a simple example of database updating based on the concept of active learning, as illustrated in Fig. 2.5(a).

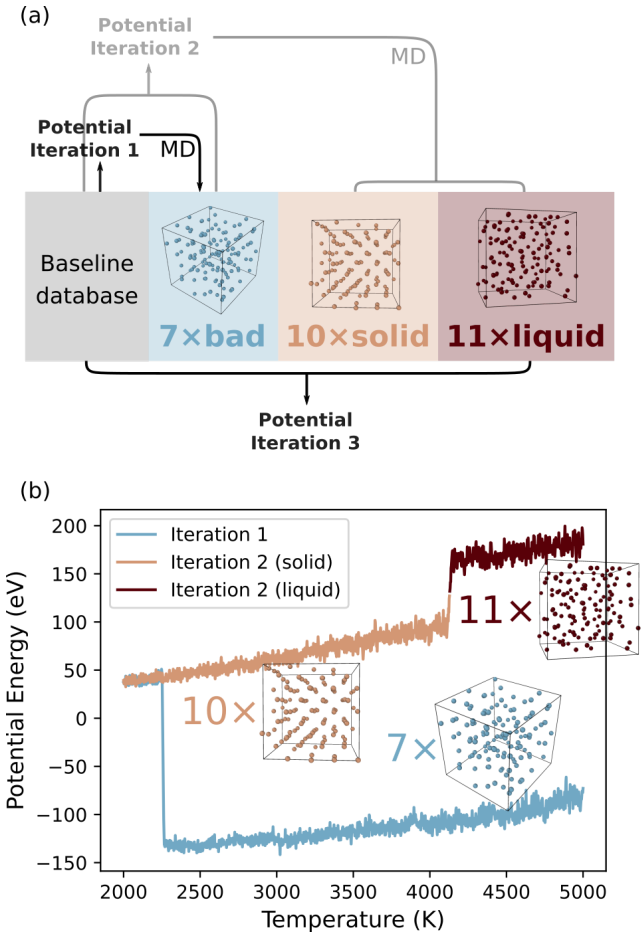

Figure 2.5: (a) Optimization strategy of ML potentials based on active learning. The main idea is to enlarge the database for the construction of new potential by adding configurations sampled with MD simulations and calculated with DFT. (b) Variation of the potential energy during an MD simulation in the NPT ensemble with increasing temperature.

Firstly, an ML potential is built upon the baseline database (Table 2.1), denoted by "Iteration 1". Then an MD simulation in NPT ensemble is carried out using this potential together with the MiLaDy-Lammps package [157]. The simulation system contains 128 atoms, and the temperature varies linearly from 2000 K to 5000 K in 150000 steps. However, this computation diverges, as illustrated by the blue line in Fig. 2.5(b), and damage to the system is observed. Clearly, this potential is poorly constructed, and additional data points are necessary for the database.

To complete the database, 7 "bad" configurations are randomly selected from the MD simulation and then recalculated using DFT with VASP 6.2.0 [158] package. Based on this database, a new potential, denoted by "Iteration 2", is constructed. As shown Fig. 2.5(b), this new potential performs much better than the previous one, holding a higher stability at high temperatures. This is actually an efficient approach to correct the stability problems in the construction of ML potentials.

Moreover, we notice that in the baseline database Table 2.1, only the configurations at 300 K, 1 000 K and 3 000 K exist, while the melting point of W is 3695 K. In response to the demand of a better performance over the whole range of temperature up to the melting point, the database can be further enlarged with another 21 configurations generated by the potential "Iteration 2", including 10 samples of bcc crystal and 11 samples of liquid phase. In this way, the database is completed by the new features that enhance the prediction of the energy landscape at finite temperatures.

Epistemic uncertainty from the predictive variance, which will be detailed in Subsection 2.1.3.3, is another strong tool for active learning [159, 160]. Computation of uncertainty (e.g., the variance of the force prediction, as subsequently formulated and illustrated in Fig. 2.8) provides quantitative estimates of the force field's accuracy for each configuration. It can then inform us which configurations have high uncertainty and should, therefore, be recalculated using DFT and added to the database. This uncertainty-driven method for selecting training points allows for the training of an accurate force field with a minimal number of relatively expensive DFT calculations.

Overall, a well-designed and maintained database is crucial for the success of machine learning potentials, as it forms the foundation upon which models are trained, tested, and deployed. To be capable of predicting the energy and forces for various types of configurations, the database should contain relevant configurations and demonstrate transferability.

2.1.2 Representation of local atomic environment

Atomic descriptor is a fundamental component for the machine learning methods applied to materials science. It serves as a representation of the LAE of an atom, considering the neighboring atoms within a defined cutoff radius, and captures crucial features necessary for accurately predicting material behaviors and properties.

In this section, we will discuss how to describe the LAE using descriptors. First, in Subsection 2.1.2.1, we will define and explain the use of local atomic descriptors in multi-scale simulations by listing various types of descriptors, with a particular focus on those that we will employ in subsequent sections. Then, in Subsection 2.1.2.2, we will introduce the method for constructing the design matrix based on these descriptors, which will be directly utilized in the regression process.

2.1.2.1 Local atomic descriptor: definition and examples

To be efficient, machine learning methods applied to materials science do not directly use the coordinates $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{N_s}) \in \mathbb{R}^{3N_s}$ of all N_s atoms within the system as input but instead employ a reduced representation of these coordinates known as a descriptor. Descriptors are functions that map the configuration space to the descriptor space. The construction of the descriptor space depends on the choice of the specific descriptor used.

We define a local descriptor centered on atom a in system s to describe the structure of N_a neighboring atoms in the neighborhood $\mathcal{V}(a)$ as follows:

$$\begin{aligned} \mathbf{D}_{s,a} : \mathbb{R}^{3\times N_a} &\to \mathbb{R}^D \\ \mathbf{q}_{N_a} &\to \mathbf{D}_{s,a}(\mathbf{q}). \end{aligned} \tag{2.2}$$

Here, $\mathbf{q}_{N_a} = (\mathbf{q}_{a_1}, \mathbf{q}_{a_2}, \ldots, \mathbf{q}_{a_{N_a}}) \in \mathbb{R}^{3N_a}$ represents the vector restricted to the neighborhood of atom a from the full atomic coordinate vector $\mathbf{q} \in \mathbb{R}^{3N_s}$. The function $\mathbf{D}_{s,a}$ projects the local environment of atom a into the descriptor space with given dimensionality D. For an extensive physical quantity that decomposes locally in an exact manner, we can construct a global descriptor of the system by summing these local descriptors:

$$\mathbf{D}_s = \sum_{a=1}^{N_s} \mathbf{D}_{s,a}. \tag{2.3}$$

This global descriptor of dimension D takes into account the entire local configurations. The sum is the simplest method for aggregating local descriptors into a global descriptor for the entire system, preserving permutation invariance; that is, the value of $\mathbf{D}_s$ does not change when the atomic ID labels a are permuted. While other aggregation methods are possible, particularly those found in the literature on GNNs [70, 71, 72, 73, 74], we will use the sum in this thesis.

More generally, the system is described by combining the local descriptor computed for each atom, which represent a transformation from the space $\mathbb{R}^{3N_s}$ to $\mathbb{R}^{D\times N_s}$. In most cases, $D \gg 3$, meaning the dimensionality of the descriptor space is much greater than that of the original configuration space. Consequently, the problem posed in the descriptor space inherently has a higher dimensionality than the original problem. The fundamental idea behind this representation is to linearize the original problem, which becomes feasible when projecting it into a sufficiently high-dimensional space. In this way, the complex topology of the data is encoded by the descriptor space, allowing the transformation of a non-linear problem into one that is sufficiently "flat" to be linearizable. The choice of projection should not be made randomly and must depend on the specific problem being studied. Descriptors must respect important symmetry properties and their components must be sufficiently non-collinear to describe a configuration in a quasi-unique manner, as the mapping $\mathbb{R}^{3N_s} \to \mathbb{R}^D$ cannot be bijective.

We will now present the main symmetry properties that a descriptor must satisfy. We denote $\mathfrak{G}$ as the symmetry group of the system under study. The most important symmetry groups in materials science include: (i) the permutations group, ensuring equivalence for exchanges of two identical particles; (ii) the translations group; and (iii) the orthogonal transformations group, encompassing rotations and reflections for crystalline systems. The representation of descriptors should remain invariant under transformations by the elements

of the system's symmetry group, i.e., for a local descriptor of the atomic environment centered on the atom a in the system s:

$$\forall \boldsymbol{g} \in \mathfrak{G}, \quad \mathbf{D}_{s,a} \circ \boldsymbol{g} = \mathbf{D}_{s,a}. \tag{2.4}$$

Here, $\circ$ indicates the composition operator. In this thesis, symmetry composition will be utilized; however, recent literature has made significant efforts towards employing equivariant methods to preserve the aforementioned physical symmetries [161, 162, 163].

The concept of a descriptor that we have discussed so far is still quite abstract. Now, we will present some examples of local atomic descriptors that satisfy the conditions mentioned above. The simplest one is the **coordination number**. The coordination number meets all the symmetry and invariance requirements of the system and provides qualitative information about the local environment of atom a. For instance, analyzing the coordination number can help detect the presence of a vacancy or an interstitial near the atom a. However, this information is not quantitative, meaning that the coordination number alone is insufficient for constructing an accurate local energy model for atom a. To develop a more precise model, we need to use a descriptor that includes more information than just the coordination number.

Another simple example of a local descriptor for quantitative analysis is the **Coulomb matrix** [164, 165, 166]. The Coulomb matrix is typically employed for small systems where electrostatic interactions play a significant role, such as in organic molecules. In this case, the system coordinates are replaced by the following matrix:

$$C_{ab} = \begin{cases} Z_a^{2.4}, & \text{if } a = b \\ \frac{Z_a Z_b}{|\boldsymbol{q}_a - \boldsymbol{q}_b|}, & \text{else} \end{cases} \tag{2.5}$$

where Z_i and $\boldsymbol{q}_i$ represent the electric charge and the position of atom i, respectively. The Coulomb matrix respects the symmetry and invariance properties of the system. Unlike the coordination number, the Coulomb matrix contains enough information to quantitatively predict various observables of the system, such as its energy [164, 165, 166]. In the literature, a wide range of more complex and varied atomic descriptors have been developed. Here, we will describe the main categories of these descriptors: (i) radial and angular descriptors, (ii) descriptors based on neural networks, (iii) spectral descriptors, and (iv) tensor descriptors.

Radial and angular descriptors. The analytical development of descriptors was pioneered by Behler and Parrinello [28, 31, 167], as well as Bartók *et al.* [33, 27, 50], and it serves as the current basis for local descriptors used in materials science. These descriptors encode the local geometry of neighboring atoms by considering radial distances and triplet angles among atoms in a system. Examples include the Behler-Parrinello symmetry function [28, 31, 168, 76, 77, 169] and the Angular Fourier Series (AFS) descriptor [33, 27, 50]. Typically, these descriptors consist of between 10 and 100 components. Additionally, Permutationally Invariant Polynomial Descriptors (PiP) [59, 60] and tabulated Gaussian Approximation Potential (tabGAP) [170] also account for bonds, angles, triangles, and higher body-order interactions, such as 4-body and 5-body.

Descriptors based on neural networks. Descriptors can also be specifically tailored for a particular system [171, 172, 61, 93], and in such cases, they can be directly created using DL

methods [63, 173, 174, 175]. In this scenario, the descriptors are automatically generated by NNs, which inherently construct symmetries and invariances. Initially introduced in 2017 [176], GNNs have become the most popular DL method for structural analysis in chemistry, biology, and drug design applications. Since then, numerous methods based on invariant or equivariant descriptions of LAEs have been proposed. The most popular include SchNet, a weighted atom-centered symmetry function and the deep tensor neural network [173, 177]; DimeNET - Directional Message Passing Neural Network [89, 90]; ALIGNN - Atomistic Line Graph Neural Network [178]; PAINN - Polarizable Atom Interaction Neural Network [82]; Nequip - Neural Equivariant Interatomic Potentials [86]; and MACE - Higher Order Equivariant Message Passing Neural Networks [87, 88, 179]. These GNN-based approaches have demonstrated remarkable performance in accurately predicting and characterizing various chemical properties, making them primarily utilized in the domains of chemistry and biology. However, their numerical cost is several orders of magnitude higher than that of traditional descriptors. Consequently, in the field of materials science where LAEs are usually simpler, traditional descriptors are predominantly used [103]. It is important to note that the analytical form of these descriptors is not explicitly known and is "hidden" within the weights of the NN. As a result, this approach does not guarantee invariance to certain symmetry operations unless those symmetries are explicitly represented in the training data.

Spectral descriptors are constructed by decomposing the local atomic density onto a basis of hyperspherical functions, the properties of which are then utilized to build invariants. The form and quality of these descriptors depend on the choice of the distribution function. The atomic density can be represented as a sum of delta functions centered on each atom in the system, leading to the bi-spectrum descriptor SO(4) [50]. Alternatively, using Gaussians of finite width results in the Smooth Overlap of Atomic Positions (SOAP) descriptor [33, 27, 50]. Using Gaussians instead of delta functions ensures that the descriptor is smooth. The two descriptors mentioned above are local descriptors, but it is also possible to develop multi-scale spectral descriptors. This is the case of solid harmonic wavelet scattering representation developed by Mallat *et al.* [56, 180, 55, 181]. Here, the invariance construction is more general and is not based on the specific properties of hyperspherical functions. By multiplying solid harmonic functions with Gaussian windows dilated at different scales, one can construct a descriptor that splits the information across scales and provides scale interaction coefficients related to multi-scale physical phenomena.

Two main families of spectral descriptors can be distinguished: (i) compact descriptors, such as the bi-spectrum SO(4), typically consisting of between 10 and 50 components, and (ii) non-compact descriptors, including SOAP and scattering representation, which range from 100 to 4000 components. The compactness of these descriptors depends, in particular, on how radial and angular information are coupled in the descriptor.

We will provide a detailed description of the bi-spectrum SO(4) descriptor, which is very often utilized throughout this thesis. Based on the decomposition of local atomic density on the basis of 4-dimensional hyperspherical functions [27, 50], there exists a bijection between the real space $\mathbb{R}^3$ and the unit hypersphere $\mathcal{S}^4 \in \mathbb{R}^4$. The atomic environment of atom a is

described by its density $\rho_a(\boldsymbol{r})$ and is decomposed as follows on hyperspherical functions:

$$\rho_a(\boldsymbol{r}) = \sum_{k \in \mathcal{V}(a)} w_k \delta(\boldsymbol{r} - \boldsymbol{r}_{ak}) \tag{2.6}$$

$$= \sum_{j=0,1/2,\cdots}^{\infty} \sum_{m,m'=-j}^{j} \boldsymbol{c}_{a,j}^{m,m'} U_j^{m,m'} \left(\theta_0^{ak}, \theta^{ak}, \phi^{ak}\right), \tag{2.7}$$

where $\mathcal{V}(a)$ is the neighborhood of the atom a and the density is zero outside this neighborhood. Here $\boldsymbol{r}_{ak} = \mathbf{q}_k - \mathbf{q}_a$ is the Cartesian coordinate of the vector from the centered atom a to is neighbor atom k, w_k is a scalar depending on the chemical species, $\boldsymbol{c}_{a,j}^{m,m'}$ is the result of the scalar product between the density function centered on atom a and the hyperspherical harmonics $U_j^{m,m'}$:

$$\boldsymbol{c}_{a,j}^{m,m'} = \sum_{k \in \mathcal{V}(a)} w_k U_j^{m,m'} \left(\theta_0^{ak}, \theta^{ak}, \phi^{ak}\right), \tag{2.8}$$

where the j values can only be positive integer or half-integer, i.e., $j = 0, \frac{1}{2}, 1, \frac{3}{2}$, etc. For the maximal value of the angular moment j_{max}, the total number of components $U_j^{m,m'}$ is $\sum_{j=0}^{j_{\max}} (2j+1)^2$. The polar angles $\theta_0^{ak}, \theta^{ak}, \phi^{ak}$ are the representation of the vector $\mathbf{r}_{ak}$ on the $\mathbb{R}^4$-sphere.

Using Eq. 2.7 and the expansion coefficients $\boldsymbol{c}_{a,j}^{m,m'}$ (also called power spectrum coefficients), we can deduce the power spectral decomposition and the bi-spectrum of the atomic density centered on the atom a. We then define the components of the bi-spectrum SO(4) $\mathbf{B}$ as follows with $j \leq j_{max}$ and $|j_1 - j_2| \leq j \leq j_1 + j_2$:

$$B_{jj_1j_2}^a = \left(\boldsymbol{c}_{a,j}^{m,m'}\right)^\dagger \boldsymbol{H}^{j_1 j_2} \left(\boldsymbol{c}_{a,j_1}^{m_1,m_1'} \otimes \boldsymbol{c}_{a,j_2}^{m_2,m_2'}\right). \tag{2.9}$$

Here, $\otimes$ denotes the tensor product and $\boldsymbol{H}^{j_1 j_2}$ are related to the Clebsch-Gordan coefficients for the SO(4) group. A detailed description of these coefficients is given by Bartók *et al.* [27, 50]. The bi-spectrum SO(4) inherently ensures translation and permutation invariance. However, rotational invariance is more subtle, enabled by the properties of hyperspherical functions and the construction of the descriptor [27, 50]. Due to Clebsch-Gordan coefficients, certain components of the bi-spectrum SO(4) are zero; hence, one can utilize the non-zero components or exclusively the diagonal components, i.e., those corresponding to $j_1 = j_2$ [182, 27, 50]. The bi-spectrum SO(4) provides a highly sensitive description of the atomic environment. Even a small difference in the Euclidean norm between $\mathbf{q}_1$ and $\mathbf{q}_2$ can lead to a significant discrepancy in terms of irreducible representations, resulting in a large difference between $\mathbf{D}_s(\mathbf{q}_1)$ and $\mathbf{D}_s(\mathbf{q}_2)$. Despite its sensitivity, the bi-spectrum SO(4) descriptor suffers from a significant drawback in its construction: it is impossible to decouple radial and angular information due to their projection onto the $\mathcal{S}^4$ sphere. Qualitatively, the bi-spectrum SO(4) offers a precise angular description but requires considerable computational time due to the recurring calculation of Clebsch-Gordan coefficients, which limits the value of j_{max}.

Tensor descriptors represent a novel approach for systematically generating invariants relative to a symmetry group. Tensor methods employ a similar "trick" to scattering representation but are not confined to using a basis of hyperspherical functions. This method allows for

the creation of rotation- and permutation-invariant descriptors using polynomial basis functions [53, 54]. Descriptor construction can encompass interactions involving N_s atoms, as seen in the work of Oord and Allen [60, 59]. Special cases of this method lead to some aforementioned descriptors. For example, when hyperspherical functions serve as the basis and are restricted to second-order tensor development, the resulting tensor descriptors yield SOAP [27]. Another example is the atomic cluster expansion (ACE) descriptor [36, 37], which leverages the invariant properties of spherical harmonics and explicitly incorporates n-body interactions. Tensor descriptors typically involve a very large number of components, ranging from 1 000 to 10 000, which often makes them challenging to use directly in practical applications.

Other types of descriptors. Finally, more "exotic" descriptors can be created by concatenating several different descriptors, resulting in what are known as hybrid descriptors [38]. A concrete example of hybrid descriptors will be provided in Subsection 2.2.3.2. Additionally, it is possible to construct more "physical" descriptors. Based on surrogate models, various physically-informed descriptors were recently proposed [58, 57, 69].

2.1.2.2 Construction of design matrix

Let us consider a database $\mathcal{DB} = \{(\mathbf{x}_m, y_m) \text{ for } m = 1, \dots M\}$ with M observations y_m (as mentioned previously, $M = M_E + M_F + M_S$) for a D-dimensional representation of the data $\mathbf{x}_m \in \mathbb{R}^D$. This representation is based on the D-dimensional descriptors elaborated in the previous Subsection 2.1.2.1. Any supervised machine learning or deep learning algorithm optimizes a function f_{ML}: $\mathbb{R}^D \times \mathbb{R}^P \to \mathbb{R}$ such that

$$f_{\text{ML}}\left(\mathbf{x}_m, \mathbf{w}\right) = y_m, \tag{2.10}$$

where $\mathbf{x}_m \in \mathbb{R}^D$ represents the m^{th} data point of the database, $\mathbf{w} \in \mathbb{R}^P$ denotes the parameters and $y_m \in \mathbb{R}$ is the m^{th} target property of the database (m^{th} element of the target vector).

Here we make a precision so as to clarify the difference between the descriptor $\mathbf{x}_m$, which is the m^{th} data point of the database $\mathcal{D} = \{\mathbf{x}_m \in \mathbb{R}^D \text{ for } m = 1 \dots M\}$, and the local atomic descriptor of the atom a from the system s, denoted by $\mathbf{D}_{s,a} \in \mathbb{R}^D$, which encodes the LAE. In the current context, the descriptor $\mathbf{x}_m$ is associated with a physical observable of the system s, such as energy, atomic forces, or virial stress. These descriptors $\mathbf{x}_m$ constitute what is typically referred to as the input space of dimension D in the artificial intelligence and machine learning community. In materials science, the local atomic descriptor $\mathbf{D}_{s,a}$ serves as an intermediate step between the atomic coordinates and the descriptors $\mathbf{x}_m$ for energy, force, and stress:

- The global energy descriptor of a system is denoted by $\mathbf{D}_s \in \mathbb{R}^D$ and computed by Eq. 2.3.
- The force descriptor $\mathbf{D}^F_{s,a}$ of the a^{th} atom in the system s is obtained by applying the gradient with respect to the Cartesian coordinates of a to the global energy descriptors:

$$\mathbf{D}^F_{s,a} = -\nabla_a \mathbf{D}_s = -\sum_{b \in \mathcal{V}(a)} \frac{\partial \mathbf{D}_{s,b}}{\partial \mathbf{q}_a} \in \mathbb{R}^{3 \times D}. \tag{2.11}$$

- The virial stress descriptor of the system s is constructed based on the definition of the virial stress tensor. For each dimension i of the local atomic descriptor $\mathbf{D}_{s,a}$, denoted by $\mathrm{D}_{s,a,i}$ with $i = 1, \ldots, D$, the virial stress tensor $\boldsymbol{\tau}_s$ can be expressed as:

$$\boldsymbol{\tau}_{s,i} = \frac{1}{2V_s} \sum_{a \in s} \sum_{b \in \mathcal{V}(a)} (\mathbf{q}_a - \mathbf{q}_b) \otimes \frac{\partial \mathrm{D}_{s,b,i}}{\partial \mathbf{q}_a}. \tag{2.12}$$

  where V_s is the volume of the system s and the dimension i of the virial stress descriptor is formulated using the Voigt notation:

$$\mathbf{D}^S_{s,i} = \left(\tau_{s,i,xx}, \tau_{s,i,yy}, \tau_{s,i,zz}, \tau_{s,i,yz}, \tau_{s,i,xz}, \tau_{s,i,xy}\right). \tag{2.13}$$

  Therefore, the virial stress descriptor of the system s is defined by $\mathbf{D}^S_s \in \mathbb{R}^{6 \times D}$.

Any row of an energy, force, or stress descriptor is of dimension D and serves as a data point $\mathbf{x}_m \in \mathbb{R}^D$ in the database.

The function f_{ML} can take a simple linear form in both the parameter and descriptor spaces, as exemplified by a linear ML model, i.e., $f_{\mathrm{ML}}(\mathbf{x}_m, \mathbf{w}) = \mathbf{w}^\top \mathbf{x}_m$. Alternatively, it can be highly non-linear, as in the case of deep learning applications, such as neural networks. A general model that is linear in the parameter space can be expressed as:

$$y_m = f_{\mathrm{ML}}(\mathbf{x}_m, \mathbf{w}) = \mathbf{w}^\top \phi(\mathbf{x}_m) = \phi(\mathbf{x}_m)^\top \mathbf{w} \tag{2.14}$$

with the notations:

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_P \end{pmatrix} \in \mathbb{R}^{P \times 1} \; ; \; \phi(\mathbf{x}_m) = \begin{pmatrix} \varphi_1(\mathbf{x}_m) \\ \varphi_2(\mathbf{x}_m) \\ \vdots \\ \varphi_P(\mathbf{x}_m) \end{pmatrix} \in \mathbb{R}^{P \times 1};$$

where $\varphi_p : \mathbb{R}^D \to \mathbb{R}$ encodes the original descriptors in such a way that the initial D-dimensional descriptor space is mapped into a P-dimensional space using a collection of P functions φ_p, simply denoted by a multi-dimensional map $\phi : \mathbb{R}^D \to \mathbb{R}^P$. In general, P is greater than D, and this technique is used to increase the dimensionality of the linear fit from D to P. This applies to quadratic, cubic or general kernel methods, where $P \sim D^2$, $P \sim D^3$ or P is usually denoted as K, respectively. With the above notations, the design matrix is defined as the matrix associated with the D-dimensional database representation $\mathcal{D} = \{\mathbf{x}_m \in \mathbb{R}^D \text{ for } m = 1 \ldots M\}$ for a given P-dimensional map ϕ that encodes the descriptors. As illustrated in Fig. 2.6, the design matrix has M rows (the number of data points) and P columns, and each row represents a data point related to an observable (energy, force component or stress component):

$$\boldsymbol{\Phi}(\mathcal{D}) = \begin{pmatrix} \varphi_1(\mathbf{x}_1) & \varphi_2(\mathbf{x}_1) & \cdots & \varphi_P(\mathbf{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(\mathbf{x}_M) & \varphi_2(\mathbf{x}_M) & \cdots & \varphi_P(\mathbf{x}_M) \end{pmatrix} \in \mathbb{R}^{M \times P} \tag{2.15}$$

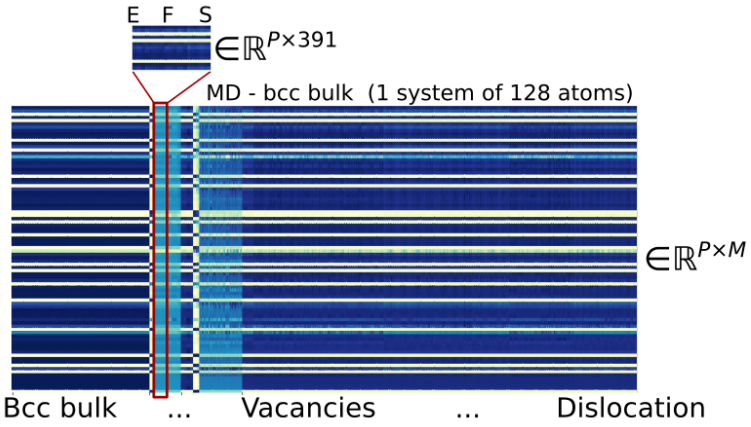

Figure 2.6: The transpose of the design matrix ($\mathbf{\Phi}(\mathcal{D})^\top \in \mathbb{R}^{P\times M}$), filled by the ϕ-encoded information of energy (E), force components (F) and stress components (S) system by system.

or in a more compact form:

$$\begin{aligned}\mathbf{\Phi}(\mathcal{D}) &= \begin{pmatrix}\phi^\top(\mathbf{x}_1)\\ \phi^\top(\mathbf{x}_2)\\ \vdots\\ \phi^\top(\mathbf{x}_M)\end{pmatrix} \in \mathbb{R}^{M\times P}\,;\\ \mathbf{\Phi}(\mathcal{D})^\top &= \begin{pmatrix}\phi(\mathbf{x}_1) & \phi(\mathbf{x}_2) & \cdots & \phi(\mathbf{x}_M)\end{pmatrix} \in \mathbb{R}^{P\times M}.\end{aligned} \tag{2.16}$$

Here are some useful elementary relations for the multiplication of the design matrix:

$$\begin{aligned}\mathbf{\Phi}(\mathcal{D})^\top\mathbf{\Phi}(\mathcal{D}) &= \begin{pmatrix}\phi(\mathbf{x}_1) & \phi(\mathbf{x}_2) & \cdots & \phi(\mathbf{x}_M)\end{pmatrix}\begin{pmatrix}\phi^\top(\mathbf{x}_1)\\ \phi^\top(\mathbf{x}_2)\\ \vdots\\ \phi^\top(\mathbf{x}_M)\end{pmatrix}\\ &= \sum_{m=1}^{M}\phi(\mathbf{x}_m)\phi^\top(\mathbf{x}_m) \in \mathbb{R}^{P\times P}\end{aligned}$$

and

$$\begin{aligned}\mathbf{\Phi}(\mathcal{D})\mathbf{\Phi}(\mathcal{D})^\top &= \begin{pmatrix}\phi^\top(\mathbf{x}_1)\\ \phi^\top(\mathbf{x}_2)\\ \vdots\\ \phi^\top(\mathbf{x}_M)\end{pmatrix}\begin{pmatrix}\phi(\mathbf{x}_1) & \phi(\mathbf{x}_2) & \cdots & \phi(\mathbf{x}_M)\end{pmatrix}\\ &= \begin{pmatrix}\phi^\top(\mathbf{x}_1)\phi(\mathbf{x}_1) & \cdots & \phi^\top(\mathbf{x}_1)\phi(\mathbf{x}_M)\\ \vdots & \ddots & \vdots\\ \phi^\top(\mathbf{x}_M)\phi(\mathbf{x}_1) & \cdots & \phi^\top(\mathbf{x}_M)\phi(\mathbf{x}_M)\end{pmatrix} \in \mathbb{R}^{M\times M}.\end{aligned}$$

Each element of the above matrix is in $\mathbb{R}$ and has the form:

$$\phi^\top(\mathbf{x}_{m_1})\phi(\mathbf{x}_{m_2}) = \sum_{p=1}^{P}\varphi_p(\mathbf{x}_{m_1})\varphi_p(\mathbf{x}_{m_2}) = \phi^\top(\mathbf{x}_{m_2})\phi(\mathbf{x}_{m_1})\,,$$

implying that the matrix $\mathbf{\Phi}(\mathcal{D})\,\mathbf{\Phi}(\mathcal{D})^{\top}$ is symmetric.

2.1.3 Regression model

A regression model is employed to establish the relationship between the atomic environment encoded in the design matrix and the target values of energy, force, and stress obtained from highly accurate DFT computations. A well-performing regression model allows machine learning potentials for atomic-scale simulations to approximate the accuracy of DFT.

Any regression that is linear in parameters can be formulated as an optimization problem of a linear system:

$$f_{\mathrm{ML}}(\mathbf{x},\mathbf{w}) \quad = \quad \phi(\mathbf{x})^{\top}\mathbf{w}\,, \tag{2.17}$$

$$f_{\mathrm{ML}}(\mathbf{x}_m,\mathbf{w}) = y_m, m = 1\ldots M \quad \Longleftrightarrow \quad \mathbf{\Phi}(\mathcal{D})\,\mathbf{w} = \mathbf{y}\,. \tag{2.18}$$

Consequently, the solution of the linear-in-parameters model is the solution of the linear system $\mathbf{\Phi}(\mathcal{D})\,\mathbf{w} = \mathbf{y}$, which can be formally written as:

$$\mathbf{w} = \left(\mathbf{\Phi}(\mathcal{D})^{\top}\mathbf{\Phi}(\mathcal{D})\right)^{-1}\mathbf{\Phi}(\mathcal{D})^{\top}\mathbf{y} = \left(\mathbf{\Phi}(\mathcal{D})^{\top}\mathbf{\Phi}(\mathcal{D})\right)^{-1}\left(\sum_{m=1}^{M} y_m\phi(\mathbf{x}_m)\right) \tag{2.19}$$

and any prediction for the data $\mathbf{x}^{\star}$ reads:

$$y(\mathbf{x}^{\star}) = \phi^{\top}(\mathbf{x}^{\star})\left(\mathbf{\Phi}(\mathcal{D})^{\top}\mathbf{\Phi}(\mathcal{D})\right)^{-1}\left(\sum_{m=1}^{M} y_m\phi(\mathbf{x}_m)\right) \tag{2.20}$$

The solution presented in Eq. 2.19 is merely a formal solution that is not typically employed in practice. The practical solutions will be detailed in Subsection 2.1.3.3. In the following subsections, we will illustrate some of the most widely used regression models in materials science.

2.1.3.1 Linear model

Linear Machine Learning (LML) fitting implies that the dimensionality of the data $\mathbf{x} \in \mathbb{R}^D$ equals the number of parameters $\mathbf{w} \in \mathbb{R}^P$, i.e., $D = P$, and the function $f_{\mathrm{ML}}(\mathbf{x},\mathbf{w}) = \mathbf{w}^{\top}\mathbf{x}$. In this case, $\varphi_p(\mathbf{x}_m) = x_m^{(p)}$ map the p^{th} component of D dimensional vector $\mathbf{x}_m \in \mathbb{R}^D$:

$$\phi(\mathbf{x}_m) = \begin{pmatrix} \varphi_1(\mathbf{x}_m) \\ \varphi_2(\mathbf{x}_m) \\ \vdots \\ \varphi_P(\mathbf{x}_m) \end{pmatrix} = \begin{pmatrix} x_m^{(1)} \\ x_m^{(2)} \\ \vdots \\ x_m^{(D)} \end{pmatrix} \in \mathbb{R}^{D\times 1}$$

and the design matrix is:

$$\mathbf{\Phi}^{\mathrm{LML}}(\mathcal{D}) = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(D)} \\ \vdots & \vdots & \ddots & \vdots \\ x_M^{(1)} & x_M^{(2)} & \cdots & x_M^{(D)} \end{pmatrix} \in \mathbb{R}^{M\times D}. \tag{2.21}$$

It should be noted that the above discussion is only applicable to cases where the data is unbiased. For the general case, a supplementary parameter $\varphi_0(\mathbf{x}) = w_0$ is added to the linear fit, which serves as a constant to represent the intercept of the linear regression. Then in this LML case, $P = D + 1$ and the fitting function can be written as $f_{\mathrm{ML}}(\mathbf{x}, \mathbf{w}, w_0) = w_0 + \mathbf{w}^\top \mathbf{x}$.

2.1.3.2 Nonlinear model

Usually, a linear regression model is insufficient to describe the nontrivial topology of the data. To address the limitations of linear regression, nonlinear models are widely employed. In this subsection, we will discuss a simple yet effective nonlinear method: quadratic machine learning (QML) fitting. In this case, $P = D^2$ and the application $\varphi_p(\mathbf{x}_m) = x_m^{(p_1)} x_m^{(p_2)}$ mixes 2 components of the D-dimensional vector $\mathbf{x}_m \in \mathbb{R}^D$:

$$\phi(\mathbf{x}_m) = \begin{pmatrix} \varphi_1(\mathbf{x}_m) \\ \varphi_2(\mathbf{x}_m) \\ \vdots \\ \varphi_P(\mathbf{x}_m) \end{pmatrix} = \begin{pmatrix} x_m^{(1)} x_m^{(1)} \\ x_m^{(1)} x_m^{(2)} \\ \vdots \\ x_m^{(1)} x_m^{(D)} \\ x_m^{(2)} x_m^{(1)} \\ \vdots \\ x_m^{(D)} x_m^{(D)} \end{pmatrix} \in \mathbb{R}^{D^2 \times 1}.$$

Then the design matrix becomes:

$$\begin{aligned} \mathbf{\Phi}^{\mathrm{QML}}(\mathcal{D}) &= \begin{pmatrix} \phi^\top(\mathbf{x}_1) \\ \phi^\top(\mathbf{x}_2) \\ \vdots \\ \phi^\top(\mathbf{x}_M) \end{pmatrix} \\ &= \begin{pmatrix} x_1^{(1)} x_1^{(1)} & x_1^{(1)} x_1^{(2)} & \cdots & x_1^{(1)} x_1^{(D)} & \cdots & x_1^{(D)} x_1^{(D)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_M^{(1)} x_M^{(1)} & x_M^{(1)} x_M^{(2)} & \cdots & x_M^{(1)} x_M^{(D)} & \cdots & x_M^{(D)} x_M^{(D)} \end{pmatrix} \in \mathbb{R}^{M \times D^2} \end{aligned} \tag{2.22}$$

The QML fitting function is symmetric $f_{\mathrm{QML}}(\mathbf{x}, \mathbf{w}) = \mathbf{x}^\top \mathbf{w} \mathbf{x}$. Following this symmetric structure, the columns of the design matrix are repeated, and the rank of the matrix, i.e., the maximum number of independent parameters in the QML model, cannot exceed $D(D+1)/2$.

It is also noteworthy that models with various functional forms can be combined. An example of this is the Quadratic Noise Machine Learning (QNML) model, where the design matrix concatenates both LML and QML models. The combined form is:

$$\mathbf{\Phi}^{\mathrm{QNML}}(\mathcal{D}) = \mathbf{\Phi}^{\mathrm{LML}}(\mathcal{D}) \oplus \mathbf{\Phi}^{\mathrm{QML}}(\mathcal{D}) \in \mathbb{R}^{M \times D} \oplus \mathbb{R}^{M \times D^2} \in \mathbb{R}^{M \times (D + D^2)} \tag{2.23}$$

A notable feature of linear-in-parameters hybrid models is that their regression parameters naturally exhibit a direct sum decomposition structure:

$$\mathbf{w}^{\mathrm{QNML}} = \mathbf{w}^{\mathrm{LML}} \oplus \mathbf{w}^{\mathrm{QML}},$$

which suggests various training strategies, such as direct optimization over the entire QNML space or sequential optimization, i.e., initially optimizing in the LML space, followed by optimization in the remaining QML space.

Linear interpolation exhibits reasonably good performance outside its fitting range. To preserve this advantageous property in other non-linear fitting approaches, we impose a precondition, set by the linear interpolation, on the quadratic regime. It means that only the deviation of the LML values from the target properties is fitted within the quadratic regime:

$$\begin{aligned} y = f_{\mathrm{LML}}\left(\mathbf{x},\mathbf{w}\right) + \Delta y &\simeq f_{\mathrm{LML}}\left(\mathbf{x},\mathbf{w}\right) + f_{\mathrm{QML}}\left(\mathbf{x},\mathbf{w}\right) = f_{\mathrm{ML}}\left(\mathbf{x},\mathbf{w}\right) , \\ E_s^{\mathrm{DFT}} = E_s^{\mathrm{LML}} + \Delta E &\simeq E_s^{\mathrm{LML}} + E_s^{\mathrm{QML}} = E_s^{\mathrm{ML}} , \end{aligned} \tag{2.24}$$

where the last equation is formulated assuming the target observable is the total energy of system s (i.e., the DFT energy), and the term E_s^{QML} exhibits a quadratic form in terms of the descriptors, $E_s^{\mathrm{QML}} = f_{\mathrm{QML}}\left(\mathbf{D}_s,\mathbf{w}\right)$.

The local atomic energy can be written as:

$$\begin{aligned} \epsilon_{s,a} &= \epsilon_{s,a}^{\mathrm{LML}} + \epsilon_{s,a}^{\mathrm{QML}} \\ \epsilon_{s,a}^{\mathrm{LML}} &= f_{\mathrm{LML}}\left(\mathbf{D}_{s,a},\mathbf{w}\right) \\ \epsilon_{s,a}^{\mathrm{QML}} &= f_{\mathrm{QML}}\left(\mathbf{D}_{s,a},\mathbf{w}\right) . \end{aligned} \tag{2.25}$$

The above QNML development incorporates a well-defined preconditioning imposed by the linear fitting. Specifically, the parameters $\mathbf{w}^{\mathrm{LML}}$ of the linear fit are first evaluated and then fixed, so as to eventually determine the parameters $\mathbf{w}^{\mathrm{QML}}$ of the quadratic fit.

Similar to the linear case [34, 35, 38], the atomic forces and the virial stress are also considered in the quadratic formalism. The force acting on an atom is computed from the total ML energy E_s^{ML} in Eq. 2.24. In the case of QNML, the descriptor associated with the force on atom a in the direction α becomes:

$$\mathbf{D}_{s,a,\alpha}^{F} = -\left(0, \sum_{b\in\mathcal{V}(a)} \nabla_{a,\alpha}\mathbf{D}_{s,b}, \sum_{b\in\mathcal{V}(a)} \left(\nabla_{a,\alpha}\mathbf{D}_{s,b}\otimes\mathbf{D}_{s,b} + \mathbf{D}_{s,b}\otimes\nabla_{a,\alpha}\mathbf{D}_{s,b}\right)\right) .$$

The force descriptors mentioned above reside within the same $\mathbb{R}^{1\times(1+D+D^2)}$ space and require a non-trivial computation of derivatives for each atomic descriptor with respect to the atomic Cartesian coordinates in the neighborhood of the central atom a within the cutoff radius.

We emphasize that the proposed QNML formalism is different from the quadratic Spectral Neighbor Analysis Potential (qSNAP) in Ref. [35]. While qSNAP potentials are fitted as a second-degree polynomial with explicit linear and quadratic forms, QNML treats only the error of the linear fit as a quadratic form of atomic descriptors. Furthermore, the QNML procedure is inspired by the nearly Gaussian shape of the error distribution in the linear fit. As showed in Fig. 2.7(b), the distribution of the noise $E_s^{\mathrm{DFT}} - E_s^{\mathrm{LML}}$ closely resembles a unimodal Gaussian distribution, whereas the absolute DFT energies in our database exhibit a bimodal shape. Such distribution of DFT data, without appropriate treatment, can lead to overfitting or highly heterogeneous parameterization. The error of the LML fit serves as the

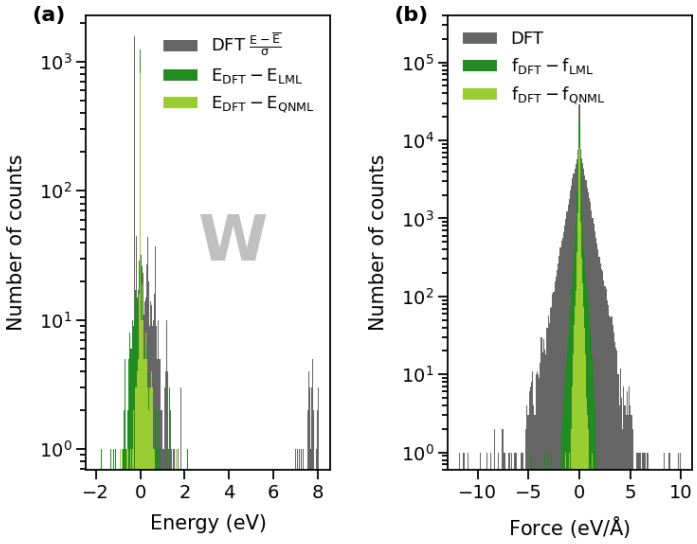

Figure 2.7: Histogram depicting (a) the distribution of DFT data for energy (in eV) and force (in eV/Å), and (b) the error noise deviation of LML and QNML force fields for bcc W [148]. The absolute DFT energy is presented in a re-scaled form $\left(E - \overline{E}\right)/\sigma$, where $\overline{E}$ is the average energy of the database and σ is the standard deviation. The error distribution for energy and force follows a Gaussian shape, with the distribution being narrower for the QNML potential.

target in the QNML formalism. To mitigate the inherent risk of overfitting with nonlinear approaches, appropriate regularization techniques, such as $\mathcal{L}^2$ ridge and Bayesian methods, are applicable [183, 100, 99] and will be discussed in the following subsection.

2.1.3.3 Loss function and error evaluation

As mentioned previously, any regression that is linear in parameters can be formulated as an optimization problem in the form of a linear system:

$$\mathbf{\Phi}\left(\mathcal{D}\right)\mathbf{w} \quad = \quad \mathbf{y}\,. \tag{2.26}$$

Solutions of this linear system can be formally given using the pseudo inversion of the design matrix $\mathbf{\Phi}(\mathcal{D})^{+} \in \mathbb{R}^{P\times M}$. For the case where $P \geq M$, meaning that we have more equations than unknowns, the entire ensemble of solutions $\mathbf{w} \in \mathbb{R}^{P}$ can be expressed as:

$$\mathbf{w} = \mathbf{\Phi}(\mathcal{D})^{+}\mathbf{y} + \left[\mathbf{I}_P - \mathbf{\Phi}(\mathcal{D})^{+}\mathbf{\Phi}(\mathcal{D})\right]\mathbf{r}, \tag{2.27}$$

where $\mathbf{r} \in \mathbb{R}^{P}$ is an arbitrary vector. Solutions exist if and only if the equality $\mathbf{\Phi}(\mathcal{D})\mathbf{\Phi}(\mathcal{D})^{+}\mathbf{y} = \mathbf{y}$ holds. If this condition is satisfied, the solution is unique if and only if $\mathbf{\Phi}(\mathcal{D})$ has full column rank, i.e., $\text{rank}(\mathbf{\Phi}(\mathcal{D})) = P$, in which case $\mathbf{I}_P - \mathbf{\Phi}(\mathcal{D})^{+}\mathbf{\Phi}(\mathcal{D})$ is a zero matrix. If solutions exist but $\mathbf{\Phi}(\mathcal{D})$ does not have full column rank, we have an indeterminate system, where the infinite solutions are determined by a random choice of the $\mathbf{r}$ vector.

In the case where $M \geq P$, which is the most common case in materials science, indicating that the number of data points is larger than the feature space dimension, there exists no

exact solution, only an approximate one, with an error $\boldsymbol{\epsilon} = \mathbf{y} - \boldsymbol{\Phi}(\mathcal{D})\mathbf{w}$. Among all these approximate solutions, we choose in general the one that minimizes this error:

$$\mathbf{w} = \underset{\mathbf{w}\in\mathbb{R}^P}{\operatorname{argmin}} \|\mathbf{y} - \boldsymbol{\Phi}(\mathcal{D})\mathbf{w}\| = \underset{\mathbf{w}\in\mathbb{R}^P}{\operatorname{argmin}}\, J(\mathbf{w}), \tag{2.28}$$

where $J(\mathbf{w})$ is the $\mathcal{L}^2$ objective function:

$$J(\mathbf{w}) = \sum_{m=1}^{M} |y_m - \hat{y}(\mathbf{x}_m, \mathbf{w})|^2 \tag{2.29}$$

with $\hat{y}(\mathbf{x}_m, \mathbf{w}) = f_{\mathrm{ML}}(\mathbf{x}_m, \mathbf{w}) = \phi^\top(\mathbf{x}_m)\mathbf{w}$ the prediction of the ML model with parameters $\mathbf{w}$. The formal solution of this optimization problem can be expressed by the left pseudo inversion of the design matrix $\boldsymbol{\Phi}_l(\mathcal{D})^+ \in \mathbb{R}^{P\times M}$ as

$$\mathbf{w} = \boldsymbol{\Phi}_l(\mathcal{D})^+\mathbf{y} = \left(\boldsymbol{\Phi}\left(\mathcal{D}\right)^\top \boldsymbol{\Phi}\left(\mathcal{D}\right)\right)^{-1} \boldsymbol{\Phi}\left(\mathcal{D}\right)^\top \mathbf{y}. \tag{2.30}$$

This solution is the orthogonal projection of $\mathbf{y}$ on the orthogonal complement of the null space $\mathcal{N}(\boldsymbol{\Phi}(\mathcal{D}))$. The projection error is given by $[\mathbf{I}_M - \boldsymbol{\Phi}(\mathcal{D})\boldsymbol{\Phi}_l(\mathcal{D})^+]\mathbf{y}$.

The solution given above in Eq. 2.30 has an apparently unique form. However, this solution that minimizes the $\mathcal{L}^2$ objective function $J(\mathbf{w})$ is one solution among an infinite number of solutions. Very often, this solution is mathematically optimal but may not be the most "physical". To approach more "physically" plausible alternative solutions, we propose the following three principal methods.

Regularization

The variation of the predicted physical observable should exhibit smoothness with respect to small variations in atomic coordinates. Rough solutions may arise, particularly when the model has high dimension P ($P \sim$ several thousands). Regularization methods are absolutely essential in the context of high-dimensional regression. As previously discussed, they help prevent overfitting on certain datasets, thereby enhancing the transferability of models. In the current context, we regularize the solution to prevent the occurrence of rough solutions with poor extrapolation capacity and to avoid ill-defined numerical inversion of the matrix $\boldsymbol{\Phi}\left(\mathcal{D}\right)^\top \boldsymbol{\Phi}\left(\mathcal{D}\right)$ in Eq.2.30.

Regularization refers to any modification made to a learning algorithm with the intent to reduce generalization error rather than training error. This technique is often grounded in the expectation that models should be smooth rather than overly complex. In the framework of linear regression, smoother functions are typically associated with smaller weight magnitudes. Therefore, we can penalize complex functions by incorporating an appropriate penalty term into the loss function that we aim to minimize:

$$J_r(\mathbf{w}) = J(\mathbf{w}) + \mathbf{w}^T\mathbf{C}_P\mathbf{w}, \tag{2.31}$$

where $\mathbf{C}_P \in \mathbb{R}^{P\times P}$ can be regarded as a matrix that imposes constraints on the parameters $\mathbf{w}$. The choice of $\mathbf{C}_P$ controls our preferences for weights. There are various methods to

select $\mathbf{C}_P$, either implicitly or explicitly. An explicit approach involves properly defining the validation error, which can be addressed by designating a validation set from the training data. Minimizing the validation error corresponds to optimizing the hyperparameters on the validation data. The most widely used choice for $\mathbf{C}_P$ is a diagonal matrix $\mathbf{C}_P = \lambda^2 \mathbf{I}_P$, where $\mathbf{I}_P$ is the $P \times P$ identity matrix. Under this condition, we prefer the parameters $\mathbf{w}$ that minimize the norm weighted by a factor of λ^2, giving the loss function:

$$J_r(\mathbf{w}) = J(\mathbf{w}) + \lambda^2 \|\mathbf{w}\|^2 \tag{2.32}$$

This regularization term is the $\mathcal{L}^2$ regularization, also known as Tikhonov regularization, proposed by Tikhonov and Phillips [184]. Using this regularization, the solution becomes

$$\mathbf{w} = \left[\mathbf{\Phi}\left(\mathcal{D}\right)^\top \mathbf{\Phi}\left(\mathcal{D}\right) + \mathbf{C}_P\right]^{-1} \mathbf{\Phi}\left(\mathcal{D}\right)^\top \mathbf{y}. \tag{2.33}$$

For $\lambda = 0$, we obtain the previous solution, whereas a large value of λ forces the weights to become smaller, leading to a smoother fit. Selecting the appropriate value for the hyperparameter λ is challenging. As we will see below, the matrix $\mathbf{C}_P$ has a clear statistical interpretation.

Weighting the fit

The "physical" solution can be produced by weighting the fit, specifically by assigning greater importance to certain columns of $\mathbf{\Phi}(\mathcal{D})$, in accordance with the physical significance of those observations in the database. Physical intuition often guides the selection of database observations that are pertinent to the target problem. Certain datasets may be preferred over others to accurately predict specific observations. For example, when aiming to accurately determine elastic constants and lattice parameters of a crystalline structure, it may be sufficient to carefully fit only a few deformed configurations around the equilibrium state while leaving the majority of data in the database untouched.

In this case, we complete the information contained in the dataset $\mathcal{D}$ and by associating the desired weights $\omega_m \in \mathbb{R}^+$ with descriptors and observations $(\mathbf{x}_m, y_m)$. The weighted database is now defined as $\mathcal{D}_\omega = \{((\mathbf{x}_m, y_m, \omega_m) \,\text{for}\, m = 1 \dots M\}$. To take into account these weights, the $\mathcal{L}^2$ objective function is slightly changed and becomes:

$$J(\mathbf{w}) = \sum_{m=1}^{M} \omega_m \left| y_m - \hat{y}_m(\mathbf{x}_m, \mathbf{w}) \right|^2 , \tag{2.34}$$

The complete loss function should incorporate the components of the database for energy (E), force (F), and stress (S). Specifically, the objective function consists of at least four components: three for energy, force, and stress, and a fourth component that includes the regularization terms $R(\mathbf{w}, \lambda)$:

$$J(\mathbf{w}) = J_E(\mathbf{w}) + J_F(\mathbf{w}) + J_S(\mathbf{w}) + R(\mathbf{w}, \lambda) \,, \tag{2.35}$$

To balance the contribution of energy, force, and stress, the simplest formulation would be:

$$\begin{aligned} J(\mathbf{w}) \;\; &= \;\; \frac{1}{2M_\mathrm{E}} \sum_{m_E=1}^{M_E} \omega_{m_E}^E \left(y_{m_E}^\mathrm{E} - \hat{y}_{m_E}^\mathrm{E}\right)^2 + \frac{1}{2M_\mathrm{F}} \sum_{m_F=1}^{M_F} \omega_{m_F}^F \left(y_{m_F}^\mathrm{F} - \hat{y}_{m_F}^\mathrm{F}\right)^2 \\ &+ \;\; \frac{1}{2M_\mathrm{S}} \sum_{m_S=1}^{M_S} \omega_{m_S}^S \left(y_{m_S}^\mathrm{S} - \hat{y}_{m_S}^\mathrm{S}\right)^2 + R(\mathbf{w}, \lambda). \end{aligned} \tag{2.36}$$

Here, M_E denotes the total number of energy data points indexed by m_E. Similarly, M_F and M_S represent the total number of force and stress data points, with m_F and m_S indexing the m_F^{th} force data point and the m_S^{th} stress data point, respectively. The objective function can be expressed as $J(\mathbf{w}, \boldsymbol{\omega})$, where $\boldsymbol{\omega}$ encompasses all database weights $(\omega_{m_E}^E, \omega_{m_F}^F, \omega_{m_S}^S)$ for energy, force, and stress.

Equation 2.34 is equivalent to the following linear system:

$$\boldsymbol{\Omega}^{1/2}\boldsymbol{\Phi}\left(\mathcal{D}\right)\mathbf{w} = \boldsymbol{\Omega}^{1/2}\mathbf{y}\,, \tag{2.37}$$

where $\boldsymbol{\Omega}$ is the weights matrix that is positive definite and diagonal:

$$\boldsymbol{\Omega} = \begin{pmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_M \end{pmatrix} \in \mathbb{R}^{M\times M}\,. \tag{2.38}$$

The corresponding solution is equivalent to that of the non-weighted problem after transforming the design matrix $\boldsymbol{\Phi}\left(\mathcal{D}\right)$ and the target vector $\mathbf{y}$ according to:

$$\begin{aligned} \boldsymbol{\Phi}\left(\mathcal{D}\right) &\rightarrow \boldsymbol{\Omega}^{1/2}\boldsymbol{\Phi}\left(\mathcal{D}\right) \\ \mathbf{y} &\rightarrow \boldsymbol{\Omega}^{1/2}\mathbf{y}\,. \end{aligned}$$

Therefore, the formal solution is given by:

$$\mathbf{w} = \left[\boldsymbol{\Phi}\left(\mathcal{D}\right)^\top \boldsymbol{\Omega}\boldsymbol{\Phi}\left(\mathcal{D}\right)\right]^{-1}\boldsymbol{\Phi}\left(\mathcal{D}\right)^\top\boldsymbol{\Omega}\mathbf{y}\,, \tag{2.39}$$

and with the $\mathcal{L}^2$ regularization $R(\mathbf{w}, \lambda) = \lambda^2\|\mathbf{w}\|^2$,

$$\mathbf{w} = \left[\boldsymbol{\Phi}\left(\mathcal{D}\right)^\top \boldsymbol{\Omega}\boldsymbol{\Phi}\left(\mathcal{D}\right) + \lambda^2\mathbf{I}_P\right]^{-1}\boldsymbol{\Phi}\left(\mathcal{D}\right)^\top\boldsymbol{\Omega}\mathbf{y}\,. \tag{2.40}$$

Bayesian interpretation

Sometimes, the "physical" solution corresponds to the Bayesian approach of linear fitting, which incorporates our prior or assumptions about the conditions satisfied by the parameters $\mathbf{w}$. The solution of linear regression can be reinterpreted from a Bayesian probabilistic perspective. From this viewpoint, the parameters are inherently uncertain due to the limited available training data and are assumed to follow a distribution that should be determined. Even the measurement of observations is subject to error, so the database $\mathcal{D} = \{(\mathbf{x}_m, y_m)|\,\text{for}\, m = 1, M\}$ can be regarded as being generated from an exact, noise-free model $f(\mathbf{x}, \mathbf{w})$ with additive Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma_M^2)$, formulated as $y = f(\mathbf{x}, \mathbf{w}) + \epsilon$. In a Bayesian framework, the quantity to determine for obtaining the optimal set of parameters is derived from the distribution function or the conditional probability $p(\mathbf{w}|\mathcal{D})$. This quantity of interest is commonly referred to as the *posterior probability*, which can be determined using Bayes' rule:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}\,, \tag{2.41}$$

$$p(\mathbf{w}|\mathbf{x},\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{x},\mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{x},\mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{x})}\,, \tag{2.42}$$

where the marginal likelihood is independent of the parameters and given by

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{x},\mathbf{w})p(\mathbf{w})d\mathbf{w}\,. \tag{2.43}$$

It is important to note that the previous equation combines the likelihood $p(\mathbf{y}|\mathbf{x},\mathbf{w})$ and the prior over the parameters $p(\mathbf{w})$ to obtain the posterior distribution. This means that our initial belief about the parameters $p(\mathbf{w})$ is corrected by the likelihood distribution of the observed data. The prior belief about the parameters assumes that their distribution is Gaussian with zero mean and a covariance matrix $\mathbf{\Sigma}_P$, i.e., $\mathbf{w} \sim \mathcal{N}(\mathbf{0},\mathbf{\Sigma}_P)$, with the probability density function

$$p(\mathbf{w}) = (2\pi)^{-P/2}\det(\mathbf{\Sigma}_P)^{-1/2}\exp\left(-\frac{1}{2}\mathbf{w}^\top\mathbf{\Sigma}_P^{-1}\mathbf{w}\right). \tag{2.44}$$

With these statistical conditions and using Bayes' rule (Eq. 2.42), it can be demonstrated [185] that the mean of the posterior distribution for the parameters $p(\mathbf{w}|\mathbf{x},\mathbf{y})$ is

$$\overline{\mathbf{w}} = \left[\sigma_M^{-2}\mathbf{\Phi}\left(\mathcal{D}\right)^\top\mathbf{\Phi}\left(\mathcal{D}\right) + \mathbf{\Sigma}_P^{-1}\right]^{-1}\mathbf{\Phi}\left(\mathcal{D}\right)^\top\mathbf{y}\,.$$

We can now clearly recognize the least squares solution of our linear model with regularization in the case where the constraint matrix is the inverse of the covariance matrix of the parameters, $\mathbf{C}_P = \mathbf{\Sigma}_P^{-1}$.

The strength of Bayesian models lies in their ability to perform marginalization, which means integrating over the parameters of the model to make predictions. In the classical approach, predictions are derived directly from the values of the function as specified by a given set of parameters. However, it is the prediction itself that is of primary importance, rather than the specific set of fitted parameters. In the Bayesian framework, marginalization allows for the integration over all possible parameter values. The probability distribution of the prediction $y_\star$ for a given input $\mathbf{x}_\star$ and a dataset $\mathcal{D}$ can be expressed as:

$$p\left(y_\star|\mathbf{x}_\star,\mathcal{D}\right) = \int p(y_\star|\mathbf{x}_\star,\mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w}\,. \tag{2.45}$$

The prediction follows a normal distribution with mean $\mu(y_\star)$ and variance $\mathbb{V}(y_\star)$:

$$y_\star \sim \mathcal{N}(\mu(y_\star),\mathbb{V}(y_\star))\,, \tag{2.46}$$

$$\mu(y_\star) = \phi(\mathbf{x}_\star)^\top\left[\sigma_M^{-2}\mathbf{\Phi}\left(\mathcal{D}\right)^\top\mathbf{\Phi}\left(\mathcal{D}\right) + \mathbf{\Sigma}_P^{-1}\right]^{-1}\mathbf{\Phi}\left(\mathcal{D}\right)^\top\mathbf{y}\,, \tag{2.47}$$

$$\mathbb{V}(y_\star) = \phi(\mathbf{x}_\star)^\top\left[\sigma_M^{-2}\mathbf{\Phi}\left(\mathcal{D}\right)^\top\mathbf{\Phi}\left(\mathcal{D}\right) + \mathbf{\Sigma}_P^{-1}\right]^{-1}\phi(\mathbf{x}_\star)\,, \tag{2.48}$$

where $\phi(\mathbf{x}) = (\varphi_1(\mathbf{x}),\ldots,\varphi_p(\mathbf{x}))^\top$ is the $P\times 1$ vector that contains the functions of the linear model at a given point $\mathbf{x}$, and we denote by $\mathbf{A}\in\mathbb{R}^{P\times P}$ the matrix $\mathbf{A} = \sigma_M^{-2}\mathbf{\Phi}\left(\mathcal{D}\right)^\top\mathbf{\Phi}\left(\mathcal{D}\right) + \mathbf{\Sigma}_P^{-1}$. This distribution incorporates uncertainty in the parameters by averaging over all possible values of $\mathbf{w}$. Classical regularization, or even maximum a posteriori predictions, does not account for how well the parameters $\mathbf{w}$ are determined. In addition, we can estimate the epistemic uncertainty from the predictive variance at a given input $\mathbf{x}_\star$ as $\sigma_M^2 + \phi(\mathbf{x}_\star)^\top\mathbf{A}^{-1}\phi(\mathbf{x}_\star)$,

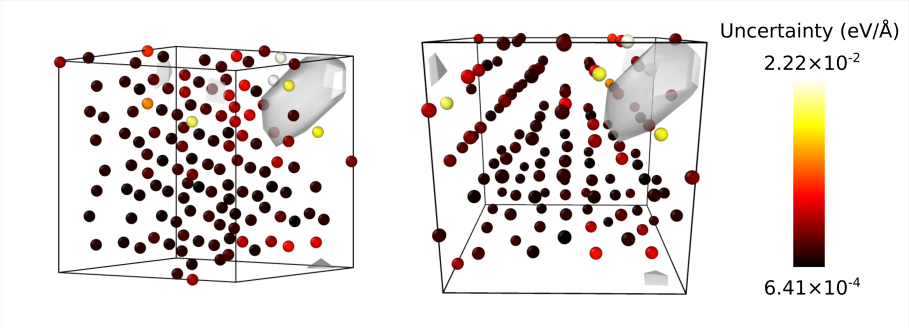

Figure 2.8: Epistemic uncertainty in force prediction, evaluated from the predictive variance, for a tungsten configuration of 126 atoms containing two vacancies (the white region). The prediction for the atoms near the defects is less certain.

as shown in Fig. 2.8. This uncertainty can be used as input for the active learning procedure introduced in Subsection 2.1.1.2.

Now we will discuss the computational cost of predicting using the aforementioned Eq. 2.46. Both prediction and variance computations require the inversion of the $P \times P$ matrix $\mathbf{A}$. Depending on the dimensions of the design matrix and which dimension is larger—either the feature space dimension P or the number of observations M—the equation above needs to be rearranged accordingly. If $P < M$, the above equations provide the most efficient way to make predictions. However, if $P \geq M$ (for example, in Gaussian processes where $P \to \infty$), the mean and variance equations for the prediction $y_\star$ of a given input $\mathbf{x}_\star$ should be rewritten as:

$$\mu(y_\star) = \phi(\mathbf{x}_\star)^\top \boldsymbol{\Sigma}_P \boldsymbol{\Phi}(\mathcal{D})^\top \left(\mathbf{K}_{MM} + \sigma_M^2 \mathbf{I}_M\right)^{-1} \mathbf{y}\,, \tag{2.49}$$

$$\begin{aligned} \mathbb{V}(y_\star) = {} & \phi(\mathbf{x}_\star)^\top \boldsymbol{\Sigma}_P \phi(\mathbf{x}_\star) \\ & - \phi(\mathbf{x}_\star)^\top \boldsymbol{\Sigma}_P \boldsymbol{\Phi}(\mathcal{D})^\top \left(\mathbf{K}_{MM} + \sigma_M^2 \mathbf{I}_M\right)^{-1} \boldsymbol{\Phi}(\mathcal{D}) \boldsymbol{\Sigma}_P \phi(\mathbf{x}_\star)\,, \end{aligned} \tag{2.50}$$

where the matrix $\mathbf{K}_{MM} = \boldsymbol{\Phi}(\mathcal{D}) \boldsymbol{\Sigma}_P \boldsymbol{\Phi}(\mathcal{D})^\top$ is an $M \times M$ matrix and consequently the time complexity is of $\mathcal{O}(M^3)$. It should also be noted that in the general case, the matrix $\mathbf{K}_{MM}$ has the form:

$$\mathbf{K}_{MM} = \begin{pmatrix} \phi^\top(\mathbf{x}_1) \\ \phi^\top(\mathbf{x}_2) \\ \vdots \\ \phi^\top(\mathbf{x}_M) \end{pmatrix} \boldsymbol{\Sigma}_P \left(\phi(\mathbf{x}_1), \cdots, \phi(\mathbf{x}_M)\right)\,, \tag{2.51}$$

which means that the elements of $\mathbf{K}_{MM}$ are structured as follows:

$$k(\mathbf{x}_m, \mathbf{x}_{m'}) = \phi(\mathbf{x}_m)^\top \boldsymbol{\Sigma}_P \phi(\mathbf{x}_{m'})\,. \tag{2.52}$$

For reasons that will become clear later, we call $k(\cdot, \cdot)$ a *covariance function* or *kernel*, which is a key quantity in Gaussian processes (GPs). A GP is, in general, a collection of random variables, any finite number of which have a joint Gaussian distribution. A GP is completely specified by its mean function and covariance function. Often, GPs are defined over time, but this is not the case in our use. In our case, the random variables represent the value of the function $f_{\mathrm{ML}}(\mathbf{x}, \mathbf{w})$ at location $\mathbf{x} \in \mathbb{R}^D$. The Bayesian linear regression model $f_{\mathrm{ML}}(\mathbf{x}, \mathbf{w}) = \phi(\mathbf{x})^\top \mathbf{w}$

with prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_P)$ provides a simple example of GP, where the covariance matrix $\mathbf{C}$ for any dataset $\mathcal{D}$ is defined as:

$$\mathbf{C} = \mathbb{E}\left(\hat{\mathbf{y}}\hat{\mathbf{y}}^\top\right) = \mathbf{\Phi}\left(\mathcal{D}\right)\mathbb{E}(\mathbf{w}\mathbf{w}^\top)\mathbf{\Phi}\left(\mathcal{D}\right)^\top = \mathbf{\Phi}\left(\mathcal{D}\right)\mathbf{\Sigma}_P\mathbf{\Phi}\left(\mathcal{D}\right)^\top \tag{2.53}$$

which corresponds to the definition of the matrix $\mathbf{K}_{MM}$.

A separate dataset consisting of a specified number of data points is randomly selected from the database for testing and validating the trained models. To evaluate their performance, we utilize the root mean square error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}(\mathbf{x}_i, \mathbf{w})|^2} \tag{2.54}$$

where N is the number of data points in the training or testing dataset, $\hat{y}$ represents the value predicted by the model for $\mathbf{x}_i$, and y denotes the associated target value obtained from the DFT calculations.

Another check to validate an ML potential is the comparison between the ML potential-predicted value and the DFT value of some basic physical properties (lattice constant a_0, bulk modulus B and the cubic elastic constants C_{11}, C_{12} and C_{44}). Table 2.2 shows the comparison for two ML potentials with different regression models (LML and QNML), demonstrating that these ML potentials have good prediction performance on the elastic properties. Besides the basic elastic properties, some other properties related to the target problem should be verified. For example, in the section 4.1.2.2, the aim is to build an ML potential for studying the vacancy properties. Therefore, the formation energy and the migration energy of mono-vacancy and di-vacancies are also compared with the DFT values (Table 4.2), to guarantee the reliability of the ML potential in predicting the vacancy behaviors.

Table 2.2: Elastic properties of bcc W provided by the developed LML and QNML potentials and their comparison with the reference DFT values, obtained from DB class 2 in Table 2.1.

	LML potential	QNML potential	DFT	Unit
a_0	3.1855	3.1854	3.1854	Å
B	303.2	304.1	304.5	GPa
C_{11}	509.8	516.7	516.6	GPa
C_{12}	200.0	197.8	198.5	GPa
C_{44}	144.5	137.2	140.2	GPa

2.2 Kernel regression

For many regression algorithms, the raw data must be explicitly transformed into feature vector representations via a user-specified feature map. In the present context, the raw data are the data points $\mathbf{x}_m$, which can be transformed using the feature map ϕ into the feature space $\mathcal{V} \subset \mathbb{R}^P$, as previously introduced. We talked about the kernel (or covariance function)

of Gaussian processes in Subsection 2.1.3.3, based on the feature map ϕ chosen for the machine learning model. However, the computation of the covariance matrix $\mathbf{K}_{MM}$ has a complexity of $\mathcal{O}(M^2P^2)$, which is computationally intensive. Therefore, we aim to identify a kernel function $k(\cdot,\cdot)$ that directly provides the value of the covariance function instead of computing Eq. 2.52. This method is called the kernel regression method.

First, we introduce the mathematical framework for this method. We cover reproducing kernel Hilbert spaces (RKHSs), which define a Hilbert space of sufficiently-smooth functions corresponding to a user-specified, symmetric, positive definite kernel $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$. A kernel k is considered positive definite if, over the input space $\mathcal{X}$, the matrix $K_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)$ is a positive definite matrix.

Definition *(Reproducing kernel Hilbert space). Let $\mathcal{H}$ be a Hilbert space of real functions f defined on an index set $\mathcal{X}$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space endowed with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and norm $\|f\|_{\mathcal{H}}=\sqrt{\langle f,f\rangle_{\mathcal{H}}}$ if there exists a function $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ with the following properties:*
1. $\forall\mathbf{x}\in\mathcal{X}$, the function $\mathbf{x}'\to k(\mathbf{x},\mathbf{x}')$ belongs to $\mathcal{H}$, and
2. k has the reproducing property: $\forall\mathbf{x}\in\mathcal{X}$, $\forall f\in\mathcal{H}$, we have $\langle f(\cdot),k(\cdot,\mathbf{x})\rangle_{\mathcal{H}}=f(\mathbf{x})$.

The RKHS uniquely determines k, and vice versa, as stated in the Moore-Aronszajn theorem theorem:

Theorem *(Moore-Aronszajn theorem [186]). Let $\mathcal{X}$ be an index set. Then for every symmetric, positive definite kernel $k(\cdot,\cdot)$ on $\mathcal{X}\times\mathcal{X}$, there exists a unique RKHS, and vice versa.*

The Moore-Aronszajn theorem states that every symmetric, positive definite kernel defines a unique RKHS. We denote $\mathcal{H}_k$ the RKHS defined by a user-specified kernel function k. In this space, every function can be reproduced by an inner product. In particular, if we set $f(\cdot)=k(\cdot,\mathbf{x}')\in\mathcal{H}_k$, the reproducing property becomes $f(\mathbf{x})=k(\mathbf{x},\mathbf{x}')=\langle k(\cdot,\mathbf{x}'),k(\cdot,\mathbf{x})\rangle_{\mathcal{H}_k}$. This equality holds for every $\mathbf{x},\mathbf{x}'\in\mathcal{X}$. Let $\phi(\mathbf{x})=k(\cdot,\mathbf{x})$, then we have

$$\forall\mathbf{x},\mathbf{x}'\in\mathcal{X},\ k\left(\mathbf{x},\mathbf{x}'\right)=\langle k(\cdot,\mathbf{x}),k(\cdot,\mathbf{x}')\rangle_{\mathcal{H}_k}=\langle\phi(\mathbf{x}),\phi(\mathbf{x}')\rangle_{\mathcal{H}_k}. \tag{2.55}$$

The feature map ϕ in ML kernels is infinite-dimensional but only requires a finite-dimensional matrix from user input according to the *representer theorem*, which states that every function in an RKHS that minimises an empirical risk functional can be written as a linear combination of the kernel function evaluated at the training points. This is a practically useful result as it effectively simplifies the optimization problem from an infinite-dimensional to a finite-dimensional one. In our case, the index set $\mathcal{X}$ is the database $\mathcal{D}=\{\mathbf{x}_m\in\mathbb{R}^D \text{ for } m=1,\dots M\}$. It is a finite set with M elements, on which we have $k(\mathbf{x}_m,\mathbf{x}_{m'})=\langle\phi(\mathbf{x}_m),\phi(\mathbf{x}_{m'})\rangle_{\mathcal{H}_k}$. Here $\phi(\mathbf{x}_m):\mathbf{x}\to k(\mathbf{x},\mathbf{x}_m)$ is a discrete function parameterized by $\mathbf{x}_m$, so a common inner product gives $k(\mathbf{x}_m,\mathbf{x}_{m'})=\sum_{i=1}^{M}k(\mathbf{x}_i,\mathbf{x}_m)\,k(\mathbf{x}_i,\mathbf{x}_{m'})$.

Notice that Eq. 2.52 is also an inner product (with respect to Σ_P). As Σ_P is positive definite, we can define $\Sigma_P^{1/2}$ so that $(\Sigma_P^{1/2})^2 = \Sigma_P$. Then defining $\psi(\mathbf{x}) = \Sigma_P^{1/2}\phi(\mathbf{x})$, we obtain a simple dot product representation of Eq. 2.52: $k\left(\mathbf{x}_m, \mathbf{x}_{m'}\right) = \psi(\mathbf{x}_m)^\top \psi(\mathbf{x}_{m'}) = \langle \psi(\mathbf{x}_m), \psi(\mathbf{x}_{m'}) \rangle$.

The kernel function is a similarity function for covariance over all pairs of data points. In the case of atomic simulations, it defines the similarity between different LAEs. This section focuses on kernel regression models for building interatomic potentials, such as Gaussian Approximation Potentials (GAPs) [33]. GAP is based on GPs and approximates the local atomic energy with a user-specified kernel. This kernel-based interatomic potential has been employed to predict the energy and forces for a wide range of materials, including metals, semiconductors, and amorphous solids [187, 188, 189, 102, 30, 190]. In the following subsections, we will detail this method, as well as the developments and optimizations we have made for it.

2.2.1 Kernel model: definition and formulation

Kernel machine learning (KML) regressions using D-dimensional descriptors $\mathbf{x} \in \mathbb{R}^D$ is a nonlinear regression model. Basically, it involves computing the kernel $k\left(\mathbf{x}_m, \mathbf{x}_{m'}\right)$ for every unordered pair of data points $(\mathbf{x}_m, \mathbf{x}_{m'})$ in the database $\mathcal{D} = \{\mathbf{x}_m \in \mathbb{R}^D \text{ for } m = 1 \dots M\}$. Instead of performing $M(M+1)$ kernel calculations, a sparse approximation is used to reduce the training complexity in the GAP approach. It is possible to select K representative points, often referred to as sparse points [33, 191, 192] in the literature, with $K \leq M$. In ML models, many selections procedures and algorithms can be used in order to ensure the "representativeness" of the sparse points with respect to the database and the type of regression.

The selection using statistical distances is the focus of the present discussion, but let us first demonstrate how to integrate the design matrix into the kernel formalism. Assume that the K sparse points have been selected among the M database points. In that case, the application $\phi(\mathbf{x}_m) = k(\mathbf{x}_m, \cdot) = k(\cdot, \mathbf{x}_m)$, where $\mathbf{x}_m \in \mathcal{D} \subset \mathbb{R}^D$ and $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is a function having the kernel properties, is defined on the subset $\mathcal{K} \subset \mathcal{D}$ consisting of the K selected sparse points $\mathcal{K} = \{\mathbf{z}_k \in \mathcal{D} \text{ for } k = 1, \dots, K\}$. $\phi(\mathbf{x}_m)$ can then be expressed by listing all its values over the domain $\mathcal{K}$, indicated by $\varphi_p(\mathbf{x}_m) = k(\mathbf{x}_m, \mathbf{z}_p)$, where p is one point among the K selected points. We have

$$\phi(\mathbf{x}_m) = \begin{pmatrix} \varphi_1(\mathbf{x}_m) \\ \varphi_2(\mathbf{x}_m) \\ \vdots \\ \varphi_K(\mathbf{x}_m) \end{pmatrix} = \begin{pmatrix} k(\mathbf{x}_m, \mathbf{z}_1) \\ k(\mathbf{x}_m, \mathbf{z}_2) \\ \vdots \\ k(\mathbf{x}_m, \mathbf{z}_p) \\ \vdots \\ k(\mathbf{x}_m, \mathbf{z}_K) \end{pmatrix} \in \mathbb{R}^{K \times 1}, \tag{2.56}$$

and the kernel-based design matrix over the database becomes:

$$\begin{aligned}\boldsymbol{\Phi}^{\mathrm{KML}}(\mathcal{D}) &= \begin{pmatrix}\phi^{\top}(\mathbf{x}_1)\\ \phi^{\top}(\mathbf{x}_2)\\ \vdots\\ \phi^{\top}(\mathbf{x}_M)\end{pmatrix} \\ &= \begin{pmatrix}k(\mathbf{x}_1,\mathbf{z}_1) & k(\mathbf{x}_1,\mathbf{z}_2) & \cdots & k(\mathbf{x}_1,\mathbf{z}_K)\\ \vdots & \vdots & \ddots & \vdots\\ k(\mathbf{x}_M,\mathbf{z}_1) & k(\mathbf{x}_M,\mathbf{z}_2) & \cdots & k(\mathbf{x}_M,\mathbf{z}_K)\end{pmatrix} \in \mathbb{R}^{M\times K}.\end{aligned} \tag{2.57}$$

In this case, the dimension of the parameters $\mathbf{w}$ is K. Moreover, the collection of $\mathbf{z}$ points from the database $\mathcal{D}$ is also part of the parametrization. Consequently, the total number of real numbers required for the parametrization of a KML model is $K + KD$.

Let us now introduce how the present formulation of kernel regression is related to GPs. In a standard noisy GP for the database $\mathcal{D}$, the design matrix has a square shape $K = M$, which means that all the data points of the database are selected. Then the prediction $y_\star$ for a new point $\mathbf{x}_\star$ is made using the key predictive equations [100]:

$$y_\star \sim \mathcal{N}(\mu(\mathbf{x}_\star), \mathbb{V}(\mathbf{x}_\star)),$$

$$\mu(y_\star) = \mathbf{k}(\mathbf{x}_\star)^{\top}(\mathbf{K}_{MM} + \sigma_M^2\mathbf{I}_M)^{-1}\mathbf{y}, \tag{2.58}$$

$$\mathbb{V}(y_\star) = k(\mathbf{x}_\star,\mathbf{x}_\star) - \mathbf{k}(\mathbf{x}_\star)^{\top}(\mathbf{K}_{MM} + \sigma_M^2\mathbf{I}_M)^{-1}\mathbf{k}(\mathbf{x}_\star), \tag{2.59}$$

with $\mathbf{k}(\mathbf{x}_\star)^{\top} = (k(\mathbf{x}_1,\mathbf{x}_\star),\ldots,k(\mathbf{x}_M,\mathbf{x}_\star))$ and $\mathbf{K}_{MM} = \boldsymbol{\Phi}^{\mathrm{KML}}(\mathcal{D}) \in \mathbb{R}^{M\times M}$. The similarity of these equations to Eq. 2.49 and Eq. 2.50 is evident. Actually, Eq. 2.49 and Eq. 2.50 can be transformed to Eq. 2.58 and Eq. 2.59 by regarding the kernel function $\mathbf{k}(\mathbf{x}_\star)$ as the map feature $\phi(\mathbf{x}_\star)$ in Eq. 2.49 and Eq. 2.50, and assuming a prior of the parameters $\mathbf{w} \sim \mathcal{N}(0, \mathbf{K}_{MM}^{-1})$. Alternatively speaking, this transformation can be achieved by applying the following replacements to Eq. 2.49 and Eq. 2.50:

$$\phi(\mathbf{x}_\star) \to \mathbf{k}(\mathbf{x}_\star), \boldsymbol{\Phi}(\mathcal{D}) \to \mathbf{K}_{MM} = \mathbf{K}_{MM}^{\top} \quad \text{and} \quad \boldsymbol{\Sigma}_P \to \mathbf{K}_{MM}^{-1}. \tag{2.60}$$

The GP prediction can be obtained from a linear fit in parameters using a generalized finite basis $y_\star = \sum_{m=1}^{M} w_m k(\mathbf{x}_\star,\mathbf{x}_m)$ with a prior $\mathbf{w} \sim \mathcal{N}(0, \mathbf{K}_{MM}^{-1})$.

The design matrix for standard kernel or GP approaches ($K = M$) is called the full kernel matrix $\mathbf{K}_{MM}$. In cases where $K < M$ sparse points are selected, the kernel matrix between all points in the database $\mathcal{D}$ and the selected points is called the partial kernel matrix, denoted by $\mathbf{K}_{MK} = \boldsymbol{\Phi}^{\mathrm{KML}}(\mathcal{D}) \in \mathbb{R}^{M\times K}$. We similarly denote the transpose of $\mathbf{K}_{MK}$ by $\mathbf{K}_{KM} = \mathbf{K}_{MK}^{\top} \in \mathbb{R}^{K\times M}$, and the square kernel matrix of the sparse points by $\mathbf{K}_{KK} = \boldsymbol{\Phi}^{K}(\mathcal{K}) \in \mathbb{R}^{K\times K}$. As K is less than M, the predictive equations can be given by preforming the same transformation:

$$\phi(\mathbf{x}_\star) \to \mathbf{k}(\mathbf{x}_\star) = (k(\mathbf{x}_1,\mathbf{x}_\star),\ldots,k(\mathbf{x}_K,\mathbf{x}_\star))^{\top}, \tag{2.61}$$

$$\boldsymbol{\Phi}(\mathcal{D}) \to \mathbf{K}_{MK} \quad \text{and} \quad \boldsymbol{\Sigma}_P^{-1} \to \mathbf{K}_{KK}. \tag{2.62}$$

on Eq. 2.47 and Eq. 2.48. This operation gives the complete expression for the mean and variance of the prediction $y_\star = \sum_{k=1}^{K} w_k k(\mathbf{x}_\star,\mathbf{z}_k)$ with a prior $\mathbf{w} \sim \mathcal{N}(0, \mathbf{K}_{KK}^{-1})$:

$$\mu(y_\star) = \mathbf{k}(\mathbf{x}_\star)^{\top}\left(\sigma_M^{-2}\mathbf{K}_{KM}\mathbf{K}_{MK} + \mathbf{K}_{KK}\right)^{-1}\mathbf{K}_{KM}\mathbf{y}, \tag{2.63}$$

$$\mathbb{V}(y_\star) = \mathbf{k}(\mathbf{x}_\star)^{\top}\left(\sigma_M^{-2}\mathbf{K}_{KM}\mathbf{K}_{MK} + \mathbf{K}_{KK}\right)^{-1}\mathbf{k}(\mathbf{x}_\star). \tag{2.64}$$

As can be seen, both approaches can be easily transformed from one to the other. This method is sometimes called a "subset of regressors" as suggested by Wahba [193] and Rasmussen [100]. The advantage of using such a partial kernel matrix is the reduction in computational complexity from $\mathcal{O}(M^3)$ to $\mathcal{O}(MK^2)$. After training is completed, the prediction of the mean and variance has complexities of $\mathcal{O}(K)$ and $\mathcal{O}(K^2)$, respectively.

As we have pointed out earlier in the case of quadratic models (QML), the KML model can also be hybridized with a linear model (LML) to form a Kernel Noise Machine Learning (KNML) model:

$$\mathbf{\Phi}^{\mathrm{KNML}}\left(\mathcal{D}\right) = \mathbf{\Phi}^{\mathrm{LML}}\left(\mathcal{D}\right) \oplus \mathbf{\Phi}^{\mathrm{KML}}\left(\mathcal{D}\right) \in \mathbb{R}^{M\times D} \oplus \mathbb{R}^{M\times K} \in \mathbb{R}^{M\times(D+K)}\,. \tag{2.65}$$

Similar to QNML, the parameters are in direct product $\mathbf{w} = \mathbf{w}^{\mathrm{LML}} \oplus \mathbf{w}^{\mathrm{KML}}$. This can be interpreted as an extension with K components of the original D-dimensional descriptor space. The total energy of the system s is the sum of the local atomic energies:

$$\begin{aligned} E_s &= \sum_{a\in s} \epsilon_{s,a}^{\mathrm{KNML}} = \sum_{a\in s} \epsilon_{s,a}^{\mathrm{LML}} + \sum_{a\in s} \epsilon_{s,a}^{\mathrm{KML}}\,, &(2.66)\\ \epsilon_{s,a}^{\mathrm{KNML}} &= (\mathbf{w}^{\mathrm{LML}})^{\top}\mathbf{D}_{s,a} + (\mathbf{w}^{\mathrm{KML}})^{\top}\mathbf{k}(\mathbf{D}_{s,a}) \\ &= (\mathbf{w}^{\mathrm{LML}} \oplus \mathbf{w}^{\mathrm{KML}})^{\top}\left[\mathbf{D}_{s,a} \oplus \mathbf{k}(\mathbf{D}_{s,a})\right]\,. &(2.67)\end{aligned}$$

From the above equations it is straightforward to see that the new local atomic descriptor becomes $\mathbf{D}_{s,a} \oplus \mathbf{k}(\mathbf{D}_{s,a})$ and the energy descriptor of system s becomes $\mathbf{x}_s \oplus \mathbf{x}'_s \in \mathbb{R}^{D+K}$ where $\mathbf{x}_s = \sum_{a\in s}\mathbf{D}_{s,a} \in \mathbb{R}^D$ and $\mathbf{x}'_s = \sum_{a\in s}\mathbf{k}(\mathbf{D}_{s,a}) \in \mathbb{R}^K$.

2.2.2 Selection of sparse points

In this section, we will talk about how to build and optimize the kernel by selecting the most representative data points of the database as the so-called sparse points. Firstly, we introduce the notion of an atomic design matrix associated with a database. The difference between the atomic design matrix and the previous design matrix is that here each line lays the local energy descriptor $\mathbf{D}_{s,a}$ of one atom a where a runs over the entire array of N_s atoms from the s^{th} system of the database, instead of the feature-mapped energy, force, or stress descriptors as in the case of the previous design matrix. The underlying local atomic database is denoted by $\mathcal{D}_a = \{\mathbf{x}_m^a \equiv \mathbf{D}_{s,a} \in \mathbb{R}^{D\times 1} \text{ for } m = 1,\ldots,M_a\}$ where M_a is the total number of LAEs of the database, and there is a bijective function that map the m^{th} component of the database into the tuple of integers (s,a), i.e., $h(m) = (s,a)$ and $h^{-1}(s,a) = m$. The atomic design matrix $\mathbf{\Phi}^a$ over the atomic database $\mathcal{D}_a$ is defined as:

$$\mathbf{\Phi}^a\left(\mathcal{D}_a\right) = \begin{pmatrix} \mathbf{x}_1^{a\top} \\ \vdots \\ \mathbf{x}_{M_a}^{a\top} \end{pmatrix} = \begin{pmatrix} x_1^{a,(1)} & x_1^{a,(2)} & \cdots & x_1^{a,(D)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M_a}^{a,(1)} & x_{M_a}^{a,(2)} & \cdots & x_{M_a}^{a,(D)} \end{pmatrix} \in \mathbb{R}^{M_a\times D}\,. \tag{2.68}$$

We have implemented five procedures for selecting the sparse points: (i) random selection, and four others based on (ii) statistical distance, (iii) normalized statistical distance, (iv) CUR decomposition, and (v) statistical distance-based CUR decomposition.

Random selection is the most straightforward and fast procedure: we simply pick K lines among the M lines of the atomic design matrix $\mathbf{\Phi}^{\mathrm{a}}(\mathcal{D}_a)$. By defining an injective function m that randomly maps the first K positive integers into the first $M \geq K$ positive integers, the set of sparse points can be written as $\mathcal{K} = \{\mathbf{z}_k = \mathbf{x}^a_{m(k)} \in \mathcal{D} \text{ for } k = 1, \ldots, K\}$. In practice, the random selection is performed within each class (e.g., the column "DB class" in Table 2.1) separately, and the number of random picks p_c within a class c is normalized by the proportion of the number of LAEs in this class (M^c_a) to the total number, i.e., $p_c = KM^c_a/M_a$.

Statistical distance selection is based on uniform sampling of the statistical distances between LAEs and a given dataset. For this purpose, let us consider a subset $\mathcal{S} \subset \mathcal{D}_a$ with $n_{\mathcal{S}}$ LAEs on which we define the statistical distance $d_{\mathcal{S}}(\mathbf{x}^a) \in \mathbb{R}$ of each point $\mathbf{x}^a \in \mathcal{D}_a$:

$$\begin{aligned} d_{\mathcal{S}}(\mathbf{x}^a) &= \left[\left(\mathbf{x}^a - \boldsymbol{\mu}_m\right)^\top \mathbf{\Sigma}_{\mathcal{S}}^{-1} \left(\mathbf{x}^a - \boldsymbol{\mu}_m\right) \right]^{\frac{1}{2}} , \\ \boldsymbol{\mu}_{\mathcal{S}} &= \frac{1}{n_{\mathcal{S}}} \sum_{s \in \mathcal{S}} \mathbf{x}^a_s , \\ \mathbf{\Sigma}_{\mathcal{S}} &= \frac{1}{n_{\mathcal{S}} - 1} \sum_{s \in \mathcal{S}} \left(\mathbf{x}^a_s - \boldsymbol{\mu}_{\mathcal{S}}\right) \left(\mathbf{x}^a_s - \boldsymbol{\mu}_{\mathcal{S}}\right)^\top , \end{aligned} \tag{2.69}$$

where $\mathbf{\Sigma}_{\mathcal{S}} \in \mathbb{R}^{D \times D}$ and $\boldsymbol{\mu}_{\mathcal{S}} \in \mathbb{R}^D$ are the sample covariance matrix and the center of set $\mathcal{S}$, respectively. For $\mathcal{S} = \mathcal{D}_a$, we obtain for any $x_m \in \mathcal{D}_a$ the Mahalanobis distance $d(\mathbf{x}^a)$. There are two important observations: (i) in the case of the Mahalanobis distance, the sample covariance matrix can be easily computed from the atomic design matrix $\mathbf{\Sigma}_{\mathcal{D}_a} = \left[\mathbf{\Phi}^a(\mathcal{D}_a)^\top \mathbf{\Phi}^a(\mathcal{D}_a) - M_a \left(\mu_i \mu_j\right)_{i,j} \right] / \left(M_a - 1\right)$ and (ii) when the data in the domain $\mathcal{D}_a$ has an underlying Gaussian distribution, the Mahalanobis distance follows a $\chi^2(D)$ distribution.

The interest in using statistical distance for selecting the most representative points stems from its foundational role in the recently introduced concept of distortion score [49]. For a specific choice of the set $\mathcal{S}$, where the determinant of the sample covariance matrix $\mathbf{\Sigma}_{\mathcal{S}}$ is minimal [194, 195], the corresponding statistical distance becomes the distortion score [49]. The distortion score of LAEs describes a statistical distance from a reference distribution in the feature space of atomic descriptors. The reference distribution derived from $\mathcal{S}$ can be constructed from the LAEs of a defect-free crystalline system at a given temperature, or from a subset of atoms of particular interest. It is pointed out by Goryaeva *et al.* [49] that when computed with respect to the distribution of the underlying bulk structures, the distortion score demonstrates a correlation with the local atomic energy, as shown in Fig. 2.9. The comparison is performed in bcc Fe, for the atomic arrays with three classes of structural defects: vacancies, self-interstitials and stacking faults. These configurations are included in the training database of the GAP for Fe [29]. The atomic energies are computed using the same potential. For all three defect classes, the determination-correlation coefficient R^2 between the distortion score $d_{\mathcal{S}}$ and local energy is higher than 80 %. Both concepts, local atomic energy and distortion score, encode the geometric information of the LAE.

The strong correlation between the distortion score described by statistical distances and the local energy suggests a physical selection of sparse points from databases designed for materials modeling of defects in crystalline solids. This conjecture enables us to select sparse

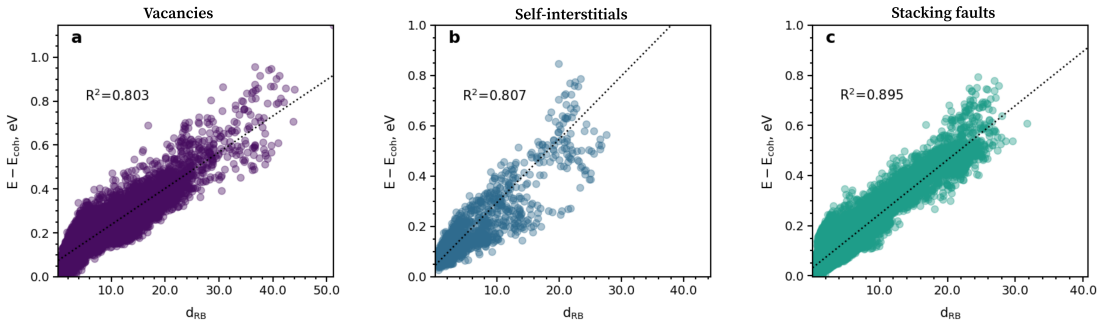

Figure 2.9: The correlation between energy per atom and the distortion score $d_{\mathcal{S}}(\mathbf{x})$ [49]. The distortion score is described via a selection of points from classes of physical interest and using MCD algorithm [196, 194, 195] that gives a robust distance d_{RB} in bcc Fe systems with: (a) vacancies; (b) self-interstitials; (c) stacking faults. Each point on the plot represents an individual atom in a simulation box. The atomic arrays are taken from the GAP potential database [29]. The correlation is performed over 103 000 LAEs, and each defect class gathers diverse instances from 0 K static relaxation to molecular dynamics simulations at various temperatures. MCD analysis is performed on the structural data represented using bispectrum SO(4) descriptor [33, 50] with the angular moment $j_{\max}$=4.5. The atomic energies are computed with the GAP potential [29].

points only based on geometrical considerations, which are representative for the entire range of atomic energies involved in the fit. With this respect, we define a linear grid of G bins between $d_{\min} = \min_{\mathbf{x}^a \in \mathcal{D}_a} d_{\mathcal{S}}(\mathbf{x}^a)$ and $d_{\max} = \max_{\mathbf{x}^a \in \mathcal{D}_a} d_{\mathcal{S}}(\mathbf{x}^a)$, with an interval of $\delta d = (d_{\max} - d_{\min})/G$. The g^{th} bin is defined as the collection of G_g points:

$$\begin{aligned} \mathcal{G}_g &= \{\mathbf{x}^a \in \mathcal{D}_a \mid x_{g+1} \geq d_{\mathcal{S}}(\mathbf{x}^a) \geq x_g\}\,, \\ x_g &= d_{\min} + (g-1)\delta d\ . \end{aligned} \tag{2.70}$$

In the selection based on **statistical distance**, we divide the range of $[d_{\min}, d_{\max}]$ into K bins, i.e., $G = K$ where K is the target number of sparse points. Then we simply select one point from each grid set $\mathcal{G}_g$ if there is more than one point in it. If there are no points within this set, it is skipped, and no point is selected. Additionally, we have implemented an alternative selection approach known as the **normalized statistical distance** method. In this case, the number of points collected from the set $\mathcal{G}_g$, denoted as K_g, is proportional to its population: $K_g \sim G_g/M_a$.

CUR matrix decomposition [48] is an enhanced data analysis method. It involves a low-rank matrix decomposition that is explicitly represented in terms of a small number of actual columns and/or rows of the data matrix. This method is particularly useful for sparse matrices and falls within the same class as traditional linear algebra techniques like singular value decomposition (SVD) and principal component analysis (PCA). However, a significant difference between CUR decomposition and these methods is the the direct construction from original data elements. Since then, CUR decomposition is interpretable by practitioners of the field from which the data are drawn, keeping the "real" meaning of the data.

In our case, performing CUR decomposition to select the rows from the atomic design matrix allows to select the sparse points from the database. The CUR decomposition of the atomic design matrix gives $\mathbf{\Phi}^a(\mathcal{D}_a) = \mathbf{CUR}$. It means that a set of columns $\mathbf{C}$ and a set of rows $\mathbf{R}$ are directly chose from the original matrix. $\mathbf{C}$ and $\mathbf{R}$ play the same role as the

matrices $\mathbf{U}$ and $\mathbf{V}$ in SVD decomposition ($\mathbf{\Phi}^a(\mathcal{D}_a) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$). Similar to standard SVD, CUR decomposition is an approximation of the original matrix. Truncated SVD is widely used, as it provides the most accurate k-rank approximation $\mathbf{\Phi}^a_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^\top$ in terms of the Frobenius norm. However, the vectors $\mathbf{U}_k$ and $\mathbf{V}_k$ themselves may lack any meaning in the field from which the data are drawn. After all, the singular vectors are mathematical abstractions that mix all the columns and rows of the original matrix, and they can be calculated for any data matrix. These abstract objects do not have any physical reality. Therefore, SVD is not applicable in the current context.

Back to CUR decomposition. In order to identify the sparse points from the matrix $\mathbf{\Phi}^a(\mathcal{D})$, we are interested in sampling the most representative rows. We denote by $\mathbf{A} = \mathbf{\Phi}^a(\mathcal{D}_a)^\top \in \mathbb{R}^{D\times M_a}$ the transpose of the atomic design matrix, from which we extract the most important column features $\mathbf{x}^a_m$. The only information that we have are: (i) $\mathbf{A} \in \mathbb{R}^{D\times M_a}$ (ii) the number of columns c that we intend to select, $1 \leq c \leq M_a$ and (iii) the order k of CUR decomposition $1 \leq k \leq r_A = \text{rank}(\mathbf{A}) \leq \min(D, M_a)$. The CUR algorithm proposed by Mahoney *et al.* [48] proceeds as follows:

1. Perform k-rank SVD decomposition $\mathbf{A} = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^\top$ and compute the normalized statistical leverage scores of the m^{th} column as:

$$\text{ls}_m = \frac{1}{k}\sum_{j=1}^{k}(\mathbf{V}_k)^2_{m,j}$$

   where $(\mathbf{V}_k)_{m,j}$ is the element (m, j) of the matrix $\mathbf{V}_k \in \mathbb{R}^{M_a\times k}$. With this normalization, we have $\text{ls}_m \geq 0$ and $\sum_{m=1}^{M_a}$, and thus that these scores form a probability distribution over the M_a columns.

2. Based on the probability $p_m = \min(1, c\text{ls}_m)$, we perform c attempts to select the columns from the entire set of M_a columns:

   $i = 0$
   for t from 1 to c **do**
   select $m \in \{1, \cdots, M_a\}$, compute $p_m = \min(1, c\text{ls}_m)$
   sample $x \sim \mathcal{U}(0, 1)$
   if $p_m \geq x$ **then**
   the column m is selected
   $i = i + 1$
   set i^{th} column of matrix $\mathbf{C} \in \mathbb{R}^{c\times D}$ as $(\mathbf{C})_{:,i} = \mathbf{A}_{:,m}$
   else
   the column m is not selected
   end if
   end for
   the final number of selected column $c' = i$

3. Return the matrix $\mathbf{C} \in \mathbb{R}^{D\times c'}$.

With this procedure, the matrix $\mathbf{C}$ contains c' columns, where $c' \leq c$ in expectation.

The CUR algorithm can also be used to perform a selection of r rows from the entire D rows of the matrix $\mathbf{A}$. This selection can be easily done by applying the above column algorithm to the transpose matrix $\mathbf{A}^\top = \mathbf{\Phi}^a(\mathcal{D}_a)$. The selection of rows for the matrix $\mathbf{A}$ actually corresponds to the selection of dimensions of the descriptor among all its D dimensions. For a high-dimensional descriptor (where D is large), this operation allows to reduce the dimensionality by retaining only the most representative dimensions. Finally, the matrix $\mathbf{A}$ is decomposed as the product of three matrices: $\mathbf{A} = \mathbf{\Phi}^a(\mathcal{D}_a)^\top = \mathbf{CUR}$ with $\mathbf{C} \in \mathbb{R}^{D\times c'}$, $\mathbf{U} \in \mathbb{R}^{c'\times r'}$ and $\mathbf{R} \in \mathbb{R}^{r'\times M_a}$. Each column of the matrix $\mathbf{C}$ represents a sparse point $\mathbf{z}_k \in \mathbb{R}^D$ and we have the number of sparse points $K = c'$.

Finally, we have implemented an additional selection method that combines the methods based on statistical distance and CUR decomposition. We replace simply the probability $p_m = \min(1, c\mathrm{ls}_m)$ in the CUR selection algorithm by $p_m = p' \exp\left(-d_{\mathcal{S}}(\mathbf{x}^a_m)^2/2\right)$ with $p' = c/\sum_{s=1}^{M_a} \exp\left(-d_{\mathcal{S}}(\mathbf{x}^a_s)^2/2\right)$. This formalism is inspired by Ref. [49], where the statistical distance is regarded as an analogy to the energy. In this way, SVD decomposition is not necessary. Similar to CUR decomposition selection, the number of selected sparse points K is not exactly equal to the user-specified c, but they are very close. This method is called **statistical distance-based (d-based) CUR decomposition**.

Figure 2.10 illustrates the distribution of sparse points among the entire database $\mathcal{D}_a$, selected by the four aforementioned methods. Here, we try to select 5 000 sparse points from a bcc W database, completed from Table 2.1, which contains a total of 48 557 LAEs ($M_a = 48\,557$). The number of sparse points actually selected by the methods based on statistical distance, normalized statistical distance, CUR decomposition, and d-based CUR decomposition is 3 607, 4 655, 4 821 and 4 934, respectively. The vertical axis in Fig. 2.10 indicates the class of the database to which a LAE $\mathbf{x}^a_m$ belongs, while the horizontal axis $d_{\mathcal{D}_a}$ denotes the statistical distance between this data point and the entire database $\mathcal{D}_a$. All data points within the database are plotted in dark blue, and the selected sparse points among them for building the kernel are represented in light blue. The histogram of data points distribution according to statistical distance ($\{d_{\mathcal{D}_a}(\mathbf{x}^a_m)$ for $m = 1, M_a\}$) and configuration class (column "DB class" in Table 2.1) is displayed at the top and right-hand side of each subfigure. We can see that the selections based on normalized statistical distance and d-based CUR decomposition are more "homogeneously" distributed in the database, and they better follow the distribution over the entire database with respect to both statistical distance and configuration class. Given their expressions, this is evident. However, CUR decomposition emphasizes the importance of "rare" data, that is, the minority of data points that are distant from the main cluster, in accurately representing the database. The classes favored by CUR decomposition (classes 18, 19, 22, 23) are actually the liquid configurations obtained from MD simulations that exhibit large statistical distances from the overall database.

Moreover, the selection of $\mathcal{K}$ can be biased by user-selected classes. Besides the general selection from the entire database that builds $\mathcal{K}$, it is possible to make an additional selection from several classes that the user considers more important for the target problem, denoted as $\mathcal{K}_c$. In this case, the final set of sparse points becomes $\mathcal{K} \cup \mathcal{K}_c$.

How is the performance of the four selection methods? To better address this question, models using these selection methods were trained and tested on the Ta-Ti-V-W high-entropy

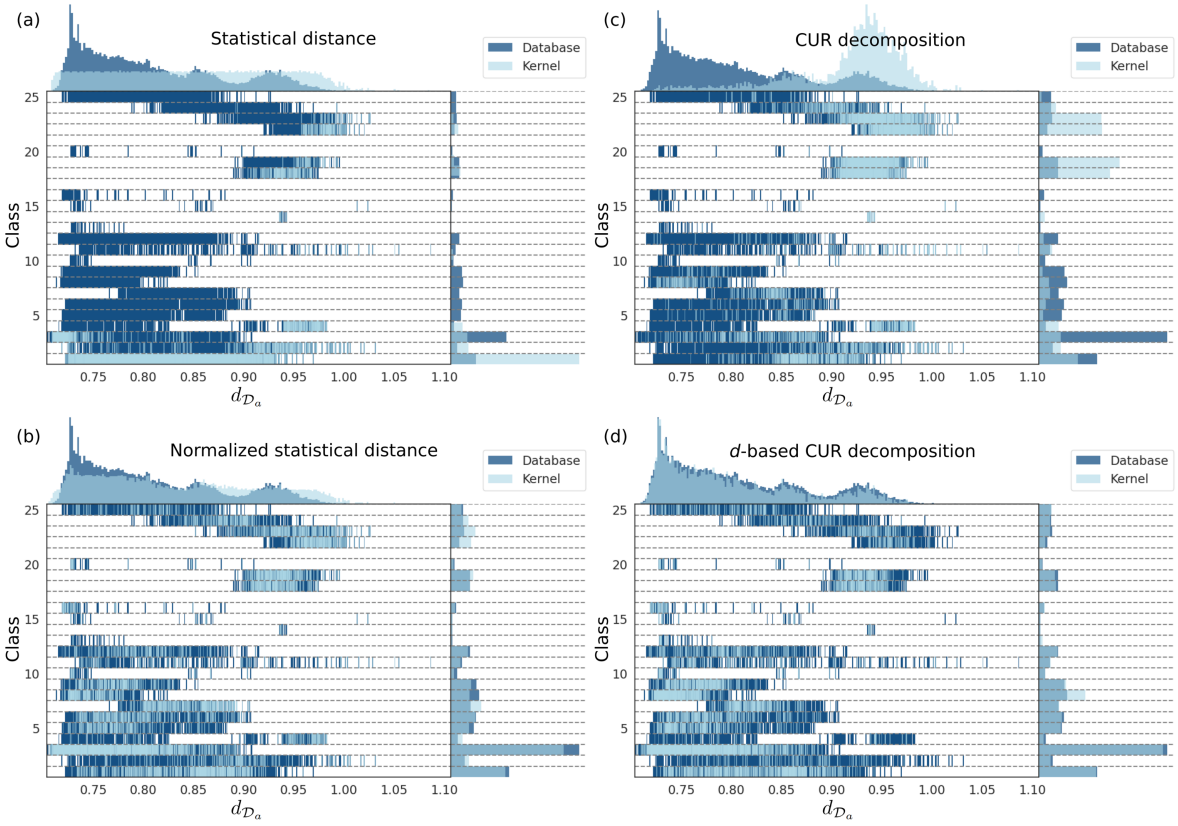

Figure 2.10: The selection of sparse points with four models using (a) statistical distance, (b) normalized statistical distance, (c) CUR decomposition and (d) statistical distance-based (d-based) CUR decomposition. The horizontal axis $d_{\mathcal{D}_a}$ denotes the statistical distance between a LAE $\mathbf{x}_m^a$ and the entire database $\mathcal{D}_a$, and the various classes of the bcc W database are indicated by the vertical axis. All data points within the database are plotted in dark blue, while the selected sparse points $\mathbf{z}_k$ in the kernel are represented in light blue. The histogram of data points distribution according to statistical distance and configuration class is also provided, displayed at the top and right-hand side of each sub-figure, respectively.

alloy system. We consider a polynomial kernel, the formulation of which is detailed in the next subsection 2.2.3.1, with $l_{\mathrm{PO}} = 0.05$, $\sigma_{\mathrm{PO}} = 0$ and $p = 4$ in Eq. 2.74. The models are trained and tested on a database containing the configurations listed in Table 4.3, wherein there are 883 896 LAEs. The RMSE in energy and force prediction, evaluated on both the training and testing datasets, is presented in Fig. 2.11. For all four methods, as the number of sparse points increases, the RMSE decreases, indicating an improvement in accuracy. However, the differences in RMSE among the methods are not significant. In terms of numerical efficiency, we notice that the complexity of CUR decomposition and of the other three methods is respectively $\mathcal{O}(MD^2)$ and $\mathcal{O}(MD)$ with D varying from 10^2 to 10^3, while the difference in accuracy between them, in terms of the RMSE, is not remarkable. Hence, the methods based on normalized statistical distance and d-based CUR provide better trade-off between the computational cost and the prediction performance. Moreover, the d-based CUR method results in a much better consistency between the distribution of the selected kernel and the entire database. As shown in Fig. 2.10(c), the original CUR decomposition tends to select outliers from the database, which cannot effectively represent the underlying physics. In contrast, the d-based CUR

decomposition captures the database in a more physical manner by respecting its distribution (Fig. 2.10(d)).

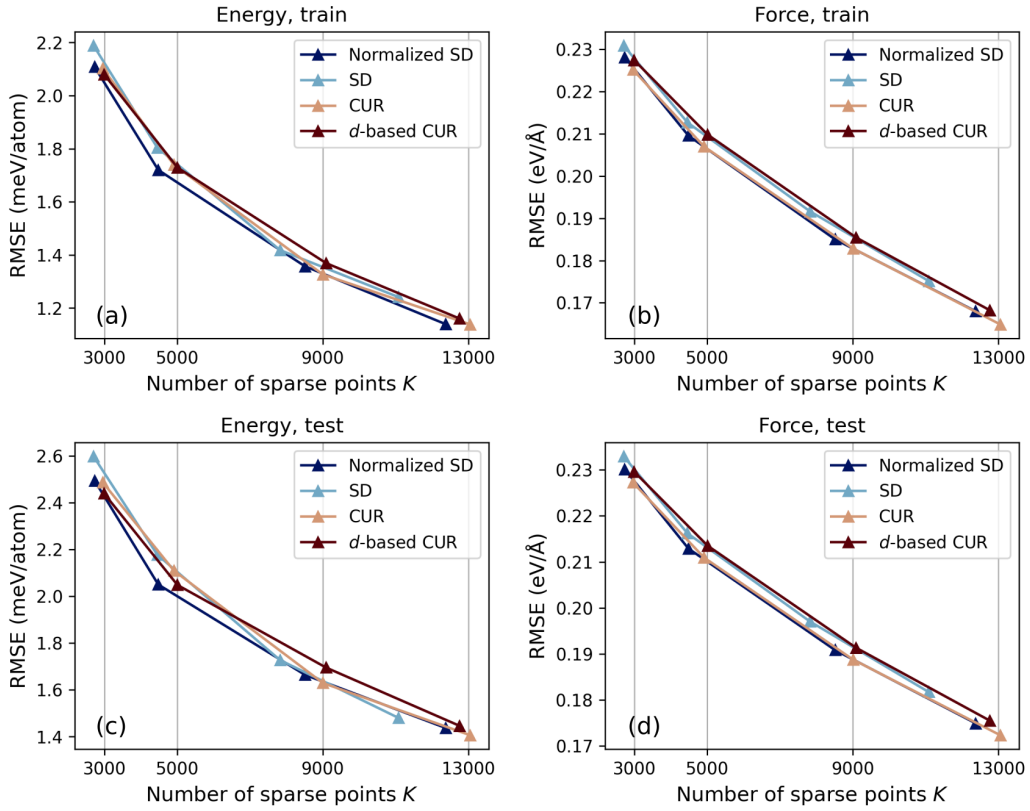

Figure 2.11: RMSE in energy and force prediction by the polynomial kernel potentials with different number of sparse points K and four selection methods: statistical distance (SD), normalized statistical distance (Normalized SD), CUR decomposition (CUR) and statistical distance-based CUR decomposition (d-based CUR). The RMSE is evaluated on the training/testing datasets of Ta-Ti-V-W high entropy alloy system. The vertical grid lines at 3 000, 5 000, 9 000 and 13 000 denote the user-specified number of sparse points.

2.2.3 Examples of kernel functions

In the upcoming subsection, we will explore various symmetric, positive definite kernel functions used in the construction of kernel regression potentials. The kernels are crucial for ensuring accuracy and efficiency of the ML model. We will cover well-known kernels such as the squared-exponential, polynomial, Mahalanobis and n-body kernels based on their mathematical formulations and applications.

2.2.3.1 Squared-exponential, polynomial and Mahalanobis kernels

Firstly, we focus on three types of kernel based on the descriptors associated with the selected sparse points. Here the k^{th} effective sparse point $\mathbf{z}_k$ is centered by being recomputed from the original local energy descriptor of the corresponding atomic environment $\mathbf{D}_k \in \mathbb{R}^{D\times 1}$:

$$\mathbf{z}_k = \frac{\mathbf{D}_k - \boldsymbol{\mu}_K}{\mathrm{Var}_K} \tag{2.71}$$

with $\boldsymbol{\mu}_K = \frac{1}{K}\sum_{k=1}^{K}\mathbf{D}_k$ and $\mathrm{Var}_K = \frac{1}{K}\sum_{k=1}^{K}|\mathbf{D}_k - \boldsymbol{\mu}_K|^2$, and we employ the normalized version of kernel:

$$k(\mathbf{x}_m, \mathbf{x}'_m) = \frac{\tilde{k}(\mathbf{x}_m, \mathbf{x}'_m)}{\sqrt{\tilde{k}(\mathbf{x}_m, \mathbf{x}_m)}\sqrt{\tilde{k}(\mathbf{x}'_m, \mathbf{x}'_m)}}\,. \tag{2.72}$$

There is a wide variety of kernel functions $\tilde{k}$ available. Two standard kernels with tunable hyperparameters are presented as follows.

Squared-exponential (SE) kernel

$$\tilde{k}(\mathbf{D}_{s,a}, \mathbf{z}_k) = \sigma_{\mathrm{SE}}^2 \exp\left(-\frac{|\mathbf{D}_{s,a} - \mathbf{z}_k|^2}{2l_{\mathrm{SE}}^2}\right) \tag{2.73}$$

Polynomial (PO) kernel

$$\tilde{k}(\mathbf{D}_{s,a}, \mathbf{z}_k) = \left(\sigma_{\mathrm{PO}}^2 + \frac{\mathbf{D}_{s,a}\cdot\mathbf{z}_k}{2l_{\mathrm{PO}}^2}\right)^p \tag{2.74}$$

Here σ, l and p are the kernel hyperparameters.

A third kernel, which is hyperparameter-free, is also available based on statistical distance in the Mahalanobis form.

Mahalanobis (MA) kernel

The Mahalanobis distance describes the statistical distance between an atomic environment $\mathbf{D}_{s,a}$ and a sparse point $\mathbf{z}_k$ with respect to $\mathcal{K}$ is:

$$\begin{aligned} d(\mathbf{D}_{s,a}, \mathbf{z}_k, \mathcal{K}) &= \left[(\mathbf{D}_{s,a} - \mathbf{z}_k)^\top \boldsymbol{\Sigma}_{\mathcal{K}}^{-1}(\mathbf{D}_{s,a} - \mathbf{z}_k)\right]^{\frac{1}{2}}\,, \\ \boldsymbol{\mu}_{\mathcal{K}} &= \frac{1}{K}\sum_{k=1}^{K}\mathbf{z}_k = \mathbf{0}\,, \\ \boldsymbol{\Sigma}_{\mathcal{K}} &= \frac{1}{K-1}\sum_{k=1}^{K}(\mathbf{z}_k - \boldsymbol{\mu}_{\mathcal{K}})(\mathbf{z}_k - \boldsymbol{\mu}_{\mathcal{K}})^\top = \frac{1}{K-1}\sum_{k=1}^{K}\mathbf{z}_k\mathbf{z}_k^\top \in \mathbb{R}^{D\times D}\,. \end{aligned} \tag{2.75}$$

Based on this concept, the Mahalanobis kernel is defined as:

$$\tilde{k}(\mathbf{D}_{s,a}, \mathbf{z}_k) = (\mathbf{D}_{s,a} - \mathbf{z}_k)^\top \boldsymbol{\Sigma}_{\mathcal{K}}^{-1}(\mathbf{D}_{s,a} - \mathbf{z}_k)\,. \tag{2.76}$$

To compute the atomic forces, we must also provide the derivatives of the kernels. In order to account for the derivatives of the kernel contribution with respect to the coordinates of atom b in the neighborhood of atom a ($b \neq a$), we need the derivatives of the local energy:

$$\nabla_b \epsilon_{s,a}^{\mathrm{KML}} = \sum_{k=1}^{K} w_k^{\mathrm{KML}} \nabla_b k(\mathbf{D}_{s,a}, \mathbf{z}_k) \tag{2.77}$$

$$\nabla_b k(\mathbf{D}_{s,a}, \mathbf{z}_k) = \frac{1}{\sqrt{\tilde{k}(\mathbf{D}_{s,a}, \mathbf{D}_{s,a})}\sqrt{\tilde{k}(\mathbf{z}_k, \mathbf{z}_k)}} \times \left[\nabla_b \tilde{k}(\mathbf{D}_{s,a}, \mathbf{z}_k) - \tilde{k}(\mathbf{D}_{s,a}, \mathbf{z}_k) \frac{\nabla_b \tilde{k}(\mathbf{D}_{s,a}, \mathbf{D}_{s,a})}{2\tilde{k}(\mathbf{D}_{s,a}, \mathbf{D}_{s,a})}\right] \tag{2.78}$$

where derivatives are computed only with respect to the target descriptors and not the sparse points. The two gradient terms to consider for the squared exponential (SE), polynomial (PO), and Mahalanobis (MA) kernels are:

$$\begin{aligned}
\nabla_b \tilde{k}(\mathbf{D}_{s,a}, \mathbf{z}_k) &= \frac{\partial \tilde{k}(\mathbf{D}_{s,a}, \mathbf{z}_k)}{\partial \mathbf{D}_{s,a}} \nabla_b \mathbf{D}_{s,a} \\
\mathrm{SE} &= -\frac{2\sigma_{\mathrm{SE}}^2}{2l_{\mathrm{SE}}^2} \nabla_b \mathbf{D}_{s,a} \cdot (\mathbf{D}_{s,a} - \mathbf{z}_k) \exp\left(-\frac{|\mathbf{D}_{s,a} - \mathbf{z}_k|^2}{2l_{SE}^2}\right) \\
\mathrm{PO} &= \frac{p}{2l_{\mathrm{PO}}^2} \nabla_b \mathbf{D}_{s,a} \cdot \mathbf{z}_k \left(\sigma_{\mathrm{PO}}^2 + \frac{\mathbf{D}_{s,a} \cdot \mathbf{z}_k}{2l_{\mathrm{PO}}^2}\right)^{p-1} \\
\mathrm{MA} &= 2\nabla_b \mathbf{D}_{s,a} \boldsymbol{\Sigma}_{\mathcal{K}}^{-1} (\mathbf{D}_{s,a} - \mathbf{z}_k)
\end{aligned}$$

and

$$\begin{aligned}
\nabla_b \tilde{k}(\mathbf{D}_{s,a}, \mathbf{D}_{s,a}) &= \frac{\partial \tilde{k}(\mathbf{D}_{s,a}, \mathbf{D}_{s,a})}{\partial \mathbf{D}_{s,a}} \nabla_b \mathbf{D}_{s,a} \\
\mathrm{SE} &= 0 \\
\mathrm{PO} &= \frac{2p}{2l_{\mathrm{PO}}^2} \nabla_b \mathbf{D}_{s,a} \cdot \mathbf{D}_{s,a} \left(\sigma_{\mathrm{PO}}^2 + \frac{\mathbf{D}_{s,a} \mathbf{D}_{s,a}}{2l_{\mathrm{PO}}^2}\right)^{p-1} \\
\mathrm{MA} &= 0\,.
\end{aligned}$$

2.2.3.2 2- and 3-body kernels

Another type of kernel is based on the geometric configurations within the neighborhood of an atom. It is natural to decompose the total energy of the system into body-ordered contributions, which can then be summed into local atomic (or site) energies. According to Glielmo *et al.* [197, 198, 199], a smooth translation- and permutation-invariant 2-body kernel between two atomic environments a and b can be defined by summing all the squared exponential distances between the relative positions of the atoms included in these environments:

$$k_{2b}(a, b) = \sum_{j \in v(a)} \sum_{i \in v(b)} \exp -\frac{|\mathbf{r}_{ja} - \mathbf{r}_{ib}|^2}{2\sigma_2^2}\,, \tag{2.79}$$

where $\mathbf{r}_{ja}$ is the vector from the position of atom a to the position of its neighbor atom j and σ_2 is a hyperparameter. It is important to note that the kernels defined as above are not rotation invariant, i.e., for rotations $\mathcal{R}$ and $\mathcal{R}'$, the equality $k(v(a), v(b)) = k(\mathcal{R}v(a), \mathcal{R}'v(b))$ does not always hold. In order to give rotational invariance, we can follow the procedure proposed by Glielmo *et al.* [198] for symmetrization using Haar integration. However, explicit symmetrization through Haar integration invariably requires the evaluation of computationally intensive functions of the atomic positions. Motivated by this observation, one could take an alternative approach and consider symmetric n-kernels defined as functions of the effective rotation-invariant degrees of freedom of n-plets of atoms [199]. In particular, for 2-body ($n = 2$) and 3-body ($n = 3$) kernels we can choose these degrees of freedom to be simply the interatomic distances in atomic pairs and triplets. The resulting kernels are formulated as:

2-body kernel

$$k_{2b}(a, b) = \sum_{j \in v(a)} \sum_{i \in v(b)} \exp - \frac{(r_{ja} - r_{ib})^2}{2\sigma_2^2} \tag{2.80}$$

and

3-body kernel

$$k_{3b}(a, b) = \sum_{\substack{j_1 > j_2 \\ j_1, j_2 \in v(a)}} \sum_{\substack{i_1 > i_2 \\ i_1, i_2 \in v(b)}} \sum_{\mathbf{P} \in \mathcal{P}_3} \exp - \frac{|(r_{j_1 a}, r_{j_2 a}, r_{j_1 j_2})^\top - \mathbf{P}(r_{i_1 b}, r_{i_2 b}, r_{i_1 i_2})^\top|^2}{2\sigma_3^2} \tag{2.81}$$

where $\mathbf{P}$ runs over all 6 permutations of 3 elements, i.e., all elements of the $\mathcal{P}_3$ permutation group.

For the multi-species case:

$$\begin{aligned} k_{2b}(a, b) &= \sum_{j \in v(a)} \sum_{i \in v(b)} \delta_{s_{aj}, s_{bi}} \exp - \frac{(r_{ja} - r_{ib})^2}{2\sigma_2^2} \\ &= \sum_{j \in v(a)} k_{2b, s_{aj}}(r_{ja}, b) \end{aligned} \tag{2.82}$$

where s_{aj} denotes an unordered set of the species of atoms a and b, and

$$\delta_{s_{aj}, s_{bi}} = \begin{cases} 1, & \text{if } s_{aj} = s_{bi} \\ 0, & \text{else} \end{cases} \tag{2.83}$$

The same generalization can also be done for the 3-body kernel. If we denote by $\mathbf{q}_{aij} = (r_{ia}, r_{ja}, r_{ij})^\top \in \mathbb{R}^3$ a tri-dimensional vector for a 3-body triangle aij, then:

$$\begin{aligned} k_{3b}(a, b) &= \sum_{\substack{j_1 > j_2 \\ j_1 \in v(a) \\ j_2 \in v(a)}} \sum_{\substack{i_1 > i_2 \\ i_1 \in v(b) \\ i_2 \in v(b)}} \sum_{\mathbf{P} \in \mathcal{P}_3} \delta_{s_{aj_1 j_2}, \mathbf{P} s_{b, i_1, i_2}} \exp - \frac{|\mathbf{q}_{aj_1 j_2} - \mathbf{P}\mathbf{q}_{bi_1 i_2}|^2}{2\sigma_3^2} \\ &= \sum_{\substack{j_1 > j_2 \\ j_1 \in v(a) \\ j_2 \in v(a)}} k_{3b, aj_1 j_2}(r_{aj_1}, r_{aj_2}, r_{j_1 j_2}, b)\,. \end{aligned} \tag{2.84}$$

The computational cost of evaluating the multi-species kernels described above does not increase with the number of species present in a given environment.

In practice, the n-plets geometries composed by atoms are sampled as sparse points. For two-body kernel, suppose that we have K_{2b} sparse points. The k^{th} sparse point samples two atoms of species (k_1, k_2) at the distance $z_{2b,k}$. The 2-body local energy on the site a can be written as:

$$\begin{aligned}
\epsilon_{2b,a} &= \sum_{j\in v(a)} \sum_{k=1}^{K_{2b}} w_{2b,k} (\delta_{s_a,s_{k_1}} \delta_{s_j,s_{k_2}} + \delta_{s_a,s_{k_2}} \delta_{s_j,s_{k_1}}) \exp - \frac{(r_{ja} - z_{2b,k})^2}{2\sigma_2^2} \\
&= \sum_{k=1}^{K_{2b}} w_{2b,k} \left[\sum_{j\in v(a)} \delta_{s_{aj},s_{k_1 k_2}} \exp - \frac{(r_{ja} - z_{2b,k})^2}{2\sigma_2^2} \right] \\
&= \sum_{j\in v(a)} \left[\sum_{k=1}^{K_{2b}} w_{2b,k} \delta_{s_{aj},s_{k_1 k_2}} k_{2b}(r_{ja}, z_{2b,k}) \right] \\
&= \sum_{j\in v(a)} K_{2b,s_{ja}}(r_{ja})
\end{aligned} \tag{2.85}$$

where the last function $K_{2b,s_{ja}}$ can be viewed as a 2-body potential:

$$K_{2b,s_{ja}}(r_{ja}) = \sum_{k=1}^{K_{2b}} w_{2b,k} \delta_{s_{aj},s_{k_1 k_2}} k_{2b}(r_{ja}, z_{2b,k}), \tag{2.86}$$

and the distance $z_{2b,k}$ is equally distributed from 0 Å to the cutoff radius for each species pair instead of sampling in the database.

The same procedure can be applied for any order of n-body term. For example, the local energy based on the 3-body kernel with multi-species can be expressed as follows:

$$\begin{aligned}
\epsilon_{3b,a} &= \sum_{\substack{j_1>j_2 \\ j_1\in v(a) \\ j_2\in v(a)}} K_{3b,s_{aj_1j_2}}(r_{aj_1}, r_{aj_2}, r_{j_1j_2}) \\
K_{3b,s_{aj_1j_2}}(r_{aj_1}, r_{aj_2}, r_{j_1j_2}) &= \sum_{k=1}^{K_{3b}} w_{3b,k} \delta_{s_{aj_1j_2},s_{k_1k_2k_3}} k_{3b}(r_{aj_1}, r_{aj_2}, r_{j_1j_2}, z_{3b,k}).
\end{aligned} \tag{2.87}$$

The kernel regression can be regarded as the linear regression of kernel LAE descriptors. Let us take the example of the 2-body kernel. As formulated in Eq. 2.85, local energy of the atom a can be expressed as a linear combination of the components of the 2-body descriptor $\mathbf{D}_{2b,s,a} \in \mathbb{R}^{K_{2b}}$. This descriptor takes into account only the radial atomic environment and thus can be a complement of the descriptors with reduced radial description. For the bi-spectrum SO(4) descriptor introduced in the section 2.1.2, a low value of the angular moment j_{max} is sufficient to ensure the accuracy of the angular description. However, this low value of j_{max} will degrade the radial description provided by the spectral descriptor. To guarantee the robustness of the radial description without increasing the value of j_{max} which is proportional

to the computational cost, a hybrid form of atomic descriptors is proposed by Goryaeva *et al.* [38] by combining the radial 2-body kernel descriptor and the bi-spectrum SO(4) with relatively low dimensionality ($j_{max} = 1.5 - 3.5$). The new descriptor is built as a direct sum $\mathbf{D}_{s,a} = \mathbf{B}_{s,a} \oplus \mathbf{D}_{2b,s,a}$ and its dimension is equal to the sum $dim(\mathbf{B}_{s,a}) + dim(\mathbf{D}_{2b,s,a})$.

Hybrid atomic descriptors can also be built upon other fast and less accurate descriptors complemented with the slow and numerically accurate ones. Following this general idea, it is often recommended to combine a many-body descriptor with the 2-body kernel descriptor, denoted as $\mathbf{D}_{s,a} = \mathbf{D}_{many\text{-}b,s,a} \oplus \mathbf{D}_{2b,s,a}$.

2.2.4 Fourier-sampled kernels

Constructing and evaluating positive definite kernel functions for large datasets poses a major challenge in machine learning, particularly in the n-body case discussed in Subsection 2.2.3.2 when n exceeds 3. In this subsection, we will elucidate the reasons for the complexity of this task and present a solution. Specifically, we will demonstrate how to approximate a kernel function by transforming samples from their spectral density by leveraging Bochner's theorem.

2.2.4.1 Invariant n-body kernels and Fourier sampling

We denote by $\mathbf{b}$ an n-plets body term which can be 2-body, 3-body, 4-body, etc. All of them can be written in unique form:

$$k_{nb}(a,b) = \sum_{\mathbf{j} \in v(a)} \sum_{\mathbf{i} \in v(b)} \sum_{\mathbf{P} \in \mathcal{P}_n} \exp - \frac{|\mathbf{x}_{\mathbf{j}a} - \mathbf{P}\mathbf{x}_{\mathbf{i}b}|^2}{2\sigma_n^2}. \tag{2.88}$$

The expressions of $\mathbf{j}$ and $\mathbf{x}_{\mathbf{j}a}$ are determined by the body-order n:

2-body

$$\begin{aligned} \mathbf{j} &\equiv j \\ \mathbf{x}_{\mathbf{j}a} &\equiv r_{ja} \end{aligned}$$

3-body

$$\begin{aligned} \mathbf{j} &\equiv (j_1, j_2) \text{ with } j_1 > j_2 \text{ and } j_1, j_2 \in v(a) \\ \mathbf{x}_{\mathbf{j}a} &\equiv (r_{j_1 a}, r_{j_2 a}, r_{j_1 j_2})^\top \in \mathbb{R}^{3\times 1} \end{aligned}$$

4-body

$$\begin{aligned} \mathbf{j} &\equiv (j_1, j_2, j_3) \text{ with } j_1 > j_2 > j_3 \text{ and } j_1, j_2, j_3 \in v(a) \\ \mathbf{x}_{\mathbf{j}a} &\equiv (r_{j_1 a}, r_{j_2, a}, r_{j_3 a}, r_{j_1 j_2}, r_{j_1 j_3}, r_{j_2 j_3})^\top \in \mathbb{R}^{6\times 1} \end{aligned}$$

The permutation $\mathbf{P}$ in Eq. 2.88 is necessary to preserve permutation symmetry, and, to ensure the symmetry of the kernel, i.e., $k_{nb}(a,b) = k_{nb}(b,a)$. The latter property is simple to prove. A permutation $\mathbf{P} \in \mathcal{P}_n$ can be regarded as a binary matrix in $\mathbb{R}^{n\times n}$ that has exactly one entry of 1 in each row and each column with all other entries 0. This matrix has two nice properties: (i) full rank ($\text{rank}(\mathbf{P}) = n$ in this case), which means that it is invertible and (ii) $\mathbf{P}^\top = \mathbf{P}^{-1}$. In those conditions, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n\times 1}$ and $\mathbf{P} \in \mathcal{P}_n$, the function $k(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{P}\mathbf{y}|^2$ is symmetric:

$$|\mathbf{x} - \mathbf{P}\mathbf{y}|^2 = |\mathbf{P}\mathbf{P}^\top\mathbf{x} - \mathbf{P}\mathbf{y}|^2 = |\mathbf{P}(\mathbf{P}^\top\mathbf{x} - \mathbf{y})|^2 = |\mathbf{P}^\top\mathbf{x} - \mathbf{y}|^2$$

then

$$\begin{aligned} k_{nb}(a,b) &= \sum_{\mathbf{j}\in v(a)} \sum_{\mathbf{i}\in v(b)} \sum_{\mathbf{P}\in\mathcal{P}_n} \exp -\frac{|\mathbf{x}_{\mathbf{j}a} - \mathbf{P}\mathbf{x}_{\mathbf{i}b}|^2}{2\sigma_n^2} \\ &= \sum_{\mathbf{j}\in v(a)} \sum_{\mathbf{i}\in v(b)} \sum_{\mathbf{P}\in\mathcal{P}_n} \exp -\frac{|\mathbf{P}^\top\mathbf{x}_{\mathbf{j}a} - \mathbf{x}_{\mathbf{i}b}|^2}{2\sigma_n^2} \\ &= \sum_{\mathbf{j}\in v(a)} \sum_{\mathbf{i}\in v(b)} \sum_{\mathbf{P}\in\mathcal{P}_n} \exp -\frac{|\mathbf{x}_{\mathbf{i}b} - \mathbf{P}\mathbf{x}_{\mathbf{j}a}|^2}{2\sigma_n^2} \\ &= k_{nb}(b,a)\,. \end{aligned}$$

The permutations can be avoided by introducing permutation invariant function. For simplicity, we consider only one type of element in the subsequent discussion. For the case of 3-body, the considered geometry is a triangle. As illustrated in Fig. 2.12(b), $x_1 = r_{j_1 a}$ and $x_2 = r_{j_2,a}$ are the distances between the neighboring atoms j_1, j_2 and the center atom a, while x_3 is the angle $\angle j_1 a j_2$. Then the permutation invariant function is written as:

$$\mathbf{q}_{\mathbf{j}a} = (x_1 + x_2, x_1 x_2, x_3)^\top \in \mathbb{R}^{3\times 1}. \tag{2.89}$$

Similarly, for the case of 4-body (Fig. 2.12(c)) we have:

$$\mathbf{q}_{\mathbf{j}a} = \left(\sum_{i=1}^{3} x_i, \sum_{i=4}^{6} x_i, \sum_{i=1}^{3} x_i^2, \sum_{i=4}^{6} x_i^2, \sum_{i=1}^{3} x_i^3, \sum_{i=4}^{6} x_i^3\right)^\top \in \mathbb{R}^{6\times 1}. \tag{2.90}$$

By utilizing the permutation invariant function, the kernel can be expressed in the following form:

$$k_{nb}(a,b) = \sum_{\mathbf{j}\in v(a)} \sum_{\mathbf{i}\in v(b)} \exp -\frac{|\mathbf{q}_{\mathbf{j}a} - \mathbf{q}_{\mathbf{i}b}|^2}{2\sigma_n^2}\,, \tag{2.91}$$

and by uniformly sampling K_{nb} points instead of sampling the n-body geometries $\mathbf{q}_{\mathbf{i}b}$, as previously presented in Subsection 2.2.3.2, the n-body local energy on the site a can be written as:

$$\epsilon_{nb,a} = \sum_{\mathbf{j}\in v(a)} \sum_{k=1}^{K_{nb}} w_{nb,k} \exp -\frac{|\mathbf{q}_{\mathbf{j}a} - \mathbf{z}_{nb,k}|^2}{2\sigma_n^2}\,. \tag{2.92}$$

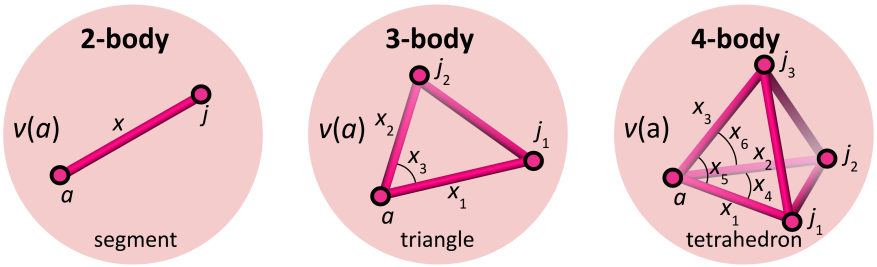

Figure 2.12: Illustration of (a) 2-body, (b) 3-body and (c) 4-body features for the permutation invariant function.

In a general n-body case, we have $\mathbf{z}_{nb,k} \in \mathbb{R}^{\nu(n)\times 1}$ where $\nu(n) = 3n-6$ if $n > 2$ and $\nu(n) = 1$ if $n = 2$ is the number of degrees of freedom of the n-plets feature. Note that to achieve the same accuracy as 2-body kernel with N sparse points, i.e., $K_{2b} = N$ and $\mathbf{z}_{nb,k} = kR_{\text{cut}}/K_{2b}$ for $k = 1, \ldots, K_{2b}$, the n-body case requires $K_{nb} = N^{\nu(n)}$ sparse points, which is computationally heavy. In practice, K_{2b} is often set on the order of magnitude of 10. Therefore, for 4-body ($\nu(4) = 6$) and 5-body kernel ($\nu(5) = 9$), K_{4b} and K_{5b} need to reach 10^6 and 10^9, respectively. Potentials with such a large number of parameters are impossible to use in MD simulations at reasonable CPU cost.

To be capable of capture high body-order features, we proposed a Fourier transformed version of n-body kernel, wherein the kernels are approximated with random projections [200]. The main idea is based on Bochner's theorem:

Theorem *(Bochner's theorem). If $f : \mathbb{R}^n \to \mathbb{C}$ is positive-definite, continuous, and satisfies $f(\mathbf{0}) = 1$, then there exists a unique probability measure μ on $\mathbb{R}^n$ such that f is the Fourier transform of μ.*

The function $f : \mathbf{x} \to \exp -\frac{|\mathbf{x}|^2}{2\sigma_n^2}$ for $\mathbf{x} \in \mathbb{R}^n$ is positive-definite, continuous, and it satisfies $f(\mathbf{0}) = 1$. According to Bochner's theorem, $f(\mathbf{x} - \mathbf{x}')$ can be written as:

$$f(\mathbf{x} - \mathbf{x}') = \int_{\mathbb{R}^n} e^{-i\boldsymbol{\omega}(\mathbf{x}-\mathbf{x}')} d\mu(\boldsymbol{\omega}) = \mathbb{E}_\mu \left[\zeta_{\boldsymbol{\omega}}(\mathbf{x})\zeta^\star_{\boldsymbol{\omega}}(\mathbf{x}')\right] \tag{2.93}$$

with μ a probability measure on $\mathbb{R}^n$, $\zeta^\star_{\boldsymbol{\omega}}$ the complex conjugate of $\zeta_{\boldsymbol{\omega}}$ and $\zeta_{\boldsymbol{\omega}}(\mathbf{x}) = e^{-i\boldsymbol{\omega}\mathbf{x}}$. We can get an unbiased approximation of this expectation by sampling $\boldsymbol{\omega}$ from μ. By drawing F samples $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_F$ independently from μ, we have:

$$f(\mathbf{x} - \mathbf{x}') = \mathbb{E}_\mu \left[\zeta_{\boldsymbol{\omega}}(\mathbf{x})\zeta^\star_{\boldsymbol{\omega}}(\mathbf{x}')\right] = \frac{1}{F}\sum_{f=1}^{F} \zeta_{\boldsymbol{\omega}_f}(\mathbf{x})\zeta^\star_{\boldsymbol{\omega}_f}(\mathbf{x}'). \tag{2.94}$$

Simplification can be made by eliminating the imaginary part. Since $e^{-i\boldsymbol{\omega}\mathbf{x}} = \cos(\boldsymbol{\omega}\mathbf{x}) - i\sin(\boldsymbol{\omega}\mathbf{x})$ and f is a real-valued function, expectation of the sin term with respect to $\mu(\boldsymbol{\omega})$ will be zero. Thus, for our purposes, we can ignore the imaginary component of the expression

and get:

$$\begin{aligned} f(\mathbf{x}-\mathbf{x}') &= \int_{\mathbb{R}^n} \cos\left(\boldsymbol{\omega}(\mathbf{x}-\mathbf{x}')\right) d\mu(\boldsymbol{\omega}) \\ &= \int_{\mathbb{R}^n} \left[\cos\left(\boldsymbol{\omega}\mathbf{x}\right)\cos\left(\boldsymbol{\omega}\mathbf{x}'\right) + \sin\left(\boldsymbol{\omega}\mathbf{x}\right)\sin\left(\boldsymbol{\omega}\mathbf{x}'\right)\right] d\mu(\boldsymbol{\omega}) \\ &= \frac{1}{F}\sum_{f=1}^{F}\left[\cos\left(\boldsymbol{\omega}_f\mathbf{x}\right)\cos\left(\boldsymbol{\omega}_f\mathbf{x}'\right) + \sin\left(\boldsymbol{\omega}_f\mathbf{x}\right)\sin\left(\boldsymbol{\omega}_f\mathbf{x}'\right)\right] \\ &= \mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{x}') \end{aligned} \tag{2.95}$$

with

$$\mathbf{z}(\mathbf{x}) = \sqrt{\frac{1}{F}} \begin{pmatrix} \cos\left(\boldsymbol{\omega}_1\mathbf{x}\right) \\ \sin\left(\boldsymbol{\omega}_1\mathbf{x}\right) \\ \vdots \\ \cos\left(\boldsymbol{\omega}_F\mathbf{x}\right) \\ \sin\left(\boldsymbol{\omega}_F\mathbf{x}\right) \end{pmatrix} \in \mathbb{R}^{2F\times 1}.$$

In practice, we take

$$\mathbf{z}(\mathbf{x}) = \sqrt{\frac{2}{F}} \begin{pmatrix} \cos\left(\boldsymbol{\omega}_1\mathbf{x} + b_1\right) \\ \vdots \\ \cos\left(\boldsymbol{\omega}_F\mathbf{x} + b_F\right) \end{pmatrix} \in \mathbb{R}^{F\times 1}$$

instead in Eq. 2.95. Here, $\mu(\boldsymbol{\omega})$ is a user-specified distribution that depends on the database we are using, and b_f is sampled from a uniform distribution.

With this technique, the above n-body kernel (Eq. 2.91) can be sampled using the Fourier random feature:

$$\begin{aligned} k_{nb}(a,b) &= \sum_{\mathbf{j}\in v(a)}\sum_{\mathbf{i}\in v(b)} \exp -\frac{|\mathbf{q}_{\mathbf{j}a} - \mathbf{q}_{\mathbf{i}b}|^2}{2\sigma_n^2} \\ &= \sum_{\mathbf{j}\in v(a)}\sum_{\mathbf{i}\in v(b)} f(\mathbf{q}_{\mathbf{j}a} - \mathbf{q}_{\mathbf{i}b}) \\ &= \frac{1}{F}\sum_{\mathbf{j}\in v(a)}\sum_{\mathbf{i}\in v(b)}\sum_{f=1}^{F}\left[\cos\left(\boldsymbol{\omega}_f\mathbf{q}_{\mathbf{j}a} + b_f\right)\cos\left(\boldsymbol{\omega}_f\mathbf{q}_{\mathbf{i}b} + b_f\right)\right] \\ &= \mathbf{z}_{\boldsymbol{\omega}}(a)^\top \mathbf{z}_{\boldsymbol{\omega}}(b) \end{aligned} \tag{2.96}$$

with

$$\mathbf{z}_{\boldsymbol{\omega}}(a) = \begin{pmatrix} \sum_{\mathbf{j}\in v(a)} z_{\boldsymbol{\omega}_1}(\mathbf{q}_{\mathbf{j}a}) \\ \sum_{\mathbf{j}\in v(a)} z_{\boldsymbol{\omega}_2}(\mathbf{q}_{\mathbf{j}a}) \\ \vdots \\ \sum_{\mathbf{j}\in v(a)} z_{\boldsymbol{\omega}_F}(\mathbf{q}_{\mathbf{j}a}) \end{pmatrix} = \sqrt{\frac{1}{F}} \begin{pmatrix} \sum_{\mathbf{j}\in v(a)} \cos\left(\boldsymbol{\omega}_1\mathbf{q}_{\mathbf{j}a} + b_1\right) \\ \sum_{\mathbf{j}\in v(a)} \cos\left(\boldsymbol{\omega}_2\mathbf{q}_{\mathbf{j}a} + b_2\right) \\ \vdots \\ \sum_{\mathbf{j}\in v(a)} \cos\left(\boldsymbol{\omega}_F\mathbf{q}_{\mathbf{j}a} + b_F\right) \end{pmatrix} \in \mathbb{R}^{F\times 1},$$

and the local energy becomes:

$$\begin{aligned}\epsilon_a &= \mathbf{w}_{\boldsymbol{\omega}}^{\top}\mathbf{z}_{\boldsymbol{\omega}}(a) = \sum_{f=1}^{F} w_{\boldsymbol{\omega},f} \sum_{\mathbf{j}\in v(a)} z_{\boldsymbol{\omega}_f}(\mathbf{q}_{\mathbf{j}a}) \\ &= \sum_{\mathbf{j}\in v(a)} \left[\sum_{f=1}^{F} w_{\boldsymbol{\omega},f} z_{\boldsymbol{\omega}_f}(\mathbf{q}_{\mathbf{j}a})\right] = \sum_{\mathbf{j}\in v(a)} F(\mathbf{q}_{\mathbf{j}a})\,. \end{aligned} \tag{2.97}$$

Equation 2.96 provides an approximation of the covariance matrix in the GP. As illustrated in Fig. 2.13(a), the covariance between the LAE of atoms a and b, denoted as $\mathrm{cov}(a,b)$, is approximated by the sampled information of dimension F for the corresponding LAEs $\mathbf{z}_{\boldsymbol{\omega}}(a)$ and $\mathbf{z}_{\boldsymbol{\omega}}(b)$. It should be noted that here $k_{nb}(a,b)$ represents the Fourier-sampled approximation of the kernel function in Eq. 2.91, which defines the covariance $\mathrm{cov}(\cdot,\cdot)$ in the input space, using random projections. Figure 2.13(b) presents the performance of the current approximation feature with $F = 10$, 200 and 400 for a covariance matrix of dimension 4000×4000, shown in the last graph. It can be observed that the performance improves with larger values of F. However, even with a small F amounting to a few tenths of M, the approximation performance remains correct.

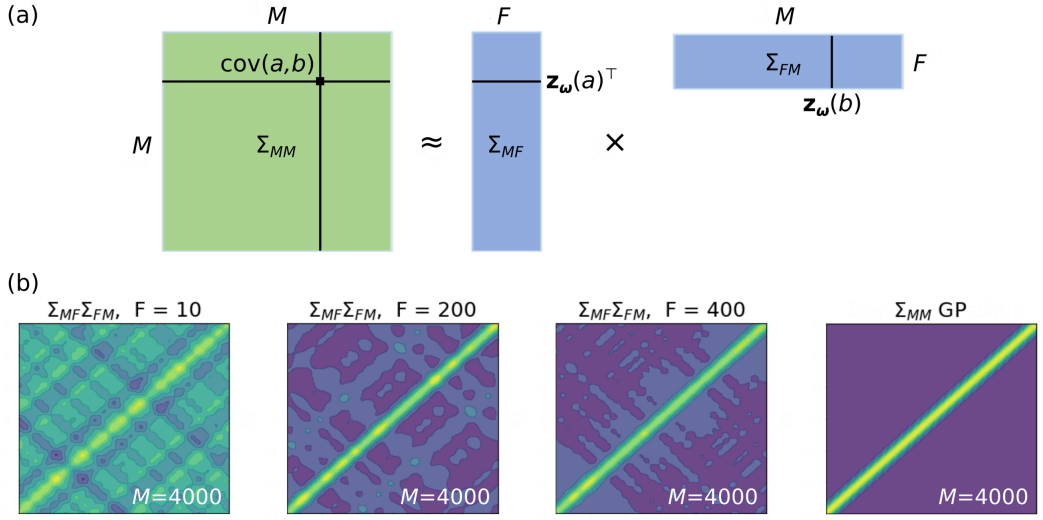

Figure 2.13: (a) Illustration of the F-dimensional Fourier-sampled approximation for the covariance (kernel) matrix and (b) its performance with $F = 10$, 200 and 400 for a covariance matrix of dimension $M = 4000$. The values of the matrix elements are colored in `viridis` scale from 1 (yellow) to 0 (blue).

2.2.4.2 Fourier-sampled n-body descriptors

As introduced in Subsection 2.2.3.2, the kernel regression can be regarded as the linear regression of kernel LAE descriptors. Adopting this concept, we propose a type of descriptors based on Fourier-sampled n-body random kernels (Subsection 2.2.4.1), denoted as $\mathbf{D}_{\mathrm{FT}nb,s,a}$ for the atom a in the system s. This so-called Fourier-sampled n-body descriptor is formulated

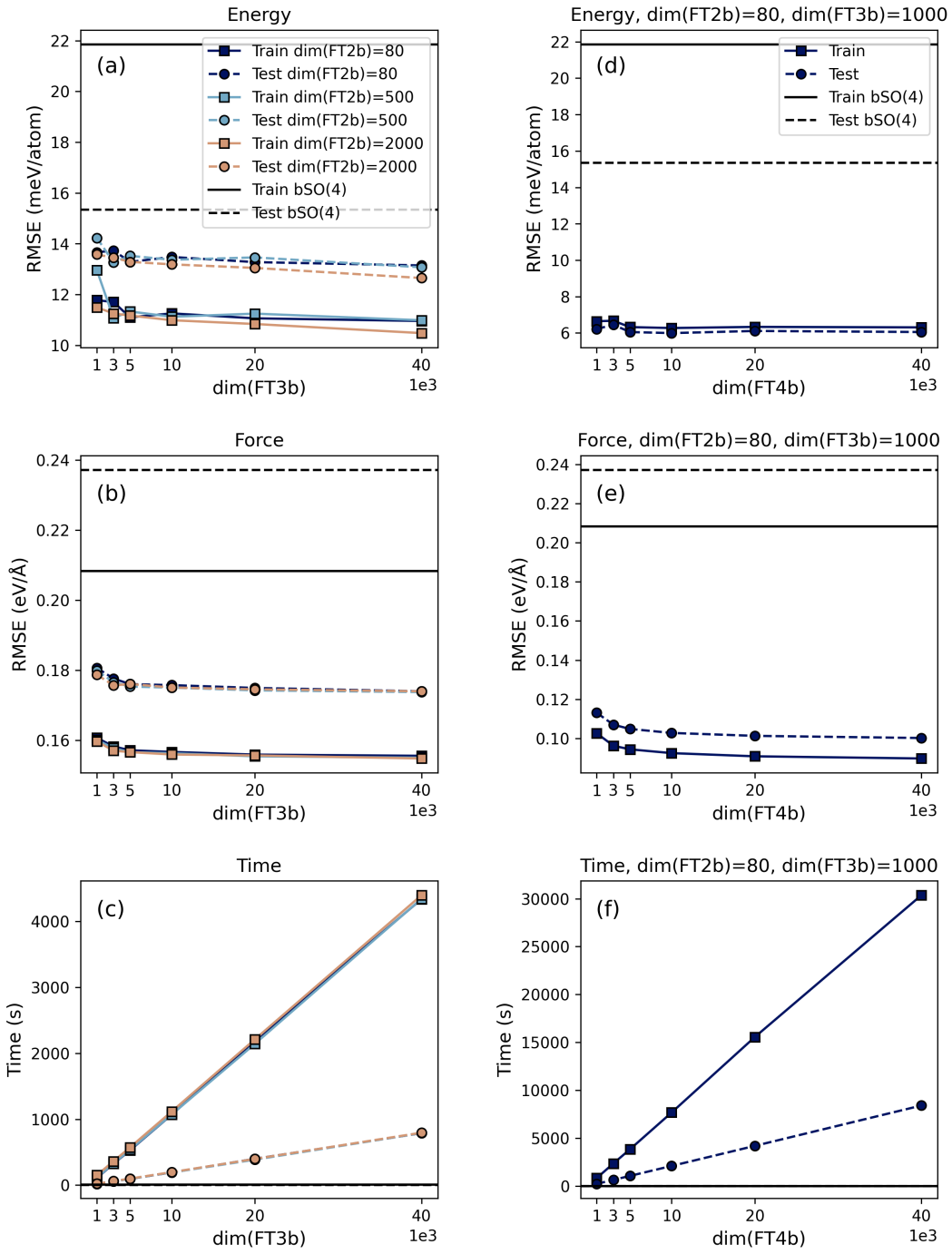

Figure 2.14: RMSE in (a, d) energy and (b, e) force prediction evaluated on the training/testing datasets and (c, f) computational time for training/testing the interatomic potential with Fourier-sampled n-body descriptors using linear regression. (a-c) Varying dim(FT3b) at a given dim(FT2b). (d-f) Varying dim(FT4b) at dim(FT2b) = 80 and dim(FT3b) = 1000. The horizontal lines in the sub-figures (a-b) and (d-e) denote the corresponding RMSE of the bispectrum SO(4) descriptor. The graphs (b) and (c) share the same legends as (a), and the graphs (e) and (f) share the same legend as (d).

as follows:

$$\mathbf{D}_{\mathrm{FT}nb,s,a} = \mathbf{z}_{\boldsymbol{\omega}}(a) = \sqrt{\frac{1}{F}} \begin{pmatrix} \sum_{\mathbf{j}\in v(a)} \cos\left(\boldsymbol{\omega}_1 \mathbf{q}_{\mathbf{j}a} + b_1\right) \\ \sum_{\mathbf{j}\in v(a)} \cos\left(\boldsymbol{\omega}_2 \mathbf{q}_{\mathbf{j}a} + b_2\right) \\ \vdots \\ \sum_{\mathbf{j}\in v(a)} \cos\left(\boldsymbol{\omega}_F \mathbf{q}_{\mathbf{j}a} + b_F\right) \end{pmatrix} \in \mathbb{R}^F . \tag{2.98}$$

For Fourier-sampled 2-, 3- and 4-body descriptors, $\boldsymbol{\omega}_f$ and $\mathbf{q}_{\mathbf{j}a}$ have dimensions of 1, 3 and 6, as expressed in the section 2.2.4.1. Theses descriptors can be combined by a direct sum $\mathbf{D}_{\mathrm{FT}nb,s,a} = \mathbf{D}_{\mathrm{FT}2b,s,a} \oplus \mathbf{D}_{\mathrm{FT}3b,s,a} \oplus \mathbf{D}_{\mathrm{FT}4b,s,a} \oplus \ldots$ with the dimension $dim(\mathbf{D}_{s,a}) = \sum_n dim(\mathbf{D}_{\mathrm{FT}nb,s,a})$, where $dim(\mathbf{D}_{\mathrm{FT}nb,s,a})$ corresponds to F in Eq. 2.98 for body-order n. The notation $dim(\mathbf{D}_{\mathrm{FT}nb,s,a})$ can be simplified as $dim(\mathrm{FT}nb)$ and FTnb stands for "Fourier transform-sampled n-body descriptor".

In Figure 2.14, we show the root-mean-square error (RMSE) of the energy and force prediction evaluated on the training/testing datasets when combining the 2- and 3-body Fourier-sampled descriptors (sub-figures (a-c)), as well as combining the 2-, 3- and 4-body Fourier-sampled descriptors (sub-figures (d-f)). All the ML potentials tested here are fitted with linear regression on the database listed in Table 2.1.

RMSE of the bispectrum SO(4) descriptor (see the section 2.1.2) with the angular moment $j_{\max} = 4$ is also plotted as horizontal lines to provide comparison. From the graphs (a-b) of Fig. 2.14, we can derive that the accuracy gains from increasing the dimensionality of 2-body Fourier-sampled descriptor are negligible. And when we fix $dim(\mathrm{FT}2b)$ and increase the dimensionality of 3-body Fourier-sampled descriptor, a relatively significant reduction of RMSE can be observed at small values of $dim(\mathrm{FT}3b)$, while at $dim(\mathrm{FT}3b) > 10\,000$, increasing $dim(\mathrm{FT}3b)$ does not seem to be efficient anymore. However, in terms of the RMSE of force prediction which is the key factor in atomic simulations, even using the 2- and 3-body hybrid Fourier-sampled descriptor with $dim(\mathrm{FT}2b) = 80$ and $dim(\mathrm{FT}3b) = 1000$ (the lowest dimensionality that makes sense) can lead to an accuracy improvement of about 25% in comparison with the bispectrum SO(4) descriptor. In this way, it is not that interesting to increase the dimensionality of 2- and 3-body descriptors.

Next, we also take the 4-body descriptors into account. In the graphs (d-e) of Fig. 2.14, we show the RMSE of the energy and force prediction by the 2-, 3- and 4-body hybrid Fourier-sampled descriptors at $dim(\mathrm{FT}2b) = 80$, $dim(\mathrm{FT}3b) = 1000$ and different values of $dim(\mathrm{FT}4b)$. Similar observation can be found: increasing $dim(\mathrm{FT}4b)$ brings about reduced RMSE but only at low values, and the magnitude of this reduction is actually not attractive when considering that the accuracy improvement in force prediction attains 55%, even at $dim(\mathrm{FT}2b) = 80$ and $dim(\mathrm{FT}3b) = dim(\mathrm{FT}4b) = 1000$, compared to the bispectrum SO(4) descriptor.

Another drawback of increasing the dimensionality of the n-body descriptors to lower the RMSE is the augmentation of the computational cost. Computational time for training and testing the linear ML potentials using the corresponding n-body descriptors is presented in the graphs (c) and (f) of Fig. 2.14, showing that the computational cost grows linearly as the dimensionality increases. Moreover, the computational cost of applying the ML potential in

MD simulations is higher with larger dimensionality of the LAE descriptors. The computational time for computing the forces on one atom at each MD step using linear ML potentials with different LAE descriptors, as well as the RMSE of force prediction, are listed in Table 2.3. The computational time of an EAM potential [22] is also provided as benchmark. Each test on the computational speed of MD simulations shown in this table is performed with MiLaDy-Lammps package on 32 processes and averaged on 100 steps within a system of 128 atoms. From this table, we can see that the 2-, 3-, and 4-body hybrid Fourier-sampled descriptor, which improves the force prediction accuracy by 55%, also results in a 370-fold increase in computational time. However, Eq. 2.97 can be represented by multidimensional ($\nu(n)$-dimensional) spline interpolation and tabulated, as $F(\mathbf{q}_{ja})$ is a function of the n-plets feature, i.e., $\nu(n)$-dimensional geometric information. Therefore, instead of computing from $\mathbf{w}_{\boldsymbol{\omega}}$, $\boldsymbol{\omega}_f$ and b_f for $f = 1 \dots F$, the local energy can be directly computed from the spline representation. This approach can drastically reduce the computational time required for force evaluation in simulations, by one order of magnitude for the 2-body case and two orders of magnitude for the 3-body case. Nowadays, packages for multidimensional spline interpolation [201] can effectively handle up to 6 dimensions, providing access to the spline interpolation for Fourier-sampled n-body descriptors up to $n = 4$.

Table 2.3: Computational time for computing the forces on one atom at each molecular dynamics (MD) step, and the root-mean-square error (RMSE) of force prediction for linear ML potentials using different LAE descriptors. For the Fourier-sampled n-body descriptors in this table, the lowest reasonable dimensionality is taken: $dim(\text{FT2b}) = 80$, $dim(\text{FT3b}) = 1000$ and $dim(\text{FT4b}) = 1000$. Note that the computational time for the Fourier-sampled n-body descriptors can be drastically reduced by one order of magnitude for 2-body and two orders of magnitude for 3-body if the spline interpolation of Eq. 2.97 is used.

Potential	MD time (s/step/atom)	Force RMSE (eV/Å)
EAM [22]	7.5×10^{-7}	
2-body Fourier-sampled	8.7×10^{-6}	0.42
Bispectrum SO(4)	1.0×10^{-4}	0.24
2,3-body Fourier-sampled	1.5×10^{-3}	0.18
2,3,4-body Fourier-sampled	3.7×10^{-2}	0.11

In the next step, we want to test the unprecedented 5-body feature by deploying the Fourier sampling scheme. However, as shown in Table 2.3, even if the Fourier sampling makes it feasible to use the 4-body descriptor, there is no advantage in terms of computational speed. As a result, to implement the 5-body feature which is more accurate but even computationally more expensive, a self-adaptive cutoff radius approach is applied. It means that different cutoff radius is set to different body-ordered n. Six different cases are considered and the corresponding RMSE of the energy and force prediction evaluated on the training/testing datasets are presented in Fig. 2.15(a, b). The graph (c) reveals the computational training/testing time for each case (left y-axis), and also the computational time of forces evaluation per atom per MD step obtained from MiLaDy-Lammps package (right y-axis). All the trials are performed with the following dimensionality: $dim(\text{FT2b}) = 80$, $dim(\text{FT3b}) = 500$, $dim(\text{FT4b}) = 3000$ and $dim(\text{FT5b}) = 5000$. We detail in the sub-figure (d) of Fig. 2.15 the 6 cases of cutoff radius combination with respect to the index of R_{cut} case. For example, for the R_{cut} case 3, the cutoff radius of 2- and 3-body Fourier-sampled descriptors is 5.3 Å (indicated at the right-hand side of the illustration), covering the first to the forth nearest neighbours (1nn to 4nn, indicated at the right-hand side of the illustration), and the cutoff radius of 4- and 5-body descriptors is

4.6 Å, only covering the first to the third nearest neighbours. The index 0 represents the case where there is no 5-body feature in the hybrid descriptor, and the case 1 denotes the "full" 2-, 3-, 4- and 5-body Fourier-sampled descriptors, which means that the cutoff radius of all components is 5.3 Å. The cases 0 and 1 are regarded as the benchmark.

Taking into account the trade-off between accuracy and efficiency, we notice that the case 5 is not interesting due to an RMSE of force prediction much larger than the other cases. The similar cases 2 and 3 provide a slight improvement of accuracy (about 10%) at a reasonable cost of computational resource, so they are possible choices when pursuing a very high accuracy. The case 4, however, is a suitable alternative of the case 0. With a loss in force RMSE about 6% (0.007 eV/Å), the ML potential with the hybrid descriptors at R_{cut} case 4 runs around 3 times faster in an MD simulation than the ML potential without 5-body descriptor (R_{cut} case 0). To be more quantitative, the computational time of force evaluation using the ML potentials corresponding to the case 0 and the case 4 is respectively 0.098 s/step/atom and 0.034 s/step/atom.

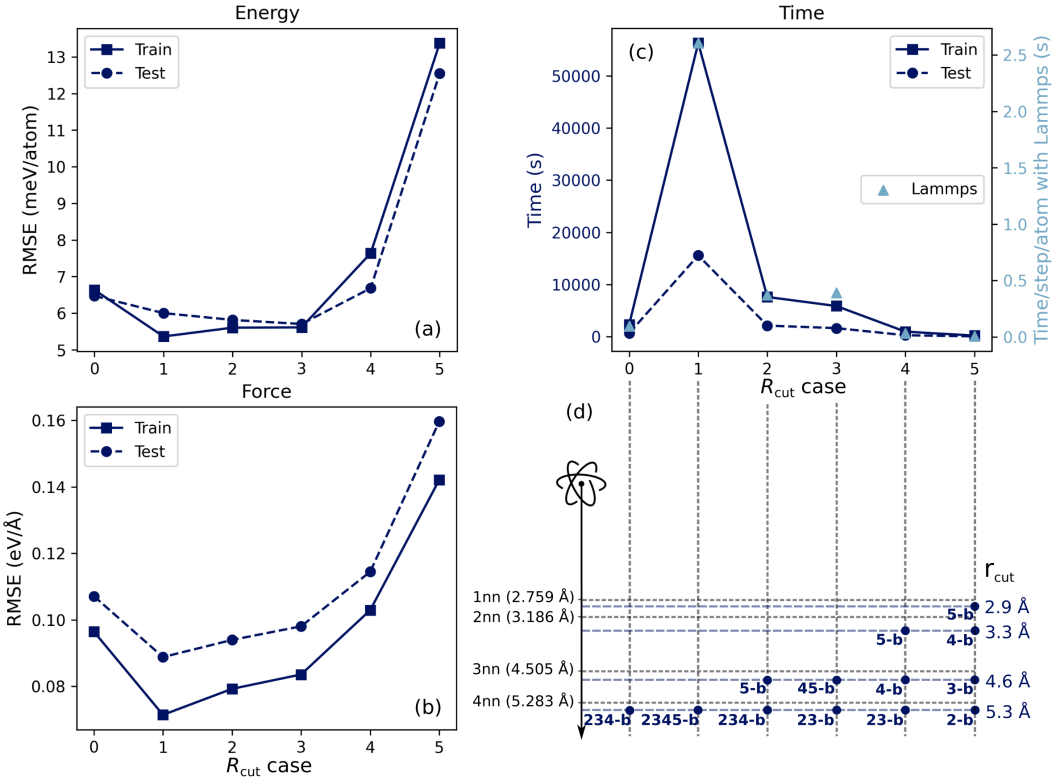

Figure 2.15: RMSE in (a) energy and (b) force prediction evaluated on the training/testing datasets and (c) computational time for training/testing the interatomic potential with Fourier-sampled n-body descriptors using linear regression, when setting different cutoff radius for different body-ordered n. We detail in the sub-figure (d) the cases of cutoff radius corresponding to the index of R_{cut} case in the sub-figures (a-c), where the options of cutoff radius are listed at the right-hand side of the graph (2.9 Å, 3.3 Å, 4.6 Å and 5.3 Å), and the distances between an atom and its first to forth nearest neighbours (nn) are given at the laft-hand side. Dimensionality of the Fourier-sampled n-body descriptors in this figure: $dim(\text{FT2b}) = 80$, $dim(\text{FT3b}) = 500$, $dim(\text{FT4b}) = 3000$ and $dim(\text{FT5b}) = 5000$.

Finally, we fix the cutoff radius of the 2-, 3- and 4-body features in the hybrid Fourier-

sampled descriptors at 5.3 Å and only vary the cutoff radius of the 5-body features. Similarly, the RMSE of the energy and force prediction, and the computational time for training/testing/employing the ML potentials are plotted to characterize its quality (Fig. 2.16). It proves that reducing the cutoff radius of the 5-body feature to a low value is not advisable. The optimal choice, balancing accuracy and efficiency, is 4.6 Å, which corresponds to the R_{cut} case 2 in Fig. 2.15.

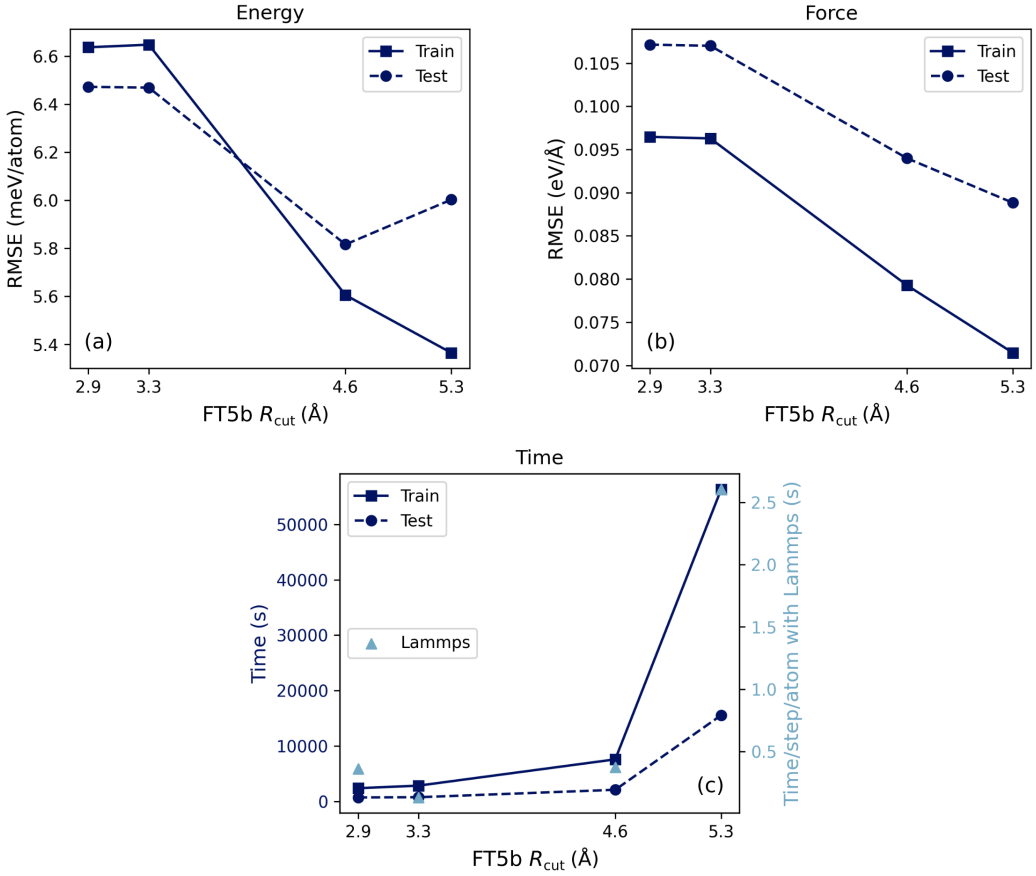

Figure 2.16: RMSE in (a) energy and (b) force prediction evaluated on the training/testing datasets and (c) computational time for training/testing the interatomic potential with Fourier-sampled n-body descriptors using linear regression, when setting the cutoff radius for 2-, 3- and 4-body Fourier-sampled descriptors at 5.3 Å and varying the cutoff radius of 5-body descriptors.

2.2.4.3 High-order kernels: incomplete but very fast kernels

In the previous section, we developed a method capable of sampling n-body kernels for large values of n. This method is systematic and robust; however, it entails a significant numerical burden. The primary outcome of the previous investigation reveals that in high-order body approximations, it is not necessary to systematically include all n-body terms. This insight paves the way for new research avenues, such as the use of incomplete high-order kernels, which can be evaluated numerically with great efficiency. Here, we outline the strategy

to follow, which closely aligns with the approach emphasized in previous work by Glielmo *et al.* [202, 198, 199].

Suppose that we have developed an n-body kernel, which includes all invariant properties. A new kernel can be obtained as follows:

$$k_{n'b}(a,b) = [k_{nb}(a,b)]^p \tag{2.99}$$

The order of the new kernel is $n' = (n-1)p+1$, because, due to the multiplication, it is possible to have n'-tuplets of atoms: p series of $n-1$ atoms and the central atom. This observation allows modeling any interaction order while incurring only the computational cost of computing the 2-body kernel ($n = 2$). For example, with $p = 4$, we can achieve a body order of 5 at minimal cost. It seems too good to be true. However, the problem is that the n'-body kernel is incomplete. The number of degrees of freedom for a complete n-body kernel is $3n-6$ for $n > 2$ (and 1 for $n = 2$), while for the output kernel from Eq. 2.99 with power p, the number of independent variables is $p(3n-3)$ for the case $n > 2$. This number is always lower than the number of degrees of freedom for a complete n'-body kernel with $n' = (n-1)p+1$, which are $3n'-6 = 3p(n-1)-3$. This indicates that the output kernel is incomplete. Let us consider an example where $n = 3$ and $p = 2$. The resulting kernel will have an order of $n' = 7$. If the output kernel were complete, it would have $3n'-6 = 15$ degrees of freedom. However, the number of independent variables is $p(3n-3) = 12$. Thus, 12 variables should describe a manifold with 15 degrees of freedom. This is a general characteristic of this method of constructing kernels: the number of independent variables is always lower than the number of degrees of freedom of a complete kernel. Regardless of the size of the database or the learning algorithm, these kernels will always have a lack of representation for certain classes of configurations, resulting in imperfect learning. However, as we have seen in Refs. [202, 198, 199], even these incomplete representations can significantly improve the fits.

Incompleteness is a serious issue in high-dimensional representations of databases. There are well-known examples of popular descriptors that have been found to be incomplete, such as the SOAP kernels [203]. Even complex neural networks, such as recent architectures based on Graph Neural Networks, can suffer from this problem [204]. However, this phenomenon can be formalized and avoided in many situations, presenting opportunities for very fast evaluation of high-order interactions.

2.3 Short-range correction: Ziegler-Biersack-Littmark potential

Some of the descriptors exhibit strange behavior at small distances, even in their hybrid form with the 2-body kernel. To make the force field applicable for the simulations of cascade or liquid in which the atoms can be very close in between, the internuclear repulsion at extremely short distances can be described by an external pair potential V_{pair}, in the form of a screened Coulomb potential.

In this case, the total energy of the system s is a sum of the pair potential and the ML

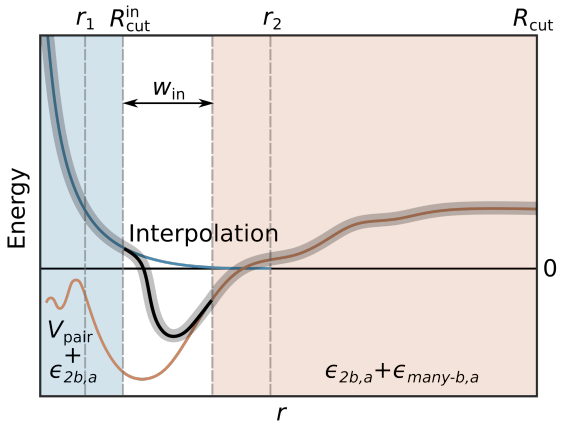

Figure 2.17: Schematic illustration of the correction for short interatomic distances. At distances shorter than an inner cutoff distance $R^{\text{in}}_{\text{cut}}$, we only consider the pair potential V_{pair} coupled with the 2-body potential $\epsilon_{2b,a}$, without any many-body interaction. At distances larger than $R^{\text{in}}_{\text{cut}} + w_{\text{in}}$, full ML potential is computed from the hybrid descriptors including both 2-body and many-body contributions. Between $R^{\text{in}}_{\text{cut}}$ and $R^{\text{in}}_{\text{cut}} + w_{\text{in}}$, interpolation is performed to ensure continuity up to the second derivatives across the entire range of distances, from 0 Å to the cut-off distance R_{cut}.

parts, which can be further divided into the 2-body and many-body interactions:

$$\begin{aligned} E_s &= \sum_{i,j\in s} V_{\text{pair}}\ (r_{ij}) + \sum_{a\in s} \epsilon^{\text{ML}}_{s,a}\left(\mathbf{w}, \mathbf{D}_{2b,s,a} \oplus \mathbf{D}_{many\text{-}b,s,a}\right) \\ &= \sum_{i,j\in s} V_{\text{pair}}\ (r_{ij}) + \sum_{a\in s} \left(\epsilon_{2b,a} + \epsilon_{many\text{-}b,a}\right). \end{aligned} \tag{2.100}$$

In practice, the many-body interaction is turned off and the 2-body interaction (Eq. 2.85) is coupled to the external pair potential for distances shorter than an inner cutoff distance $R^{\text{in}}_{\text{cut}}$. Only the differences in energies and forces between the external pair potential and the training data need to be reproduced by the 2-body kernel. The screened Coulomb potential fully dictates the short-range dynamics as desired, and the 2-body ML potential only needs to be trained to predict energies and forces close to zero for short interatomic distances. Between $R^{\text{in}}_{\text{cut}}$ and $R^{\text{in}}_{\text{cut}} + w_{\text{in}}$, a buffer region of width w_{in} (a possible choice is 0.4 Å) is established for a smooth interpolation to ensure continuity up to the second derivatives between the pair potential coupled with the 2-body potential and the pure ML parts from the hybrid descriptors. These features are illustrated in Fig. 2.17.

The function V_{pair} is identical to the universal Ziegler-Biersack-Littmarck (ZBL) potential [205]:

$$V_{\text{pair}}\ (r_{ij}) = \frac{1}{4\pi\varepsilon_0}\frac{Z_i Z_j e^2}{r_{ij}}\psi\left(r_{ij}/a\right) f_{\text{cut}}\left(r_{ij}\right), \tag{2.101}$$

where Z_i and Z_j are the atomic number of the atom i and j, and

$$a = \frac{0.46848}{Z_i^{0.23} + Z_j^{0.23}}. \tag{2.102}$$

The screening function $\psi(x)$ should be specifically refitted for the target system, as the ZBL potential may differ significantly from the VASP data. For example, in the case of W-W

repulsion [206], $\psi(x)$ is refitted using the all-electron DFT-DMol data from Ref. [207] as

$$\begin{aligned}\psi(x) = \quad & 0.32825\exp(-2.54931x) + 0.09219\exp(-0.29182x) \\ & +0.58110\exp(-0.59231x)\,.\end{aligned} \tag{2.103}$$

The screened Coulomb potential is forced to zero by the cutoff function

$$f_{\text{cut}}(r) = \begin{cases} 1, & r \leqslant r_1 \\ 1 - \chi^3\left(6\chi^2 - 15\chi + 10\right), & r_1 < r < r_2 \\ 0, & r \geqslant r_2 \end{cases} \tag{2.104}$$

where $\chi = (r - r_1)/(r_2 - r_1)$. The cutoff range is chosen as $r_1 = 1\,\text{Å}$ and $r_2 = 2.2\,\text{Å}$. The characteristic distances here should satisfy the following relationship: $r_1 < R_{\text{cut}}^{\text{in}} < R_{\text{cut}}^{\text{in}} + w_{\text{in}} < r_2$, leaving the region from $R_{\text{cut}}^{\text{in}} + w_{\text{in}}$ outward to be fully machine-learned. In practice, V_{pair} can be tabulated to accelerate the training and the application of the ML potential.

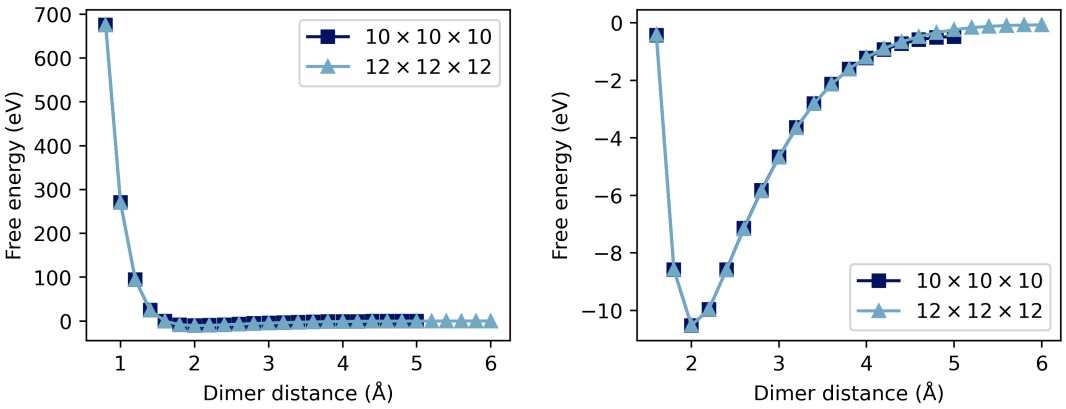

Figure 2.18: Free energy from DFT calculations of dimer at different distances in the box of size $10 \times 10 \times 10$ and $12 \times 12 \times 12$ (in Å). The sub-figure at right is the local zooming of the sub-figure at left.

We emphasize that, to construct a model with the ZBL correction for short interatomic distances, it is necessary to include specific configurations in the database, where atoms are positioned at very short distances from each other. In the case of tungsten, we compute with VASP 6.2.0 [153] the configurations of dimers, positioned at $(0, 0, 0)$ et $(d, 0, 0)$ with d the interatomic distance ranging from 0.8 Å to 5.0 Å for the boxes of size $10 \times 10 \times 10$ and from 0.8 Å to 6.0 Å for the boxes of size $12 \times 12 \times 12$ (in Å), at an interval of 0.2 Å. Considering the periodic boundary conditions, repetition occurs when the distance between dimers exceeds half the length of the box. As presented in Fig. 2.18, the results of free energy can be considered as independent of the box size. We added therefore the 27 configurations obtained within the boxes of size $12 \times 12 \times 12$ into the database of tungsten.

2.4 Conclusion of the chapter

In this chapter, we talk about how to build and optimize the machine learning potentials, based on its three key factors of construction: database, representation of LAEs and regression model. As shown in Fig. 2.19, the LAEs consisting of the atomic coordinates $\mathbf{q}_m$ in the database are described by a user-specified descriptor of local energy, from which the descriptors of global energy, atomic forces and virial stress are derived ($\mathbf{x}_m$). The descriptors are further encoded using a feature map $\phi(\mathbf{x}_m)$ to "flatten" the topology of the descriptor space, enabling linear regression of the elements.

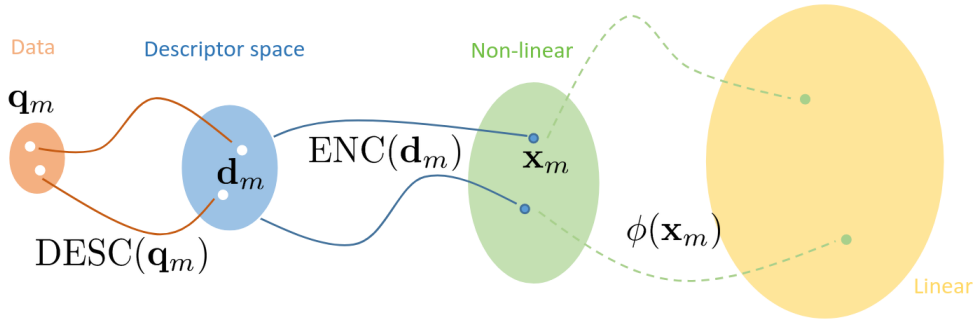

Figure 2.19: Structure of machine learning potential in 4 bubbles.

Within this framework, the author contributed to the `MiLaDy` package by:

- Database: completing the database of tungsten, molybdenum and Ta-Ti-V-W high entropy alloys with the active learning method (Subsection 2.1.1.2).
- Representation: developing a novel Fourier-sampled n-body descriptor, which enables the capture of high body-order geometrical features in the LAE (Subsection 2.2.4.2) and leads to a 55 % improvement in force prediction accuracy compared to the widely used bispectrum SO(4) descriptor.
- Regression: implementing (i) various sparse points selection methods (Subsection 2.2.2) for the kernel regression model, and identifying the recommended ones in terms of the accuracy-cost trade-off; (ii) implementing the short-range correction using the ZBL potential (Section 2.3).

Moreover, based on the aforementioned techniques, ML potentials for tungsten and Ta-Ti-V-W high entropy alloys were trained and tested with precise calibration of hyperparameters. These potentials will be employed in Chapter 4 to provide accurate predictions of material properties at high temperatures.

Chapter 3

Enhanced sampling methods for free energy landscape

Contents

3.1 Free energy estimation: general perspectives 66
3.1.1 Preliminaries in statistical physics 67
3.1.1.1 Canonical ensemble 67
3.1.1.2 Sampling in the canonical ensemble 68
3.1.1.3 Harmonic approximation and its limitations 69
3.1.2 Including anharmonicity in free energy 70
3.1.2.1 Thermodynamic integration-based calculations 71
3.1.2.2 Free energy perturbation and adaptive sampling 74
3.2 Accelerated Bayesian adaptive biasing force method 77
3.2.1 Bayesian adaptive biasing force: theoretical framework and implementation 77
3.2.2 Optimizations of BABF method 81
3.2.2.1 Reference system: choices and numerical instabilities filter 81
3.2.2.2 Acceleration of convergence: reweighting sampling 86
3.2.3 Validation: comparison with MD simulations and TI 87
3.3 Bound BABF method: sampling a metastable state 91
3.3.1 Motivation: sampling free energy landscape of defects 91
3.3.2 Constrained configuration space exploration without bias 93
3.3.3 Parallelization based on shared bias 97
3.4 Fluid-phase free energy computations 99
3.5 Conclusion of the chapter 101

An essential task of atomic simulations in materials science is the computation of free energy. Free energy is a crucial physical quantity that connects the microstructure of materials to their macroscopic properties. From the free energy landscape, we can derive thermodynamic properties such as thermal expansion, elastic constants, thermal conductivity, etc. Moreover, the free energy profile of defects can be used to predict their behavior under complex conditions, such as high temperatures or irradiation, thereby determining the material's performance under these conditions.

Obtaining an accurate free energy landscape is not an easy task. While DFT-based methods offer high accuracy, they are computationally intensive. A practical approach is to utilize the machine learning (ML) force fields presented in Chapter 2, which achieve the accuracy of *ab initio* calculations while surpassing their computational speeds by up to more than six orders of magnitude. However, compared to traditional EAM potentials, ML potentials are slower by a factor of several tens to four orders of magnitude. Consequently, free energy sampling remains challenging, particularly at high temperatures where atomic dynamics exhibit significant statistical fluctuations. To address this issue, more efficient sampling methods are necessary to enable the integration of ML potentials into dynamic simulations.

In this chapter, we discuss the general perspectives of free energy sampling in Section 3.1, covering background knowledge in statistical physics and classical computational techniques, alongside recent developments. Subsequently, in Section 3.2, we present a highly efficient method for sampling anharmonic free energy. This method builds upon the existing approaches introduced in Section 3.1 and incorporates Bayesian reasoning. In Section 3.3, we extend this method to enable sampling of specific metastable states, allowing for investigation into the free energy landscape of defects. Finally, in Section 3.4, we demonstrate that our approach is also applicable to liquid systems. This chapter focuses on the methodology of free energy computations. Here, we detail our progress in sampling methods, covering theoretical formulation, algorithm implementation, optimizations, and validation.

3.1 Free energy estimation: general perspectives

This section outlines the theoretical framework of free energy sampling and the methodological advancements in numerical computations. In Subsection 3.1.1, we provide a brief introduction to the thermodynamic ensemble under consideration and the stochastic dynamics for sampling the phase space within it. Furthermore, we emphasize the significance of anharmonic contributions to the free energy by elucidating the limitations of the harmonic approximation. This subsection is optional for experts in statistical physics. Subsection 3.1.2 primarily discusses two widely-used and well-established methods in free energy computations: thermodynamic integration and free energy perturbation. Furthermore, a review of the existing literature on the utilisation of adaptive biasing techniques is conducted. The approaches presented in this section provide the basis of our enhanced methods.

3.1.1 Preliminaries in statistical physics

Calculating the free energy requires the knowledge of the system's Hamiltonian. Assuming the Hamiltonian is separable, it can be decomposed into two contributions:

$$\mathcal{H}(\mathbf{q},\mathbf{p}) = E_k(\mathbf{p}) + U(\mathbf{q})\,. \tag{3.1}$$

The first term and the second term on the right-hand side represent the microscopic kinetic energy and the microscopic potential energy of the system, respectively. In this chapter, we denote $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{N_s}) = (q_1, q_2, \cdots, q_{3N_s})$ and $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_{N_s}) = (p_1, p_2, \cdots, p_{3N_s})$ the atomic coordinate vector and the atomic momentum vector of the system containing $3N_s$ atoms, with $\mathbf{p}_i = (p_{3i-2}, p_{3i-1}, p_{3i})$ and $\mathbf{q}_i = (q_{3i-2}, q_{3i-1}, q_{3i})$ the coordinate and the momentum of atom i. A microstate $(\mathbf{q},\mathbf{p})$ is defined by an element in the phase space $\mathcal{E} = \mathcal{Q}\times\mathcal{P} \subset \mathbb{R}^{3N_s}\times\mathbb{R}^{3N_s}$.

3.1.1.1 Canonical ensemble

In many physical scenarios, systems are in contact with a heat bath, which is commonly referred to as a thermostat. A canonical ensemble is the statistical ensemble that represents the possible states of a system in thermal equilibrium with such a thermostat. The canonical ensemble is often referred to as the NVT ensemble because the number of particles N_s, the volume V, and the temperature T are fixed. The canonical ensemble allows heat exchange with the thermostat, leading to fluctuations in the system's energy around an average value, while maintaining a constant temperature. In this case, the microscopic configurations are distributed according to the canonical measure, which is defined as follows:

$$\mu(d\mathbf{q}d\mathbf{p}) = Z^{-1}\left[\frac{1}{h^{3N_s}C}\exp\left(-\beta\mathcal{H}(\mathbf{q},\mathbf{p})\right)\right]d\mathbf{q}d\mathbf{p}\,, \tag{3.2}$$

where $\beta = (k_{\mathrm{B}}T)^{-1}$, k_{B} is the Boltzmann constant, h is the Planck constant, C is a degeneracy factor depending on the statistical nature of the atoms involved, and the dimensionless normalization constant Z is called the partition function:

$$Z = \frac{1}{h^{3N_s}C}\int_{\mathcal{Q}\times\mathcal{P}}\exp\left(-\beta\mathcal{H}(\mathbf{q},\mathbf{p})\right)d\mathbf{q}d\mathbf{p}\,. \tag{3.3}$$

For indistinguishable atoms following Bose-Einstein statistics, as is the case here, the degeneracy factor is

$$C = \prod_{\alpha} N_\alpha!\,, \tag{3.4}$$

where N_α is the number of atoms of species α and we have $N_s = \sum_\alpha N_\alpha$. As the Hamiltonian $\mathcal{H}$ is separable, the canonical measure is of the tensorized form $\mu(d\mathbf{q}d\mathbf{p}) = \nu(d\mathbf{q})\kappa(d\mathbf{p})$, where ν and κ are the two following probability measures:

$$\nu(d\mathbf{q}) = Z_\nu^{-1}\exp\left(-\beta U(\mathbf{q})\right)d\mathbf{q}\,,\ Z_\nu = \int_{\mathcal{Q}}\exp\left(-\beta U(\mathbf{q})\right)d\mathbf{q} \tag{3.5}$$

and

$$\kappa(d\mathbf{p}) = \left(\frac{\beta}{2\pi}\right)^{3N_s/2} \prod_{i=1}^{N_s} m_i^{-3/2} \exp\left(-\frac{\beta}{2}\mathbf{p}^\top \mathbf{M}^{-1}\mathbf{p}\right) d\mathbf{p}\,. \tag{3.6}$$

Here, m_i is the mass of particle i and $\mathbf{M}$ is the mass matrix:

$$\mathbf{M} = \begin{pmatrix} m_1\mathbf{I}_3 & 0 & \cdots & 0 \\ 0 & m_2\mathbf{I}_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_{N_s}\mathbf{I}_3 \end{pmatrix} \in \mathbb{R}^{3N_s \times 3N_s}\,. \tag{3.7}$$

Under the canonical measure μ, the coordinate $\mathbf{q}$ and the momentum $\mathbf{p}$ are independent random variables. Thus, sampling $(\mathbf{q},\mathbf{p}) \sim \mu(d\mathbf{q}d\mathbf{p})$ can be realized by independently sampling $\mathbf{q} \sim \nu(d\mathbf{q})$ and $\mathbf{p} \sim \kappa(d\mathbf{p})$. It is straightforward to sample from κ since the momenta are Gaussian random variables. The actual difficulty lies in sampling from ν.

3.1.1.2 Sampling in the canonical ensemble

The canonical measure Eq. 3.2 can be sampled by the techniques of stochastic dynamics $t \to (\mathbf{q}_t, \mathbf{p}_t)$, which are ergodic for the canonical measure, in the sense that the expectation of a given observable

$$\mathbb{E}(\mathcal{O}) = \int_{\mathcal{Q}\times\mathcal{P}} \mathcal{O}(\mathbf{q},\mathbf{p})\mu(d\mathbf{q}d\mathbf{p}) \tag{3.8}$$

can be obtained as an ergodic limit

$$\mathbb{E}(\mathcal{O}) = \lim_{T\to+\infty} \frac{1}{T}\int_0^T \mathcal{O}(\mathbf{q}_t,\mathbf{p}_t)dt \tag{3.9}$$

over a single realization of the stochastic dynamics. The dynamics we use are chosen solely for their ergodic property (Eq. 3.9), and should therefore be regarded as a sampling mean. The physical relevance of the evolution is not of concern, as our primary interest is in time-independent equilibrium properties. The Langevin dynamics satisfies the ergodic theorem [149] and is described by the following evolution equations for a system of N_s atoms:

$$\begin{cases} d\mathbf{q}_t &= \mathbf{M}^{-1}\mathbf{p}_t\, dt\,, \\ d\mathbf{p}_t &= -\nabla_{\mathbf{q}} U(\mathbf{q}_t)\, dt - \gamma\mathbf{M}^{-1}\mathbf{p}_t\, dt + \sigma\, dW_t\,, \end{cases} \tag{3.10}$$

where $\gamma, \sigma > 0$ verify $\sigma^2\beta = 2\gamma$, and the last term represents a standard $3N_s$-dimensional Brownian motion, with W_t denoting a $3N_s$-dimensional Wiener process. It satisfies the property $W_{t+u} - W_u \sim \mathcal{N}(\mathbf{0}, t\mathbf{I}_{3N_s})$, i.e., $W_t - W_0 = \int_0^t dW_t$ is normally distributed with zero mean vector and $t\mathbf{I}$ variance matrix.

Since the challenging task is sampling the configurational part ν (Eq. 3.5) of the canonical measure, we could consider a dynamics restricted to the configurational space, such as the overdamped Langevin dynamics:

$$d\mathbf{q}_t = -\nabla_{\mathbf{q}} U(\mathbf{q}_t)\, dt + \sqrt{2\beta^{-1}}\, dW_t\,. \tag{3.11}$$

Under reasonable assumptions [149], this dynamics satisfies the ergodic theorem

$$\mathbb{E}(\mathcal{O}) = \int_{\mathcal{Q}} \mathcal{O}(\mathbf{q})\nu(d\mathbf{q}) = \lim_{T\to+\infty} \frac{1}{T}\int_0^T \mathcal{O}(\mathbf{q}_t)dt\,. \tag{3.12}$$

Intuitively, each term in Eq. 3.11 can be interpreted as follows: the gradient force $-\nabla_{\mathbf{q}}U(\mathbf{q}_t)$ ensures a decrease in energy, thereby ensuring the sampling of energetically favorable configurations, while the random noise term supplies some energy to maintain the correct temperature.

3.1.1.3 Harmonic approximation and its limitations

For a canonical ensemble, thermodynamic variables are directly related to the canonical partition function Z of the system. In particular, the Helmholtz free energy of a system is defined as $F = -k_{\mathrm{B}}T\ln Z$. To compute this value, the remaining difficulty is to calculate the expressions for $E_k(\mathbf{p})$ and $U(\mathbf{q})$ in the Hamiltonian Eq. 3.1. The microscopic kinetic energy of a system with N_s particles can be expressed as follows:

$$E_k(\mathbf{p}) = \frac{1}{2}\mathbf{p}^\top\mathbf{M}^{-1}\mathbf{p}\,. \tag{3.13}$$

In the harmonic approximation, the system is close to a minimum of the potential energy, corresponding to the coordinates $\mathbf{q}_0$. Without loss of generality, we assume that the potential depends only on the coordinates $\mathbf{q}$. We then denote $U_0 = U(\mathbf{q}_0)$ the value of the potential at the minimum. The harmonic approximation involves performing a Taylor expansion of the potential $U(\mathbf{q})$ up to the second order, around the equilibrium position $\mathbf{q}_0$:

$$U(\mathbf{q}) = U_0 + \boldsymbol{\nabla}U(\mathbf{q}_0)\cdot(\mathbf{q}-\mathbf{q}_0) + \frac{1}{2}(\mathbf{q}-\mathbf{q}_0)\cdot\mathbf{H}\cdot(\mathbf{q}-\mathbf{q}_0) + o(\gamma\|\mathbf{q}-\mathbf{q}_0\|^3) \tag{3.14}$$

where $\boldsymbol{\nabla}U(\mathbf{q}_0) \in \mathbb{R}^{3N_s}$ is the gradient of the potential U with respect to the coordinates $\mathbf{q}$ evaluated at $\mathbf{q}_0$, $\|\cdot\|$ is the Euclidean norm and γ is a constant that ensures homogeneity. $\mathbf{H} = [H_{ij}] \in \mathbb{R}^{3N_s\times 3N_s}$ is the Hessian matrix of the potential energy with respect to $\mathbf{q}$ evaluated at $\mathbf{q}_0$:

$$H_{ij} = \left.\frac{\partial^2 U(\mathbf{q})}{\partial\, q_i \partial\, q_j}\right|_{\mathbf{q}=\mathbf{q}_0}\,. \tag{3.15}$$

We define the harmonic approximation by the second-order expansion of Eq. 3.14, which describes the potential energy around a minimum of the energy landscape, as a quadratic form of $\mathbf{q}-\mathbf{q}_0$.

The harmonic framework is a local approximation of the energy landscape around a given minimum. For small displacements $\mathbf{q}-\mathbf{q}_0$, a minimum can still be accurately described using this approximation. However, for larger displacements $\mathbf{q}-\mathbf{q}_0$, e.g., in the case of finite-temperature systems, the assumptions of this approximation may no longer be valid. Many finite-temperature phenomena cannot be understood without taking into account the anharmonic terms, i.e., the terms of order greater than 2, in the expansion of the potential energy Eq. 3.14. An example is the temperature dependence of the thermal expansion of crystalline materials, which remains independent of temperature within the framework of the harmonic approximation, as the equilibrium distance between atoms in such a model is independent of

temperature [208]. Another physical quantity that requires phonon anharmonicity as a key factor is thermal conductivity. In a perfect crystal of an insulator, harmonic phonons would not be scattered, resulting in theoretically infinite thermal conductivity at all temperatures.

3.1.2 Including anharmonicity in free energy

Due to the inadequacy of the harmonic approximation in describing volume-dependent thermal effects, the quasi-harmonic approximation [209] has been proposed. This approach assumes that the harmonic approximation holds for each value of the lattice constant, which implies that the free energy becomes volume-dependent. From a phonon perspective, the quasi-harmonic approximation assumes that phonons are volume-dependent but non-interacting. This assumption breaks down when phonon-phonon interactions need to be taken into account, which occurs in various situations, such as at high temperatures and during structural phase transitions. Therefore, we will move directly beyond the quasi-harmonic approximation and pursue a fully anharmonic free energy calculation.

To account for full anharmonicity, the complete Hamiltonian $\mathcal{H}(\mathbf{q},\mathbf{p})$ of the system, without any approximation of the potential energy $U(\mathbf{q})$, should be used. We recall the definition of the Helmholtz free energy for a system at temperature T:

$$
\begin{aligned}
F &= -\beta^{-1}\ln\left[\frac{1}{h^{3N_s}C}\int_{\mathcal{Q}\times\mathcal{P}}\exp\left(-\beta\mathcal{H}(\mathbf{q},\mathbf{p})\right)d\mathbf{q}d\mathbf{p}\right] \\
&= -\beta^{-1}\left[\ln\left(\frac{V^{N_s}}{h^{3N_s}C}\int_{\mathcal{P}}\exp\left(-\beta E_k(\mathbf{p})\right)d\mathbf{p}\right)+\ln\left(\frac{1}{V^{N_s}}\int_{\mathcal{Q}}\exp\left(-\beta U(\mathbf{q})\right)d\mathbf{q}\right)\right] \\
&= \underbrace{\left[\beta^{-1}\ln\left(C\prod_{i=1}^{N_s}\left(\Lambda_i^3/V\right)\right)\right]}_{F_{\mathrm{ig}}}+\underbrace{\left[-\beta^{-1}\ln\left(\frac{1}{V^{N_s}}\int_{\mathcal{Q}}\exp\left(-\beta U(\mathbf{q})\right)d\mathbf{q}\right)\right]}_{F_{\mathrm{ex}}} \\
&= -\beta^{-1}\ln\left[\frac{1}{C\prod_{i=1}^{N_s}\Lambda_i^3}\int_{\mathcal{Q}}\exp\left(-\beta U(\mathbf{q})\right)d\mathbf{q}\right],
\end{aligned}
\tag{3.16}
$$

where V is the supercell volume and $\Lambda_i = h/\sqrt{2\pi k_{\mathrm{B}}Tm_i}$ is the thermal de Broglie wavelength. The quantity F_{ig} is referred to as the ideal gas free energy (i.e., the free energy of a corresponding system in which the atoms are not subject to interatomic interactions), and F_{ex} is called the excess free energy. The latter quantity is zero for an ideal gas, for which the potential energy is zero ($U(\mathbf{q}) = 0$). The free energy F as given in Eq. 3.16 implicitly includes anharmonic effects. It is important to note that, despite the emergence of efficient methods, its direct calculation remains numerically challenging today. The ideal gas contribution of the free energy, F_{ig}, is easy to calculate through direct integration over the momentum space. However, evaluating the excess contribution, F_{ex}, is difficult due to the complexity of the energy landscapes of the systems under study. In the following, we will compute the differences in either the free energy or the excess free energy. The distinction between F and F_{ex} is unimportant, as we will always consider free energy differences between two systems with the same volume, atomic composition and temperature. The free energy differences computed in solids represent exactly the anharmonic contribution when the reference system is chosen as

the corresponding harmonic solid. Various numerical methods will be introduced to compute these free energy differences.

3.1.2.1 Thermodynamic integration-based calculations

The anharmonic free energy contribution, which is crucial for deriving the high-temperature properties, can be directly evaluated by thermodynamic integration (TI) from a suitable reference system [108, 109, 105]. In TI, the first derivative of free energy with respect to the reaction coordinate is firstly estimated using a sampling algorithm and then integrated. A reaction coordinate is an application defined on the configurational space $\mathcal{Q}$

$$\xi : \mathcal{Q} \to \mathbb{R}^m \,. \tag{3.17}$$

The reaction coordinate characterizes the transformation of interest. For example, it can be defined as the distance between two groups of atoms, which serves as a measure of binding energy. Defining the manifolds $\Sigma(z) = \{\mathbf{q} \in \mathcal{Q} \mid \xi(\mathbf{q}) = z\}$ where $z \in \mathbb{R}^m$ allows decomposing the configuration space as

$$\mathcal{Q} = \bigcup_{z \in \mathbb{R}^m} \Sigma(z) \tag{3.18}$$

where $m \leq 3N_s$. The corresponding partition function for a system with N_s atoms in the canonical ensemble can then be written as:

$$Z(N_s, V, T, z) = \frac{1}{C \prod_{i=1}^{N_s} \Lambda_i^3} \int_{\Sigma(z)} \exp\left[-\beta U(\mathbf{q})\right] \delta_{\xi(\mathbf{q})-z} d\mathbf{q} \,, \tag{3.19}$$

where $\delta_{\xi(q)-z}$ is the Dirac mass at z with $\xi(q)$ as the argument. Then the Landau free energy is defined as $A(z) = -\beta^{-1} \ln Z(N_s, V, T, z)$ (up to an addictive constant), and the difference in Landau free energy between two states represented by $\xi(\mathbf{q}) = a$ and $\xi(\mathbf{q}) = b$ becomes:

$$\Delta A = A(a) - A(b) = -\beta^{-1} \ln \left[\frac{\int_{\Sigma(a)} \exp\left[-\beta U(\mathbf{q})\right] \delta_{\xi(\mathbf{q})-a} d\mathbf{q}}{\int_{\Sigma(b)} \exp\left[-\beta U(\mathbf{q})\right] \delta_{\xi(\mathbf{q})-b} d\mathbf{q}} \right] . \tag{3.20}$$

Detailed procedure of TI in the reaction coordinate case can be found in Section 3.2 of Ref. [149]. Here, we briefly provide the expression for the case of an internal reaction coordinate. This is done to outline the generality and the difficulty associated with the TI approach. TI simply consists of integrating the derivative of $A(z)$ along a smooth curve $\mathcal{C}(a, b)$ whose endpoints are a and b:

$$\Delta A = \int_{\mathcal{C}(a,b)} \nabla A(z)^\top dz = \int_{\mathcal{C}(a,b)} \left[\frac{\int_{\Sigma(z)} \mathbf{f}(\mathbf{q}) \exp\left[-\beta U(\mathbf{q})\right] \delta_{\xi(\mathbf{q})-z} d\mathbf{q}}{\int_{\Sigma(z)} \exp\left[-\beta U(\mathbf{q})\right] \delta_{\xi(\mathbf{q})-z} d\mathbf{q}} \right]^\top dz \,, \tag{3.21}$$

where $u^\top v$ stands for the scalar product of vectors $u, v \in \mathbb{R}^m$ and $\mathbf{f}(\mathbf{q})$ is the local mean force. Its expression is quite involved and requires to introduce the $m \times m$ Gram matrix:

$$G = (\nabla \xi)^\top \nabla \xi \,, \tag{3.22}$$

with $G_{\alpha,\beta} = \nabla \xi_\alpha \cdot \nabla \xi_\beta$. The components of the local mean force are, for $\alpha \in \{1, \ldots, m\}$,

$$f_\alpha = \sum_{\gamma=1}^{m} G^{-1}_{\alpha,\gamma} \nabla \xi_\gamma \cdot \nabla U - \beta^{-1} \operatorname{div} \left(\sum_{\gamma=1}^{m} G^{-1}_{\alpha,\gamma} \nabla \xi_\gamma \right) \tag{3.23}$$

with $1 \leq \alpha \leq m$. The difficulty in implementing the TI approach lies in the ability to compute the local mean force, which involves second derivatives of the reaction coordinates. This task is non-trivial, especially for reaction coordinates based on bond-orientational order parameters. TI along an external parameter $\zeta \in \Lambda \subset \mathbb{R}$, such as temperature, pressure, volume or number of atoms, can be cast in this general set-up. It requires defining extended coordinates within an extended space $(\zeta, \mathbf{q}) \in \Lambda \times \mathcal{Q}$ together with an extended potential energy $U(\zeta, \mathbf{q})$. Since the "reaction coordinate" $\zeta \equiv \xi(\zeta, \mathbf{q})$ is orthogonal to the remaining coordinates $\mathbf{q}$, the local mean force simply becomes $\partial_\zeta U(\zeta, \mathbf{q})$.

In the following sections, we will exclusively consider the simple alchemical setting, based on a general potential energy $U(\zeta, \mathbf{q})$ that linearly mixes the potential energy of the given target system $U(\mathbf{q})$ and that of a reference system $U_{\text{ref}}(\mathbf{q})$ through a coupling parameter ζ:

$$U(\zeta, \mathbf{q}) = \zeta U(\mathbf{q}) + (1 - \zeta) U_{\text{ref}}(\mathbf{q}) \,. \tag{3.24}$$

When periodic boundary conditions are applied during sampling, the coordinate $\mathbf{q}$ is defined on the $3N_s$-dimensional torus $\mathcal{Q} = \mathbb{T}^{3N_s} \subset \mathbb{R}^{3N_s}$, i.e., the configuration space with periodic boundary conditions. Here, the coupling parameter ζ, similar to a reaction coordinate, has real values inside the $[0, 1]$ range. The canonical partition function in this case can be derived as:

$$Z(N_s, V, T, \zeta) = \frac{1}{C \prod_{i=1}^{N_s} \Lambda_i^3} \int_{\mathcal{Q}} \exp\left[-\beta U(\zeta\,, \mathbf{q})\right] d\mathbf{q} \,. \tag{3.25}$$

This partition function gives access to the probability of finding the generalized system in a state characterized by ζ:

$$P_0(\zeta) = \frac{Z(N_s, V, T, \zeta)}{\int_0^1 Z(N_s, V, T, \tilde{\zeta}) d\tilde{\zeta}} \,. \tag{3.26}$$

The associated Landau free energy reads:

$$A(\zeta) = -\beta^{-1} \ln P_0(\zeta) = -\beta^{-1} \ln Z(N_s, V, T, \zeta) + B \,, \tag{3.27}$$

where the quantity B is independent of ζ. Then, the derivative of the free energy can be deduced:

$$A'(\zeta) = \frac{\int_{\mathcal{Q}} \partial_\zeta U(\zeta, \mathbf{q}) \exp\left[-\beta U(\zeta, \mathbf{q})\right] d\mathbf{q}}{\int_{\mathcal{Q}} \exp\left[-\beta U(\zeta, \mathbf{q})\right] d\mathbf{q}} = \left\langle \frac{\partial U(\zeta, \mathbf{q})}{\partial \zeta} \right\rangle_\zeta \,, \tag{3.28}$$

where $\langle \cdot \rangle_\zeta$ denotes the ensemble average. Integration of Eq. 3.28 between $\zeta = 0$ and $\zeta = 1$ gives the difference in free energy:

$$\Delta A = A(1) - A(0) = \int_0^1 \langle U(\mathbf{q}) - U_{\text{ref}}(\mathbf{q}) \rangle_\zeta d\zeta \,, \tag{3.29}$$

where $A(0)$ and $A(1)$ represent the free energy of the reference system with potential $U_{\text{ref}}(\mathbf{q})$ and the target system with potential energy $U(\mathbf{q})$, respectively. Based on Eq. 3.29, determination of free energy of any given system is subject to two requirements: the existence of reference system for which free energy can be calculated numerically or analytically, as well as a reversible artificial pathway between the system of interest and the reference crystal, which allows to carry out the thermodynamic integration of potential energy. Here the coupling parameter ζ provides a smooth transition from a known reference state to a realistic system.

TI-based calculations of thermodynamic properties including the effect of anharmonicity from accurate electronic structure calculations, i.e., *ab initio* calculations, were initiated in 2001 [110, 111]. Such *ab initio* brute-force TIs are often computationally prohibitive in practice, because it is necessary to sample too many configurations along the path of integration [42]. Accuracy of TI is highly dependent on $\partial U(\zeta, \mathbf{q})/\partial\zeta$, which is strongly curved in nature. As a result, the discretization of ζ often involves up to tens of steps, and for each discretized value of ζ, a canonical sampling of configurations should be launched. Due to the high computational requirements, several improvements have been proposed to make TI-based methods more feasible and, in particular, amenable to electronic structure calculations. Notably, upsampled thermodynamic integration using Langevin dynamics (UP-TILD) method was developed [112], which enhances the performance of TI by combining the DFT calculations using "reduced" DFT parameters (energy cutoff of kinetic energy and k-points sampling of the Brillouin zone) with an almost configuration-independent offset with respect to the fully converged energy:

$$\begin{aligned} \Delta A &= \int_0^1 \left[\langle U_{\text{low}}^{\text{DFT}} - U_{\text{ref}}^{\text{qh}} \rangle_\zeta + \langle \Delta U \rangle_\zeta^{\text{UP}} \right] d\zeta \\ &= \int_0^1 \left[\langle U_{\text{low}}^{\text{DFT}} - U_{\text{ref}}^{\text{qh}} \rangle_\zeta + \frac{1}{N} \sum_{i=1}^{N} (U_{\text{high}}^{\text{DFT}} - U_{\text{low}}^{\text{DFT}}) \right] d\zeta \,. \end{aligned} \tag{3.30}$$

Here $U_{\text{low}}^{\text{DFT}}$ is given by the "reduced" DFT calculations and $U_{\text{ref}}^{\text{qh}}$ is the DFT quasi-harmonic energy of the quasi-harmonic reference system. These two energies are required for every atomic configuration during the Langevin dynamics run associated with a specific ζ. $U_{\text{high}}^{\text{DFT}}$ is calculated using fully converged DFT parameter, but only a small number of configurations ($N \approx 5$) are required to evaluate this term. Based on UP-TILD, an improved version, referred to as two-stage upsampled thermodynamic integration using Langevin dynamics (TU-TILD), was developed to further accelerate the convergence of the calculation [113], wherein TI is split into two stages, first from the quasi-harmonic system to an intermediate potential, and then from the intermediate potential to the exact DFT Hamiltonian:

$$\Delta A = \int_0^1 \left[\langle U^{\text{pot}} - U_{\text{ref}}^{\text{qh}} \rangle_{\zeta_1} \right] d\zeta_1 + \int_0^1 \left[\langle U_{\text{low}}^{\text{DFT}} - U^{\text{pot}} \rangle_{\zeta_2} + \langle \Delta U \rangle_{\zeta_2}^{\text{UP}} \right] d\zeta_2 \,, \tag{3.31}$$

where the second term of the right-hand side is computed using the aforementioned UP-TILD method. This method is more efficient compared to the UP-TILD method because the time-consuming DFT MD runs are only conducted during the second stage, and the convergence of TI in this stage is much faster than in the UP-TILD approach since the configuration space sampled by the intermediate potentials is closer to the DFT sampling. The TU-TILD scheme has recently been applied in combination with Moment Tensor Potentials (MTPs) [114], a class of ML potentials that have demonstrated both accuracy and efficiency, serving here as an intermediate thermodynamic state and an intermediate potential [42, 43, 44]. However, these methods assume that the points sampled from the intermediate reference distribution faithfully represent the target distribution associated with the exact Hamiltonian. This assumption can be true for high quality reference or intermediate potentials, such as the MTPs within the ML framework, but is not very reliable when using simple reference systems such as harmonic or quasi-harmonic Hamiltonians [42]. In general, when ML potentials are chosen as intermediate potentials, the free energy from the reference should be computed with great accuracy. However, this can be a crude task because, at least for crystalline materials, the ML

force fields are from several tens times slower up to four orders of magnitude slower in terms of CPU times than, for example, traditional EAM potentials. For this reason, the method proposed in this chapter aims to reduce the computational cost.

3.1.2.2 Free energy perturbation and adaptive sampling

Free energy perturbation (FEP) [115] is another powerful computational method used to calculate free energy difference between two states of a system. One of the key advantages of FEP is its computational efficiency. Without dependence on the designation of ζ, it typically requires fewer simulations than TI, making it a more attractive option for large systems or when computational resources are limited. The FEP method consists of sampling the reference distribution and estimating the free energy difference via the logarithm of a partition function ratio [115]. In the alchemical case, we can deduce from Eq. 3.27 that

$$\begin{aligned}\Delta A &= A(1) - A(0) = -\beta^{-1} \ln \frac{Z_1}{Z_0} \\ &= -\beta^{-1} \ln \left[Z_0^{-1} \frac{1}{\prod_{i=1}^{N_s} \Lambda_i^3} \int_{\mathcal{Q}} \exp\left(-\beta U(\mathbf{q})\right) d\mathbf{q} \right] \\ &= -\beta^{-1} \ln \left[\int_{\mathcal{Q}} \exp\left[-\beta(U(\mathbf{q}) - U_{\text{ref}}(\mathbf{q}))\right] \nu_0(d\mathbf{q}) \right] . \end{aligned} \tag{3.32}$$

where we denote $Z_\zeta = Z(N_s, V, T, \zeta)$ as defined in Eq. 3.25, and for a given ζ,

$$\nu_\zeta(d\mathbf{q}) = Z_\zeta^{-1} \left[\frac{1}{\prod_{i=1}^{N_s} \Lambda_i^3} \exp\left(-\beta U(\zeta, \mathbf{q})\right) \right] d\mathbf{q} \,,\; Z_\zeta = \frac{1}{\prod_{i=1}^{N_s} \Lambda_i^3} \int_{\mathcal{Q}} \exp\left(-\beta U(\zeta, \mathbf{q})\right) d\mathbf{q} \tag{3.33}$$

is the canonical measure on the configurational space for a separable Hamiltonian, as explained in Subsection 3.1.1.1. An approximation of Eq. 3.32 is obtained by generating configurations $\mathbf{q}_s \sim \nu_0$, using sampling techniques such as Markov chain methods (e.g., Monte Carlo (MC) simulations) and stochastic dynamics (e.g., the Langevin process), and averaging the quantities $\exp\left[-\beta(U(\mathbf{q}_s) - U_{\text{ref}}(\mathbf{q}_s))\right]$, leading to the estimator:

$$\widehat{\Delta A} = -\beta^{-1} \ln \left[\frac{1}{M} \sum_{s=1}^{M} \exp\left[-\beta(U(\mathbf{q}_s) - U_{\text{ref}}(\mathbf{q}_s))\right] \right] , \; \mathbf{q}_s \sim \nu_0 \,. \tag{3.34}$$

Alternatively, the free energy difference can also be computed from an average with respect to the measure ν_1 by regarding $\Delta A = -\left[A(0) - A(1)\right]$. In this case, the ratio Z_0/Z_1 is estimated and we have:

$$\widehat{\Delta A} = \beta^{-1} \ln \left[\frac{1}{M} \sum_{s=1}^{M} \exp\left[\beta(U(\mathbf{q}_s) - U_{\text{ref}}(\mathbf{q}_s))\right] \right] , \; \mathbf{q}_s \sim \nu_1 \,. \tag{3.35}$$

In practice, the accuracy of the FEP method is controlled by the degree of overlap between the reference and target distributions [116]. A straightforward idea is to divide the transition into a series of independent smaller "windows", i.e., $\Delta A = \sum_{i=1^N-1} A(\zeta_{i+1}) - A(\zeta_i)$

with $\zeta_1 = 0$ and $\zeta_N = 1$. This setup actually returns to the concept of TI. An enhanced method to improve the accuracy of FEP involves sampling from a third distribution that exhibits better overlapping properties with both the target and reference distributions. This importance-sampling approach for the calculation of free energy difference traditionally employs an auxiliary biasing potential and is dubbed umbrella sampling [118, 119], precisely because the sampling distribution specified by the biasing potential should cover simultaneously the region of configuration space relevant to both the target and reference systems. Establishing such a biasing potential that provides good overlapping properties is a challenging task. To achieve this, the well-established framework of adaptive methods serves as a source of inspiration [120, 121, 129, 130, 123, 131, 124, 125, 117].

The main motivation for the development of the adaptive sampling methods is to overcome the obstacles arising during the stochastic process due to free energy barriers [115, 149]. The principle is to modify the potential U applied on the atoms during the simulation, based on an approximation $A_\star$ of the free energy A associated with some reaction coordinate of interest ξ. The corresponding biased potential $U_{A_\star}$ is defined as:

$$U_{A_\star}(\zeta, \mathbf{q}) = U(\zeta, \mathbf{q}) - A_\star(\zeta)\,, \tag{3.36}$$

where $A_\star$ is referred to as the biasing potential added to the original potential. The biasing potential is adjusted on-the-fly, and it is necessary to provide some rule to update the biasing term $A_\star$ in such a way that $A_\star$ tends towards the free energy A in the longtime limit. When the method is constructed carefully, it can be proved that if $A_\star$ converges (which is usually nontrivial), then it converges to A, up to an additive constant [149].

Let dW_t^ζ and $dW_t^\mathbf{q}$ denote the standard Wiener process in Λ and $\mathcal{Q}$. There are two types of adaptive methods, depending on the way the bias is updated.

Adaptive Biasing Potential (ABP) method in the extended space $\Lambda \times \mathcal{Q}$ updates the biasing potential $A_\star$, indicated as A_t for a given time t. The biased overdamped Langevin dynamics reads:

$$\begin{cases} d\zeta_t = \alpha(A_t'(\zeta_t) - \partial_\zeta U(\zeta_t, \mathbf{q}_t))\,dt + \sqrt{2\alpha\beta^{-1}}\,dW_t^\zeta, \\ d\mathbf{q}_t = -\nabla_\mathbf{q} U(\zeta_t, \mathbf{q}_t)\,dt + \sqrt{2\beta^{-1}}\,dW_t^\mathbf{q}\,, \\ dA_t(\zeta^\star) = -\eta_t \beta^{-1} \delta^\epsilon_{\zeta^\star - \zeta_t}\,dt\,, \end{cases} \tag{3.37}$$

where the factor α allows to control the time scale associated with ζ_t and to decouple it from $\mathbf{q}_t$. The function δ^ϵ corresponds to a Gaussian distribution with zero mean and variance ϵ, which converges to the Dirac delta function as ϵ approaches 0. The function η_t is strictly positive and decays to 0. It controls the speed at which the biasing potential is updated and is the crucial simulation parameter of the ABP method. We can intuitively interpret that the ABP method penalizes the regions in the reaction coordinate space that have been previously visited by increasing the potential energy in these areas, as illustrated in Fig. 3.1. Consequently, this technique encourages the exploration of unvisited regions. In this framework, if A_t converges in the longtime limit, then it converges to A. Furthermore, the convergence speed of A_t in the ABP method is exponential. The challenge in implementing this approach lies in the tuning of the decaying function η_t. If the decay is too rapid, the biasing potential will converge too slowly, whereas if the decay is too slow, the biasing potential will experience significant

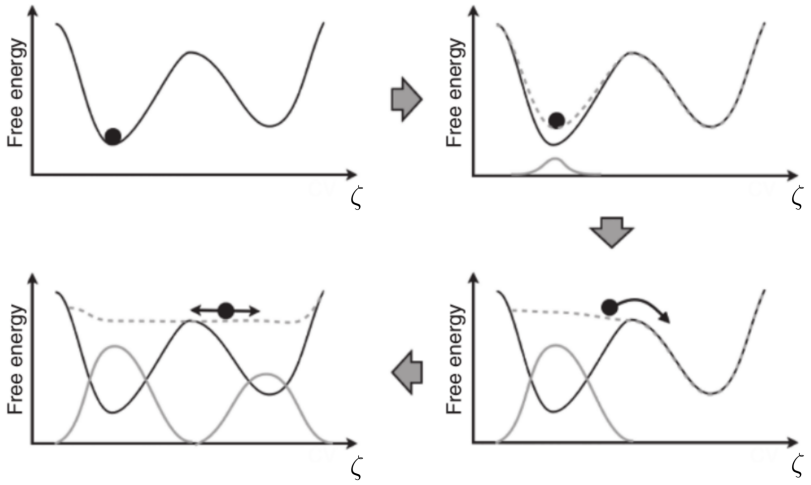

Figure 3.1: Iterative construction of the biasing potential $A_\star$ to sample the free energy profile along a reaction coordinate ζ, from Bussi *et al.* [210]. The real potential U is shown in dark gray and the biasing potential $A_\star(\zeta)$ in light gray.

fluctuations and will not stabilize. It is necessary to identify an appropriate balance between these two extremes.

In terms of convergence, the Adaptive Biasing Force (ABF) method is more reliable and possesses fewer simulation parameters [149].

Adaptive Biasing Force (ABF) method in the extended space $\Lambda \times \mathcal{Q}$ updates the derivative $A'_\star$ at $\zeta^\star \in \Lambda$, which is denoted as $A'_t(\zeta^\star)$ for a given time t in the following discussion. The ABF dynamics takes the following form:

$$\begin{cases} d\zeta_t = \alpha(A'_t(\zeta_t) - \partial_\zeta U(\zeta_t, \mathbf{q}_t))\, dt + \sqrt{2\alpha\beta^{-1}}\, dW_t^\zeta\,, \\ d\mathbf{q}_t = -\nabla_\mathbf{q} U(\zeta_t, \mathbf{q}_t)\, dt + \sqrt{2\beta^{-1}}\, dW_t^\mathbf{q}\,, \\ A'_t(\zeta^\star) = \frac{\int_0^t \partial_\zeta U(\zeta_s, q_s)\delta^\epsilon_{\zeta^\star - \zeta_s} ds}{\int_0^t \delta^\epsilon_{\zeta^\star - \zeta_s} ds}\,. \end{cases} \tag{3.38}$$

It is rigorously proved that the biasing force A'_t converges to the free energy A in the long time limit [134]. The interpretation of the ABF method should be slightly different from that of the ABP approach. Here, a bias is added directly to the forces acting on the ζ coordinate, counteracting the mean force along the transition direction to prevent the system from being trapped by free energy barrier, as well as reducing the statistical variance [145].

ABF-based techniques have been widely developed and applied over the past two decades [135, 136, 137, 140, 138, 211]. As an improvement of the traditional MD- and MC-based sampling methods with proven convergence [134], ABF algorithms enhance remarkably the sampling efficiency [129, 130, 131, 132, 133]. In the upcoming sections, we will present a series of ABF-based methods for the fast evaluation of the anharmonic free energy of computationally intensive force fields. These methods build upon the previously developed approaches introduced in this section and demonstrate a significant improvement in computational efficiency compared to those earlier approaches.

3.2 Accelerated Bayesian adaptive biasing force method

In this section, we present an accelerated ABF-based method that enables the calculation of anharmonic free energy with DFT accuracy at a very reasonable computational cost. Specifically, our approach is approximately 100 times faster than traditional TI for achieving an accuracy of 0.1 meV/atom, and about 800 times faster than conventional MD simulations for obtaining the bulk modulus (the second derivative of free energy) at high temperatures (see subsection 3.2.3). Therefore, it allows for the use of ML force fields that provide DFT-level accuracy, noting they are typically computationally slower by several tens to four orders of magnitude compared to traditional EAM potentials.

Our methods mix the reference and target Hamiltonians without the specific requirement of good overlapping properties between the reference and target distributions, and they recover the right statistics resorting to Bayes formula, without further approximations. The Bayesian reasoning framework [143, 144, 145, 212] for sampling the distribution associated to an extended Hamiltonian allows to systematically reduce the statistical variance of the estimated free energy difference when the converged biasing force is frozen, compared with the other standard estimators (FEP, TI, thermodynamic-occupation and weighted-histogram estimators, see Ref. [145]). Note that this Bayesian approach can be viewed as a particular and efficient FEP method because the biasing mean force is computed through umbrella sampling and because the implemented Bayes formula corresponds to a FEP equation in which the sampled reference distribution is a biased marginal probability distribution.

Here, a Bayesian adaptive approach is used for the first time to compute the anharmonic free energy of a crystalline solid [213]. In Subsection 3.2.1, we detail the method, including the theoretical framework and the algorithm implementation. In Subsection 3.2.2, two new features are proposed to enhance the robustness and efficiency of the method: one concerns the reference system for TI, and the other involves a weighting technique for sampling. In Subsection 3.2.3, we validate the current approach by comparing it with MD simulations for bulk modulus calculations at high temperatures. Moreover, we demonstrate the significant reduction in computational cost achieved by our method compared to MD and TI.

3.2.1 Bayesian adaptive biasing force: theoretical framework and implementation

We recall that for a system containing N_s atoms, the alchemical TI is performed using a general potential energy $U(\zeta, \mathbf{q})$ that linearly mixes the potential energy of the given target system $U(\mathbf{q})$ and of the reference system $U_{\text{ref}}(\mathbf{q})$ through a coupling parameter ζ:

$$U(\zeta, \mathbf{q}) = \zeta U(\mathbf{q}) + (1 - \zeta) U_{\text{ref}}(\mathbf{q}) \,. \tag{3.39}$$

Here, $\mathbf{q}$ is defined in the $3N_s$-dimensional torus $\mathbb{T}^{3N_s} \subset \mathbb{R}^{3N_s}$ since periodic boundary conditions are applied. The interval of $\zeta \in [0, 1]$ is discretized, in our study into 201 equally spaced discrete values. Beyond this simple numerical setup, some strategies have proven useful in variance minimization and convergence acceleration, including adoption of a nonlinear dependence of the potential on the coupling parameter [115], as well as modification of discretization

method [149]. Besides, the free energy difference can be directly computed from a single very long simulation in which the temporal reaction coordinate $\zeta(t)$ progresses linearly from 0 to 1 [214]. Although sometimes advantageous, it is unnecessary to employ such methods in our case for which accuracy and efficiency are both guaranteed by the Bayesian formalism.

To numerically calculate the integral in Eq. 3.29, we propose an optimized ABF algorithm, namely the Bayesian ABF (BABF) method, based on the Bayesian estimator proposed in Ref. [143]. In the alchemical setting, the biased potential $U_{A_\star}$ in Eq. 3.36 is written as:

$$U_{A_\star}(\zeta, \mathbf{q}) = U(\zeta, \mathbf{q}) - A_\star(\zeta), \tag{3.40}$$

where $A_\star$ is the biasing potential added to the extended potential, and it will be adaptively constructed by averaging and integrating the derivative of the extended Hamiltonian with respect to the coupling parameter.

Let $P_{A_\star}(\zeta, \mathbf{q})$ denote the joint probability of the extended state $(\zeta, \mathbf{q})$ in the extended ensemble with biasing potential $A_\star$. We have

$$P_{A_\star}(\zeta, \mathbf{q}) = \frac{\exp\left[-\beta U_{A_\star}(\zeta, \mathbf{q})\right]}{\iint_{\mathbb{T}^{3N_s}\times[0,1]} \exp\left[-\beta U_{A_\star}(\tilde{\zeta}, \tilde{\mathbf{q}})\right] d\tilde{\zeta} d\tilde{\mathbf{q}}}. \tag{3.41}$$

Then the mean force $A'(\zeta)$ formalized by Eq. 3.28 can be transposed in the extended ensemble associated with the biasing potential $U_{A_\star}(\zeta, \mathbf{q})$:

$$\begin{aligned} A'(\zeta) &= \frac{\int_{\mathbb{T}^{3N_s}} \partial_\zeta U(\zeta, \mathbf{q}) \exp[-\beta U_{A_\star}(\zeta, \mathbf{q})] d\mathbf{q}}{\int_{\mathbb{T}^{3N_s}} \exp[-\beta U_{A_\star}(\zeta, \mathbf{q})] d\mathbf{q}} \\ &= \frac{\int_{\mathbb{T}^{3N_s}} \partial_\zeta U(\zeta, \mathbf{q}) P_{A_\star}(\zeta, \mathbf{q}) d\mathbf{q}}{\int_{\mathbb{T}^{3N_s}} P_{A_\star}(\zeta, \mathbf{q}) d\mathbf{q}}. \end{aligned} \tag{3.42}$$

The above equality indicates that it is possible to compute the mean force in the biased extended system. We show how to do it through Bayesian reasoning.

The marginal probability of ζ and $\mathbf{q}$ can be respectively expressed as

$$P_{A_\star}(\zeta) = \int_{\mathbb{T}^{3N_s}} P_{A_\star}(\zeta, \mathbf{q}) d\mathbf{q} \tag{3.43}$$

and

$$P_{A_\star}(\mathbf{q}) = \int_0^1 P_{A_\star}(\zeta, \mathbf{q}) d\zeta. \tag{3.44}$$

Then, the conditional probabilities of ζ for a given $\mathbf{q}$ and of $\mathbf{q}$ for a given ζ are respectively given by $p_{A_\star}(\zeta|\mathbf{q}) = P_{A_\star}(\zeta, \mathbf{q})/P_{A_\star}(\mathbf{q})$ and $p_{A_\star}(\mathbf{q}|\zeta) = P_{A_\star}(\zeta, \mathbf{q})/P_{A_\star}(\zeta)$. The two equivalent expressions of the joint probability below

$$p_{A_\star}(\mathbf{q}|\zeta) P_{A_\star}(\zeta) = p_{A_\star}(\zeta|\mathbf{q}) P_{A_\star}(\mathbf{q}) \tag{3.45}$$

allow to formulate the Bayes relation, expressing the conditional probability of $\mathbf{q}$ given ζ as a function of that of ζ given $\mathbf{q}$:

$$p_{A_\star}(\mathbf{q}|\zeta) = \frac{p_{A_\star}(\zeta|\mathbf{q}) P_{A_\star}(\mathbf{q})}{P_{A_\star}(\zeta)}. \tag{3.46}$$

The two equivalent expressions in Eq. (3.45) also allow to cast the mean force in Eq. (3.42) into the two respective forms:

$$A'(\zeta) = \int_{\mathbb{T}^{3N_s}} \partial_\zeta U(\zeta, \mathbf{q}) p_{A_\star}(\mathbf{q}|\zeta) d\mathbf{q} \tag{3.47}$$

$$= \frac{\int_{\mathbb{T}^{3N_s}} \partial_\zeta U(\zeta, \mathbf{q}) p_{A_\star}(\zeta|\mathbf{q}) P_{A_\star}(\mathbf{q}) d\mathbf{q}}{\int_{\mathbb{T}^{3N_s}} p_{A_\star}(\zeta|\mathbf{q}) P_{A_\star}(\mathbf{q}) d\mathbf{q}}, \tag{3.48}$$

where the marginal probability of ζ cancels in Eq. (3.47) and appears as the denominator of Eq. (3.48) in the form $P_{A_\star}(\zeta) = \int_{\mathbb{T}^{3N_s}} p_{A_\star}(\zeta|\mathbf{q}) P_{A_\star}(\mathbf{q}) d\mathbf{q}$. Equation (3.48) corresponds to the expectation form of Bayes formula.

If the distribution $P_{A_\star}(\mathbf{q})$ can be sampled, the above average can be well approximated owing to the ergodic theorem. Given a sequence of N points $\{\mathbf{q}_s\}_{1\leq s\leq N}$ sampled from the probability distribution $P_{A_\star}(\mathbf{q})$, the mean force is estimated as:

$$A'(\zeta) = \frac{\sum_{s=1}^{N} \partial_\zeta U(\zeta, \mathbf{q}_s) p_{A_\star}(\zeta|\mathbf{q}_s)}{\sum_{s=1}^{N} p_{A_\star}(\zeta|\mathbf{q}_s)}, \tag{3.49}$$

where the conditional probabilities of ζ given the sampled points are directly calculated through numerical quadrature and the relation below:

$$p_{A_\star}(\zeta|\mathbf{q}) = \frac{\exp\left[-\beta U_{A_\star}(\zeta, \mathbf{q})\right]}{\int_0^1 \exp[-\beta U_{A_\star}(\tilde{\zeta}, \mathbf{q})] d\tilde{\zeta}}. \tag{3.50}$$

A simple way to sample the distribution $P_{A_\star}(\mathbf{q})$ consists in implementing an overdamped Langevin dynamics:

$$d\mathbf{q}_t = \nabla_{\mathbf{q}} \left\{\beta^{-1} \ln\left[P_{A_\star}(\mathbf{q}_t)\right]\right\} dt + \sqrt{2\beta^{-1}} dW_t. \tag{3.51}$$

The first term of the right-hand side of Eq. 3.51 can be written as an effective force field:

$$\nabla_{\mathbf{q}} \left\{\beta^{-1} \ln\left[P_{A_\star}(\mathbf{q})\right]\right\} = -\int_0^1 \nabla_{\mathbf{q}} U(\zeta, \mathbf{q}) p_{A_\star}(\zeta|\mathbf{q}) d\zeta = \mathbb{F}_{A_\star}(\mathbf{q}) .$$

In practice, we prevent the center of mass of the system from drifting with respect to the fixed reference lattice. To achieve this, we introduce a projection $\mathbf{P}$ that shifts the center of mass from any position to its initial position. This operator being symmetric and idempotent, we have $\mathbf{P} = \mathbf{P}\mathbf{P}^\top$. Consequently, the projected Langevin dynamics of the system reads:

$$d\mathbf{q}_t = \mathbf{P}\mathbb{F}_{A_\star}(\mathbf{q}_t) dt + \sqrt{2\beta^{-1}} \mathbf{P} dW_t. \tag{3.52}$$

By definitions in Eq. 3.25 and Eq. 3.27, we have $\forall z \in [0, 1]$,

$$\ln \frac{P_{A_\star}(\zeta = z)}{P_{A_\star}(\zeta = 0)} = \ln \frac{\int_{\mathbb{T}^{3N_s}} \exp\left[-\beta U_{A_\star}(z, \mathbf{q})\right] d\mathbf{q}}{\int_{\mathbb{T}^{3N_s}} \exp\left[-\beta U_{A_\star}(0, \mathbf{q})\right] d\mathbf{q}}$$

$$= \beta[(A_\star(z) - A_\star(0)) - (A(z) - A(0))]. \tag{3.53}$$

Therefore, once we obtain a uniform sampling of $P_{A_\star}(\zeta)$ over $\zeta \in [0, 1]$, we have $A_\star(z) - A_\star(0) = A(z) - A(0)$ which gives the difference in free energy from the difference of biasing potentials.

To implement the numerical algorithm, Eq. 3.52 is discretized. The $(n+1)^{th}$ molecular dynamics move can be realized based on the first n steps with the following algorithm:

1. $A'_n(\zeta)$ is computed as:

$$A'_n(\zeta) = \frac{\sum_{s=1}^{n-1} \partial_\zeta U(\zeta, \mathbf{q}_s) p_{A_s}(\zeta|\mathbf{q}_s) w(s)}{\sum_{s=1}^{n-1} p_{A_s}(\zeta|\mathbf{q}_s) w(s)}, \tag{3.54}$$

where $\nabla_\zeta U(\zeta, \mathbf{q}_s)$, being exactly equal to $U(\mathbf{q}) - U_{ref}(\mathbf{q})$, is easy to evaluate and the weighting function $w(s)$, detailed in the following section, is introduced to accelerate the initial convergence.

2. The free energy $A_n(\zeta)$ for the step n is determined as:

$$A_n(\zeta) = \int_0^\zeta A'_n(\tilde{\zeta}) d\tilde{\zeta} + A_n(0). \tag{3.55}$$

3. The corresponding conditional probability of ζ for a given $\mathbf{q}_n$ is calculated as:

$$p_{A_n}(\zeta|\mathbf{q}_n) = \frac{\exp\left[-\beta U_{A_n}(\zeta, \mathbf{q}_n)\right]}{\int_0^1 \exp\left[-\beta U_{A_n}(\tilde{\zeta}, \mathbf{q}_n)\right] d\tilde{\zeta}}. \tag{3.56}$$

4. The effective force field is obtained by the equation:

$$\mathbb{F}_{A_n}(\mathbf{q}_n) = -\int_0^1 \nabla_{\mathbf{q}} U(\zeta, \mathbf{q}_n) p_{A_n}(\zeta|\mathbf{q}_n) d\zeta. \tag{3.57}$$

5. Integration of the dynamic equation is carried out to obtain:

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \mathbf{P}\mathbb{F}_{A_n}(\mathbf{q}_n)\delta t + \sqrt{2\beta^{-1}\delta t} B_n, \tag{3.58}$$

where $B_n \sim \mathcal{N}(\mathbf{0}, \mathbf{P})$ is a normal deviate with zero mean and $\mathbf{P}$ variance.

Once the new positions are determined, iteration continues until the desired accuracy of $A_n(\zeta)$ is achieved. Based on Eq. 3.53, we have $\lim_{n\to+\infty} \Delta A_n = \Delta A$. The iterative process is concisely illustrated by the flowchart in Fig. 3.2.

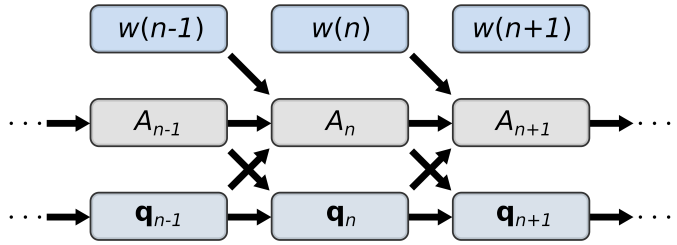

Figure 3.2: Bayesian iterative flowchart of the present free energy sampling: the forward step integrates the whole history of the Langevin dynamics. The weight $w(n)$ is inserted at each integration step in order to accelerate the convergence of A_n and its form is a user choice as discussed in Subsection 3.2.2.2.

In practical applications, we analyze the sampling stability on the fly by monitoring the running estimate of the marginal probability at step n:

$$P_{A_n}(\zeta) \quad = \quad \frac{1}{n}\sum_{s=1}^{n} p_{A_s}(\zeta|\mathbf{q}_s)\,, \tag{3.59}$$

which is the mean value of the probability of ζ conditioned on the sampled states $\mathbf{q}_s$ during the Langevin dynamics and given the successive biasing potentials A_s, with $1 \leq s \leq n$. As a probability density, the integral of $P_{A_n}(\zeta)$ over $\zeta \in [0,1]$ is 1. This distribution should be uniform in ζ in order to ensure that the difference in free energy of the target and reference systems becomes equal to the corresponding difference in the biasing potentials yielded by the adaptive Langevin dynamics, as expected from Eq. (3.53).

For assessing the relative efficiency of sampling using different reference systems and weighting functions, we also monitor the Kullback–Leibler (KL) divergence of $P_{A_n}(\zeta)$ from the uniform distribution:

$$D_{\mathrm{KL}}(P_{A_n}||\mathbf{1}_{[0,1]}) = \int_0^1 P_{A_n}(\zeta) \ln \frac{P_{A_n}(\zeta)}{\mathbf{1}_{[0,1]}(\zeta)} d\zeta. \tag{3.60}$$

This statistical pseudo-norm measures how close the probability distribution $P_{A_n}(\zeta)$ is to the uniform distribution $\mathbf{1}_{[0,1]}(\zeta)$. It is 0 for identical distributions and infinite for non overlapping distributions. The closer to zero the KL divergence is, the better the adaptive sampling procedure has converged. The reason is that the biasing potential integrated from a well-converged biasing force yields a reliable estimate of the free energy $A(\zeta)$. The recorded marginal probability of ζ, $P_{A_n}(\zeta)$ from Eq. 3.59, should thus converge towards $P_A(\zeta)$ that corresponds to the uniform probability distribution over $[0,1]$ range, which is $\mathbf{1}_{[0,1]}(\zeta)$.

3.2.2 Optimizations of BABF method

In this subsection, we enrich the Bayesian adaptive approach with two new features. First, we eliminate the numerical instabilities of the reference Hamiltonian via a singular value decomposition (SVD) filter to improve sampling robustness. This technique is detailed in Subsection 3.2.2.1, where a comparison between two different reference systems is also provided. Second, we introduce in Subsection 3.2.2.2 a simple weighting scheme for the biasing force to enhance the initial speed of convergence.

3.2.2.1 Reference system: choices and numerical instabilities filter

Hoover and Ree's single-occupancy cell (SOC) method [215, 216] is one of the first methods proposed to obtain a reference system in the form of artificial solid. According to the SOC method, each of the N_s atoms in the system is confined in its own cell of volume V/N_s at all densities to ensure that no melting occurs and the solid-phase thermodynamic properties of this artificial solid hold true. Though unphysical, this method was successfully applied to the calculation of the free energy of hard-spheres [215] using computer simulations. Hoover *et al.* [217, 218] also suggested an alternative method which involved a two-step calculation [109]. In the first step, the given solid is approximated as a harmonic crystal by cooling it to a sufficiently low temperature and then the free energy of this solid can be analytically calculated using lattice dynamics. The second step involves using the result from the first step as a reference to calculate the free energy at a given higher temperature. However, both methods have some drawbacks. For instance, numerical integration for the SOC method is complicated

for unstable solids where phase transitions may occur; cooling used to obtain a harmonic crystal may not be reversible, etc.

In order to improve the performance of the method, other useful references have been proposed, such as Einstein crystal [219, 220], harmonic solid [221, 222], Morse potential [107], Lennard-Jones system [111], inverse power potential [111, 110], EAM [110, 223], MTP [114] and linear combination of the potentials above [110]. In this study only reference systems with analytically known free energy are considered. We test the use of Einstein and harmonic approximations and further optimize the harmonic reference. We demonstrate that such simple references are sufficient to provide accurate and efficient calculations with the BABF approach.

Einstein approximation. The Einstein crystal has been used extensively as a reference system in free energy calculations since the works of Broughton *et al.* [224] and Frenkel *et al.* [219], due to the simplicity of its formalism and implementation. In Einstein approximation (EA), it is assumed that the atoms oscillate independently about their mechanical equilibrium positions denoted by $\mathbf{q}_0 \in \mathbb{T}^{3N_s}$, i.e., the lattice vibrations are represented by N_s three-dimensional classical distinguishable oscillators, yielding $3N_s$ oscillators with an average frequency ω_i for $i = 1, \cdots, 3N_s$ and atomic mass m_i. The Hamiltonian of the Einstein system has the following potential energy:

$$\begin{aligned} U(\mathbf{q}) &= U(\mathbf{q}_0) + U_{\mathrm{EA}}(\mathbf{q}, \mathbf{q}_0) \\ U_{\mathrm{EA}}(\mathbf{q}, \mathbf{q}_0) &= \frac{1}{2}\sum_{i=1}^{3N_s} m_i \omega_i^2 (q_i - q_{i,0})^2. \end{aligned} \tag{3.61}$$

For the special case where all atoms have the same mass $m_i = m$ and vibrate at the same frequency $\omega_i = \omega$, the free energy can be written as:

$$F_{\mathrm{EA}} = 3N_s \beta^{-1} \ln \beta\hbar\omega. \tag{3.62}$$

As previously emphasized, our aim is to construct a reversible path from the non-interacting Einstein crystal to the crystal of interest. This implies that the lattice of the reference system at mechanical equilibrium should coincide with that of the target crystal. This condition can be ensured by harmonically coupling (via springs) the center of mass of the Einstein crystal to that of the crystal of interest. According to Ryckaert *et al.* [225, 109], Eq. 3.62, for a constrained system, can be written as:

$$F_{\mathrm{EA,\,c}} = (3N_s - 3)\beta^{-1} \ln \beta\hbar\omega. \tag{3.63}$$

The main drawback of EA is that the Einstein frequency needs to be calibrated very carefully, because even a small variation in frequency can produce a large difference in the free energy of the reference and target crystals and result in computational overhead. Using Eq. (3.63), this calibration can be performed through the harmonic free energy calculation of a small system:

$$\omega(\beta) = \frac{1}{\beta\hbar} \exp\left[\frac{\beta F^{\mathrm{bulk}}(\beta)}{3N_s - 3}\right], \tag{3.64}$$

taking into account the temperature dependence of ω. For instance, the full anharmonic free energy $F^{\text{bulk}}(\beta)$ can be quickly estimated in a short simulation at inverse temperature β using a system with a small number of atoms (in the case of bcc lattice, a $2 \times 2 \times 2$ simulation cell with 16 atoms may be sufficient) and then, through the estimated value of $\omega(\beta)$, scaled to a larger system.

Harmonic approximation. Alternatively, the reference system can be built upon the harmonic approximation (HA), which involves a second-order Taylor expansion of the potential energy around the mechanical equilibrium (subsection 3.1.1.3):

$$\begin{aligned} U(\mathbf{q}) &= U(\mathbf{q}_0) + U_{\text{HA}}(\mathbf{q}, \mathbf{q}_0) \\ U_{\text{HA}}(\mathbf{q}, \mathbf{q}_0) &= \frac{1}{2} \sum_{i,j=1}^{3N_s} H_{ij} (q_i - q_{i,0}) (q_j - q_{j,0}) \end{aligned} \tag{3.65}$$

where the Hessian matrix of the potential energy at the minimum defines the force constants $H_{ij} = \frac{\partial^2 U(\mathbf{q})}{\partial q_i \partial q_j}|_{\mathbf{q}=\mathbf{q}_0}$. Moreover, the force field deriving from the HA potential has a computational complexity of $\mathcal{O}(N_s^2)$, as can be inferred from the double summation appearing in Eq. (3.65). However, the quadratic complexity can be turned linear since the atomic interactions are ignorable beyond some cutoff radius R_{cut} and therefore the force constants H_{ij} are zero above $2R_{\text{cut}}$ interatomic distance, i.e., the sum in Eq. 3.65 can be reorganized as:

$$U_{\text{HA}}(\mathbf{q}, \mathbf{q}_0) = \frac{1}{2} \sum_{i=1}^{3N_s} \sum_{j \in v(i)} H_{ij} (q_i - q_{i,0}) (q_j - q_{j,0}) \ , \tag{3.66}$$

for which $v(i)$ is the collection of neighbour atoms of the i^{th} atom within $2R_{\text{cut}}$ distance. These summations have $\mathcal{O}(N_s n)$ computational complexity, where n is the average number of neighbour atoms within $v(i)$ (in general for large systems $n \ll N_s$).

In the present case the HA development is performed around a minimum $\mathbf{q}_0$ of the potential energy, which means that HA contribution in Eq. (3.65) is a positive semi-definite quadratic form and so matrix $\mathbf{H} = [H_{ij}]$ is symmetric and non-negative. Defining the reference system based on an approximation accounting for all vibration frequencies of the physical system provides an important advantage: the degree of overlap between the reference and target distribution is expected to be high. Besides, the free energy difference between the target and reference systems should be relatively small, at least at reasonable temperatures lower than one third of the melting point.

The partition function, from which the harmonic free energy is deduced, is obtained by writing the Hamiltonian in normal coordinates and solving the decoupled equations of motion. The normal modes correspond to the eigenvectors of the dynamical matrix $D_{ij} = H_{ij}/\sqrt{m_i m_j}$. Note that the dynamical matrix has only real values in the present study because we treat only the Γ-HA modes, i.e., those in the center of the Brillouin zone of the crystal.

Being also symmetric non-negative, the matrix $\mathbf{D}$ is diagonalizable with real non-negative eigenvalues whose square roots define the frequencies of the normal modes. We sort the frequencies in descending order by convention: $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_{3N_s-3} > 0$. Here, frequencies are strictly positive except the three last ones that are exactly zero due to the periodic

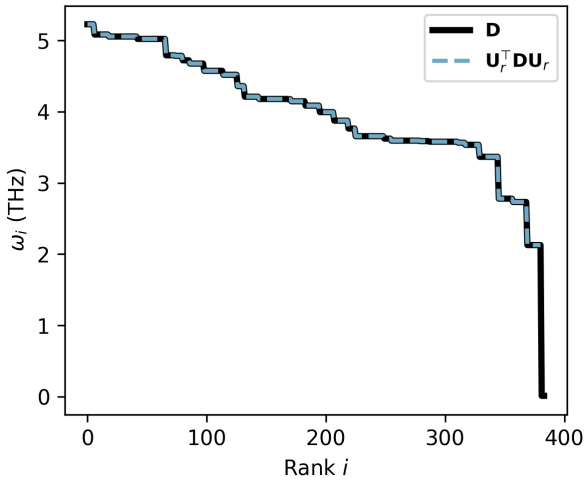

Figure 3.3: Phonon frequencies ω_i derived from the eigenvalues ω_i^2 of the dynamical matrix $\mathbf{D} \in \mathbb{R}^{384\times384}$ of 128 atoms in pure bcc W at equilibrium and the filtered matrix $\mathbf{U}_r^\top \mathbf{D}\mathbf{U}_r$, where $r = 381 = 3 \times 128 - 3$ and $\mathbf{U}_r \in \mathbb{R}^{384\times381}$ is the matrix of the first r left singular vectors of $\mathbf{D}$.

boundary conditions (PBC) that are applied to our atomistic simulations. The null-space normal modes characterize the three translation symmetries of the overall system in the three-dimensional physical space. When free boundaries are imposed on the system, the potential energy becomes rotationally invariant, resulting in two rotational symmetries and two additional null eigenvalues.

In the limit of high temperature, the phonons (normal modes) behave as independent and Boltzmann-distributed classical oscillators, so the free energy becomes:

$$F_{\mathrm{HA}} = \beta^{-1} \sum_{i=1}^{3N_s-3} \ln\left(\beta\hbar\omega_i\right). \tag{3.67}$$

Removal of numerical instabilities: SVD filter. It should be noted that employing directly the HA potential energy in Eq. (3.65) as the reference system may lead to unstable and non-converging dynamics in some cases, especially at saddle points separating local energy minima, for which negative eigenvalues of $\mathbf{D}$ exist.

The sampling instability in a physically stable state, like bulk tungsten (W), stems from the fact that the 3 translation eigenvalues are not exactly zero as they should be. These lowest eigenvalues correspond to the overall displacement of the system along 3 directions through the periodic boundaries.

In exact arithmetic, the null-space that reflects the translation symmetries does not affect the dynamics because the center of mass is shifted to its initial position at each step, which means that the displacement vector $\mathbf{q} - \mathbf{q}_0$ remains orthogonal to the null-space. However, in finite-precision arithmetic, rounding errors in the evaluated eigenvectors result in a non-zero projection of the displacement vector onto the computed null-space. As soon as a translation

eigenvalue becomes slightly negative, it generates forces that favor the atomic motion away from the unstable equilibrium position. Hence, excessive atomic displacements may occur whenever the dynamics is dominated by the harmonic potential.

To address this problem, we develop and test another numerical scheme, in which we force the translation eigenvalues of the dynamical matrix to be strictly equal to zero so that the gradient of the harmonic potential in Eq. (3.65) is null along translation symmetries and the dynamics is always orthogonal to this null-space. To achieve this goal, a numerical filter based on SVD is proposed. The SVD decomposition of the real dynamical matrix $\mathbf{D} = [D_{ij}] \in \mathbb{R}^{3N_s \times 3N_s}$ is given by:

$$\begin{aligned} \mathbf{D} &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \\ \mathbf{U} &= \left(U_1 \dots U_{3N_s}\right) \end{aligned} \tag{3.68}$$

with $U_i \in \mathbb{R}^{3N_s \times 1}$ are the $3N_s$ left singular vectors whilst $\mathbf{\Sigma}$ and $\mathbf{V}$ contain the singular values and the right singular vectors. We filter the spurious modes associated with the 3 lowest eigenvalues by projecting the constant matrix $\mathbf{D}$ into the subspace spanned by the first $3N_s - 3$ right singular vectors $\mathbf{U}_r = (U_1 \dots U_r) \in \mathbb{R}^{3N_s \times r}$ where $r = 3N_s - 3$. The projector ensuring this transformation is $\mathbf{U}_r\mathbf{U}_r^\top$. In this way, $\mathbf{H} = [D_{ij}\sqrt{m_i m_j}]$ in Eq. 3.65 is replaced by $\mathbf{H}^{\mathrm{SVD}} = [D_{ij}^{\mathrm{SVD}}\sqrt{m_i m_j}]$ where $\mathbf{D}^{\mathrm{SVD}} = \mathbf{U}_r\mathbf{U}_r^\top\mathbf{D}\mathbf{U}_r\mathbf{U}_r^\top$. The bottleneck here is the SVD decomposition whose computational complexity is cubic but which, fortunately, should be made only once, as a separate calculation before the sampling procedure begins. Note that evaluating the filtered potential energy has the same scalability as that for evaluating the non-filtered potential energy. The computational complexity is linear when the null-space components of the position vector are filtered out prior to calling the neighboring list defined in Eq. (3.66).

For the perfect bulk tungsten where no negative eigenvalues exist apart from the 3 translation eigenvalues, the applications of the SVD filter and the $\mathbf{P}$ projector in Eq. (3.52) are equivalent. However, SVD filtering is more general. Using the SVD-filtered potential is expected to improve the stability of mean force sampling by preventing the reference dynamics from diverging. It should therefore be useful in sampling unstable transition states.

Comparison between Einstein and harmonic approximations. Here we compare the convergence of the biasing force towards the mean force using the BABF method for which the reference system is based on the Einstein or harmonic approximations. The target physical system consists of W atoms described by an EAM potential [22]. To assess finite size effects, we first performed two preliminary simulations in systems containing 128 and 1024 atoms at 3 400 K. The variation of the lattice constant in BABF calculations with those systems is of the order of 10^{-4} Å. The associated difference in bulk modulus is 1 GPa (the relative error is 0.37%). Considering this error to be acceptable, we conclude that the small system size of 128 atoms does not limit the accuracy of the BABF calculations. Therefore, all simulations in the following part of the paper are carried out with 128 atoms, unless otherwise specified.

The behavior of the BABF method using the standard HA reference (denoted by HA) and the SVD-filtered harmonic reference (denoted by HA-SVD) is first briefly discussed. As previously stated in Subsection 3.2.2.1, using the projection operator $\mathbf{P}$ defined in Eq. 3.52, which shifts the center of mass from any position to its initial position at each dynamic step,

is equivalent to filtering the HA reference via SVD decomposition in this case (perfect bulk crystal). Hence, either of the two methods is able to stabilize the dynamics. This equivalence is demonstrated by the overlapping curves of recorded marginal probability $P_{A_n}(\zeta)$ at $n = 2\,000$ for HA and HA-SVD references displayed in Fig. 3.4(a). Since the SVD filter is a more general approach, the standard HA reference will not be investigated anymore in the following sections. We rather focus on the SVD-filtered harmonic reference, denoted by HA-SVD, and thus compare EA and HA-SVD references implemented in BABF simulations. The recorded marginal probability $P_{A_n}(\zeta)$ at $n = 2\,000$ for the two references are also displayed in Fig. 3.4(a), from which we observe a more uniform sampling when HA-SVD, rather than EA, is used as reference. We further assess the sampling performance by computing the Kullback–Leibler (KL) divergence of $P_{A_n}(\zeta)$ from the uniform distribution (see Eq. (3.60)).

As illustrated in Fig. 3.4(b) where the KL divergence is plotted as a function of n, the convergence level achieved after 80 000 Langevin steps using the Einstein reference necessitates three times more steps than when using the SVD-filtered harmonic reference instead. This trend is explained by the fact that the harmonic solid is more supple than the Einstein solid, as the harmonic Hamiltonian includes information from all possible pairwise interactions, while the Einstein formalism considers a set of independent but identical harmonic oscillators.

Additional profiles of the probability distributions $P_{A_n}(\zeta)$ are shown in Fig. 3.4(c-e) at increasing n values and using two types of weighting functions $w(n)$ in Eq. (3.54). The linear weighting function $w(n) = n/N$ where N denotes the total number of steps is implemented in addition to the constant function $w(n) = 1$ used previously. For both weighting functions, the distribution $P_{A_n}(\zeta)$ flattens faster with the filtered harmonic reference than with the Einstein reference, which indicates a better sampling performance of the former. In contrast, the use of the latter results in excessively high probability on both sides, especially around $\zeta = 1$. We further observe from the graphs (b-e) of Fig. 3.4 that convergence is significantly improved by the use of the linear weighting function. We next investigate the impact of the weighting function on the convergence behavior of the BABF method with HA-SVD reference.

3.2.2.2 Acceleration of convergence: reweighting sampling

As emphasized in Fig. 3.4, one should notice that the first several iterations of the BABF method give relatively biased sampling of ζ. Hence, a weighting function reevaluated at each step is set in Eq. 3.54 to reduce the contribution of the configurations initially sampled and then improve the sampling efficiency. In Fig. 3.4, results of the simplest linear weighting function $w(n) = n/N$ for $n = 1, \ldots, N$ are compared with the case without weighting function ($w(n) = 1$). Application of weighting results in significant improvement of sampling efficiency. To reach the convergence level $D_{\mathrm{KL}}(P_{A_n}||\mathbf{1}_{[0,1]}) < 10^{-3}$ with harmonic reference, 20 000 steps of Langevin dynamics are sufficient through linearly weighting the sampling, while 80 000 steps are still not enough without the use of weighting function. Consequently, it is of great interest to figure out the effects of the weighting function.

Several weighting functions varying from 0 to 1 are tested in order to measure the impact of the functional form. Figure 3.5 illustrates the results of employing the various weighting

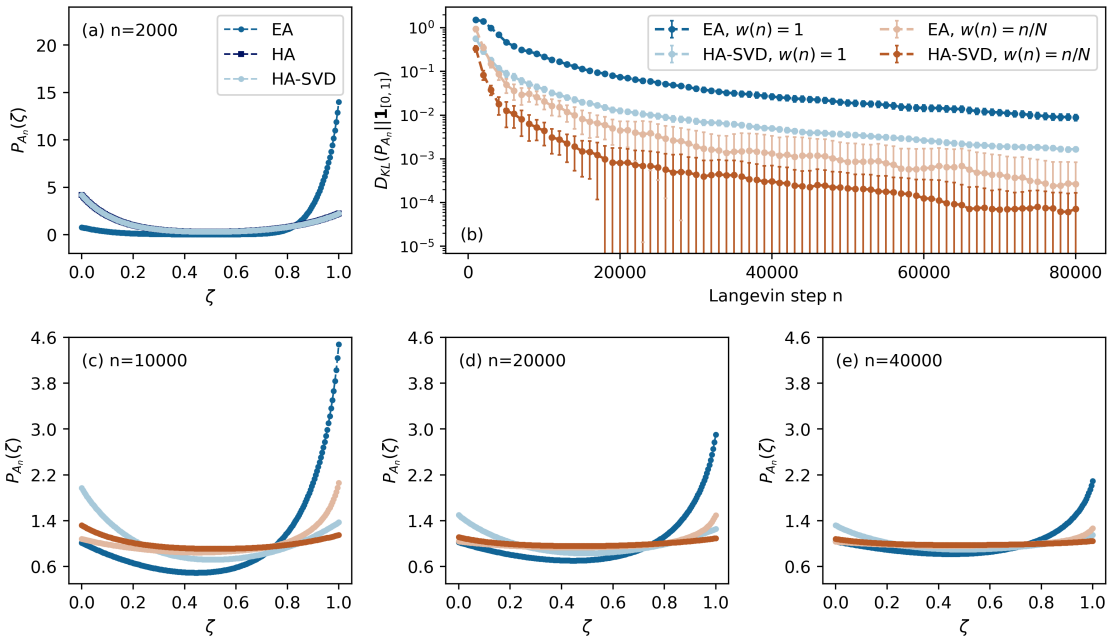

Figure 3.4: (a) Distribution of the estimated marginal probability $P_{A_n}(\zeta)$ at Langevin step $n = 2\,000$ of BABF calculations using references with Einstein approximation (EA), standard harmonic approximation (HA) and SVD-filtered harmonic approximation (HA-SVD). (b) Variation of KL divergence for the BABF calculations with EA reference and HA-SVD reference. (c-e) Distribution of $P_{A_n}(\zeta)$ at Langevin step (c) $n = 10\,000$, (d) $n = 20\,000$, and (e) $n = 40\,000$ of the calculations shown in (b). No weighting function ($w(n) = 1$) is applied in (a), while two cases of weighting function ($w(n) = 1$ and $w(n) = n/N$) are considered in (b-e), where N is the total number of Langevin steps. Legends for (c-d) are the same as that in (b). All the calculations in this figure are performed in a bcc W system of 128 atoms at 3 400 K using WEAM4 potential [22].

functions that have been selected. The different shapes of these functions are illustrated in Fig. 3.5(a). To evaluate the importance of the first n samples, we display the following characteristic function $w(n)/\sum_{s=1}^{n} w(s)$ in Fig. 3.5(b), while the corresponding KL divergence is shown in Fig. 3.5(c). The weighting function with larger value of $w(n)/\sum_{s=1}^{n} w(s)$ at $n \ll N$ leads to faster convergence (Fig. 3.5(b-c)) since it lowers the contribution of the initial sampling. From this analysis it can be concluded that an appropriate choice of weighting function, such as $w(n) = \left[\sin\left(\frac{n\pi}{2N} - \frac{\pi}{2}\right) + 1\right]\left(\frac{n}{N}\right)^2$, can further improve the sampling significantly in terms of convergence speed: it requires only 6 000 steps to attain $D_{KL}(P_{A_n}||\mathbf{1}_{[0,1]}) < 10^{-3}$, which is three times faster than using the linear weighting function.

3.2.3 Validation: comparison with MD simulations and TI

In this subsection we compare the performance of the present BABF method with classical molecular dynamics (MD) simulations. The direct comparison on free energy calculations is out of scope. It is well known that direct MD simulations are unable to estimate the free energy of the system. Here, we instead perform an indirect comparison on the second derivative of

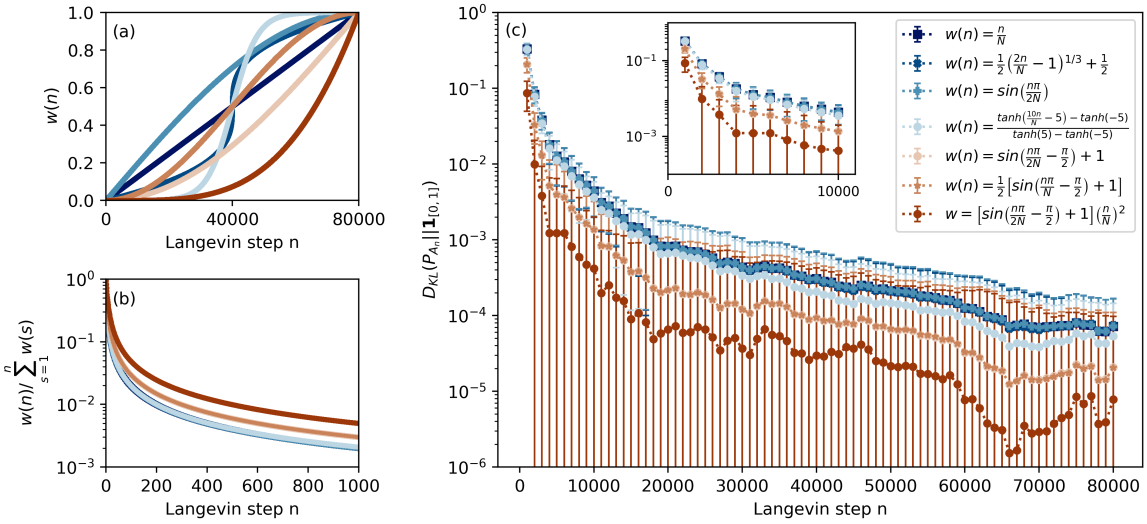

Figure 3.5: (a) Shape of different weighting functions. (b) The value of $w(n)/\sum_{s=1}^{n} w(s)$, which determines the convergence speed. (c) Variation of KL divergence for different weighting functions. Legends for (a) and (b) are the same as that in (c). All the results in this figure are obtained by BABF calculations with HA-SVD reference based on the same bcc W system and the same EAM potential as in Fig. 3.4.

the free energy with respect to the volume of the system, i.e., the isothermal bulk modulus B^T.

The previous BABF method with both EA and HA-SVD references is used to compute the free energy at 20 temperatures ranging from 10 K to 3 800 K. At each temperature, a series of deformed systems is considered, allowing to extract the bulk modulus from the second derivative of the free-energy versus volume curve. Details of the procedure are illustrated in Fig. 3.6. The bulk moduli from MD-based simulations are obtained by measuring the change in average stress tensor when the cell volume undergoes a finite deformation [226] with the ELASTIC_T package available in LAMMPS [227]. The high temperatures 3 000 K, 3 400 K, 3 600 K and 3 800 K, for which the thermal fluctuations are large, have been selected. The bulk modulus computations using both methods are made within a bcc W system modeled with WEAM4 potential [22]. A system containing 128 atoms is used in BABF calculations, while 16 000 atoms are needed in MD simulations to avoid excessive fluctuations.

The relative error is recorded along the BABF and MD simulations for the system at 3 400 K and plotted in Fig. 3.7(a). We notice that the BABF calculation with EA reference exhibits a significantly larger initial relative error than that obtained with HA-SVD reference. BABF calculations with both references converge quickly. Specifically, the relative error of BABF calculation decreases to less than 1% from step 8 000 with EA reference and from step 2 000 using HA-SVD reference. In contrast, MD simulation starts with a relatively slight error but always suffers severe fluctuations so that a large number of MD steps are required to obtain a reliable time average. In the current case, 130 000 MD steps are necessary. Thus, the much smaller number of integration steps (here, 20 000 for BABF+HA-SVD and 130 000 for

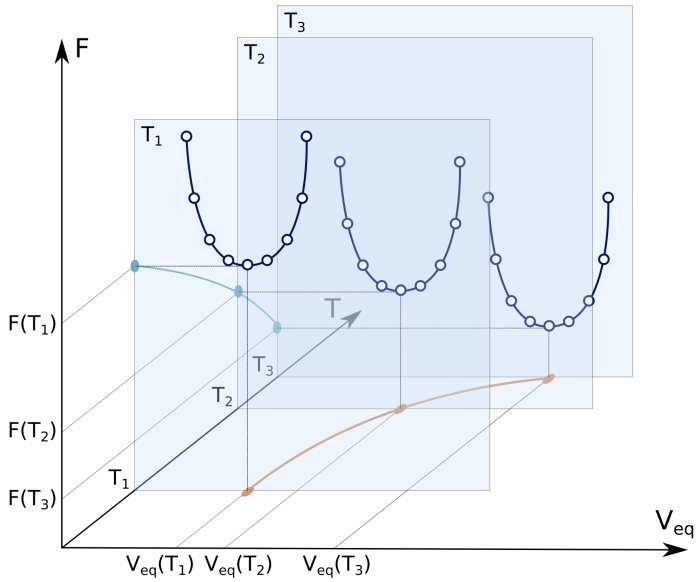

Figure 3.6: Illustration of thermodynamic properties calculation with the BABF method, where V_{eq} and F denote the equilibrium volume and the Helmholtz free energy, respectively. At temperature T, we compute the free energy for a series of systems with different volumes as presented by the dark blue curves. The value of the lattice constant corresponding to V_{eq}, i.e., the free energy minimum, is the lattice constant at this temperature, from which the volumetric thermal expansion can be obtained (dark orange curve). Bulk modulus is proportional to the second derivative of the free energy with respect to the volumetric strain. Temperature dependence of the free energy minimum (light blue curve) allows the entropy evaluation.

MD), combined with the smaller system size (128 atoms for BABF+HA-SVD and 16 000 for MD), results in a hundredfold increase in computational efficiency for the BABF approach.

We also provide the comparison between the efficiency of the present BABF method with the standard TI from a very recent publication of Jung *et al.* [44]. In Section E of the Supplementary Information of Ref. [44], Jung *et al.* declare that the standard TI is run for 50 000 steps to obtain statistically very well converged average energies for a given coupling parameter λ (denoted as ζ in our work), and at least 20 λ values are needed to perform a λ-based TI, resulting in a total of $20 \times 50\,000 = 1\,000\,000$ steps. As presented in Supplementary Fig. 4 of Ref. [44], 50 000 steps of MD simulation at a given λ allow the TI-estimated free energy to converge within 0.1 meV/atom. Here we also show a convergence test of time-steps in Fig. 3.8 for the BABF method. The test is performed with HA-SVD reference, in a bcc W system of 128 atoms at 3 400 K using WEAM4 potential [229]. Comparison between the BABF method and the standard TI is detailed as follow:

- As revealed in Fig. 3.8(b), the BABF method also requires 50 000 steps to achieve the accuracy of 0.1 meV/atom. However, with the Bayes formula, we do not need to perform the dynamics for each descritized coupling parameter. Therefore, 50 000 steps are enough using BABF calculation instead of 1 000 000 steps with standard TI.
- One should notice that the standard TI calculations shown in Supplementary Fig. 4 of

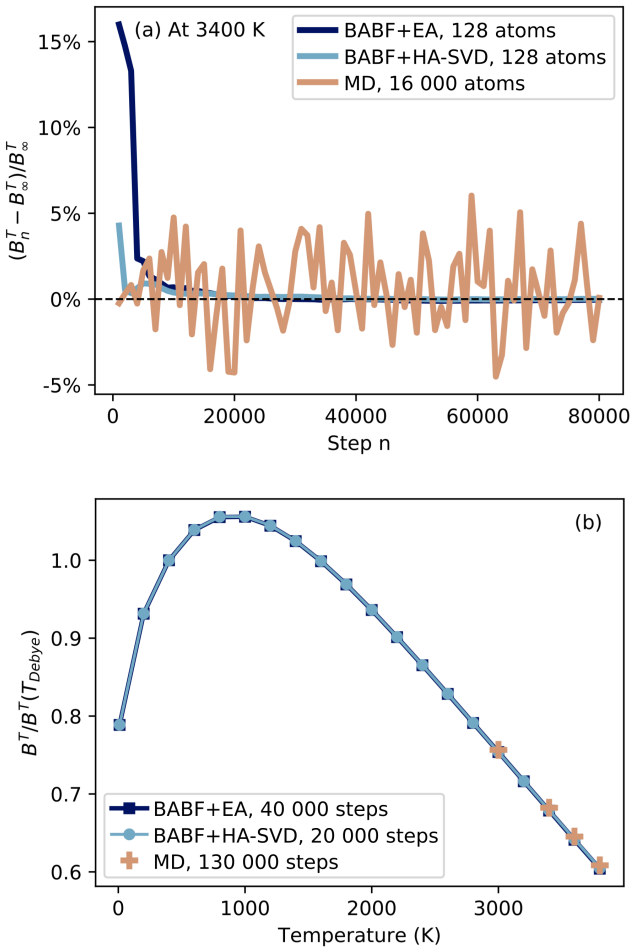

Figure 3.7: Comparison of BABF method and MD simulation. (a) Runtime relative error $(B_n^T - B_\infty^T)/B_\infty^T$ at 3 400 K with B_∞^T the fully converged value. (b) Rescaled bulk modulus as function of temperature. We take the ratio of the bulk modulus at temperature T to its value at the Debye temperature T_{Debye}, equal to 400 K for W [228]. The dark blue and light blue lines are computed with BABF method without weighting function using EA and HA-SVD references, respectively. The dark orange line is obtained from MD simulations. The number of atoms in the bcc W system and the number of iteration steps to achieve the convergence for each method are indicated in the legends. All the curves are obtained with WEAM4 potential [22].

Ref. [44] are performed in a Mg system containing 490 atoms (Supplementary Table 11), while our BABF calculation in Fig. 3.8 is conducted in a W system of 128 atoms. Hence, the BABF test is actually based on a smaller system compared with the standard TI.

- The exact temperature of the standard TI calculations shown in Supplementary Fig. 4 of Ref. [44] is not given. However, from Supplementary Table 11, we can see that the highest temperature for the calculations of Mg is 989 K. Our BABF calculation in Fig. 3.8 is performed in a system at 3 400 K, which is much higher than the simulation temperature of the standard TI.

Therefore, even at higher temperatures and using a system that is almost four times smaller,

the BABF method achieves a 20-fold increase in computational efficiency compared to the standard TI. Figure 3.8(a) reveals that with the BABF method, 30 000 Langevin steps are enough to attain a relative error of free energy less than 0.1% at high temperature (3 400 K). Moreover, the BABF method is more stable than standard TI and FEP. With the SVD-filtered harmonic reference, the BABF method can even be applied in the presence of free energy barriers and can be useful in, for example, the study of interstitial atoms.

The results of BABF with EA and HA-SVD references are consistent with those obtained from MD simulations (Fig. 3.7(b)). The obtained agreement validates the further application of the optimized BABF method to the calculations of elastic properties and thermal expansion. From the cost analysis, we conclude that the present method is fast and robust for estimating the free energy in crystalline solids and can be applied to sample not only fast standard force fields but also numerically heavy machine learning potentials [230, 168, 103, 231].

To summarise, the BABF approach is validated by the calculations of thermodynamic properties of W, in comparison to the traditional MD simulations at temperatures above 3 000 K. Its accuracy is also proved by comparison with experimental results at temperatures below 2 100 K, which will be shown in the next chapter (Fig. 4.3). With sufficient efficiency to apply the ML potentials, this accelerated BABF method makes the fast and accurate investigation of free energy at extremely high temperatures feasible. In Chapter 4, we will demonstrate the applications of this method in predicting the high-temperature thermodynamic properties of tungsten and Ta-Ti-V-W high-entropy alloys.

3.3 Bound BABF method: sampling a metastable state

In this section, we put forward an extension of the previously presented BABF method, which makes it possible to sample a given metastable state. This technique facilitates the study of the free energy landscape of defects, thereby providing insights into the mechanisms governing defect behaviors. In Subsection 3.3.1, we present the rationale for developing this enhanced methodology, exemplifying a scenario in which the existing BABF approach is insufficient to effectively address the problem at hand. Next, we outline the challenges encountered when applying the BABF method to this problem, along with our proposed approach to address these issues. The formulation, implementation, and validation of our solution are elaborated in Subsection 3.3.2. To further accelerate the sampling process, we employ a parallelization scheme detailed in Subsection 3.3.3.

3.3.1 Motivation: sampling free energy landscape of defects

Point defects, such as vacancies, have a significant influence on diffusion and microstructural evolution in solids. Understanding these defect-driven processes becomes particularly challenging at high temperatures, as they depend crucially on complex free energy surfaces. This task is quite difficult, even with DFT calculations. For instance, in tungsten, vacancy

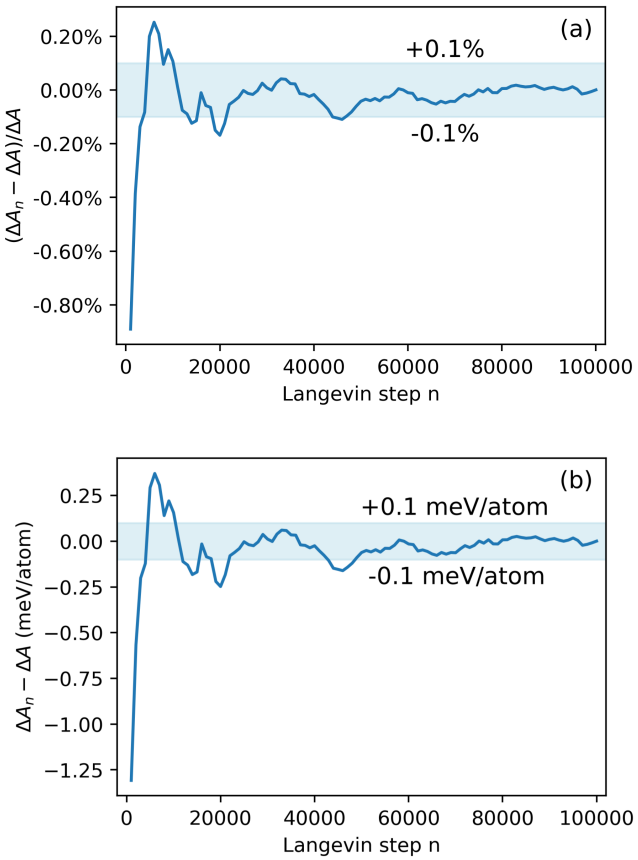

Figure 3.8: (a) Relative error of $\Delta A = A(1) - A(0)$ and (b) convergence of ΔA. ΔA_n is the free energy difference estimated at Langevin step n and ΔA is the well converged free energy difference. The BABF calculation is performed with HA-SVD reference, in a bcc W system of 128 atoms at 3 400 K using WEAM4 potential [229].

loops and voids have been experimentally observed during annealing treatments following heating [232]. However, previous DFT calculations cannot explain the existence of vacancy clusters. DFT results indicate that di-vacancies are strongly unstable in second nearest-neighbor (2NN) configurations, and either unstable or stable in first nearest-neighbor (1NN) configurations, depending on the parameters of the calculations [233, 234, 235, 236, 237, 229, 238, 239, 240]. Nevertheless, we assert that the previous calculations were not sufficiently converged. The Fermi surface of W is very complicated and a large mesh of k-points is required. The binding energies calculated by DFT for the di-vacancies are plotted in Fig. 3.9 as a function of the number of k-points used. For a cubic cell of $(128 - 1)$ atoms, i.e., $4a_0 \times 4a_0 \times 4a_0$, where a_0 is the equilibrium bcc W lattice parameter, we find that the calculations tend to converge for k-points grid larger than $6 \times 6 \times 6$ with an electronic smearing parameter σ set to 0.3 eV. For the smaller σ-value of 0.1 eV, the DFT-calculated energies are unreliable, even for the 6^3 mesh. Therefore, previous calculations [237, 235, 233, 229, 239, 240] probably did not reach convergence due to their choice of the number of k-points, resulting in parameter-dependent trends that are not supported by experimental evidence.

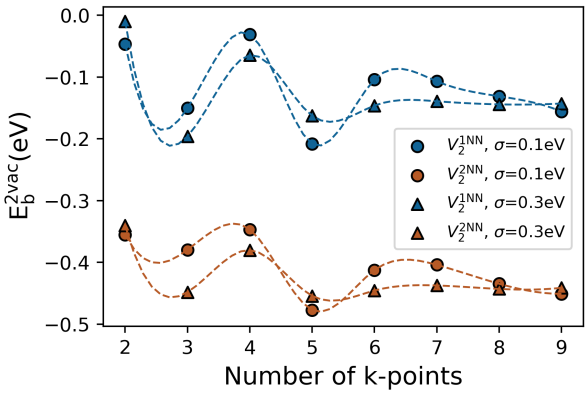

Figure 3.9: DFT-based binding energy for 1NN and 2NN di-vacancies at smearing parameter σ of 0.1 eV and 0.3 eV, as a function of the number of k-points. (The calculations are conducted using VASP 6.2.0 [158] package with AM05 [241, 242, 243] exchange-correlation functional. Methfessel and Paxton smearing method is utilized to approximate the Fermi-Dirac statistics [244].

We find that 1NN and 2NN di-vacancies in W are unstable and repulsive at 0 K for three exchange-correlation functionals, namely the Perdew-Zunger functional [245, 246] from the local-density approximation (LDA), the Perdew-Burke-Ernzerhof (PBE) functional [247, 248, 249] from the generalized gradient approximation (GGA) and the Armiento-Mattson (AM) functional [241, 242, 243] also from the GGA. Merely extrapolating the energy landscape constructed from electronic structure calculations at 0 K to high temperatures suggests that there should be no vacancy clustering in W, in contradiction with the experimental observations. Any mechanistic model describing the formation of voids must necessarily take into account the instability of 1NN and 2NN di-vacancies and concomitantly involve the agglomeration of vacancies into intermediate nano-clusters as experimentally observed at high temperatures.

To understand this phenomenon, we need to quantify the temperature effects on the formation and binding free energy landscape of small vacancy-type defects. Free energy sampling using the BABF method presented in Section 3.2 seems to be a good choice. However, sampling a specified defect configuration based on Langevin dynamics is challenging: it should be able to sample a given metastable state without allowing the system to transition from one free energy basin to another. For this reason, we develop the bound BABF method, which enables constrained exploration of the free energy landscape without bias.

3.3.2 Constrained configuration space exploration without bias

The anharmonic vibrational contribution plays an important role in the free energy landscape of a configuration. The free energy of a system containing N_s atoms at its equilibrium volume and temperature T includes three components:

$$F(N_s, T) = F_{\text{vib}}(N_s, T) + F_{\text{con}}(T) + E_0(N_s, T)\,. \tag{3.69}$$

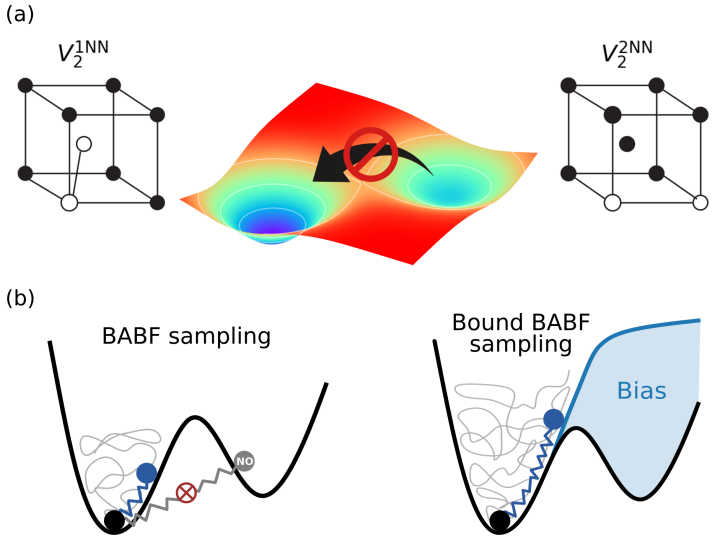

Figure 3.10: (a) Difficulty in sampling a given vacancy configuration: vacancy migration corresponding to the transition between two metastable basins, such as 1NN and 2NN di-vacancy configurations in bcc lattice (the vacancies are indicated by empty circles). (b) Schematic illustration of the original BABF sampling and the bound BABF sampling. Compared to the original approach, bound BABF sampling inhibits unfavorable transitions between free energy basins by artificially increasing the energy barrier. This additional potential energy confines sampling to a single metastable state, but introduces a bias in the free energy estimation.

Among them, the configurational free energy F_{con} depends on the number of equivalent configurations for a specified defect type, energy minimization in LAMMPS [227] gives the potential energy at equilibrium E_0, and the vibrational term F_{vib} is the most difficult to evaluate.

To compute the vibrational free energy, the TI is performed using a general potential energy $U(\zeta, \mathbf{q})$ that linearly mixes the potential energy of the target system $U(\mathbf{q})$ (e.g., ML potentials), and the potential energy of the reference system $U_{\text{HA}}(\mathbf{q})$ (here we choose the SVD-filtered harmonic approximation as recommended in Subsection 3.2.2.1), through a coupling parameter ζ: $U(\zeta, \mathbf{q}) = (1-\zeta)U^{HA}(\mathbf{q}) + \zeta U(\mathbf{q})$. The associated Landau free energy is defined as $A(\zeta) = -\beta^{-1} \ln Z(N_s, V, T, \zeta) + B$ where $Z(N_s, V, T, \zeta)$ is the partition function for the canonical ensemble. Then the target vibrational free energy F_{vib} can be calculated from the free energy difference between the target system and the harmonic reference system $\Delta A = A(1) - A(0)$ as $F_{\text{vib}} = F_{\text{vib}}^{\text{HA}} + \Delta A$. As mentioned previously, in harmonic approximation, the phonons are regarded as independent and Boltzmann-distributed classical oscillators with frequencies $\omega_i, i = 1, \ldots, N_m$, that are the eigenvalues of the Hessian matrix. For a system containing N_s atoms at metastable states corresponding to a local energy minimum, the number of non-zero normal modes $N_m = 3N_s - 3$, so the free energy becomes:

$$F_{\text{vib}}^{\text{HA}} = \beta^{-1} \sum_{i=1}^{3N_s-3} \ln\left[1 - \exp\left(-\beta\hbar\omega_i\right)\right]. \tag{3.70}$$

Based on the BABF calculation, we can efficiently obtain the free energy through the sampling driven by Langevin dynamics. However, the thermodynamic micro-state associated

with a vacancy configuration corresponds to a metastable basin of attraction. The difficulty in estimating its free energy through Langevin-based sampling is that the dynamical system easily evolves towards other metastable basins by crossing over surrounding energy barriers beyond a certain temperature, as shown in Fig. 3.10(a). This entails either the migration of the mono-vacancy or the interconversion of 1NN and 2NN di-vacancies.

To prevent the system from escaping its thermodynamic micro-state, we constrain each atom neighbouring the vacancy defect, i.e., we push the atom back to its initial basin by applying the restoring force

$$\mathbb{F}^B(\mathbf{r}) = -\frac{\mathbf{r}}{\|\mathbf{r}\|}\frac{C}{\delta}\phi\left(\frac{\|\mathbf{r}\| - R_c - \delta}{\delta}\right) \tag{3.71}$$

on the atomic displacement vector $\mathbf{r}$ with $\|\mathbf{r}\|$ denoting its Euclidean norm. The non-negative real function ϕ is even:

$$\phi(x) = \max\left[0, (1+\cosh x)^{-1} - (1+\cosh 1)^{-1}\right]. \tag{3.72}$$

Expressing the restoring force as the negative gradient of a potential energy $\mathbb{F}^B(\mathbf{r}) = -\nabla_{\mathbf{r}} E^B(\mathbf{q})$, the additional biasing potential E^B acts as an additional energy barrier further trapping the system inside its basin, as illustrated in Fig. 3.10(b). In this case, the general potential energy becomes

$$U^B(\zeta, \mathbf{q}) = (1-\zeta)U^{HA}(\mathbf{q}) + \zeta\left[U(\mathbf{q}) + E^B(\mathbf{q})\right]. \tag{3.73}$$

Constrained potential $U^B(\zeta, \mathbf{q})$ allows the system to stay confined in the original basin of the sampling. Nevertheless, propagating the Langevin dynamics with the extra potential energy $E^B(\mathbf{q})$ results in a sampling bias that must be removed from the free energy estimator. In the next paragraph, we show that it is possible to cancel the influence of the supplemental potential by proposing an unbiased procedure to estimate the free energy in a given basin.

Based on mean force estimation formula given in Eq. 3.48, we can estimate $A^{B\prime}(\zeta)$ in the constrained extended ensemble:

$$A^{B\prime}(\zeta) = \frac{\int_{\mathbb{T}^{3N_s}} \partial_\zeta U^B(\zeta, \mathbf{q}) \pi_{A^B_\star}(\zeta|\mathbf{q}) \Pi_{A^B_\star}(\mathbf{q}) d\mathbf{q}}{\int_{\mathbb{T}^{3N_s}} \pi_{A^B_\star}(\zeta|\mathbf{q}) \Pi_{A^B_\star}(\mathbf{q}) d\mathbf{q}}, \tag{3.74}$$

where $\Pi_{A^B_\star}(\mathbf{q})$ is the marginal probability associated to $\mathbf{q}$ in the constrained energy landscape. This probability can be easily sampled thanks to ergodic theorem. This expression involves also the following conditional probability:

$$\pi_{A^B_\star}(\zeta|\mathbf{q}) = \frac{\exp\left[\beta(A^B_\star(\zeta) - U^B(\zeta, \mathbf{q}))\right]}{\int_0^1 \exp\left[\beta(A^B_\star(\tilde{\zeta}) - U^B(\tilde{\zeta}, \mathbf{q}))\right] d\tilde{\zeta}}, \tag{3.75}$$

Mean force estimation proposes in Eq. 3.74 can be computed by using standard BABF method. However, this mean force is biased by the constrained potential.

We are interested in the estimation of the unbiased mean force given in equation (3.48). This estimation can be rewritten as an expectation on $\Pi_{A^B_\star}(\mathbf{q})$ distribution sampled in the

constrained extended ensemble:

$$\begin{aligned} A'(\zeta) &= \frac{\int_{\mathbb{T}^{3N_s}} \partial_\zeta U(\zeta,\mathbf{q}) P_{A^B_\star}(\zeta,\mathbf{q}) d\mathbf{q}}{\int_{\mathbb{T}^{3N_s}} P_{A^B_\star}(\zeta,\mathbf{q}) d\mathbf{q}} \\ &= \frac{\int_{\mathbb{T}^{3N_s}} \partial_\zeta U(\zeta,\mathbf{q}) \frac{P_{A^B_\star}(\zeta,\mathbf{q})}{\Pi_{A^B_\star}(\mathbf{q})} \Pi_{A^B_\star}(\mathbf{q}) d\mathbf{q}}{\int_{\mathbb{T}^{3N_s}} \frac{P_{A^B_\star}(\zeta,\mathbf{q})}{\Pi_{A^B_\star}(\mathbf{q})} \Pi_{A^B_\star}(\mathbf{q}) d\mathbf{q}}, \end{aligned} \tag{3.76}$$

where $P_{A^B_\star}(\zeta,\mathbf{q}) \propto \exp\left[\beta(A^B_\star(\zeta) - U(\zeta,\mathbf{q}))\right]$ is the joint probability associated to the unbiased system in the constrained extended ensemble. Then, in order to simplify the notation, we chose to denote by $p_{A^B_\star}(\zeta|\mathbf{q}) = \frac{P_{A^B_\star}(\zeta,\mathbf{q})}{\Pi_{A^B_\star}(\mathbf{q})}$ the conditional unbiased probability of ζ for constrained bias $\mathbf{q}$ distribution. This conditional probability can be estimated in the extended constrained ensemble and lead to the present formulation:

$$\begin{aligned} p_{A^B_\star}(\zeta|\mathbf{q}) &= \frac{P_{A^B_\star}(\zeta,\mathbf{q})}{\Pi_{A^B_\star}(\mathbf{q})} \\ &= \frac{\exp\left[\beta(A^B_\star(\zeta) - U(\zeta,\mathbf{q}))\right]}{\int_0^1 \exp\left[\beta(A^B_\star(\tilde{\zeta}) - U^B(\tilde{\zeta},\mathbf{q}))\right] d\tilde{\zeta}}. \end{aligned} \tag{3.77}$$

The new two-estimations scheme for free energy calculations allows to sample a target metastable state in the extended constrained biasing potential $U^B(\zeta,\mathbf{x}) - A^B_\star(\zeta)$. An estimation of the constrained biased free energy based on mean force integration of Eq. 3.74 allows to calculate an unbiased mean force based on the importance sampling formula given in Eq. 3.76. The thermodynamic integration of this unbiased mean force allows us to recover the unbiased free energy difference ΔA associated with the metastable basin.

The bound BABF procedure is detailed step by step below:

1. The mean forces are estimated as:

$$A'_n(\zeta) = \frac{\hat{\mathrm{E}}_n[p_{A^B}(\zeta|\mathbf{q})\partial_\zeta U(\zeta,\mathbf{q})]}{\hat{\mathrm{E}}_n[p_{A^B}(\zeta|\mathbf{q})]} = \frac{\sum_{s=1}^{n-1} \partial_\zeta U(\zeta,\mathbf{q}_s) p_{A^B_s}(\zeta|\mathbf{q}_s) w(s)}{\sum_{s=1}^{n-1} p_{A^B_s}(\zeta|\mathbf{q}_s) w(s)}, \tag{3.78}$$

$$A^{B\prime}{}_n(\zeta) = \frac{\hat{\mathrm{E}}_n[\pi_{A^B}(\zeta|\mathbf{q})\partial_\zeta U^B(\zeta,\mathbf{q})]}{\hat{\mathrm{E}}_n[\pi_{A^B}(\zeta|\mathbf{q})]} = \frac{\sum_{s=1}^{n-1} \partial_\zeta U^B(\zeta,\mathbf{q}_s) \pi_{A^B_s}(\zeta|\mathbf{q}_s) w(s)}{\sum_{s=1}^{n-1} \pi_{A^B_s}(\zeta|\mathbf{q}_s) w(s)}, \tag{3.79}$$

   with $\partial_\zeta U(\zeta,\mathbf{q}_s) = U(\mathbf{q}_s) - U^{HA}(\mathbf{q}_s)$ and $\partial_\zeta U^B(\zeta,\mathbf{q}_s) = U(\mathbf{q}_s) + U^B(\mathbf{q}_s) - U^{HA}(\mathbf{q}_s)$.

2. The free energies $A_n(\zeta)$ and $A^B_n(\zeta)$ for the step n are determined as:

$$A_n(\zeta) = \int_0^\zeta A'_n(\tilde{\zeta}) d\tilde{\zeta} + A_n(0), \tag{3.80}$$

$$A^B_n(\zeta) = \int_0^\zeta A^{B\prime}{}_n(\tilde{\zeta}) d\tilde{\zeta} + A^B_n(0). \tag{3.81}$$

3. The corresponding conditional probabilities are calculated as:

$$p_{A_n^B}(\zeta|\mathbf{q}_n) = \frac{\exp\left[\beta(A_n^B(\zeta) - U(\zeta,\mathbf{q}_n))\right]}{\int_0^1 \exp\left[\beta(A_n^B(\zeta) - U^B(\zeta,\mathbf{q}_n))\right]d\tilde{\zeta}}, \tag{3.82}$$

$$\pi_{A_n^B}(\zeta|\mathbf{q}_n) = \frac{\exp\left[\beta(A_n^B(\zeta) - U^B(\zeta,\mathbf{q}_n))\right]}{\int_0^1 \exp\left[\beta(A_n^B(\zeta) - U^B(\zeta,\mathbf{q}_n))\right]d\tilde{\zeta}}. \tag{3.83}$$

4. The effective force field is obtained by the equation:

$$\mathbb{F}_{A_n}(\mathbf{q}_n) = -\int_0^1 \nabla_{\mathbf{q}} U^B(\zeta,\mathbf{q}_n)\pi_{A_n^B}(\zeta|\mathbf{q}_n)d\zeta. \tag{3.84}$$

5. Integration of the dynamic equation is carried out to obtain:

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \mathbf{P}\mathbb{F}_{A_n}(\mathbf{q}_n)\delta t + \sqrt{2\beta^{-1}\delta t}B_n. \tag{3.85}$$

We show in Fig. 3.11 a simple test performed in a configuration containing one vacancy (thus 127 atoms) at 3 000 K with the ML potential (PBE functional), demonstrating in the sub-figure (a) that the bound BABF method achieves the accuracy of ± 0.1 meV/atom within 5×10^4 Langevin steps, and 2×10^4 Langevin steps are enough to attain a relative error of free energy less than $\pm 1\%$. The bias removal procedure is validated by the same test, where we prove that by using the present correction framework, the bias can be removed within a relative error $\frac{|\Delta A - \Delta A^B|}{\Delta A}$ of $1.2 \times 10^{-4}\%$ for 6×10^4 steps (Fig. 3.11(b)). Obviously, the force $\mathbb{F}^B$ is applied once (at about step 12 000) in the whole simulation.

3.3.3 Parallelization based on shared bias

To reach the ergodic limit, sampling can be accelerated by propagating several replicas of the system simultaneously [250, 144]. P replicas of the system are created and each of them is allocated to a distinct group of processors. Those replicas are simulated with independent Wiener processes on a parallel computer architecture, while the evaluation of the mean force is shared by all replicas: Eq. 3.74 and Eq. 3.76 are approximated by

$$A^{B\prime}(\zeta) = \frac{\sum_{p=1}^P \int_{\mathbb{T}^{3N_s}} \partial_\zeta U^B(\zeta,\mathbf{q}^p)\pi_{A_\star^B}(\zeta|\mathbf{q}^p)\Pi_{A_\star^B}(\mathbf{q}^p)d\mathbf{q}^p}{\sum_{p=1}^P \int_{\mathbb{T}^{3N_s}} \pi_{A_\star^B}(\zeta|\mathbf{q}^p)\Pi_{A_\star^B}(\mathbf{q}^p)d\mathbf{q}^p}, \tag{3.86}$$

$$A'(\zeta) = \frac{\sum_{p=1}^P \int_{\mathbb{T}^{3N_s}} \partial_\zeta U(\zeta,\mathbf{q}^p)p_{A_\star^B}(\zeta|\mathbf{q}^p)\Pi_{A_\star^B}(\mathbf{q}^p)d\mathbf{q}^p}{\sum_{p=1}^P \int_{\mathbb{T}^{3N_s}} p_{A_\star^B}(\zeta|\mathbf{q}^p)\Pi_{A_\star^B}(\mathbf{q}^p)d\mathbf{q}^p}. \tag{3.87}$$

In this study, we often adopted this approach by using 32 replicas, with each replica computed on 8 processors when employing the ML potentials.

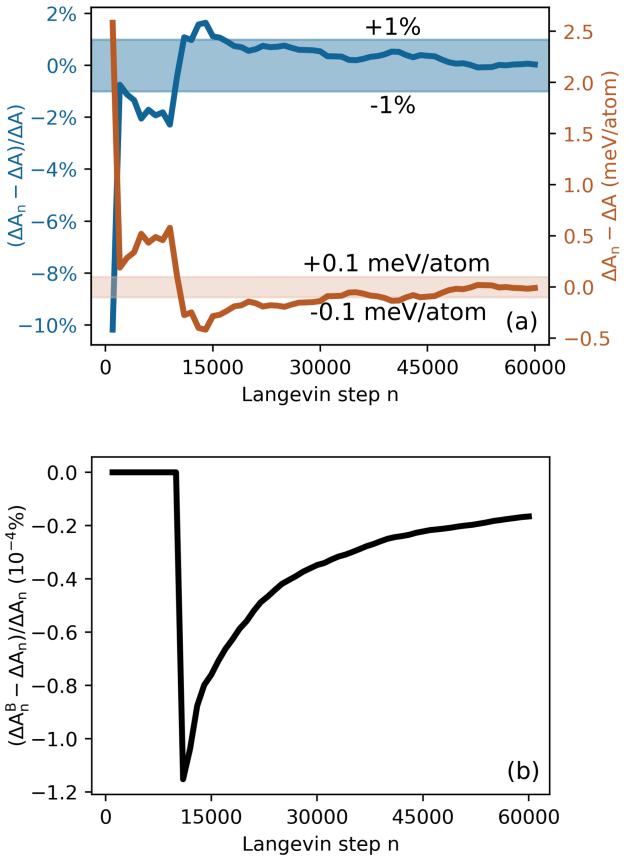

Figure 3.11: (a) Runtime relative errors (blue) and absolute errors (dark orange) during the BABF free energy calculation, where ΔA_n is the free energy difference estimated at step n and ΔA is the converged value. (b) Runtime relative errors of free energy difference between the biased value ΔA_n^B and the rescaled (debiased) value ΔA_n, when the force $\mathbb{F}^B$ is applied. All the errors are evaluated in a configuration containing one vacancy at 3 000 K, using QNML potential with PBE functional.

To implement the numerical algorithm, the bound BABF procedure is discretized: the $(n+1)$th molecular dynamics move of replica p can be realized based on the first n steps of this replica. Combining with the subsection 3.3.2, the algorithm can be detailed step by step as follows:

1. The unbiased and biased mean forces are estimated as:

$$A'_n(\zeta) = \frac{\sum_{p=1}^{P}\sum_{s=1}^{n-1}\partial_\zeta U(\zeta,\mathbf{q}_s^p)p_{A_s^B}(\zeta|\mathbf{q}_s^p)}{\sum_{p=1}^{P}\sum_{s=1}^{n-1}p_{A_s^B}(\zeta|\mathbf{q}_s^p)}, \tag{3.88}$$

$$A^{B\prime}_n(\zeta) = \frac{\sum_{p=1}^{P}\sum_{s=1}^{n-1}\partial_\zeta U^B(\zeta,\mathbf{q}_s^p)\pi_{A_s^B}(\zeta|\mathbf{q}_s^p)}{\sum_{p=1}^{P}\sum_{s=1}^{n-1}\pi_{A_s^B}(\zeta|\mathbf{q}_s^p)}, \tag{3.89}$$

where $\partial_\zeta U(\zeta,\mathbf{q}_s^p) = U(\mathbf{q}_s^p) - U^{HA}(\mathbf{q}_s^p)$ and $\partial_\zeta U^B(\zeta,\mathbf{q}_s^p) = U(\mathbf{q}_s^p) + U^B(\mathbf{q}_s^p) - U^{HA}(\mathbf{q}_s^p)$.

2. The unbiased and biased free energies for the step n are evaluated as:

$$A_n(\zeta) = \int_0^\zeta A'_n(\tilde{\zeta})d\tilde{\zeta} + A_n(0), \tag{3.90}$$

$$A_n^B(\zeta) = \int_0^\zeta {A^B}'_n(\tilde{\zeta})d\tilde{\zeta} + A_n^B(0). \tag{3.91}$$

3. The corresponding conditional probabilities are calculated as:

$$p_{A_n^B}(\zeta|\mathbf{q}_n^p) = \frac{\exp\left[\beta(A_n^B(\zeta) - U(\zeta, \mathbf{q}_n^p))\right]}{\int_0^1 \exp\left[\beta(A_n^B(\zeta) - U^B(\zeta, \mathbf{q}_n^p))\right]d\tilde{\zeta}}, \tag{3.92}$$

$$\pi_{A_n^B}(\zeta|\mathbf{q}_n^p) = \frac{\exp\left[\beta(A_n^B(\zeta) - U^B(\zeta, \mathbf{q}_n^p))\right]}{\int_0^1 \exp\left[\beta(A_n^B(\zeta) - U^B(\zeta, \mathbf{q}_n^p))\right]d\tilde{\zeta}}. \tag{3.93}$$

4. The effective force field is obtained by the equation:

$$\mathbb{F}_{A_n}(\mathbf{q}_n^p) = -\int_0^1 \nabla_{\mathbf{q}}^p U^B(\zeta, \mathbf{q}_n^p)\pi_{A_n^B}(\zeta|\mathbf{q}_n^p)d\zeta. \tag{3.94}$$

5. Integration of the Langevin dynamic equation is carried out to obtain:

$$\mathbf{q}_{n+1}^p = \mathbf{q}_n^p + \mathbf{P}\mathbb{F}_{A_n}(\mathbf{q}_n^p)\delta t + \mathbf{P}\sqrt{2\beta^{-1}\delta t}B_n^p. \tag{3.95}$$

where $\mathbf{P}$ denotes the projection that shifts the center of mass of the system to its initial position and B_n^p is a normally distributed deviate representing a $3N$-dimensional Wiener process.

3.4 Fluid-phase free energy computations

In addition to crystalline solids, the Bayesian ABF method proposed in Subsection 3.2 can also be applied to liquids. However, the selection of a reference system for a liquid is more complex than for a solid. A common choice for a reference system is the ideal gas (ig). Yet, a direct transformation path between an ideal gas and a liquid can intersect the liquid-vapor coexistence line, leading to a liquid-vapor phase transition that is usually accompanied by hysteresis [251]. To avoid passing through the two-phase region and ensure a smooth variation of thermodynamic quantities along the path, a commonly employed approach is to carry out the process in two stages:

$$\begin{aligned} \Delta A_{\text{ig}\to\text{liq}} &= \Delta A_{\text{ig}\to\text{pot}} + \Delta A_{\text{pot}\to\text{liq}} \\ &= \int_0^1 \langle U_{\text{pot}} - U_{\text{ig}}\rangle_\zeta d\zeta + \int_0^1 \langle U_{\text{liq}} - U_{\text{pot}}\rangle_\zeta d\zeta\,, \end{aligned} \tag{3.96}$$

where $U_{\text{ig}} = 0$ since there are no interactions between the atoms in an ideal gas, and a purely repulsive intermediate reference U_{pot} is required. Given that such systems do not

undergo a liquid-vapor transition, both of the free energy differences $\Delta A_{\text{ig}\to\text{pot}}$ $\Delta A_{\text{pot}\to\text{liq}}$ can be computed using the BABF method.

For selecting a reference system, the aforementioned ZBL potential formulated in Subsection 2.3 can be a suitable choice. Another option is the Uhlenbeck-Ford (UF) model [252, 253], which is a purely repulsive pair potential characterized by a single parameter. An advantage of the UF model is that only the liquid phase is stable within it, thereby preventing hysteresis associated with phase transformations. The potential energy of the UF model is given by

$$U_{\text{UF}} = -\beta^{-1} \sum_{i<j}^{N_s} p \ln \left\{ 1 - \exp\left[-\left(\frac{r_{ij}}{\sigma} \right)^2 \right] \right\} \tag{3.97}$$

where r_{ij} is the interatomic distance between the atoms i and j, σ is a length-scale parameter, and p a non-negative scaling factor that controls the strength of the interaction. The free energy of the UF model is represented as

$$F_{\text{UF}} = F_{\text{ig}} + \Delta A_{\text{ig}\to\text{UF}}\,, \tag{3.98}$$

and the last term can be expanded as

$$\Delta A_{\text{ig}\to\text{UF}} = \beta^{-1} \sum_{n=1}^{\infty} \frac{\tilde{B}_{n+1}(p)}{n} x^n\,. \tag{3.99}$$

Here, $x \equiv b\rho$ with $b \equiv \frac{1}{2}\left(\pi\sigma^2\right)^{(3/2)}$ and $\rho = N_s/V$ the number density. Although the reduced virial coefficients $\tilde{B}_{n+1}(p)$ can be computed with absolute precision, we prefer to use the accurate numerical representation of the UF free energy, expressed in terms of splines, that is available in Ref. [252]. It takes p and x as input and gives the value of $\Delta A_{\text{ig}\to\text{UF}}$.

As $U_{\text{ig}}(\mathbf{q}) = 0$, the extended potential takes the very simple form $\zeta U(\mathbf{q})$. This simple simulation problem is used to conduct a comparative analysis of the relative performance of the BABF and ABF-bin methods. The ABF-bin method directly samples $(\zeta, \mathbf{q})$ to compute Eq. 3.42:

$$A'(\zeta) = \frac{\int_{\mathbb{T}^{3N_s}} U(\mathbf{q}) P_{A_\star}(\zeta, \mathbf{q}) d\mathbf{q}}{\int_{\mathbb{T}^{3N_s}} P_{A_\star}(\zeta, \mathbf{q}) d\mathbf{q}}\,. \tag{3.100}$$

This method consists in considering the dynamics

$$\left\{ \begin{array}{rcl} d\zeta_t &=& \alpha(A'_t(\zeta) - U(\mathbf{q}_t))\, dt + \sqrt{2\alpha\beta^{-1}}\, dW_t^{\zeta}, \\ d\mathbf{q}_t &=& -\nabla_{\mathbf{q}} U(\mathbf{q}_t)\, dt + \sqrt{2\beta^{-1}}\, dW_t^{\mathbf{q}}\,. \end{array} \right. \tag{3.101}$$

The interval $[0, 1]$ is equally divided into M sub-interval ($M = 200$ for the present calculation), denoted as $z^m = [\frac{m-1}{M}, \frac{m}{M}[$ for $m = 1, \ldots, M-1$ and $z^M = [\frac{M-1}{M}, 1]$. Then, the mean force is evaluated at the mid-point of each interval $\zeta^m = \frac{2m-1}{2M}$. It should be noted that the superscript m is used to index the grid of the discretization of the coupling parameter, while the subscript n denotes the values sampled from the dynamics (Eqs. 3.101) at step n. We introduce the indicator function $\mathbf{1}_{\zeta_s \in \zeta^m}$ that takes value one if ζ_s belongs to the bin of ζ^m (i.e., $\zeta_s \in z^m$) and zero otherwise. By the ergodic theorem, the mean force evaluated at ζ^m and step n is computed as:

$$A'_n(\zeta^m) = \frac{\sum_{s=1}^{n-1} U(\mathbf{q}_s) \mathbf{1}_{\zeta_s \in \zeta^m}}{\sum_{s=1}^{n-1} \mathbf{1}_{\zeta_s \in \zeta^m}}\,. \tag{3.102}$$

To validate the application of our methods to liquids, we conducted a computation on a system containing 128 atoms of molten tungsten at 4 000 K, with the UF parameters $p = 1$ and $\sigma = \sqrt{(2)^{2/3}/\pi}$ [252]. Both the BABF method and the ABF-bin method are utilized to estimate $\Delta A_{\mathrm{ig}\to\mathrm{UF}}$, as shown in Fig. 3.12. The results are then compared with the value obtained from the spline representation in Ref. [252], indicated by the dashed line. The values of $\Delta A_{\mathrm{ig}\to\mathrm{UF}}$ given by spline representation, BABF and ABF-bin are 2.17 eV, 2.20 eV at 200 000 steps and 2.01 eV at 500 000 steps, respectively. Figure 3.12 presents the runtime estimate of $\Delta A_{\mathrm{ig}\to\mathrm{UF}}$ during the dynamics of the two ABF methods, in absolute value (sub-figure (a)) and relative difference compared to the spline representation (sub-figure (b)). The BABF calculation at step 200 000 has converged, with the result being sufficiently close to the spline value, showing a discrepancy of only 0.234 meV/atom (1.38%). This difference may also arise from the fitting error of the spline representation. However, the ABF-bin method converges slowly and does not converge even at step 500 000. This observation demonstrates the advantage of the BABF method in terms of convergence. It should also be noted that sampling liquid configurations requires more integration steps compared to sampling solid configurations.

Once the free energy difference $\Delta A_{\mathrm{ig}\to\mathrm{UF}}$ is determined, the remaining work is simple. To obtain the full anharmonic free energy of a liquid via $F_{\mathrm{liq}} = F_{\mathrm{ig}} + \Delta A_{\mathrm{ig}\to\mathrm{UF}} + \Delta A_{\mathrm{UF}\to\mathrm{liq}}$, the last term $\Delta A_{\mathrm{UF}\to\mathrm{liq}}$ can be directly computed using the BABF method, with the UF model serving as the reference system. The free energy of the ideal gas is given by

$$F_{\mathrm{ig}} = N_s\beta^{-1}\left(\ln\rho - 1 + \sum_{\alpha} c_\alpha \ln c_\alpha\right) + 3\beta^{-1}\sum_{i=1}^{N_s}\ln\Lambda_i\,, \tag{3.103}$$

with c_α the concentration of species α and Λ_i the thermal de Broglie wavelength.

3.5 Conclusion of the chapter

The computation of free energy differences is a crucial and active research field in materials science, as well as in computational statistical physics, chemistry, and biology. This chapter presents how to investigate the free energy profile of a target system, along the thermodynamic path from a reference system with accessible free energy, through phase space sampling. Despite significant progress in recent decades, calculating anharmonic free energy at finite temperatures from precise electronic structure calculations (DFT) remains an ongoing challenge. As a result, high-temperature properties and performance of materials cannot be accurately predicted.

As an alternative of DFT calculations, ML potentials approach the DFT accuracy at a computational cost several orders of magnitude lower than the DFT but several orders of magnitude higher than the traditional EAM potentials. With existing sampling methods, it is still a crude task to employ the ML potentials, especially at high temperatures. Within this context, the author proposed:

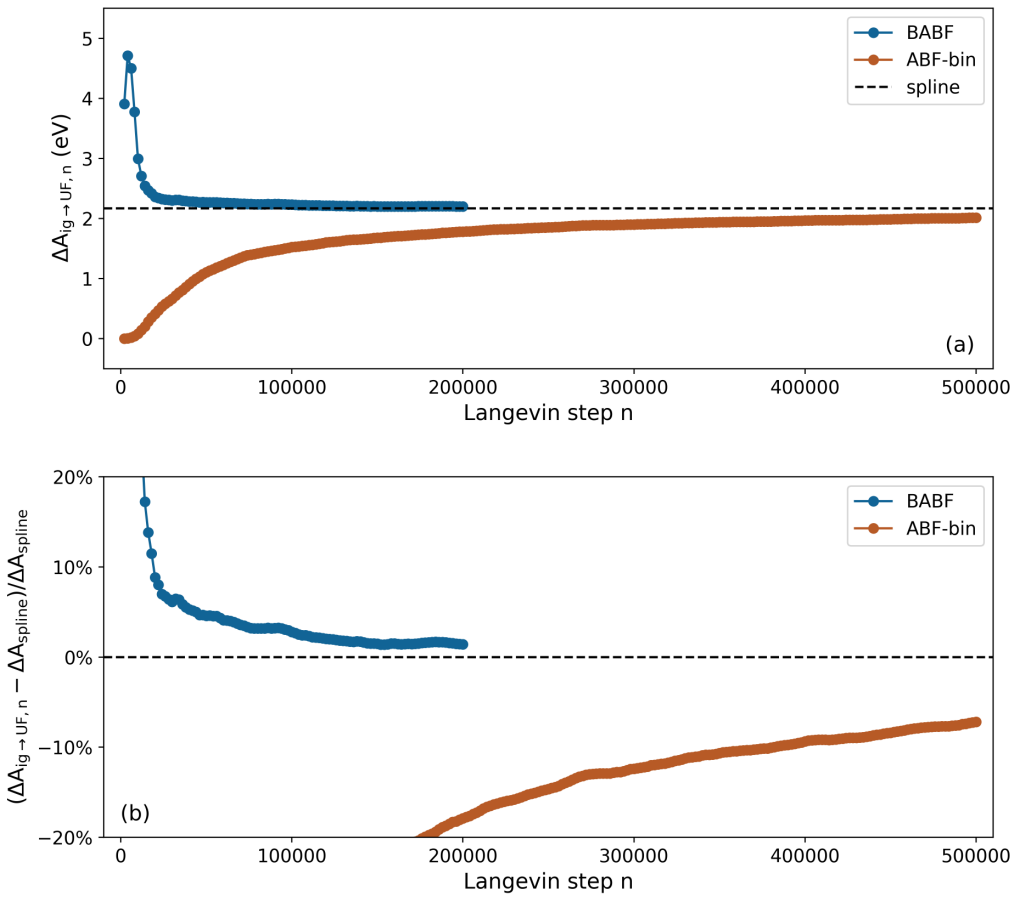

Figure 3.12: Comparison between the BABF method and the ABF-bin method, in computing the free energy difference between the UF model and the ideal gas $\Delta A_{\mathrm{ig}\to\mathrm{UF}} = A_{\mathrm{UF}} - A_{\mathrm{ig}}$, for a molten tungsten system of 128 atoms. (a) Runtime estimate of $\Delta A_{\mathrm{ig}\to\mathrm{UF}}$ during the dynamics of BABF and ABF-bin. The horizontal line denotes the reference value 2.17 eV given by the spline representation in Ref. [252]. (b) Runtime relative difference between the two ABF methods and the spline representation, constrained within the range of [-20%, 20%].

- Accelerated Bayesian adaptive biasing force (BABF) method (Section 3.2), a robust approach to compute the fully anharmonic free energy of crystalline solids, that is approximately 100 times faster than traditional TI for achieving an accuracy of 0.1 meV/atom, and about 800 times faster than conventional MD simulations for obtaining the bulk modulus (the second derivative of free energy) at high temperatures.

- Bound BABF method (Section 3.3), an unbiased free energy estimator for a specified metastable state, based on constrained sampling that prevents transitions between different energy basins.

- BABF method for fluid-phase free energy sampling built on a two-stages scheme (Section 3.4).

Enhanced numerical efficiency of these methods allows for the use of ML potentials, making it feasible to explore the high-temperature free energy landscape with DFT accuracy at a reasonable computational cost. These methods provide access to exact thermodynamic properties, including those derived from the second derivatives of the free energy, as well as the behavior of defects up to the melting point with DFT accuracy. Relevant applications of these methods incorporating ML potentials will be presented in the next chapter.

Chapter 4

Applications

Contents

4.1 Applications to bcc tungsten . . . 106
4.1.1 High-temperature thermodynamic properties . . . 107
4.1.1.1 Performance of EAM potentials . . . 109
4.1.1.2 Prediction of elastic properties based on existing ML force fields . . . 110
4.1.1.3 Improvement of ML force fields . . . 112
4.1.1.4 Prediction model of elastic properties . . . 113
4.1.2 Free energy landscape of vacancies . . . 114
4.1.2.1 Experimental observations of void formation . . . 116
4.1.2.2 ML force fields for vacancy study . . . 117
4.1.2.3 Binding free energy of di-vacancies . . . 120
4.2 Applications to bcc Ta-Ti-V-W high-entropy alloys . . . 123
4.2.1 ML force fields for Ta-Ti-V-W HEAs . . . 124
4.2.2 Finite-temperature elastic properties . . . 126
4.2.3 Finite-temperature formation free energy of vacancy . . . 127
4.3 Correlation between harmonic and anharmonic contributions . . . 129
4.4 Conclusion of the chapter . . . 133

As previously mentioned, the computational cost (CPU time) per Langevin step for machine learning (ML) force fields, as presented in Chapter 2, is significantly lower than that of equally accurate DFT calculations. However, it is several tens to four orders of magnitude higher than that of empirical EAM potentials. The accelerated Bayesian adaptive biasing force (BABF) method, detailed in Chapter 3, offers an effective solution to this efficiency problem by achieving computational speeds several hundred times faster than traditional MD simulations. This computational framework, which combines ML force fields with a fast and robust free energy sampling method, enables the exploration of high-temperature properties of materials with DFT accuracy at temperatures up to their melting point. This chapter presents the applications of this computational framework to bcc tungsten (W) in Section 4.1 and Ta-Ti-V-W high-entropy alloys (HEAs) in Section 4.2. We predict thermodynamic properties such as linear expansion, bulk modulus and elastic constants for W in Subsection 4.1.1. The free energy profile of vacancies is rigorously investigated using the bound BABF method to explain the observation of voids in W, as shown in Subsection 4.1.2. The same study is carried out for the HEA system, examining the elastic properties in Subsection 4.2.2 and vacancy formation in Subsection 4.2.3. For each topic, appropriate ML force fields are constructed based on the techniques presented in Chapter 2, and then employed within a corresponding BABF method. Throughout this chapter, all the BABF calculations are performed using the SVD-filtered harmonic approximation (HA-SVD) reference unless otherwise stated. In addition to the direct prediction of material properties, we demonstrate in Section 4.3 an intriguing correlation between the harmonic and anharmonic contributions to the free energy, observed via the BABF calculations for a bunch of ML potentials. This discovery allows for a rapid estimation of anharmonicity without the need for sampling.

4.1 Applications to bcc tungsten

Tungsten (W) is a highly promising material for the first-wall near the divertor area in fusion reactors because of its exceptional physical properties at high temperatures [254, 255]. The divertor target plates are the most thermally loaded in-vessel components in a fusion reactor, where the plasma-facing components (PFCs) experience high heat fluxes due to intense plasma bombardment, radiation, and nuclear heating. In the heat flux profile specified for the ITER divertor targets, the peak heat flux is assumed to be $10\,\mathrm{MW/m^2}$ for the quasi-stationary operation (2 h) and $20\,\mathrm{MW/m^2}$ for slow transient events ($<$10 s) [256, 257]. 3D thermo-hydraulic simulations indicate that the maximum temperature can exceed 2000 K under a heat flux of $20\,\mathrm{MW/m^2}$ [258]. Tungsten possesses the highest melting point among all metals. In the work package Divertor of EUROfusion Consortium, seven different design concepts of PFCs are currently under development for high heat flux application (see Table 2 of Ref. [258]). All of them are composed of tungsten armor (blocks or tiles) and a cooling pipe at the center serving as a heat sink. The high melting point and high temperature strength of tungsten enables it to mechanically withstand the extreme temperatures generated in the reactor, ensuring structural integrity under severe heat fluxes. Furthermore, its high thermal conductivity is crucial for efficient heat dissipation, preventing overheating and enhancing the overall thermal management of the reactor. These attributes collectively make tungsten a standout choice for PFCs in fusion energy systems.

4.1.1 High-temperature thermodynamic properties

Despite the aforementioned huge industrial interest of W, its high-temperature mechanical properties remain perplexing. As presented in Fig. 4.1, there are no direct measurements of the bulk modulus of the bcc phase of W at temperatures higher than 2 100 K, while numerical calculations of a system at such temperatures suffer from severe statistical fluctuations. In this subsection, our main application concerns the thermodynamic properties of crystalline tungsten.

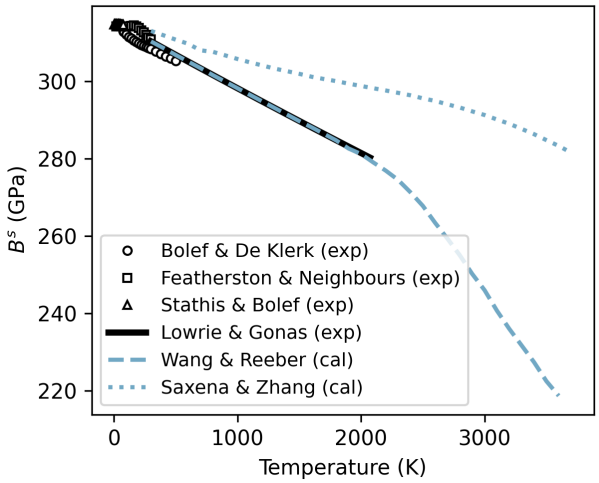

Figure 4.1: Adiabatic bulk modulus B^S of W. Experimental values are obtained from four different experiments in the temperature range from 4.2 K to 2 073.15 K [259, 260, 261, 262], while the calculated values are extended up to the melting point by two models [263, 264].

The performance of the accelerated BABF method, as discussed in Section 3.2, has been demonstrated through comparisons with MD simulations for calculating the bulk modulus of W at temperatures higher than 3000 K. Here, the temperature dependence of elastic constants for W is investigated using this approach. We conduct BABF calculations with various interatomic potentials across a wide temperature range (10 K to 3 800 K) and compare the results with the experimental values, available in the range between 4.2 K and 2 073.15 K [261, 262].

Firstly, we detail how to compute the bulk modulus B and elastic constants C_{ij} from the relationship between free energy and deformations. Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_6)$ represent the strain tensor with its six independent components:

$$\begin{pmatrix} \varepsilon_1 & \varepsilon_6/2 & \varepsilon_5/2 \\ \varepsilon_6/2 & \varepsilon_2 & \varepsilon_4/2 \\ \varepsilon_5/2 & \varepsilon_4/2 & \varepsilon_3 \end{pmatrix}. \tag{4.1}$$

For the bulk modulus, only volumetric strains $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon$ need to be considered, and the rest (ε_4, ε_5 and ε_6) are set to zero. At a given temperature T, we can determine a set of free energies $F_i = F(V_i)$ by performing BABF calculations at several volumes V_i around the expected equilibrium volume. For the case of tungsten, eleven values of ε equally distributed in the interval $[-0.8\%, 0.8\%]$ are selected, yielding ten distinct volumes. We then make a

least-squares fit of the set (V_i, F_i) to the Birch-Murnaghan equation of state:

$$F(V) = F_0 + \frac{9V_0B}{16}\left\{\left[\left(\frac{V_0}{V}\right)^{2/3} - 1\right]^3 B' + \left[\left(\frac{V_0}{V}\right)^{2/3} - 1\right]^2 \left[6 - 4\left(\frac{V_0}{V}\right)^{2/3}\right]\right\} \tag{4.2}$$

with V_0 the equilibrium volume, B the isothermal bulk modulus and B' the pressure derivative of the bulk modulus. V_0 corresponds to the minimum of the free energy F, while B and B' can be derived from the fit. In practice, we often perform a first BABF computation to estimate V_0, followed by a second computation with V_i expanded around V_0 to obtain the bulk modulus. The temperature dependence of V_0 also provides the linear thermal expansion.

At a given temperature T, when the applied strain $\boldsymbol{\varepsilon}$ is small, the free energy F of a system with volume V can be expanded around $\boldsymbol{\varepsilon} = \mathbf{0}$:

$$F(\boldsymbol{\varepsilon}) = F(\mathbf{0}) - P(V_0)\Delta V + \frac{V_0}{2}\sum_{i,j=1}^{6} C_{ij}\varepsilon_i\varepsilon_j + \mathcal{O}[\varepsilon_i^3] \tag{4.3}$$

where $P(V_0)$ is the pressure of the undeformed lattice at volume V_0, ΔV is the change in the volume of the lattice due to the strain, C_{ij} are the elastic constants in Voigt notations, and $\mathcal{O}[\varepsilon_i^3]$ notation indicates that the neglected terms in the polynomial expansion are cubic and higher powers of ε_i. For cubic lattices, there are only three independent elastic constants, C_{11}, C_{12} and C_{44}. The bulk modulus can be related to the elastic constants by the formula

$$B = (C_{11} + 2C_{12})/3\,. \tag{4.4}$$

The shear modulus $C' = (C_{11} - C_{12})/2$ and the modulus C_{44} can be evaluated from Eq. 4.3. For the shear modulus, we apply volume-conserving strains such that $\Delta V = 0$ in the following form:

$$\boldsymbol{\varepsilon} = (\varepsilon, -\varepsilon, \varepsilon^2/(1-\varepsilon^2), 0, 0, 0)\,, \tag{4.5}$$

and Eq. 4.3 becomes

$$\Delta F(\boldsymbol{\varepsilon}) = F(\boldsymbol{\varepsilon}) - F(\mathbf{0}) = V_0(C_{11} - C_{12})\varepsilon^2 + \mathcal{O}(\varepsilon^4)\,. \tag{4.6}$$

We employ a similar technique for determining the modulus C_{44} by considering volume-conserving strains, formulated as:

$$\boldsymbol{\varepsilon} = (0, 0, \varepsilon^2/(4-\varepsilon^2), 0, 0, \varepsilon)\,, \tag{4.7}$$

then

$$\Delta F(\boldsymbol{\varepsilon}) = F(\boldsymbol{\varepsilon}) - F(\mathbf{0}) = V_0C_{44}\varepsilon^2/2 + \mathcal{O}(\varepsilon^4)\,. \tag{4.8}$$

Same as the case of bulk modulus, we take eleven values of ε equally distributed in the interval $[-0.8\%, 0.8\%]$ and compute the corresponding free energy for each strain using the BABF method. The values of $C_{11} - C_{12}$ and C_{44} can then be determined by fitting $\Delta F/V_0$ to a quadratic form of ε.

It should be noted that the elastic moduli calculated in this manner are the isothermal quantities. To compare with experiments, the adiabatic results directly obtained from experiments should be converted to the isothermal results using the following relations:

$$\begin{aligned} C_{11}^T &= C_{11}^S - B^S + B^T \,, \\ B^T &= \frac{C_P B^S}{C_P + TV\alpha^2 B^S} \,, \\ C'^T &= C'^S \,,\ C_{44}^T = C_{44}^S \,, \end{aligned} \tag{4.9}$$

where the superscripts T and S represent the isothermal and adiabatic quantities, C_P and α denote the experimental isobaric heat capacity and the volumetric thermal expansion, respectively [265, 266, 263, 261, 262]. Here, for each temperature, the elastic properties are rescaled by the corresponding values at the Debye temperature T_{Debye}, equal to 400 K for W [228].

The adiabatic elastic constants of W were successively measured by pure continuous wave techniques (from 77 K to 500 K [260]), pulse-echo techniques (from 4.2 K to 300 K [261] and from 297.15 to 2073.15 K [262]) and sampled continuous-wave techniques (from 4.2 K to 77 K [259]). To develop a high-temperature model which cannot be experimentally achieved, an empirical $cB\Omega$-model [267] of self-diffusion was applied to determine the bulk modulus of W [268]. With this model, Wang and Reeber evaluated the tungsten bulk modulus from self-diffusion, thermal expansion and specific heat data over a wide range of temperature (300 - 3600 K) [263]. Besides, Gustafson provided a polynomial expression of isothermal bulk modulus for W through evaluating the available experimental data [269], based on which Saxena and Zhang gave an estimation up to 3700 K [264].

4.1.1.1 Performance of EAM potentials

In this work, we first perform BABF calculations to compute the linear thermal expansion, bulk modulus, and three elastic constants at various temperatures using five different traditional EAM potentials: WDD by Derlet *et al.* [21], WEAM2 and WEAM4 by Marinica *et al.* [22] (denoted as EAM2 and EAM4 in Ref. [22]), WJW by Juslin and Wirth [270]), as well as WMB by Mason *et al.* [271], commonly used for atomic-scale modeling of bcc W. As reported in Fig. 4.2, none of the five EAM potentials can correctly reproduce the experimental curves. The curves of WDD and WMB strongly deviate from that of the experiments. WEAM2 and WEAM4 provide an opposite trend up to 800 K compared to the experiments: for C_{11}^T, C_{12}^T and C_{44}^T, the computed values augment while the measured values decrease with increasing temperature. WJW potential can describe the decreasing tendency of elastic constants. Nevertheless, the slope is too large, and abnormal fluctuation occurs at high temperature due to the potential instability. There are several reasons why these potentials cannot correctly reproduce the temperature dependence of elastic constants in W: (i) firstly the formalism is too basic to take into account the fact that W is a metal for which the Fermi level lays into a pseudo-gap [22, 272]. This fact enhances the angular characters of the bonds, which cannot be reproduced by the radial many-body EAM force fields. (ii) For all these potentials the database used for fitting is very poor, without much data beyond 0 K (except WEAM2 and WEAM4 for which a few W liquid configurations are included). The relatively simple physical model of the EAM formalism, along with the limitations in the fitting information, does not

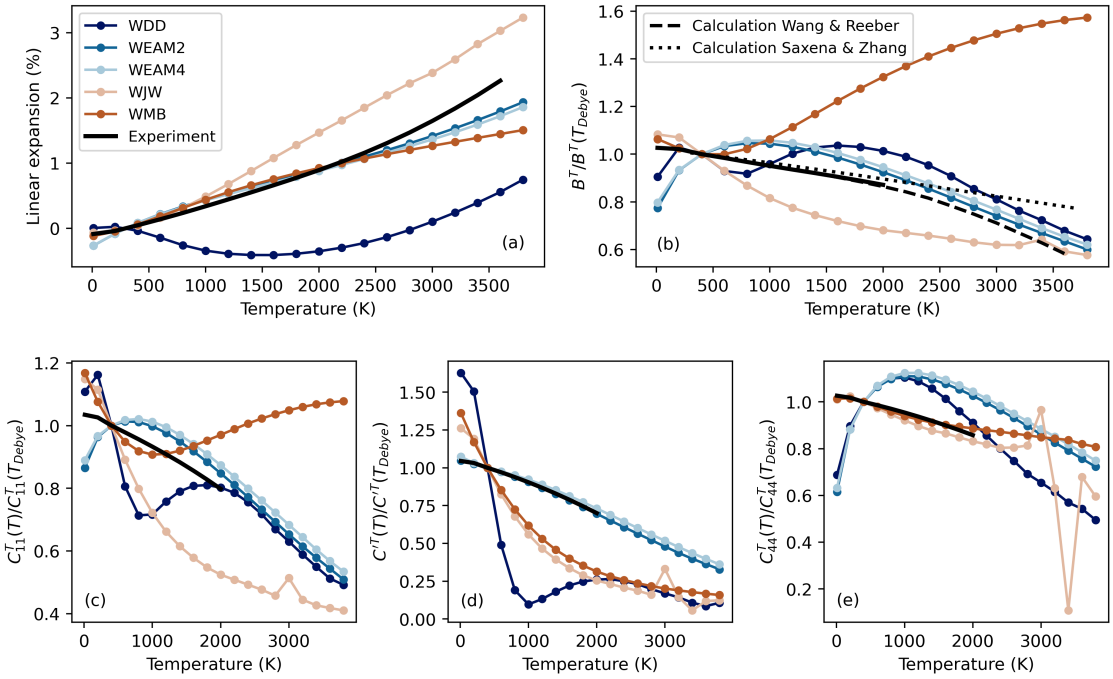

Figure 4.2: (a) Linear thermal expansion, (b) rescaled bulk modulus and (c-e) rescaled elastic constants of bcc W, from 0 K to the melting point, computed employing BABF-HA-SVD method and five different traditional EAM potentials: WDD [21], WEAM2 [22], WEAM4 [22], WJW [270], and WMB [271]. We take the ratio of the elastic properties at temperature T to its value at the Debye temperature $T_{\mathrm{Debye}} = 400\,\mathrm{K}$ [228]. In all panels, the circles denote the results of BABF-HA-SVD calculations with respective force fields whilst the black lines and dash/dotted black lines are experimental [262] and calculated values [263, 264], respectively. All the subplots share the legend provided in (a) and (b).

allow for accurate predictions of high-temperature properties of W, such as the evolution of the bulk modulus with temperature (Fig. 4.2(b)). Therefore, in the next subsection, we will turn to the class of ML potentials, which provide a remarkable improvement of accuracy compared with the empirical potentials [40, 273, 274] especially for W [30, 206, 240].

4.1.1.2 Prediction of elastic properties based on existing ML force fields

As previously stated, the traditional EAM force fields fail to reproduce thermo-elasticity of bcc W. Employing the fast and robust BABF method developed in the present study, we will run through the numerically heavy but accurate ML force fields. During the past several years, various types of ML interatomic potentials have been developed for W, which are most commonly based on the kernel methods, including the GAP [30, 206], linear potential [38, 240] and quadratic noise potential [240]. Moreover, the framework of deep learning and neural network is also used to construct the potential model for W [275].

In this subsection, we perform the BABF calculations of elastic properties and thermal

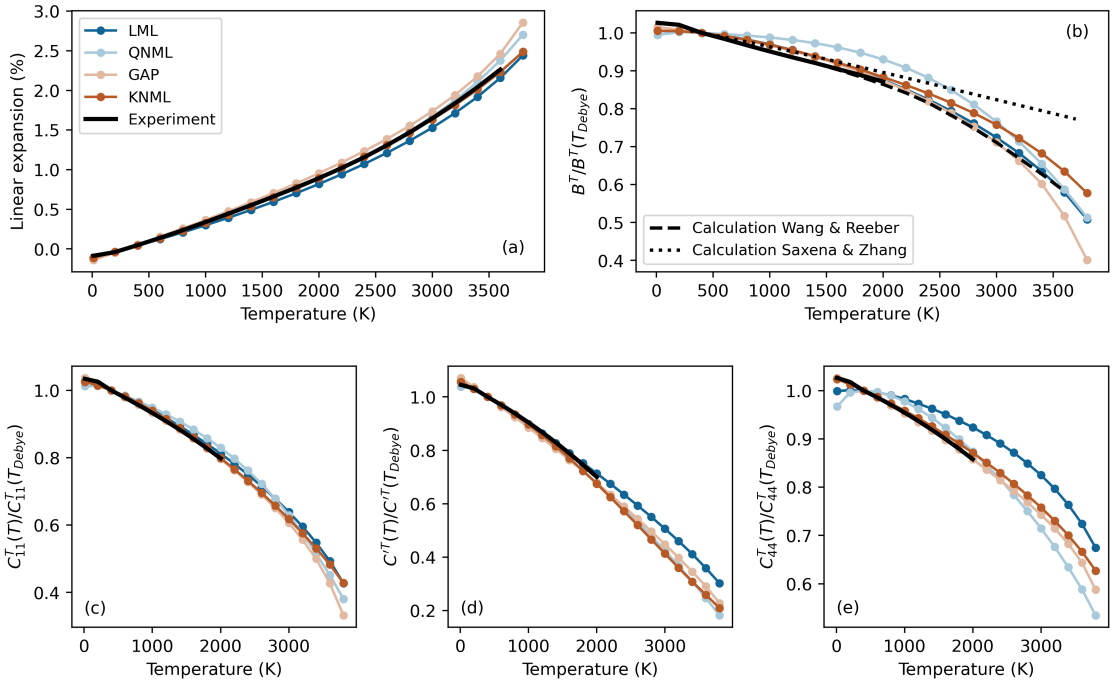

Figure 4.3: The same elastic properties of bcc W as in Fig. 4.2, evaluated using ML potentials with different formalism: LML [240], QNML [240], GAP [30], and KNML. In all panels, the circles denote the results of BABF-HA-SVD calculations with respective force fields whilst the black lines and dash/dotted black lines are experimental [262] and calculated values [263, 264], respectively. All the subplots share the legends provided in (a) and (b).

expansion in bcc W using different types of ML potentials and compare our results with the experimental values [261, 262]. Firstly, we investigate the existing recent ML potentials for W, which are constructed with linear formalism (LML) and quadratic noise formalism (QNML) [240]. These potentials are designed for the simulation of point and extended defects, and the underlying database contains some finite-temperature W systems such as bulk and liquid state. The numerical speed of these ML potentials, which approach the accuracy of DFT calculations, is much slower than that of the EAM potentials. The linear expansion and elastic properties given by both ML potentials (Fig. 4.3) are in much better agreement with the experimental results compared to those from the EAM potentials (Fig. 4.2). However, the rescaled bulk modulus calculated with QNML starts deviating from the experimental curve around 1000 K (Fig. 4.3(b)). In addition, the calculations of the elastic constants, e.g., C_{44} in Fig. 4.3(e), reveal the limitations of the two ML potentials for the calculations of high-temperature elasticity.

Therefore, the widely used model GAP [30] for bcc W is considered. High efficacy and robustness of the present BABF-HA-SVD approach make it feasible to employ this computationally heavy force field, of which the numerical cost per Langevin step is about 15 times higher than that of the LML and QNML potentials. As shown in Fig. 4.3, the previously mentioned deficiencies of LML and QNML are dealt with using the GAP. Nevertheless, the linear expansion obtained with GAP deviates from the experimental curve at high tempera-

tures. This can be explained by the lack of configurations above 1000 K in the database of GAP. In the next subsection, we present the strategy to improve the accuracy of predicting these high-temperature properties by adopting the aforementioned concept of active learning for the optimization of ML force fields.

4.1.1.3 Improvement of ML force fields

Based on the concept of active learning described in Subsection 2.1.1.2, the learning capacity of machine learning force fields can be increased by changing the formalism while the transferability can be optimized by increasing the size and the morphologies of the atomic environments in the database. Firstly, we note that no configuration above 1000 K exists in the training database of the GAP force field, while in that of LML and QNML, only the configurations at 300 K, 1000 K and 3000 K exist. However, the results of the GAP formalism appear to be very accurate up to 2 000 K compared to experimental data, and they are more reliable than the numerically efficient but lower-capacity models yielding the LML and QNML potentials. Therefore, we propose to enlarge the database by collecting information from a wider range of temperatures and simultaneously increasing the learning capacity of the ML model. Here, the bispectrum SO(4) descriptor [169] previously used in linear and quadratic noise regression to construct the LML and QNML force fields is similarly employed in the higher-capacity kernel-based models. This descriptor consists of $D = 55$ components with a cutoff radius of $R_{\text{cut}} = 5.0$Å and an angular moment cutoff $j_{\text{max}} = 4$. The Machine Learning Dynamics (`MiLaDy`) package is used to construct the associated force field.

To complete the database, MD simulation in NPT ensemble is carried out using the MiLaDy-Lammps module [157]. The simulation system contains 128 atoms, and the temperature varies linearly from 100 K to 5 000 K in 245 000 steps. From the MD simulation, 38 configurations are randomly selected and then recalculated with DFT using VASP 6.2.0 [158] and the PAW pseudopotential for W that accounts for 14 valence electrons $[\text{Xe}4f^{14}]5s^25p^66s^15d^5$ (known in VASP database as `W_sv`). The exchange-correlation energy is evaluated using the PBE parametrization [249, 276] of GGA. The DFT setup is exactly the same as in the database from Ref. [240]. In this work, the training dataset is composed of the MD-generated configurations and the original database of the previous LML and QNML potentials [240].

We use the normalized polynomial kernel formulated in the subsection 2.2.3.1:

$$\begin{aligned} k(\mathbf{D}_{s,a}, \mathbf{z}_k) &= \frac{\tilde{k}(\mathbf{D}_{s,a}, \mathbf{z}_k)}{\sqrt{\tilde{k}(\mathbf{D}_{s,a}, \mathbf{D}_{s,a})}\sqrt{\tilde{k}(\mathbf{z}_k, \mathbf{z}_k)}}, \\ \tilde{k}(\mathbf{D}_{s,a}, \mathbf{z}_k) &= \left(\sigma^2 + \frac{\mathbf{D}_{s,a} \cdot \mathbf{z}_k}{2l^2}\right)^p. \end{aligned} \tag{4.10}$$

Here σ, l and p are the kernel hyperparameters. In this framework, $\sigma = 0$ because after the linear fitting, the difference between the DFT target values and the LML prediction follows a normal distribution with zero mean [240]. After many trials the other two hyperparameters are set to $l = 0.05$ and $p = 4$. The $K = 3\,615$ sparse points $\mathbf{z}_k \in \mathbb{R}^D$ are selected via sampling the Mahalanobis distance of the fitting database, as presented in the subsection 2.2.2. We compute the statistical distance of each local atomic descriptor $\mathbf{D}_{s,a}$ for all atoms of the database that

contains in total $M = 47\,277$ data points. We recall that the Mahalanobis distance of the m^{th} local atomic environment with respect to the covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{D\times D}$ of all M local descriptors contained in the database can be written as:

$$\begin{aligned} d(\mathbf{D}_m) &= \left[(\mathbf{D}_m - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{D}_m - \boldsymbol{\mu}) \right]^{\frac{1}{2}} , \\ \mathbf{\Sigma} &= \frac{1}{M-1} \sum_{m=1}^{M} (\mathbf{D}_m - \boldsymbol{\mu})(\mathbf{D}_m - \boldsymbol{\mu})^\top , \\ \boldsymbol{\mu} &= \frac{1}{M} \sum_{m=1}^{M} \mathbf{D}_m . \end{aligned} \tag{4.11}$$

The statistical distance $d(\mathbf{D}_m)$ for each point is computed to measure the distance between this point and the distribution of the whole database. Instead of directly sampling the statistical distance $d(\mathbf{D}_m)$, we sample $d^p(\mathbf{D}_m)$ with a very low power p, here equal to 0.05. Then the interval $I_D = [\min_m(d^p(\mathbf{D}_m)), \max_m(d^p(\mathbf{D}_p))]$ is equally divided into several segments, and one point is selected as kernel sparse point in each sub-interval if possible. Firstly, from the entire database for which $M = 47\,277$, $K = 2\,812$ sparse points are collected from 4 000 sub-intervals of I_D. Secondly, we make another dense selection, in addition to the previous one, in the portion of the descriptor space that represents the primarily important components of the database for the physics of our study. In our case, since we intend to have good elastic constants at high temperatures, we intensify the collection of the configurations related to elastic deformations and finite-temperature MD simulations. Consequently, we perform the same procedure only for three special classes of the database, including the bulk systems under elastic deformations at 0 K, and the perfect bulk systems sampled from MD simulations at 300 K and 3 000 K [240]. From those classes consisting of 5 142 data points we extract 803 sparse points from 1000 sub-intervals.

4.1.1.4 Prediction model of elastic properties

Using the above KNML formalism and the fourth order polynomial kernel, we train a new potential on the new database oriented towards high-temperature properties of bcc W. The present KNML potential is validated by giving almost the same results at reasonable low temperatures ($< 2\,000$ K) as the GAP (Fig. 4.3), which is also based on the fourth order polynomial kernel. The difference between the results of KNML and of GAP, especially at high temperatures, is also reasonable considering difference in database, selection of sparse points and fitting. In terms of computational speed, our KNML potential offers a 6-fold increase over GAP potential. Gratifyingly, the predictive ability of the KNML potential is satisfying. All the properties computed with this potential (Fig. 4.3) are in excellent agreement with the experimental curves. The lattice expansion in Fig. 4.3(a) closely follows the experimental value up to the melting point. This fact is surprising because the present formulation does not take into account the electronic entropy. Perhaps, the present agreement does not exclude an error compensation in the estimation within GGA-PBE exchange-correlation functional, i.e., the finite-temperature effects are amplified by the same amount as the electronic free energy.

Slight discrepancy is observed for all elastic constants at temperatures below the Debye

temperature of W. This behavior is foreseeable because our free energy calculation method, BABF, is based on classical mechanics statics and cannot account for the quantized zero-point energy of phonons in bcc W. This work focuses on the high-temperature properties of bcc W, while the measurable impact of quantized phonons at low temperature has been investigated in a large number of recent studies [277, 278, 279]. At the intermediate temperature between 1000 K and 2000 K, a discrepancy in bulk modulus between the experimental observation and the present KNML potential-based calculation can also be observed. However, this difference is lower than the 0 K difference between the bulk modulus value from the DFT GGA-PBE approach (304.5 GPa) [240] and the low-temperature experimental value (314.73 GPa) [259]. The overall consistency between the experimental measurements and the present computational results is remarkable in terms of the temperature dependence of the elastic properties, which provides a novel perspective on the contribution of electronic entropies and the role of exchange correlation functional.

Based on the results calculated with the accurate KNML potential, a polynomial model is proposed for the rescaled isothermal bulk moduli and elastic constants of bcc W from 0 K to the melting temperature:

$$\begin{aligned}
\frac{B^T(T)}{B^T(T_{\text{Debye}})} &= -4.434\times10^{-12}T^3 - 2.082\times10^{-9}T^2 - 4.042\times10^{-5}T + 1.013\,,\\
\frac{C_{11}^T(T)}{C_{11}^T(T_{\text{Debye}})} &= -3.018\times10^{-13}T^3 - 2.209\times10^{-8}T^2 - 6.875\times10^{-5}T + 1.029\,,\\
\frac{C'^T(T)}{C'^T(T_{\text{Debye}})} &= 5.800\times10^{-12}T^3 - 5.166\times10^{-8}T^2 - 1.103\times10^{-4}T + 1.054\,,\\
\frac{C_{44}^T(T)}{C_{44}^T(T_{\text{Debye}})} &= -2.592\times10^{-12}T^3 - 1.343\times10^{-10}T^2 - 6.616\times10^{-5}T + 1.026\,.
\end{aligned}$$

This model with proven correctness could be useful to predict the elastic behaviors of bcc W at extremely high temperature which cannot be reached by the experiments.

4.1.2 Free energy landscape of vacancies

Point defects, such as vacancies, have a significant influence on diffusion and microstructural evolution in solids. The presence of vacancies facilitates the movement of atoms, thereby enhancing diffusion rates, which is critical in processes like sintering, phase transformations, and creep. Additionally, the accumulation and interaction of these point defects can lead to the formation of dislocations, grain boundaries, and other microstructural features. These changes in microstructure can further affect the mechanical properties, thermal stability, and overall performance of the material, particularly under high-temperature conditions and during long-term operation. Understanding the effects of temperature on defects is particularly important in materials like tungsten, known for its high-temperature applications.

As in many metallic systems, vacancy clusters in tungsten can be directly observed using transmission electron microscopy (TEM). However, there is still no consensus regarding the mechanism of their formation. There is no experimental evidence for the existence of vacancy

clusters at temperatures lower than a few hundred Kelvin. No clusters were observed even under the conditions of vacancy ballistic migration, where vacancies are displaced during the ballistic regime of TEM due to collisions between electrons and atoms near the vacancy. Despite the diffusion of vacancies induced by electron irradiation, vacancy aggregation has not been reported [279]. Electrical resistometry and ion field microscopy have been used to study tungsten samples previously heated to high temperatures close to the melting point and then quenched to retain the excess vacancies [232, 280, 281, 282]. Vacancy loops and voids were observed to appear during subsequent annealing treatments in the temperature range from 800 K to 1000 K [232]. Such experimental observations raise a natural question regarding the formation and clustering mechanism of vacancies in this material. Formation of voids implies firstly the formation of small clusters as di-vacancy. Di-vacancy in tungsten, as in other bcc metals of the group VI, such as Mo, has an unusual energy landscape [233, 237]. As mentioned in Subsection 3.3.1, previous DFT calculations state that di-vacancies are strongly unstable in 2NN configurations, and either unstable or stable in 1NN configurations, depending on the details of the calculations. We attribute these parameter-dependent trends to the insufficient convergence of the calculations, which resulted from an inadequate density of the k-points mesh used.

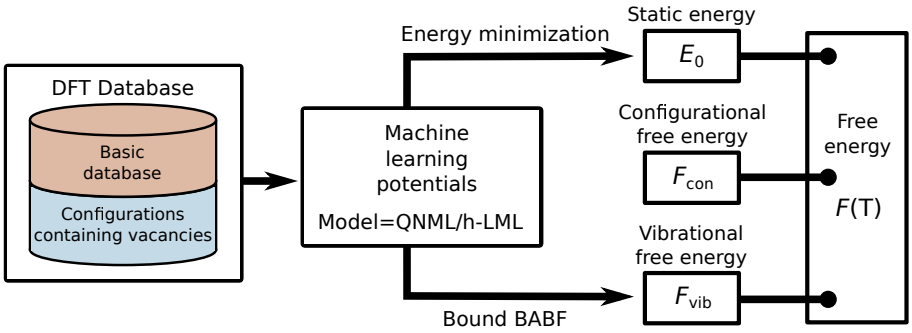

Figure 4.4: Workflow implemented in the present study. The MLPs are constructed using databases specifically tailored for vacancy studies, based on two GGA functionals, AM and PBE. The model is fitted using the bispectrum SO(4) descriptor [169] and its hybrid version, employing both quadratic and linear formalisms. For each configuration of interest (mono-vacancy, 1NN and 2NN di-vacancies), energy minimization through molecular statics provides the static energy E_0. The vibrational free energy F_{vib} is obtained via bound BABF calculations. By combining E_0 and F_{vib} with the configurational free energy F_{con}, which depends only on the vacancy type, the free energy profile can be determined at different temperatures, allowing for the further calculation of formation and binding free energies for vacancies.

In this subsection, the formation of vacancy clusters in bcc W is systematically investigated. We first present in Subsection 4.1.2.1 the experimental TEM observations from Kazuto Arakawa at Shimane University, demonstrating the existence of a threshold temperature below which void formation is inhibited and above which voids are stabilized. Understanding this phenomenon is crucial for comprehending the temperature effects on the mechanical properties of W due to void formation. This can be achieved by quantifying the temperature effects on the formation and binding free energy landscape of small vacancy-type defects. To achieve this goal, we face three major challenges. Firstly, despite recent advances in free energy calculations using TI [112, 113], the aforementioned dependence of the convergence rate on the number of k-points makes it unfeasible to sample with an entirely DFT-computed force field within reasonable CPU cost. Secondly, the differences in binding energies are very small, requiring

the use of a highly accurate free energy calculation method. Thirdly, even with an accurate sampling method, it should be able to sample a given metastable state without allowing the system to transition from one free energy basin to another. To address these challenging problems, we employ ML force fields that achieve DFT accuracy, trained on sets of converged DFT calculations using k-point meshes of appropriate density. The construction of the ML potentials will be detailed in Subsection 4.1.2.2. In particular, configurations containing vacancies are additionally added to the database. For fast sampling and robust free energy estimation of a specified metastable state with the ML force fields, we rely mainly on the bound BABF approach proposed in Section 3.3. Evaluation of the formation and binding free energy for di-vacancies will be elaborated in Subsection 4.1.2.3.

The present workflow is illustrated in Fig. 4.4. Beforehand, three ML potential variants are trained on databases specific to vacancy properties. For each potential, four thermodynamic configurations, i.e., the bulk crystal, the mono-vacancy and the 1NN and 2NN di-vacancies (cell configurations can be found in Fig. 3.10(a)), are investigated in the canonical ensemble to obtain the vibrational free energy. The full free energy profile is therefore accessible by adding the configurational free energy and the static energy at equilibrium computed from energy minimization. The electronic free energy is omitted here since its contribution to the vacancy formation free energy is negligible up to 2 000 K [283]. We compute the formation free energy for each defect type from the corresponding free energy landscape and further deduce the binding free energy of 1NN and 2NN di-vacancies at different temperatures from 10 K to 2 000 K. The binding free energy results show that the interaction of di-vacancies changes from repulsion to attraction at a critical temperature, and this conversion occurs at a lower temperature for the 1NN configuration than for the 2NN one. We demonstrate that in W, finite temperature excitations play an important role in the physics of defects, in particular in determining their aggregation properties.

4.1.2.1 Experimental observations of void formation

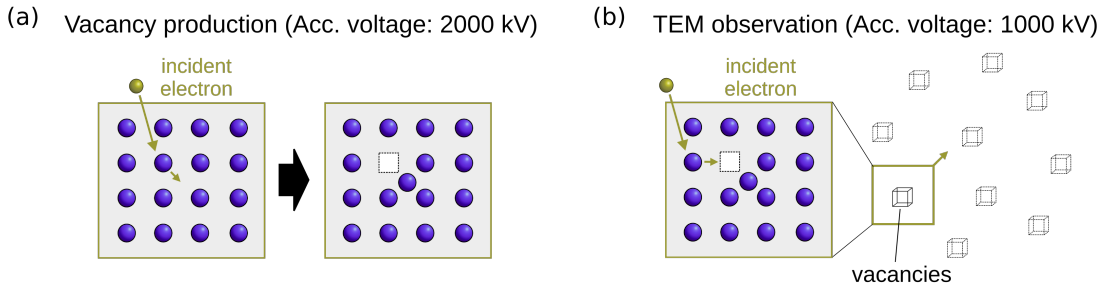

Figure 4.5: (a) Vacancy (Frenkel pair) production induced by the incident electrons under the beam (acceleration voltage: 2000 kV; beam intensity: 3.3×10^{22} $m^{-2}s^{-1}$; temperatures: 628, 665, 813, 850, 924, 1034, 1145, 1182, 1219 K). (b) Vacancy migration driven by an acceleration voltage of 1000 kV. No Frenkel pair production at this stage.

The experimental methods in this study are similar to those used in Ref. [279]. Specimens were prepared from the W (110) TEM sample at purity of 99.9999 mass per cent. For vacancy production (Fig. 4.5), Frenkel pairs were created in the thin foil specimens via knock-on dis-

placement by high-energy electron irradiation in a high-voltage electron microscope (Hitachi H-3000). Nine temperatures were considered: 628, 665, 813, 850, 924, 1034, 1145, 1182 and 1219 K. The acceleration voltage was 2 000 kV, the beam flux was 3.3×10^{22} $m^{-2}s^{-1}$ (1.0×10^{-4} displacements per atom per second, dpa/s), and the dose was 0.54 dpa for the three lowest temperatures and 0.80 dpa for the remaining temperatures.

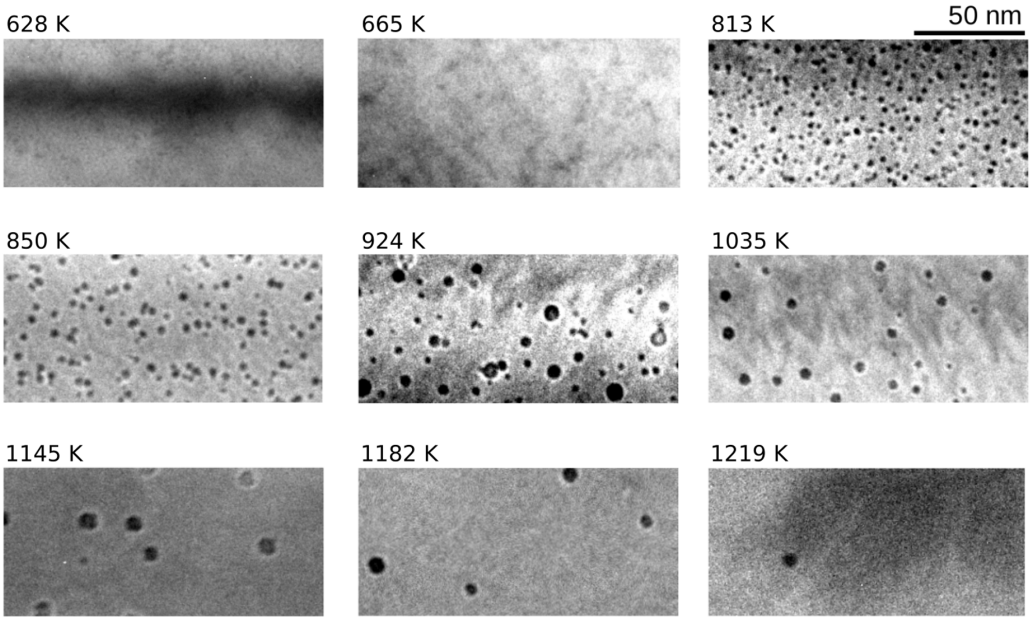

Figure 4.6: TEM images of samples at different temperatures provided by Kazuto Arakawa. There is no vacancy cluster at 628 K and 665 K, while at a temperature equal to or higher than 813 K, vacancy clusters are observed.

For the vacancy observations, electron beam is used to induce the vacancy mobility, with acceleration voltages of 1000 kV, which is below the threshold for point defect generation in tungsten [284]. The specimen thickness is 80 nm, measured with equal-thickness fringes. The observations were carried out using the bright-field imaging with a reflection of $\mathbf{g} = 200$. As shown in Fig. 4.6, voids (vacancy clusters) were observed only at temperatures equal to or higher than 813 K; no voids were observed at 628 K and 665 K. Figure 4.7 provides the void density at different temperatures. The density at 665 K and 813 K is below the observation limit of the microscope, indicating no observation of voids at these temperatures. The absence of voids formed from moving vacancies suggests a repulsive interaction between vacancies at temperatures lower than a threshold value between 665 K and 813 K. Above this temperature, voids are produced and stabilized, implying that the vacancies become attractive to each other.

4.1.2.2 ML force fields for vacancy study

The associated force fields are built with `MiLaDy` package [38, 240]. Content of ML database has a strong impact on the accuracy and transferability of the potential. Here we detail in Table 4.1 the list of configurations in the database used for fitting the potentials for W. In particular, with the aim of modeling vacancies in W, we include in the database the configurations

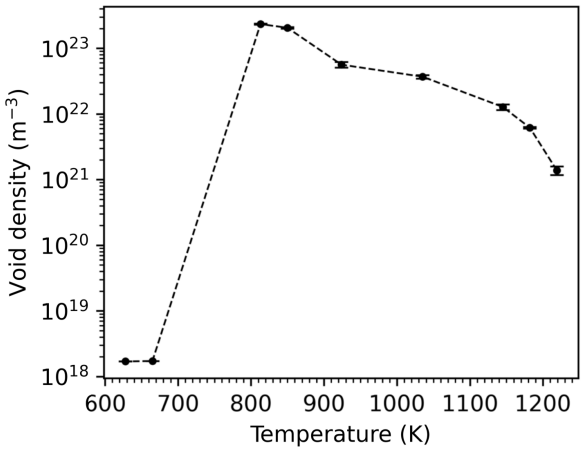

Figure 4.7: Void density at different temperatures. Suppression of void formation at low temperature ($< 665 - 813\,\mathrm{K}$) indicates the repulsive interaction between vacancies.

containing vacancies, collected from the nudged elastic band (NEB) method and the projected average force integrator (PAFI) calculations [285]. The NEB calculations are performed for

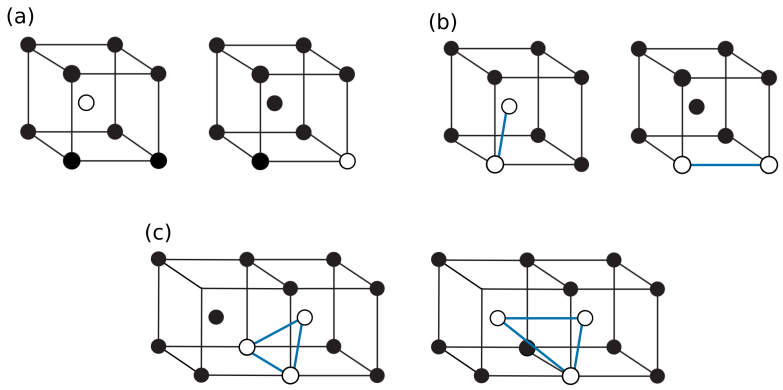

Figure 4.8: Configuration pairs containing (a) 1 vacancy, (b) 2 vacancies and (c) 3 vacancies that define the transition paths in NEB and PAFI calculations for the database construction. The difference between the two configurations in each pair results from a single vacancy migrating to its 1NN lattice site.

finding the atomic configurations associated with a minimum energy path (MEP) of the transition between two states at 0 K. As illustrated in Fig. 4.8, we consider the transition paths between two configurations both containing 1, 2 or 3 vacancies. These transitions correspond to the migration of a single vacancy to its 1NN lattice site. Note that the two configurations in graphs (a) and (c) of Fig. 4.8 are equivalent with boundary conditions, so the transition paths should be symmetric, while this is not the case in graph (b). At finite temperatures, the minimum free energy path (MFEP) is computed instead of the FEP, again between the two states of the configuration pairs in Fig. 4.8. The configurations along the MFEP can be derived using the PAFI method. Initially, the MEP is obtained via the NEB technique, after which a series of hyperplanes perpendicular to the MEP are defined, as illustrated in Fig. 4.9. Constrained MD are then performed within each hyperplane, involving a thermalization pro-

cess followed by sampling using overdamped Langevin dynamics. This procedure allows the system to move towards a minimum free energy state. By applying this method across a series of hyperplanes along the MEP, the MFEP is accurately determined.

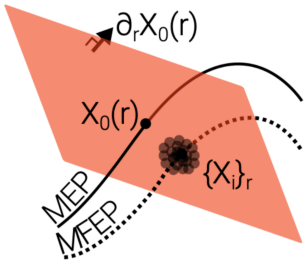

Figure 4.9: Illustration of the sampling process by PAFI [286]. An estimate of the minimum free energy position is obtained by repeating the thermalization and constrained MD several times within the hyperplanes $\partial_r \mathbf{X}_0(r)$ (shown in orange) perpendicular to the MEP $\mathbf{X}_0(r)$ with respect to some reaction coordinate r.

Table 4.1: List of configurations for the database of W oriented to vacancy studies.

System W(N_d,ϵ)	Temperature in K					Total
	0	875.0	1750	2625	3500	
bulk, $\epsilon = +0\%$	1	10	10	10	10	41
bulk, $-5\% \leq \epsilon \leq 5\%$	1000	0	0	0	0	1000
1 vacancy, $\epsilon = 0\%$	10	10	10	10	10	50
2 vacancies, 1NN	10	10	10	10	10	50
2 vacancies, 2NN	10	10	10	10	10	50
2 vacancies, 3NN	10	10	10	10	10	50
NEB 1 vacancy, 1NN	7	0	0	0	0	7
NEB 2 vacancies, 1NN	7	0	0	0	0	7
NEB 3 vacancies, 1NN	7	0	0	0	0	7
PAFI 1 vacancy, 1NN	0	5	5	5	5	20
PAFI 2 vacancies, 1NN	0	3	3	3	3	12
PAFI 3 vacancies, 1NN	0	3	3	0	0	6
MD liquid						5
Elastic constants	13	0	0	0	0	13
Molecular dynamics	0	12	12	12	12	48
Total	1075	73	73	70	70	1368

All the configurations in the database are recalculated with DFT using VASP 6.2.0 [158]. $(4a_0)^3$ cells with $6 \times 6 \times 6$ k-points grid are used at $\sigma = 0.3\,\text{eV}$, which has demonstrated convergence in Fig. 3.9. The exchange-correlation energy is evaluated using the GGA-PBE [249] and GGA-AM [241] functionals, resulting in the creation of two separate databases. GGA-PBE is a widely used and universal functional applicable to a wide range of systems including metals. It generally performs better than LDA but systematically overestimates the lattice constants [287, 288]. GGA-AM has been shown to be more accurate than PBE in reproducing the lattice constants of solids [241]. Therefore, both functionals are employed in this study. The other DFT setup is exactly the same as in the database from Ref. [240]. Bispectrum SO(4) descriptor [169] with dimension $D = 55$, $R_{\text{cut}} = 5.3$ Å and the maximum angular moment $j_{\text{max}} = 4.0$ is used to describe the atomic environment, and the QNML formalism is applied for both GGA functionals. Another potential is created based on the PBE database

by linearly fitting a hybrid descriptor (bispectrum SO(4) with $j_{\max} = 5.0$ and 2-body kernel). As presented in the subsection 2.2.3.2, the 2-body kernel is a smooth translation- and permutation- invariant kernel between two atomic environments, which can be defined by summing all the squared exponential distances between the relative atomic positions included in these two atomic neighborhoods [289]. We adopt here the formulation in Eq. 2.85, where $K_{2b} = 40$ interatomic distances are considered. The cutoff distance for applying the ZBL correction (Section 2.3), i.e., the inner cutoff $R_{\text{cut}}^{\text{in}}$ for the hybrid descriptor, is set to 1.2 Å, to avoid unphysical behaviors at low interatomic distances. In this way, three different ML potentials are built for the following study:

- QNML+AM: bispectrum SO(4) descriptor, AM database, QNML regression;
- QNML+PBE: bispectrum SO(4) descriptor, PBE database, QNML regression;
- h-LML+PBE: hybrid descriptor (bispectrum SO(4) + 2-body kernel), PBE database, LML regression.

For each of them, we compare some key quantities related to the properties of vacancy-type defects between the DFT calculations and the ML potential predictions. Comparison is presented in Table 4.2, where E_{1f}, E_{2f}^{1NN} and E_{2f}^{2NN} represent the formation energy at 0 K of mono-vacancy, di-vacancies at 1NN and 2NN, respectively. E_{1m} and E_{2m} denote the migration energy for mono-vacancy and di-vacancies at 0 K. The obtained agreement validates the data-based fitting of potentials.

Table 4.2: Comparison between DFT calculations and ML potentials for key quantities in W: lattice constant, bulk modulus, elastic constants, formation energies and migration energies. Two exchange-correlation functionals are considered: PBE and AM.

	PBE			AM		Unit
	DFT	QNML	h-LML	DFT	QNML	
a_0	3.18569	3.18568	3.18568	3.15073	3.15070	Å
B	304.7	304.6	304.6	327.9	327.6	GPa
C_{11}	515.4	513.4	513.4	559.6	554.4	GPa
C_{12}	199.3	200.3	200.2	212.0	214.2	GPa
C_{44}	139.9	139.9	139.8	151.4	151.5	GPa
E_{1f}	3.20	3.19	3.23	3.53	3.52	eV
E_{2f}^{1NN}	6.56	6.52	6.52	7.20	7.18	eV
E_{2f}^{2NN}	6.89	6.84	6.84	7.50	7.49	eV
E_{1m}	1.72	1.63	1.63	1.78	1.72	eV
E_{2m}	1.30	1.20	1.13	1.37	1.38	eV

4.1.2.3 Binding free energy of di-vacancies

The formation energy of a configuration containing n vacancies at temperature T can be computed as

$$F_{nf}(T) = F(N_s - n, T) - \frac{N_s - n}{N_s} F(N_s, T), \tag{4.12}$$

with $F(N_s, T) = F_{\text{vib}}(N_s, T) + F_{\text{con}}(T) + E_0(N_s, T)$ the free energy of a system containing N_s atoms at its equilibrium volume and temperature T. The configurational free energy is expressed as $F_{\text{con}}(T) = -\beta \ln N_{\text{con}}$, where N_{con} is the number of equivalent configurations for one defect type (N_{con} equals to 1 for bulk and mono-vacancy, 8 for 1NN di-vacancies and 6 for 2NN di-vacancies). Energy minimization in LAMMPS [227] gives the static energy $E_0(N_s, T)$, and the vibrational free energy $F_{\text{vib}}(N_s, T)$, including the temperature-dependent anharmonic contribution, can be obtained by the bound BABF approach (Section 3.3). The relation 4.12 allows us to access the formation free energy for the 3 types of target defects (mono-vacancy: F_{1f}, di-vacancies at 1NN: F_{2f}^{1NN} and di-vacancies at 2NN: F_{2f}^{2NN}) at any finite temperature. Then, the binding free energy for both types of di-vacancies can be evaluated from

$$F_{2b}^{\text{1NN}}(T) = 2F_{1f}(T) - F_{2f}^{\text{1NN}}(T)\,, \tag{4.13a}$$

$$F_{2b}^{\text{2NN}}(T) = 2F_{1f}(T) - F_{2f}^{\text{2NN}}(T)\,. \tag{4.13b}$$

Within this theoretical framework, we utilize the bound BABF method to calculate the free energies of the three aforementioned defective configurations and the reference bulk crystal without defects over a temperature range from 10 K to 2000 K. In particular, the reference bulk crystal at equilibrium corresponds to a stable state where no transitions occur during sampling. Therefore, applying an additional energy barrier is unnecessary, and the BABF method alone suffices. For the bound BABF calculations of vacancy structures, the scaling and cutoff parameters in Eq. 3.71 are chosen as $\delta = 0.5\,\text{Å}$, $C = 25\,\text{eV}$ and $R_c = 2.23\,\text{Å}$. To be exhaustive, we use the three different ML potentials as the target potential energy $U(\mathbf{q})$ in the linearly mixed general potentials Eq. 3.39 and Eq. 3.73.

The discussion below exclusively considers the results obtained using ML potentials employing the QNML formalism, as the results associated with the h-LML formalism show consistent trends with those of the QNML formalism. The binding free energies for the 1NN and 2NN di-vacancies deduced from Eqs. 4.13 are presented in Fig. 4.10(a) and Fig. 4.10(b), respectively. They include the anharmonic contributions. For comparison, we also display the reference free energies computed using the harmonic approximation (HA), where phonon frequencies are derived from the Hessian of a system with the temperature-dependent lattice constant obtained from the corresponding BABF calculations. This harmonic free energy is exactly the reference free energy $A(0)$ that equals to Eq. 3.67. Obviously, the differences among the results from different ML potentials are insignificant, and all of them show the same trends. Both 1NN and 2NN di-vacancies are repulsive at low temperatures and they become attractive as temperature increases. The critical temperature of repulsion/attraction conversion is denoted as T_c, at which the binding free energy reaches zero. For all the ML potentials considered here, this critical temperature remains much higher for di-vacancies at 2NN than at 1NN. At temperatures higher than around 400 K, the di-vacancies system can be stabilized in 1NN configuration, while to have stable 2NN configuration the temperature needs to reach around 1400 K. Anyway, the agreement among the different ML potentials confirms that increasing the temperature and quenching the W solid can increase the di-vacancy concentrations. 1NN configuration is always more stable at temperatures lower than 2 000 K, however, even 2NN configuration can be stabilized by high temperature.

Comparing the anharmonic and harmonic results enables us to assess the importance of the

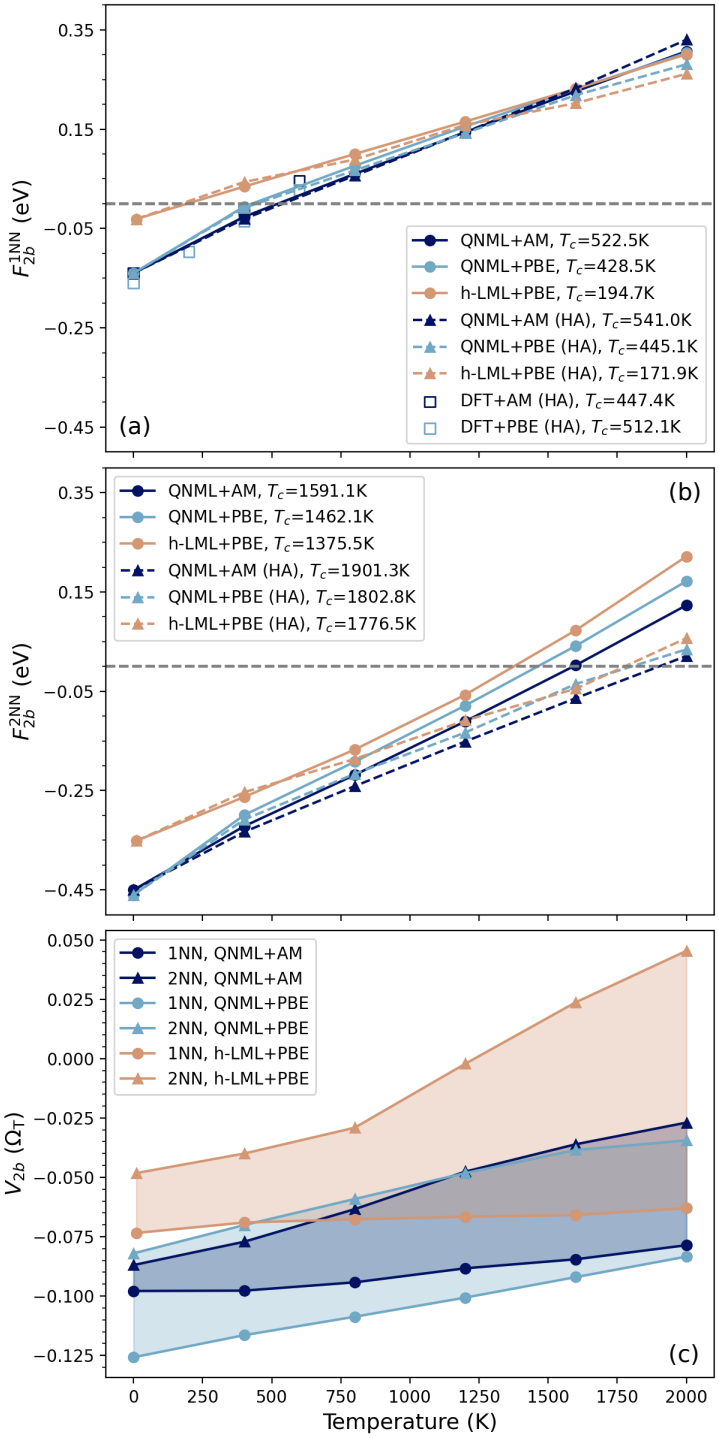

Figure 4.10: Binding free energy including anharmonic contributions from 0 K to 2 000 K of (a) 1NN di-vacancies and (b) 2NN di-vacancies computed from the bound BABF method with three machine learning potentials, harmonic approximation at temperature-dependent lattice constants given by ML potentials and DFT calculations. T_c denotes the temperature at which the binding free energy is zero. (c) Binding volume of the di-vacancies in Ω_T, the equilibrium atomic volume at temperature T.

anharmonic contribution in the free energy. For both 1NN and 2NN di-vacancy configurations, the anharmonicity becomes significant only at high temperatures, while the differences at low temperatures are negligible. Specifically, for 1NN di-vacancies, the critical temperature T_c falls within a temperature range where the anharmonic effects are slight, resulting in very close agreement between harmonic and anharmonic results. In this case, we perform the full DFT calculations based on the same temperature-dependent lattice constants at $T = 0\,\mathrm{K}$, 200 K, 400 K and 600 K from the BABF calculations, to obtain the exact values from DFT. In light of the negligible anharmonicity at these temperatures, the system is assumed to be purely harmonic and its free energy is calculated analytically by means of Eq. 3.67, with the phonon frequencies obtained from DFT. In the present DFT calculations, a reference bcc super-cell containing 128 tungsten atoms, a $6 \times 6 \times 6$ shifted k-point grid of the Monkhorst-Pack scheme and a plane wave cutoff energy of 500 eV are used. As indicated in Fig. 4.10(a), the critical temperature T_c equals to 447.4 K for AM and 512.1 K for PBE, which closely aligns with the results obtained from ML potentials using the same functional.

Finally, we also deduce the binding volume which is analogous to the binding free energy:

$$V_{2b}^{1\mathrm{NN}}(T) = 2V_{1f}(T) - V_{2f}^{1\mathrm{NN}}(T)\,, \tag{4.14a}$$

$$V_{2b}^{2\mathrm{NN}}(T) = 2V_{1f}(T) - V_{2f}^{2\mathrm{NN}}(T)\,, \tag{4.14b}$$

with

$$V_{nf}(T) = V(N_s - n, T) - \frac{N_s - n}{N_s} V(N_s, T)\,. \tag{4.15}$$

where $V(N_s, T)$ is the equilibrium volume corresponding to the free energy minimum at temperature T for a system of N_s atoms. This quantity has previously been considered temperature-independent and is now being investigated for the first time. Fig. 4.10(c) reveals that 2NN configuration always has larger binding volume than 1NN configuration.

In this subsection, we demonstrated that temperature is the crucial factor influencing the formation of di-vacancies and larger clusters. This study represents the first calculation of the formation free energy, including the anharmonic contribution, for mono-vacancy and di-vacancies in W at finite temperatures. Both 1NN and 2NN di-vacancies are repulsive at low temperatures. Via the BABF-based calculations, we demonstrate that the increasing temperature stabilizes all repulsive configurations approximately in the range of 500 K (1NN) to 1500 K (2NN), and thus augments the concentration of di-vacancies. For higher temperatures, even 2NN di-vacancies are stable. These results provide a quantitative explanation for the experimental observations shown in Subsection 4.1.2.1. Moreover, the temperature-dependence of binding volume is revealed for the first time. With the present calculations, we can depict a scenario for the formation of vacancy clusters in W.

4.2 Applications to bcc Ta-Ti-V-W high-entropy alloys

High-entropy alloys (HEAs) were discovered and synthesized in 2004 by two separate groups of scientists: Cantor *et al.* [290] and Yeh *et al.* [291]. Since then, rapid developments in this

emerging field of materials science have been observed. Experimental studies have revealed that HEAs exhibit unique microstructures and properties. Their high yield strength and hardness at elevated temperatures make HEAs promising materials for high-temperature applications. Additionally, their superior irradiation properties compared to pure elements and conventional alloys position HEAs as potential candidates for structural elements in future fusion and fission reactors [292, 293, 294, 295, 296, 297]. For example, Ref. [295] reports that there are no signs of radiation-induced dislocation loops under radiation damage of 8 dpa, at room temperature and 1073 K, in bcc-based HEA $W_{38}Ta_{35}Cr_{16}V_{11}$. In this section, we focus on the bcc Ta-Ti-V-W HEAs. The choice of Ta-Ti-V-W systems is based on the prediction that solid solutions within this alloy class exhibit some of the lowest ordering temperatures [297], implying that the formation of disordered solid solutions is possible at lower temperatures than in other alloy systems. This theoretical prediction is experimentally validated in Ref. [298], where the Ta-Ti-V-W alloy is identified as the sole quaternary alloy in the Cr-Ta-Ti-V-W system to exclusively form a single bcc phase. Moreover, this composition is worth studying because systems composed of these elements have demonstrated high tolerance to radiation damage and maintain sustainable mechanical properties after irradiation [297].

4.2.1 ML force fields for Ta-Ti-V-W HEAs

To begin the study, the fist goal is to develop the ML potential capable of investigating the properties of Ta-Ti-V-W alloys across a wide range of compositions and temperatures. Therefore, the DFT database contains the results for various compositions of the form $A_{40}B_{20}C_{20}D_{20}$, $A_{10}B_{30}C_{30}D_{30}$ and $A_{40}B_{40}C_{10}D_{10}$, where A, B, C, D cover all possible combinations of Ta, Ti, V and W atoms. The list of configurations used to develop ML potentials for the Ta-Ti-V-W system is summarized in Table 4.3. DFT calculations are performed using VASP 6.2.0 [158] with exchange and correlation treated in GGA-PBE [249, 276] parametrizations. We use the PAW pseudopotential without semi-core electron contributions and do not take magnetism into account. The same DFT parameters are taken as those used in Ref. [297]. The plane-wave cutoff energy is set to 400 eV. The Brillouin zone is sampled according to the Monkhorst–Pack scheme [299] with $4 \times 4 \times 4$ k-points mesh for the $(4a_0)^3$ bcc supercells. This computational setup is used to generate representative structures of Ta-Ti-V-W alloys to serve as the database for the ML force fields.

We employ the same type of hybrid descriptor as used in Subsection 4.1.2.2, combining the widely-used bispectrum SO(4) descriptor with the fast and simple 2-body kernel descriptor, with an general outer cutoff distance $R_{\text{cut}} = 4.9$ Å and an inner cutoff distance $R_{\text{cut}}^{\text{in}} = 1.4$ Å, below which the atomic interactions are described by the ZBL potential. The maximum angular moment $j_{\max}$ for the bispectrum SO(4) is set to 4.5, and for the 2-body kernel, $K_{2b} = 40$ distances are sampled for each possible type of atom pair (Ta-Ta, Ta-Ti, Ta-V, Ta-W, Ti-Ti, Ti-V, Ti-W, V-V, V-W, and W-W). The simple LML formalism (Section 2.1.3.1) is used for the regression. Thus, the present ML potential is denoted as h-LML in the following discussion.

4.2.2 Finite-temperature elastic properties

Table 4.3: Details of the different classes of DFT database used for the development of ML potentials for the Ta-Ti-V-W system. A short description of the calculations is given in the "Content" column. The number of atoms in the cell used for DFT calculations is presented in the "Atoms per cell" column. The "Properties to fit" column corresponds to the fitted properties for training/testing (energy E, force F, stress S). The "$M_E + M_F + M_S$" column corresponds to the sum of a number of energies (M_E), forces (M_F), and stresses (M_S) in the train/test data that are used to fit/test the potential for each considered class. Finally, the total number of configurations for each considered class that have been used for the training and testing purposes are given in the "Configurations train/test" column.

DB class	Content	Atoms per cell	Properties to fit	$M_E + M_F + M_S$ train/test	Configurations train/test
1	Structure optimization (0 K)	128	E	29/0	29/0
2	Elasticity - small strain	128	ES	10 584/3 528	1 512/504
3	Elasticity - large strain	128	ES	15 792/5 040	2 256/720
4	Equation of state	128	E	90/36	90/36
5	MD - bcc bulk (1000 K, 1500 K, 2500 K)	128	EF	20 6448/96 186	528/246
6	MD - liquid (1000 K, 1500 K, 2500 K)	100	EF	85 039/41 138	277/134
7	MD - liquid/bcc interface (1000 K, 1500 K, 2500 K)	112	EF	27 440/10 633	80/31
8	Mono-vacancy	127	E	114/46	114/46
9	Mono-SIA	129	E	230/107	230/107
10	Di-vacancies	126	E	13/7	13/7
Total				345 779/156 721	5 129/1 831

The elastic properties of alloys are of paramount importance in many applications. To predict these properties for Ta-Ti-V-W HEAs at finite temperatures, the BABF method introduced in Section 3.2 is employed. Same as the case of W in Subsection 4.1.1, this method is used to obtain the anharmonic free energies for various deformations of the system at different temperatures. The elastic constants are then computed from the second derivatives of the free energy with respect to strains. The elastic constants are calculated for disordered $[\mathrm{TaTiV}]_{100-x}\mathrm{W}_x$ pseudobinary alloys at three tungsten concentrations: 10, 25 and 40 at. % W. For each concentration, we use a system of 128 atoms, which is too small to be regarded as cubic crystal. Therefore, we compute nine elastic constants and the values adopting cubic symmetry are approximated as

$$\overline{C_{11}} = \frac{1}{3}(C_{11} + C_{22} + C_{33})\,, \overline{C_{12}} = \frac{1}{3}(C_{12} + C_{13} + C_{23})\,, \overline{C_{44}} = \frac{1}{3}(C_{44} + C_{55} + C_{66})\,. \quad (4.16)$$

The 9 elastic constants are calculated following the steps below:

1. Perform MD in NPT ensemble with h-LML potential to get a relaxed configuration as well as V_0 at each temperature.
2. Apply the strains $\boldsymbol{\varepsilon}$ listed in Table 4.4 on the relaxed configuration obtained in step 1 to construct the deformed structures. We took $\varepsilon = \pm 0.2\%, \pm 0.4\%, \pm 0.6\%, \pm 0.8\%$ for each type of deformation.
3. Using the BABF method, calculate the free energy for the relaxed configuration ($F(\mathbf{0})$) and the free energy for the deformed configurations ($F(\boldsymbol{\varepsilon})$).
4. The corresponding free energy changes $\Delta F/V_0 = [F(\boldsymbol{\varepsilon}) - F(\mathbf{0})]/V_0$ derived from Eq. 4.3 are listed in Table 4.4. Fit $\Delta F/V_0$ to a quadratic form of ε and get the nine elastic constants from the quadratic coefficient. Notice that the last six deformations in Table 4.4 conserve the volume, whereas the first three are not volume-conserving. However, the target pressure for the NPT relaxation in step 1 was set to zero, representing a state without an external load. The pressure fluctuation of such a relaxed metal system $P(V_0)$ is very slight compared to the elastic constants, which are on the order of magnitude of 10^{10} to 10^{11} Pa.

Table 4.4: Strains and free energy changes for computing the elastic constants.

Elastic constant	ε	$\Delta F/V_0$
C_{11}	$(\varepsilon, 0, 0, 0, 0, 0)$	$C_{11}\varepsilon^2/2 - P(V_0)\varepsilon + \mathcal{O}[\varepsilon^3]$
C_{22}	$(0, \varepsilon, 0, 0, 0, 0)$	$C_{11}\varepsilon^2/2 - P(V_0)\varepsilon + \mathcal{O}[\varepsilon^3]$
C_{33}	$(0, 0, \varepsilon, 0, 0, 0)$	$C_{11}\varepsilon^2/2 - P(V_0)\varepsilon + \mathcal{O}[\varepsilon^3]$
C_{44}	$(\varepsilon^2/(4-\varepsilon^2), 0, 0, \varepsilon, 0, 0)$	$C_{44}\varepsilon^2/2 + \mathcal{O}[\varepsilon^4]$
C_{55}	$(0, \varepsilon^2/(4-\varepsilon^2), 0, 0, \varepsilon, 0)$	$C_{55}\varepsilon^2/2 + \mathcal{O}[\varepsilon^4]$
C_{66}	$(0, 0, \varepsilon^2/(4-\varepsilon^2), 0, 0, \varepsilon)$	$C_{66}\varepsilon^2/2 + \mathcal{O}[\varepsilon^4]$
C_{12}	$(\varepsilon, -\varepsilon, \varepsilon^2/(1-\varepsilon^2), 0, 0, 0)$	$(C_{11} + C_{22} - 2C_{12})\varepsilon^2/2 + \mathcal{O}[\varepsilon^4]$
C_{13}	$(\varepsilon, \varepsilon^2/(1-\varepsilon^2), -\varepsilon, 0, 0, 0)$	$(C_{11} + C_{33} - 2C_{13})\varepsilon^2/2 + \mathcal{O}[\varepsilon^4]$
C_{23}	$(\varepsilon^2/(1-\varepsilon^2), \varepsilon, -\varepsilon, 0, 0, 0)$	$(C_{22} + C_{33} - 2C_{23})\varepsilon^2/2 + \mathcal{O}[\varepsilon^4]$

The approximated elastic constants C_{11}, C_{12} and C_{44} for disordered $[\mathrm{TaTiV}]_{100-x}\mathrm{W}_x$ obtained with the h-LML potential are shown as a function of temperature in Fig. 4.11. We

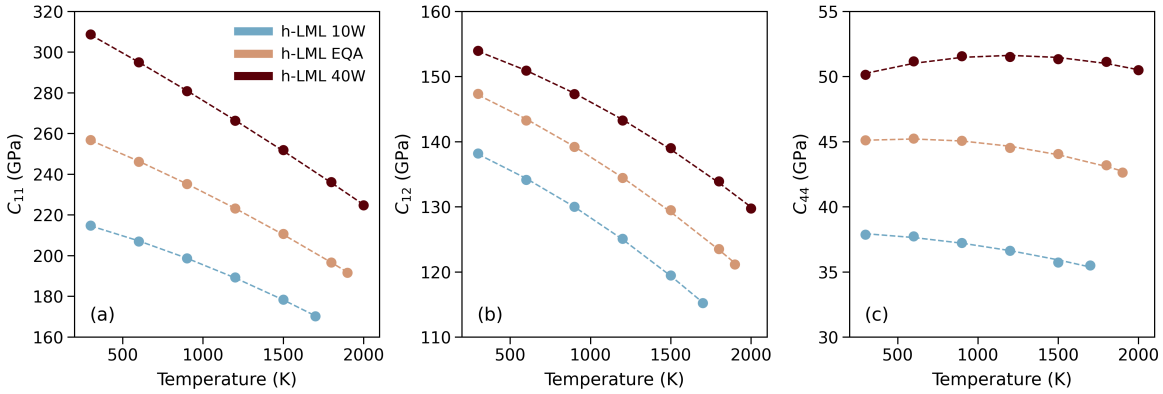

Figure 4.11: Prediction of elastic constants from the BABF anharmonic free energy calculations using the h-LML potential: (a) C_{11}, (b) C_{12} and (c) C_{44} of disordered $[TaTiV]_{100-x}W_x$ pseudobinary alloys with three different concentrations of W, namely 10, 25 and 40 at. %, as a function of temperature.

denote by "EQA" the equiatomic TaTiVW alloy at 25 at. % W. The maximum simulated temperature for each composition depends on the corresponding melting point. The trend of increasing elastic constants with higher concentrations of W is consistently observed across all temperatures. Like many metallic materials, the elastic constants C_{11}, C_{12} as well as the bulk modulus $B = (C_{11} + 2C_{12})/3$, exhibit lower values at higher temperatures for all concentrations of W, implying that the material exhibits lower resistance to deformation when subjected to normal stress. However, the temperature dependency of C_{44} is not significant, in contrast to the decreasing tendency observed in all its components: Ta [300], Ti [301], V [302] and W [303]. This indicates that the ability of alloys to withstand shear forces is less influenced by temperature changes compared to pure metals. It is important to note that the elastic constants predicted using the h-LML potential change smoothly as a function of W concentration. This stability shows that the h-LML potential is stable and can be reliably applied to predict the properties of Ta-Ti-V-W alloys at elevated temperatures.

4.2.3 Finite-temperature formation free energy of vacancy

The properties of vacancies in alloys depend not only on the composition of the alloy but also on the local environment of a defect [304, 305, 306]. Given the computational costs of DFT calculations, an alternative approach to evaluating the formation energies at 0 K in multicomponent alloys is through molecular statics (MS) calculations employing ML force fields. To obtain better statistics, MS calculations using h-LML potential are performed on 432 structures, each generated by removing one of the atoms from the non-defected structure containing 432 atoms displayed in a $6\times6\times6$ bcc supercell. The MS calculations are carried out for disordered pseudobinary alloy structures at 10, 25 (equiatomic) and 40 at. % W in order to study the influence of W concentration on vacancy properties. The increase of W content in the pseudobinary HEAs leads to a corresponding increase in the mean mono-vacancy formation energy (E_{1f}). The value obtained for the alloy with 10 at. % W (2.33 eV) approaches the E_{1f} for pure bcc V (2.51 eV) [21] and bcc Ti (2.03 eV) [307]. For the equiatomic case, E_{1f} equals

2.65 eV. When the W concentration is increased to 40 at. %, the mean value of E_{1f} reaches 3.06 eV, approaching the value for pure bcc W (3.56 eV) [21].

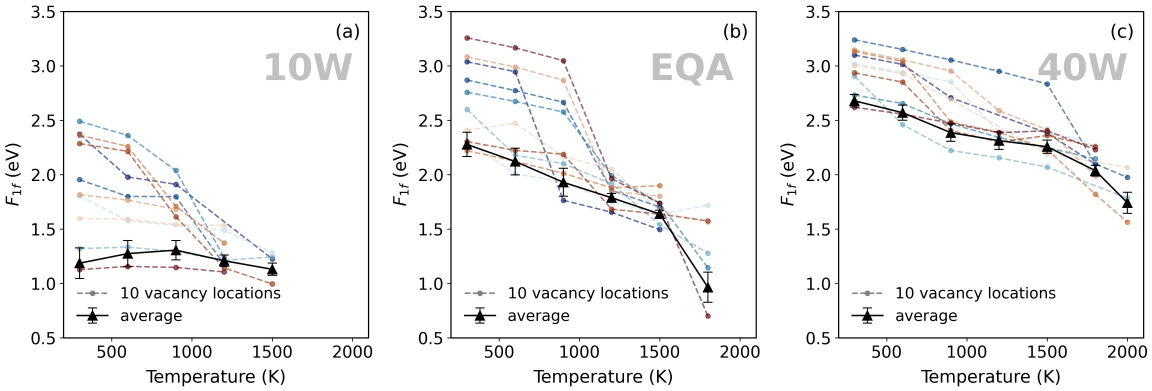

Figure 4.12: Formation free energies computed for the vacancy in disordered $[\mathrm{TaTiV}]_{100-x}\mathrm{W}_x$ pseudobinary alloys with three different concentrations of W, namely 10, 25 and 40 at. %, as a function of temperature using h-LML potential. For each composition, a reference structure is first created, and 10 calculations are performed by randomly removing one atom from the reference structure, resulting in 10 configurations with a single vacancy at different locations. These calculations are represented by colorful dashed lines, and an exponential average of them is shown in black. The maximum temperature reported in panels (a), (b) and (c) increases because the melting point of the HEA is lower with a smaller W concentration.

At finite temperatures, the same computational framework combining bound BABF calculations and ML potentials, as described in Subsection 4.1.2, is applied to compute the mean formation free energy of a mono-vacancy. As the properties of a vacancy in alloys depend on both the composition of the alloy and the local environment of the defect, adoption of Eq. 4.12 is an approximation which amounts to averaging the chemical potential over the composition. This gives the formation free energy of a vacancy at location i as

$$F_{1f}(T,i) = F^i_{vac}(T) - \frac{(N_s-1)}{N_s} F_{ref}(T)\,, \tag{4.17}$$

where F^i_{vac} is the free energy of the alloy configuration containing a vacancy at location i, and F_{ref} is the free energy of the reference structure containing N_s atoms without defects. The above definition of formation free energy is exact only in the limit of averaging the chemical potential over the composition. Otherwise, it can be seen as an arbitrary limit in the reference of formation free energy. Indeed, the free energies involved here include the anharmonic vibrational contribution but excludes the combinatorial contribution. Evaluating the latter contribution would involve extensive sampling of the alloy configurations and vacancy locations, incurring a high computational cost of approximately 1280 core-hours per configuration, which is beyond our computational capacity. We therefore limited our investigations to a series of 10 randomly selected vacancy locations ($N_{loc} = 10$) for three compositions of the disordered alloy and a range of up to 7 temperatures. An estimate of the free energy of mono-vacancy formation for a given composition and reference disordered structure is then

obtained by exponential averaging over the N_{loc} vacancy locations:

$$F_{1f}(T) = -\beta^{-1} \ln \sum_{i=1}^{N_{loc}} \frac{1}{N_{loc}} \exp\left[-\beta F_{1f}(T,i)\right] . \tag{4.18}$$

Note that the quantity $\exp\left[-\beta F_{1f}(T)\right]$ is an unbiased estimate of the vacancy concentration at temperature T given the alloy configuration: its average over all random subsets of vacancy locations gives the expected vacancy concentration. Estimating the free energy of vacancy formation resorting to the exponential average in Eq. 4.18 amounts to implementing a Widom test-particle method where a difference in vibrational free energy to suppress an atom is substituted for the usual difference in potential energy, and where a single random distribution of atoms over the lattice sites is considered in the reference system. Based on Eq. 4.18, we show in Fig. 4.12 the temperature dependence of mono-vacancy formation free energies for the disordered $[\mathrm{TaTiV}]_{100-x}\mathrm{W}_x$ pseudobinary alloys with 10, 25 and 40 at. % W, for temperatures up to a maximum determined by the corresponding melting point of each composition. For some vacancy locations, we observe very frequent vacancy migrations at elevated temperatures that lead to transitions among different free energy basins, thereby hindering the sampling of a single metastable state. Even within the bound BABF method, the high frequency of calling the restoring force results in error accumulation. Consequently, these problematic samplings are excluded, leading to some missing data points in Fig. 4.12, while the remaining data is unbiased. At lower temperatures, the ranges of vacancy formation free energies for the alloys under consideration align with the E_{1f} values computed at 0 K. As temperature increases, the average value of $F_{1f}(T)$ remains nearly stable for the alloy with 10 at. % W and decreases for alloys with 25 and 40 at. % W, reaching 1.13 eV, 1.64 eV, and 2.26 eV at 1500 K, respectively. It should be noted that 10 random vacancy locations are insufficient to provide a statistically reliable average of the formation free energy, yielding only a rough estimate. Nevertheless, interesting trends can still be identified. Notably, the range of $F_{1f}(T)$ at a given temperature (the difference between the maximum and minimum values among the 10 random vacancy locations) decreases with increasing temperature. This observation suggests that the dependence of the formation free energy on the local atomic environment of the vacancy diminishes at higher temperatures, as vibrational entropy becomes more significant.

4.3 Correlation between harmonic and anharmonic contributions

In this section, we present an interesting correlation between the harmonic and anharmonic contributions to the free energy, which is found with the help of the fast and robust BABF method detailed in Section 3.2. This observation allows for a rapid estimation of the fully anharmonic free energy based solely on the computationally inexpensive harmonic approximation. We present the workflow using bcc W, which has been thoroughly discussed in this thesis, and the results will also be validated in the bcc Fe system. The present workflow can be divided into the following steps:

1. Random creation of ML potentials: 500 ML potentials are randomly constructed with `MiLaDy` package, based on the database detailed in Table 2.1, using random weights

specified for each database class. It means that the value of ω_m in Eq. 2.34 is randomly chosen from $[0, 1000]$ for each class of the database, except those determine the basic thermodynamic properties (lattice constant, bulk modulus and elastic constants). Here, the weight ω_m for the global energy of the configurations in DB classes 2 and 3 is fixed to 10^7. Besides aiming to ensure the correct prediction of basic quantities, this large value is also due to the fact that the global energies are relatively small in these classes, given that the systems contain only 2 or 4 atoms. The potentials are built in the same way as the LML potential in Ref. [148], employing the bispectrum SO(4) descriptor and the linear regression.

2. CUR selection of ML potentials: out of the total 500 potentials generated in the first step, the 106 most "influential" potentials are selected using the CUR decomposition method presented in Subsection 2.2.2. The lattice constant, bulk modulus and three elastic constants at 0 K given by these 109 potentials are plotted in Fig. 4.13, demonstrating the basic correctness of the selected potentials, at least in terms of thermodynamic stability. The predicted properties are distributed within a reasonable interval, e.g., 0.003 Å for the lattice constant, and are not far from the DFT value of 3.1854 Å. Therefore, all the following calculations are performed based on the equilibrium configuration of 128 atoms with the DFT-computed lattice constant.

3. BABF free energy calculation: for each selected potential, the free energy difference $\Delta A = A(1) - A(0)$ between the harmonic reference system with Landau free energy $A(0) = F_{\mathrm{HA}}$ and the real system associated with the target anharmonic free energy $A(1)$ is computed using the BABF method, at 300 K, 1000 K, 2000 K and 3000 K. The anharmonic contribution ΔA is illustrated in Fig. 4.14 as a function of the harmonic free energy F_{HA} given by the same potential.

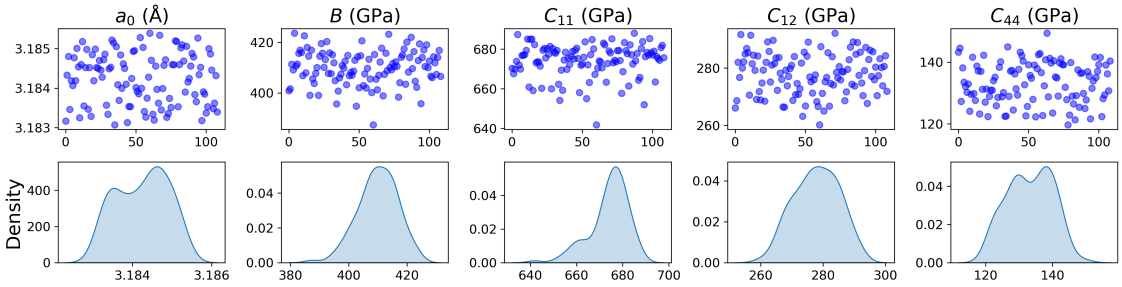

Figure 4.13: Distribution of the values of lattice constant, bulk modulus and elastic constants C_{11}, C_{12} and C_{44} at 0 K, computed from the selected ML potentials for W. The horizontal axes denote the indices of the selected potentials, and the vertical axes of the graphs in the first row present the values of the corresponding thermodynamic quantities, whose distribution densities are shown in the second row.

Figure 4.14 shows a significant linear correlation between the harmonic and anharmonic contributions for the ML potentials. These potentials are generated using completely random database weighting, resulting in parameters that are entirely different from each other. The coefficients of determination R^2 for the four temperatures in increasing order are 0.860, 0.862, 0.870 and 0.881. We fit this F_{HA}-ΔA correlation to the linear form $\Delta A = \alpha F_{\mathrm{HA}} + \beta$ at each temperature, and the temperature-dependence of the slope α and the intercept β are plotted

in Fig. 4.15. Both the slope α and the intercept β are fitted to the cubic form:

$$\begin{aligned}\alpha(T) &= -6.256\times10^{-13}T^3 + 2.320\times10^{-8}T^2 - 3.519\times10^{-4}T + 2.495\times10^{-3}\,,\\ \beta(T) &= -2.830\times10^{-10}T^3 - 2.866\times10^{-5}T^2 + 1.466\times10^{-2}T - 2.134\,.\end{aligned}$$

In this way, a rough estimate of the fully anharmonic free energy at finite temperature can be directly given by the fast and numerically inexpensive harmonic approximation. Here for W, we have

$$\Delta A(T, F_{\mathrm{HA}}) = \alpha(T)F_{\mathrm{HA}} + \beta(T)\,, \tag{4.19}$$

of which the RMSE at 300 K, 1000 K, 2000 K and 3000 K is respectively 0.0013 eV, 0.014 eV, 0.051 eV and 0.10 eV, corresponding to the per-atom values 1.03×10^{-5} eV/atom, 1.10×10^{-4} eV/atom, 4.01×10^{-4} eV/atom and 8.02×10^{-4} eV/atom.

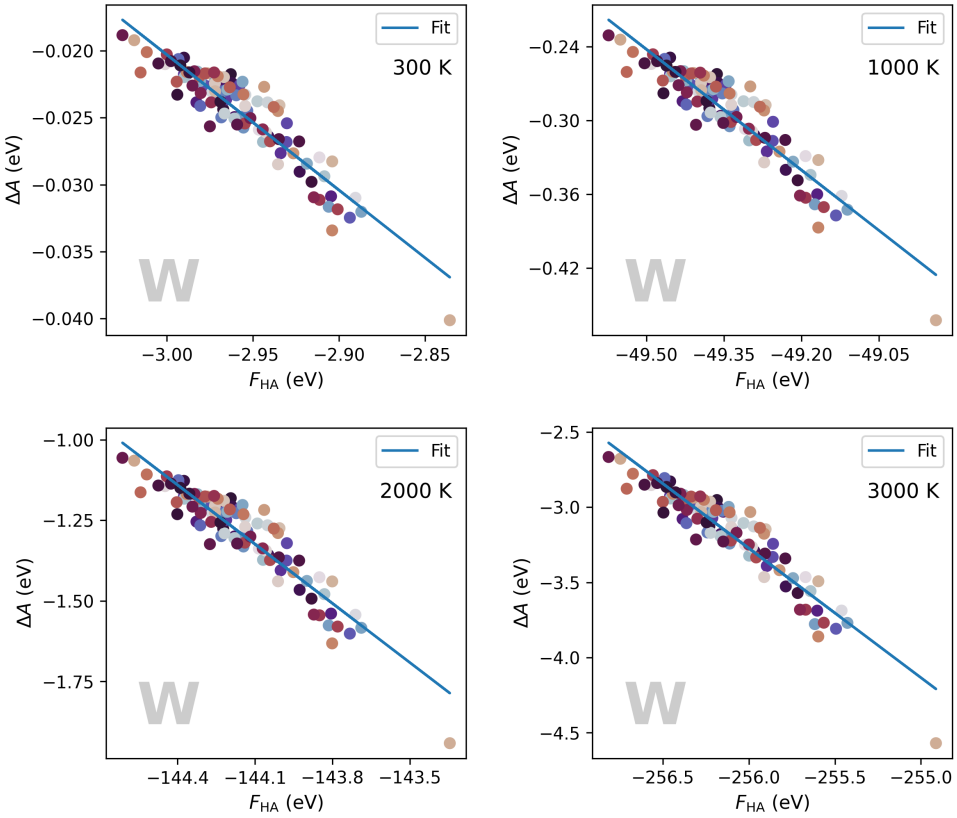

Figure 4.14: Correlation between the free energy from the harmonic approximation (F_{HA}) and the anharmonic contribution (ΔA) for W at 300 K, 1000 K, 2000 K and 3000 K. Each color represents a selected potential. A linear correlation is observed with R^2 values ranging from 0.86 to 0.88, and the fitted model Eq. 4.19 is plotted in blue. The relative positions of the potentials in this correlation are independent of temperature.

Equation 4.19 provides an approximate solution. We also notice that the relative positions of the potentials in the sub-figures of Fig. 4.14 remains unchanged (although the scale is changed), indicating that they are temperature-independent. This observation suggests that in the investigation of ML potentials, such as the optimization of database weighting, once the anharmonic free energy of two potentials is computed, the anharmonic free energy of any

other potential can be accurately determined from its harmonic approximation and the linear combination of the free energy profiles of the two known potentials. This approach is even more precise than the prediction model given in Eq. 4.19.

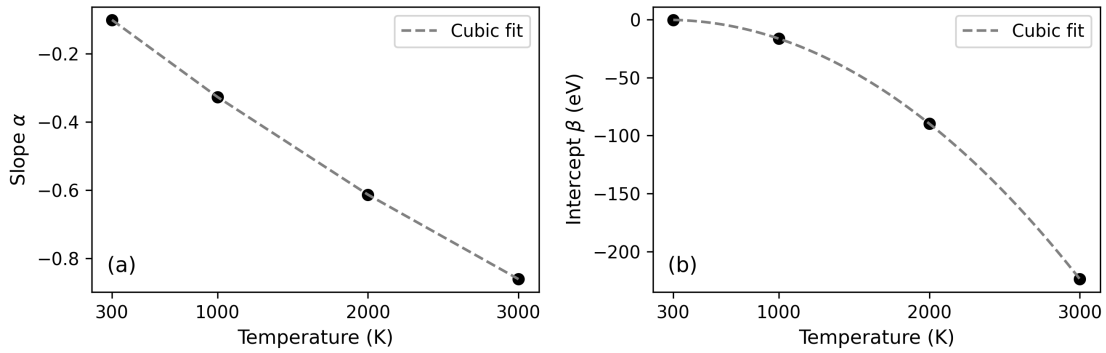

Figure 4.15: (a) Slope α and (b) intercept β of the linear correlation $\Delta A = \alpha F_{\mathrm{HA}} + \beta$ at 300 K, 1000 K, 2000 K and 3000 K, derived from the data shown in Fig. 4.14. A cubic fit is performed for both coefficients as functions of temperature, and the fitted results are indicated by the dashed lines.

To validate this method, the same procedure is applied to bcc Fe at 300 K, 700 K, 1100 K and 1600 K. Similarly, figure 4.16 displays the distribution of the values of lattice constant, bulk modulus and three elastic constants at 0 K, computed from the 103 selected ML potentials for Fe. The fitted model for Fe can be expressed as:

$$\begin{aligned} \alpha(T) &= 7.196 \times 10^{-12} T^3 - 8.329 \times 10^{-8} T^2 - 5.205 \times 10^{-4} T - 9.869 \times 10^{-3}\,, \\ \beta(T) &= -1.244 \times 10^{-8} T^3 - 2.185 \times 10^{-5} T^2 + 1.201 \times 10^{-2} T - 1.265\,. \end{aligned}$$

The RMSE for Fe is respectively 0.0022 eV (1.76×10^{-5} eV/atom) at 300 K, 0.012 eV (9.09×10^{-5} eV/atom) at 700 K, 0.028 eV (2.16×10^{-4} eV/atom) at 1100 K and 0.057 eV (4.46×10^{-4} eV/atom) at 1600 K.

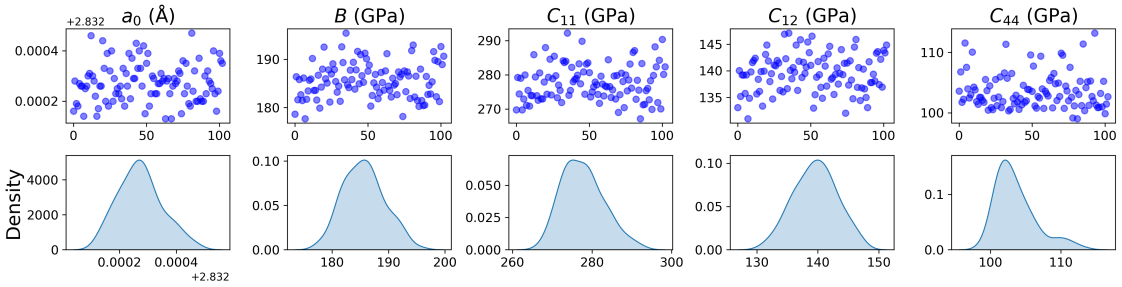

Figure 4.16: Distribution of the values of lattice constant, bulk modulus and elastic constants C_{11}, C_{12} and C_{44} at 0 K, computed from the selected ML potentials for Fe. The horizontal axes denote the indices of the selected potentials, and the vertical axes of the graphs in the first row present the values of the corresponding thermodynamic quantities, whose distribution densities are shown in the second row.

The finding of this shortcut to the anharmonic free energy is made possible by the high efficiency of the BABF method. By establishing surrogate models between the harmonic and

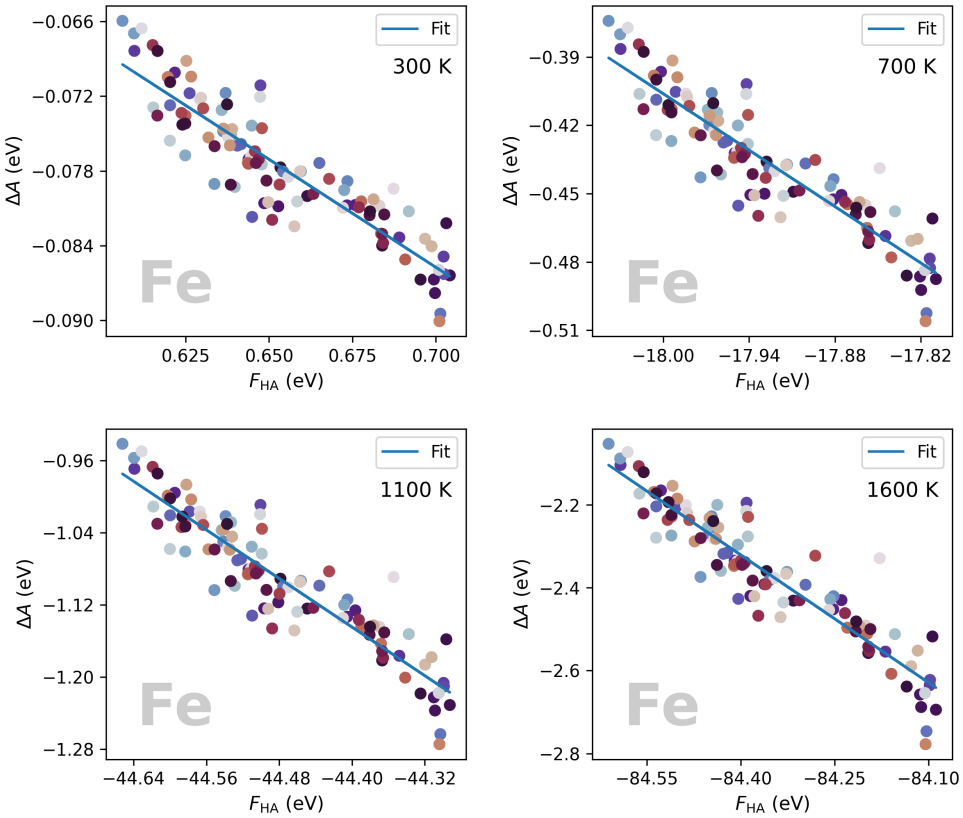

Figure 4.17: Correlation between the free energy from the harmonic approximation (F_{HA}) and the anharmonic contribution (ΔA) for Fe at 300 K, 700 K, 1100 K and 1600 K. Each color represents a selected potential. Similar to the case of W, a linear correlation is observed with R^2 values ranging from 0.81 to 0.87, where the relative position of each potential is independent of temperature. The fitted model is plotted in blue.

anharmonic contributions to the free energy profile, this observation allows for the estimation of anharmonicity directly from the harmonic approximation without any sampling, thereby significantly reducing computational cost. This approach can be applied to explore the free energy-related properties of materials, as well as for the rapid calibration of the numerically expensive ML potentials.

4.4 Conclusion of the chapter

This chapter focuses on the applications of the computational framework that integrates ML techniques from Chapter 2 with adaptive sampling methods for free energy estimation from Chapter 3. Within this framework, ML force fields improve the accuracy of atomic-scale simulations to a level comparable to that of DFT calculations. However, their implementation in sampling simulations is often restricted by computational costs, which are several orders of magnitude higher than those of traditional potentials. The fast and robust BABF methods help

overcome these limitations and enable the exploration of high-temperature free energy profiles with DFT-level accuracy. This capability facilitates the investigation of material properties at elevated temperatures.

Predicting the high-temperature properties and behaviors of defects in metallic materials represents a challenging yet essential task in a number of fields, including nuclear energy, aeronautics, and high-temperature manufacturing. Therefore, the author applied the aforementioned computational framework to address the following unresolved issues:

- For bcc W (Section 4.1), (i) thermodynamic properties such as linear expansion, bulk modulus, and elastic constants are predicted at various temperatures up to the melting point. The computations are validated by perfect consistency with experimental results at temperatures below 2100 K. (ii) Binding free energy for di-vacancies is investigated at finite temperatures to explain the temperature dependence of void formation observed in experiments.
- For Ta-Ti-V-W HEAs (Section 4.2), a new material under design, (i) the elastic constants of alloys with various compositions are calculated at different temperatures up to their melting points. (ii) The finite-temperature formation free energy of mono-vacancy is studied for different local atomic environments.
- For bcc W and bcc Fe, a correlation between harmonic and anharmonic contributions to the free energy is identified, offering a shortcut from the simple harmonic approximation to the full anharmonicity at finite temperatures without sampling (Section 4.3).

All the work above relies on the high accuracy of the ML force fields and the efficiency of the BABF-based sampling methods. Within these applications, the two-part computational framework is also validated by the experimental observations and proven to be applicable to multi-component systems.

Chapter 5

Conclusion and outlook

This thesis presents a comprehensive computational framework that effectively integrates machine learning force fields with advanced sampling and free energy estimation methods to predict finite-temperature properties of materials with accuracy comparable to DFT calculations. This framework addresses enduring challenges in computational materials science, specifically the efficient sampling of free energy surfaces at finite temperatures and the accurate estimation of free energy. The DFT-level accuracy of this framework is ensured by ML potentials, which, however, are constrained by their high computational costs. In comparison to empirical potentials, these are slower by a factor of several tens to four orders of magnitude. To make the investigation of free energy profiles with ML potentials feasible even at elevated temperatures, this thesis concentrates on two primary objectives: (i) developing ML potentials with an enhanced accuracy-efficiency trade-off, and (ii) devising faster and more robust sampling methods for free energy computations. As illustrated in Fig. 5.1, the advancements regarding these two objectives are elaborated in Chapter 2 and Chapter 3, respectively. The applications of the overall computational framework, which combines these novel techniques, are presented in Chapter 4.

The second chapter lays the groundwork for constructing and optimizing ML potentials, focusing on three key factors: the database, the representation of local atomic environments, and the regression model. The contributions in this chapter are multifaceted:

- Database enhancement: databases for tungsten, molybdenum, and Ta-Ti-V-W high-entropy alloys were completed based on the concept of active learning.
- Novel descriptors: a new Fourier-sampled n-body descriptor was introduced, significantly improving the representation of LAEs. This descriptor captures higher-order geometrical features, resulting in a 55 % improvement in force prediction accuracy compared to the widely used bispectrum SO(4) descriptor.
- Advanced regression techniques: various sparse points selection methods were implemented for the kernel regression model, optimizing the accuracy-cost trade-off. Additionally, short-range corrections using the ZBL potential were incorporated to prevent

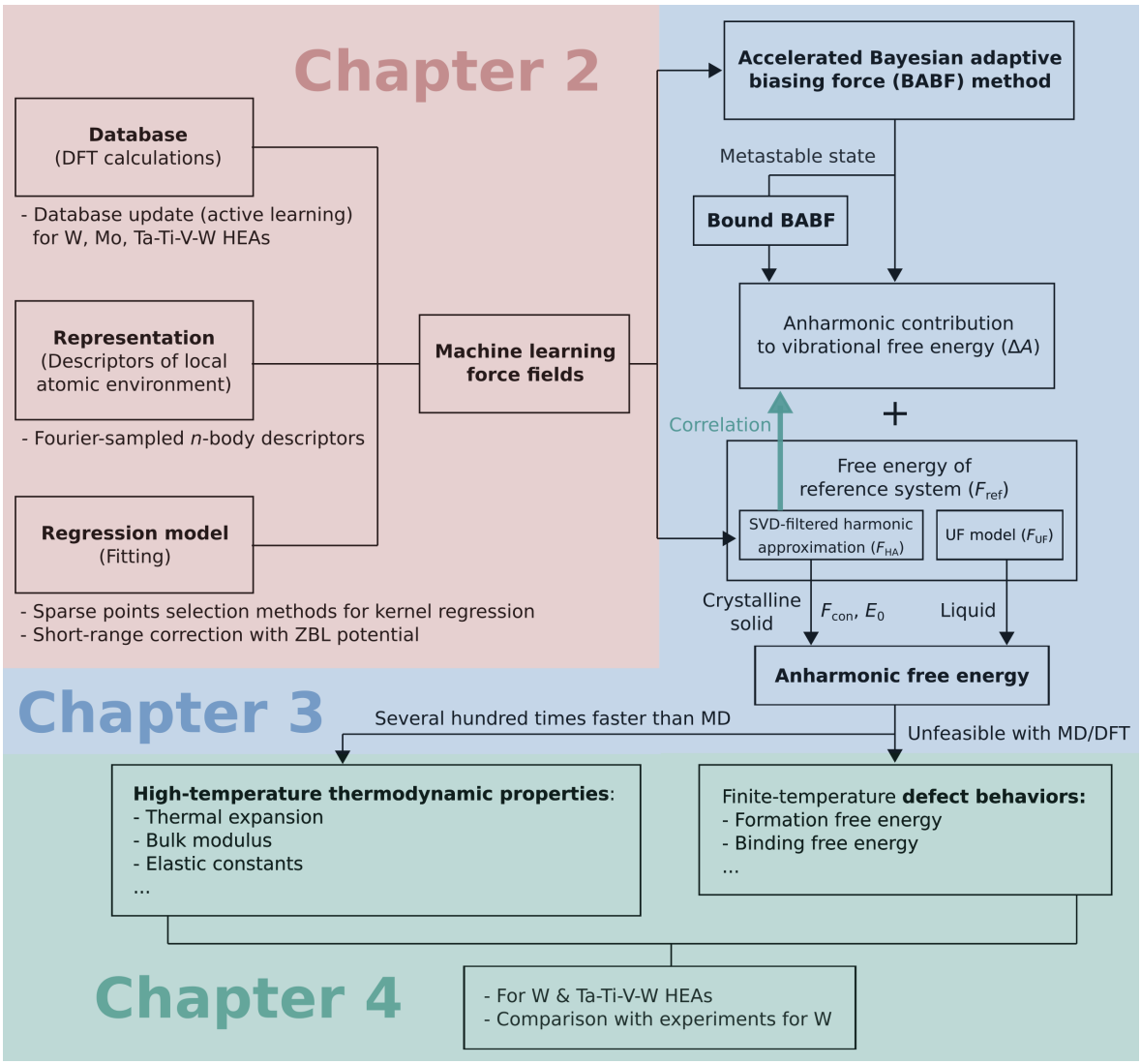

Figure 5.1: Overall workflow and achievements of the current thesis.

unphysical behavior at very short interatomic distances.

The advancements in this chapter led to the development of highly accurate ML potentials for tungsten and Ta-Ti-V-W high entropy alloys, which are employed in the subsequent part of this work. With the development of computational techniques, such as faster and larger computer clusters, these ML force fields can also be directly applied in MD simulations, for which we have developed the MiLaDy-Lammps plugin for Lammps [227]. This application allows for the investigation of physical microstructural evolution, thereby building a comprehensive mechanism from a dynamic perspective.

The third chapter delves into the computation of free energy differences, with a particular emphasis on anharmonic contributions. The extent of anharmonicity is a pivotal aspect of materials science, particularly at finite temperatures. The chapter tackles the challenges asso-

ciated with the generally high computational costs of ML potentials, proposing novel methods to overcome these hurdles:

- Accelerated Bayesian adaptive biasing force (BABF) method: this method provides a robust approach for computing the anharmonic free energy of crystalline solids, achieving an accuracy of 0.1 meV/atom approximately 100 times faster than traditional TI methods, and it is roughly 800 times faster than conventional MD simulations for determining the bulk modulus (the second derivative of free energy) at high temperatures ($> 3000\,\mathrm{K}$).
- Bound BABF Method: an unbiased free energy estimator for metastable states was introduced, using constrained sampling to prevent transitions between different energy basins, ensuring accurate free energy calculations for specified configurations.

Moreover, a two-stage BABF method was demonstrated feasible for fluid-phase systems, and its accuracy and efficiency were validated by computing the free energy difference between the ideal gas and the UF liquid model. The enhanced efficiency of these methods enables the practical application of ML potentials to the exploration of finite-temperature free energy profiles with DFT-level accuracy at a reasonable computational cost.

The fourth chapter demonstrates the practical applications of the developed computational framework, integrating the ML techniques with the free energy estimation and adaptive sampling methods from the previous chapters. The framework was applied to predict the finite-temperature properties of several materials at temperatures up to their melting points, yielding significant findings:

- Bcc tungsten: the framework predicted thermodynamic properties such as linear expansion, bulk modulus, and elastic constants up to the melting point. These predictions were validated against experimental results, showing excellent consistency. Additionally, the binding free energy of di-vacancies at finite temperatures was investigated, offering insights into the temperature dependence of void formation observed in experiments, which cannot be explained by existing DFT calculations.
- Ta-Ti-V-W high entropy alloys: for this new material under design, the elastic constants of alloys with various compositions were calculated at different temperatures up to their melting points. The formation free energy of mono-vacancies was also studied, offering valuable data for material design and optimization.
- Correlation between harmonic and anharmonic contributions: For bcc tungsten and bcc iron, a correlation was identified between harmonic and anharmonic free energy contributions. This finding offers a shortcut from the simple harmonic approximation to full anharmonicity at finite temperatures, significantly reducing computational effort.

These applications validate the accuracy and efficiency of the computational framework, demonstrating its capability to predict material properties with high precision and its applicability to multi-component systems.

The current approach can be further improved in the following aspects. For the ML potentials, we plan to (i) employ the spline interpolation for the Fourier-sampled n-body descriptors, as this representation can drastically reduce the numerical time required for force evaluation in simulations by one order of magnitude for the 2-body case and two orders of magnitude for the 3-body case; and (ii) implement the incomplete but very fast high-order kernels as explained in Subsection 2.2.4.3. In regard to the estimation of free energy from adaptive sampling, we recognize that the initialization process significantly influences the overall convergence performance. In the next work, we will adopt the multi-state Bennett acceptance ratio (MBAR) estimator [308] for the initial sampling. This approach involves performing a series of simulations at different values of an external thermodynamic parameter (ζ in this work). The MBAR estimator provides an initial estimate for the mean force by solving a large system of nonlinear equations, which can be addressed using a stochastic approximation method [309]. This estimator is optimal because its variance is lower than that of a large class of estimators [308, 145]. We anticipate that this optimization of the initialization process will greatly enhance the convergence speed.

In conclusion, the research presented in this thesis has established a powerful and efficient computational framework that combines ML potentials with advanced sampling methods for the estimation of free energy. This framework addresses key challenges in predicting the properties of materials at high temperatures, achieving unprecedented efficiency for the same level of accuracy. It opens new avenues for exploring the complex behaviors of materials under extreme conditions, paving the way for new discoveries and innovations in the field. Beyond materials research, this computational workflow is also applicable in chemistry and biology, such as calculating the free energy of ligand-protein binding processes for drug design [310]. We are enthusiastic about exploring further applications in these domains.

Résumé étendu en français

Les propriétés et le comportement des matériaux dans des conditions extrêmes sont essentiels pour les systèmes énergétiques tels que les réacteurs de fission et de fusion. Cependant, prédire avec précision les propriétés des matériaux à haute température reste un défi. Les mesures directes de ces propriétés sont limitées par les instruments expérimentaux, et les simulations à l'échelle atomique basées sur des champs de force (également connus sous le nom de "potentiels interatomiques") empiriques sont souvent peu fiables en raison d'un manque de précision. Ce problème peut être résolu à l'aide de techniques d'apprentissage statistique, qui ont récemment vu leur utilisation exploser en science des matériaux. Les champs de force construits par apprentissage statistique (aussi appelé machine learning, ML) atteignent le degré de précision des calculs *ab initio* ; cependant, leur mise en œuvre dans les méthodes d'échantillonnage est limitée par des coûts de calcul élevés, généralement supérieurs de plusieurs ordres de grandeur à ceux des champs de force traditionnels.

Pour surmonter cette limitation, cette thèse présente un cadre de calcul complet qui intègre efficacement des champs de force par apprentissage statistique avec des méthodes avancées d'échantillonnage et d'estimation de l'énergie libre pour prédire les propriétés à température finie des matériaux avec une précision comparable aux calculs DFT. Ce cadre répond aux défis persistants de la science numérique des matériaux, en particulier l'échantillonnage efficace des surfaces d'énergie libre à des températures finies et l'estimation précise de l'énergie libre. La précision de ce cadre au niveau *ab initio* est assurée par les potentiels ML, qui sont toutefois limités par leurs coûts de calcul élevés. Afin de rendre possible le calcul des profils d'énergie libre avec des champs de force ML même à des températures élevées, deux objectifs sont poursuivis dans cette thèse : (i) développer des champs de force par apprentissage statistique avec un meilleur compromis précision-efficacité et (ii) créer des méthodes accélérées d'échantillonnage de l'énergie libre afin de faciliter l'utilisation de champs de force d'apprentissage statistique coûteux en termes de calcul. Les avancées concernant ces deux objectifs sont élaborées respectivement dans le Chapitre 2 et le Chapitre 3. Les applications de ce cadre de calcul, qui combine ces nouvelles techniques, sont présentées dans le Chapitre 4.

Pour le premier objectif, nous améliorons la construction des champs de force par apprentissage statistique en nous concentrant sur trois facteurs clés : (i) la base de données contenant l'énergie, les forces et les contraintes des configurations calculées par *ab initio*, (ii) le descripteur représentant les environnements atomiques locaux, et (iii) le modèle de régression permettant

de déterminer les paramètres reliant le descripteur à l'énergie, dont la dérivée correspond aux forces. La définition de ces concepts, ainsi que les optimisations que nous proposons dans le cadre de la régression par processus gaussien, sont présentées en détail dans le deuxième chapitre :

- Amélioration de la base de données : les bases de données pour le tungstène, le molybdène et les alliages à haute entropie Ta-Ti-V-W ont été complétées sur la base du concept d'apprentissage actif.
- Nouveaux descripteurs : un nouveau descripteur n-corps basés sur des noyaux échantillonnés par la transformée de Fourier a été introduit, améliorant de manière significative la représentation des environnements atomiques locaux. Ce descripteur capture des caractéristiques géométriques d'ordre supérieur, ce qui se traduit par une amélioration de 55 % de la précision de la prédiction de la force par rapport au descripteur à bispectre SO(4) largement utilisé.
- Techniques de régression avancées : diverses méthodes de sélection de points épars (sparse points) ont été mises en œuvre pour le modèle de régression par noyau, afin d'optimiser le compromis précision/coût. En outre, des corrections à courte portée utilisant le potentiel ZBL ont été incorporées pour éviter un comportement non physique à de très courtes distances interatomiques.

Les avancées réalisées dans ce chapitre ont permis de développer des potentiels ML très précis pour le tungstène et les alliages à haute entropie de Ta-Ti-V-W, qui sont utilisés dans la partie suivante de ce travail. Avec le développement des techniques de calcul, telles que les clusters de calcul plus rapides et plus puissants, ces champs de force ML peuvent également être directement appliqués dans les simulations de dynamique moléculaire (DM), pour lesquelles nous avons développé le plugin MiLaDy-Lammps pour le logiciel communautaire de DM Lammps. Cette application permet d'étudier l'évolution physique de la microstructure et ainsi de construire un mécanisme complet d'un point de vue dynamique.

Pour le deuxième objectif, nous développons un schéma d'échantillonnage bayésien rapide et robuste pour estimer l'énergie libre anharmonique, qui est cruciale pour comprendre les effets de la température sur les solides cristallins, à l'aide d'une méthode de force de biais adaptative améliorée. En se concentrant sur le calcul précis des différences d'énergie libre, le troisième chapitre aborde les défis liés aux coûts de calcul généralement élevés des potentiels par apprentissage statistique, en proposant de nouvelles méthodes pour surmonter ces obstacles. Ces approches effectuent une intégration thermodynamique à partir d'un système de référence harmonique, où les instabilités numériques associées aux fréquences nulles sont éliminées.

- Méthode BABF (Bayesian adaptive biasing force) accélérée : basée sur l'échantillonnage bayésien, cette méthode offre une approche robuste pour calculer l'énergie libre anharmonique des solides cristallins, qui est environ 100 fois plus rapide que les méthodes d'intégration thermodynamique traditionnelles pour atteindre une précision de 0.1 meV/atome, et environ 800 fois plus rapide que les simulations DM conventionnelles pour déterminer le module d'élasticité isostatique (la dérivée seconde de l'énergie libre) à des températures élevées (> 3000 K).

- Méthode Bound BABF : un estimateur d'énergie libre sans biais pour les états métastables a été introduit, basé sur un échantillonnage contraint afin d'empêcher les transitions entre différents bassins d'énergie, garantissant ainsi des calculs d'énergie libre précis pour des configurations spécifiées.

La méthode d'échantillonnage proposée améliore considérablement la vitesse de convergence et la précision globale. En outre, il a été démontré qu'une méthode BABF en deux étapes était réalisable pour les systèmes en phase fluide, et sa précision et son efficacité ont été validées en calculant la différence d'énergie libre entre le gaz idéal et le modèle de liquide Uhlenbeck-Ford. L'efficacité accrue de ces méthodes permet l'application pratique des potentiels ML à l'exploration des profils d'énergie libre à température finie avec une précision de niveau DFT à un coût de calcul raisonnable.

La méthode BABF accélérée permet de prédire les propriétés thermodynamiques des matériaux métalliques à des températures inaccessibles expérimentalement, jusqu'à leur point de fusion, avec une précision des calculs *ab initio* grâce à l'utilisation de champs de force d'apprentissage statistique. L'extension de cette méthode, Bound BABF, permet d'échantillonner un état métastable spécifique sans transition entre différents bassins d'énergie, fournissant ainsi l'énergie libre de formation et de liaison d'une configuration défectueuse. Le quatrième chapitre démontre les applications pratiques du cadre de calcul développé, en intégrant les techniques de ML avec l'estimation de l'énergie libre et les méthodes d'échantillonnage adaptatif des chapitres précédents. Le cadre de calcul a été appliqué pour prédire les propriétés à température finie de plusieurs matériaux à des températures jusqu'à leur point de fusion, ce qui a donné des résultats significatifs :

- Tungstène cubique centré : le cadre a prédit les propriétés thermodynamiques telles que l'expansion linéaire, le module de masse et les constantes élastiques jusqu'au point de fusion. Ces prédictions ont été validées par rapport aux résultats expérimentaux, montrant une excellente cohérence. En outre, l'énergie libre de liaison des bilacunes à des températures finies a été étudiée, permettant d'élucider la dépendance de la formation des cavités à la température ainsi que le mécanisme de ce phénomène, qui ne peut pas être expliqué par les calculs *ab initio* existants.
- Alliages à haute entropie de Ta-Ti-V-W : pour ce nouveau matériau en cours de conception, les constantes élastiques d'alliages de diverses compositions ont été calculées à différentes températures jusqu'à leur point de fusion. Le profil d'énergie libre des monolacunes a également été étudié pour la première fois, offrant des données précieuses pour la conception et l'optimisation des matériaux.
- Corrélation entre les contributions harmoniques et anharmoniques : pour le tungstène cubique centré et le fer cubique centré, une corrélation a été identifiée entre les contributions harmoniques et anharmoniques à l'énergie libre. Cette découverte permet d'établir un lien direct entre la simple approximation harmonique et l'anharmonicité totale à des températures finies, réduisant ainsi de manière significative les efforts de calcul nécessaires.

Ces applications valident la précision et l'efficacité numérique du cadre de calcul proposé,

démontrant sa capacité à prédire les propriétés des matériaux avec une grande précision et son applicabilité à des systèmes multi-composants. Ce cadre de calcul ouvre de nombreuses perspectives pour une prédiction fiable des propriétés des matériaux à température finie.

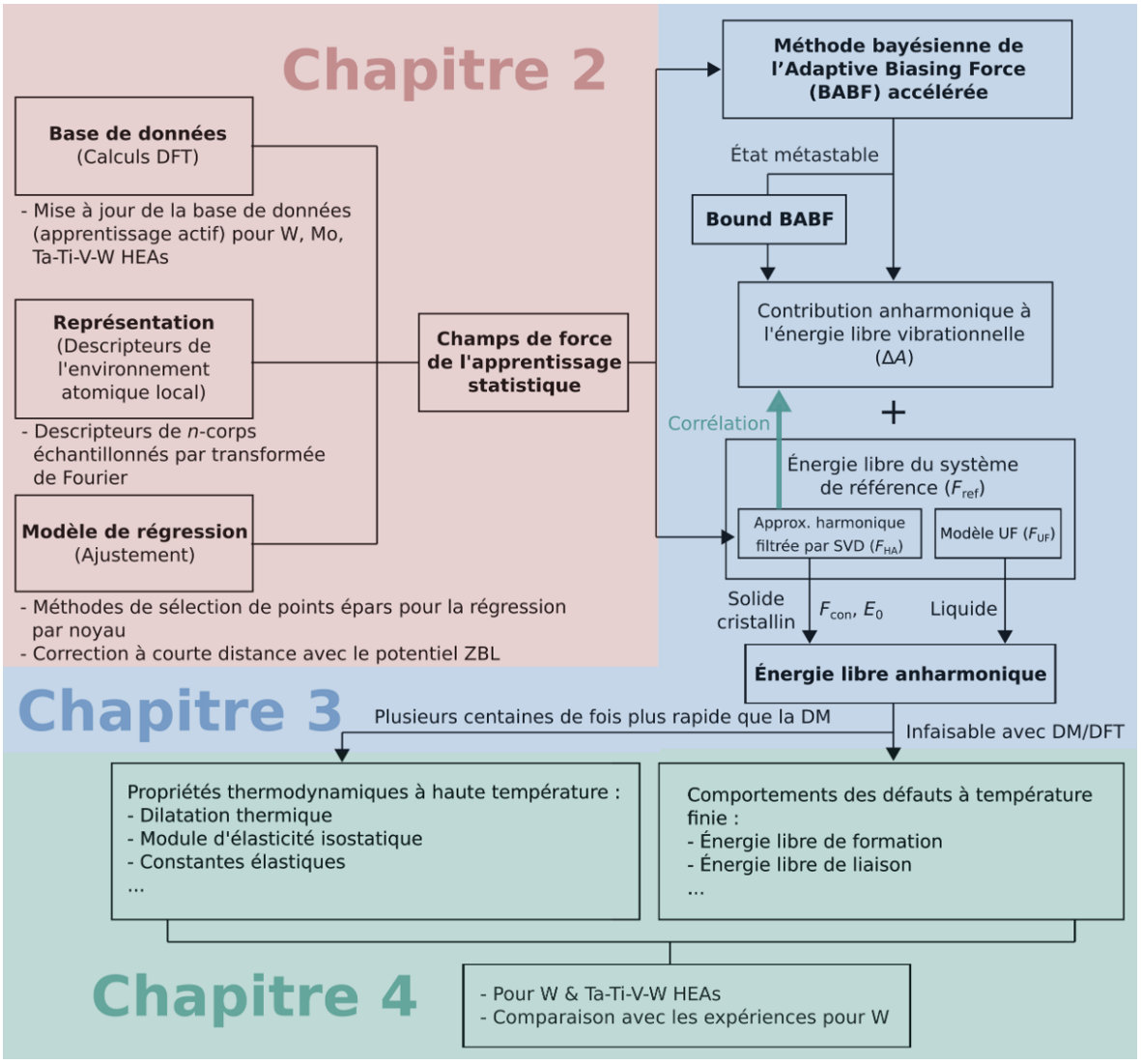

Figure 5.2: Résumé illustratif de cette thèse.

En conclusion, les travaux présentés dans cette thèse ont établi un cadre de calcul puissant et efficace, combinant des champs de force issus de l'apprentissage statistique et des méthodes d'échantillonnage avancées pour l'estimation de l'énergie libre. Ce cadre permet de relever les principaux défis liés à la détermination des propriétés des matériaux à température finie, tout en atteignant une efficacité sans précédent pour un même niveau de précision. La précision et l'efficacité numérique de ce cadre ouvrent de nombreuses perspectives pour la prédiction fiable des comportements complexes des matériaux à haute température, ouvrant ainsi la voie à de nouvelles découvertes et innovations dans le domaine des matériaux sous conditions extrêmes, qui sont essentielles dans les développements liés à l'industrie de l'énergie atomique (fission et fusion). Au-delà de la recherche sur les matériaux, ce processus de calcul est également applicable en chimie et en biologie, par exemple pour le calcul de l'énergie libre des processus

de liaison ligand-protéine dans la conception de médicaments. Nous sommes enthousiastes à l'idée d'explorer de nouvelles applications dans ces domaines.

References

[1] M. Gilbert, K. Arakawa, Z. Bergstrom, M. Caturla, S. Dudarev, F. Gao, A. Goryaeva, S. Hu, X. Hu, R. Kurtz, A. Litnovsky, J. Marian, M.-C. Marinica, E. Martinez, E. Marquis, D. Mason, B. Nguyen, P. Olsson, Y. Osetskiy, D. Senor, W. Setyawan, M. Short, T. Suzudo, J. Trelewicz, T. Tsuru, G. Was, B. Wirth, L. Yang, Y. Zhang, and S. Zinkle, "Perspectives on multiscale modelling and experiments to accelerate materials development for fusion," J. Nucl. Mater., vol. 554, p. 153113, 2021.

[2] J. Friedel, "Electronic structure of primary solid solutions in metals," Adv. Phys., vol. 3, p. 446, 1954.

[3] F. Ducastelle and F. Cyrot-Lackmann, "Moments developments and their application to the electronic charge distribution of d bands," J. Phys. Chem. Solids, vol. 31, pp. 1295–1306, 1970.

[4] M. W. Finnis and J. E. Sinclair, "A simple empirical N-Body potential for transition metals," Philos. Mag. A, vol. 50, p. 45, 1984.

[5] M. C. Desjonquères and D. Spanjaard, Concepts in Surface Physics. New York: Springer-Verlag, 1993.

[6] M. S. Daw and M. I. Baskes, "Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals," Phys. Rev. B, vol. 29, p. 6443, 1984.

[7] G. J. Ackland and R. Thetford, "An improved N-body semi-empirical model for body-centred cubic transition metals," Philos. Mag. A, vol. 56, p. 15, 1987.

[8] G. J. Ackland, M. I. Mendelev, D. J. Srolovitz, S. Han, and A. V. Barashev, "Development of an interatomic potential for phosphorus impurities in α-iron," J. Phys.: Condens. Matter, vol. 16, no. 27, p. S2629, 2004.

[9] M. Mrovec, D. Nguyen-Manh, D. G. Pettifor, and V. Vitek, "Bond-order potential for molybdenum: Application to dislocation behavior," Phys. Rev. B, vol. 69, no. 9, p. 094115, 2004.

[10] M. Mrovec, R. Gröger, A. G. Bailey, D. Nguyen-Manh, C. Elsässer, and V. Vitek, "Bond-order potential for simulations of extended defects in tungsten," Phys. Rev. B, vol. 75, no. 10, p. 104119, 2007.

[11] T. Ahlgren, K. Heinola, N. Juslin, and A. Kuronen, "Bond-order potential for point and extended defect simulations in tungsten," J. Appl. Phys., vol. 107, no. 3, p. 033516, 2010.

[12] M. Mrovec, D. Nguyen-Manh, C. Elsässer, and P. Gumbsch, "Magnetic Bond-Order Potential for Iron," Phys. Rev. Lett., vol. 106, no. 24, p. 246402, 2011.

[13] L. Malerba, M.-C. Marinica, N. Anento, C. Björkas, H. Nguyen, C. Domain, F. Djurabekova, P. Olsson, K. Nordlund, A. Serra, D. Terentyev, F. Willaime, and C. S. Becquart, "Comparison of empirical interatomic potentials for iron applied to radiation damage studies," J. Nucl. Mater., vol. 406, no. 1, p. 19, 2010.

[14] P. A. Gordon, T. Neeraj, and M. I. Mendelev, "Screw dislocation mobility in bcc metals: a refined potential description for α-fe," Philos. Mag., vol. 91, no. 30, p. 3931, 2011.

[15] M. I. Baskes, "Modified embedded-atom potentials for cubic materials and impurities," Phys. Rev. B, vol. 46, p. 2727, 1992.

[16] M. S. Daw and M. I. Baskes, "Semiempirical, quantum mechanical calculation of hydrogen embrittlement in metals," Phys. Rev. Lett., vol. 50, p. 1285, 1983.

[17] V. Rosato, M. Guillope, and B. Legrand, "Thermodynamical and structural properties of f.c.c. transition metals using a simple tight-binding model," Philos. Mag. A, vol. 59, p. 321, 1989.

[18] J. Luo and B. Legrand, "Multilayer relaxation at surfaces of body-centered-cubic transition metals," Phys. Rev. B, vol. 38, p. 1728, 1988.

[19] M. S. Daw, S. M. Foiles, and M. I. Baskes, "The embedded-atom method: a review of theory and applications," Mater. Sci. Reports, vol. 9, no. 7-8, p. 251, 1993.

[20] M.-C. Marinica, F. Willaime, and J.-P. Crocombette, "Irradiation-induced formation of nanocrystallites with $c15$ laves phase structure in bcc iron," Phys. Rev. Lett., vol. 108, p. 025501, 2012.

[21] P. M. Derlet, D. Nguyen-Manh, and S. L. Dudarev, "Multiscale modeling of crowdion and vacancy defects in body-centered-cubic transition metals," Phys. Rev. B, vol. 76, p. 054107, 2007.

[22] M.-C. Marinica, L. Ventelon, M. R. Gilbert, L. Proville, S. L. Dudarev, J. Marian, G. Bencteux, and F. Willaime, "Interatomic potentials for modelling radiation defects and dislocations in tungsten," J. Phys.: Cond. Mater., vol. 25, no. 39, p. 395502, 2013.

[23] M. I. Baskes, "Modified embedded-atom potentials for cubic materials and impurities," Phys. Rev. B, vol. 46, pp. 2727–2742, Aug 1992.

[24] D. G. Pettifor, "New many-body potential for the bond order," Phys. Rev. Lett., vol. 63, pp. 2480–2483, Nov 1989.

[25] A. Van Duin, S. Dasgupta, F. Lorant, and W. Goddard, "Reaxff: A reactive force field for hydrocarbons," J. Phys. Chem. A, vol. 105, no. 41, pp. 9396–9409, 2001.

[26] T.-R. Shan, B. D. Devine, T. W. Kemper, S. B. Sinnott, and S. R. Phillpot, "Charge-optimized many-body potential for the hafnium/hafnium oxide system," Phys. Rev. B, vol. 81, p. 125328, 2010.

[27] A. P. Bartók, Gaussian Approximation Potential : an interatomic potential derived from first principles Quantum Mechanics. PhD thesis, University of Cambridge, 2009.

[28] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett., vol. 98, p. 146401, 2007.

[29] D. Dragoni, T. D. Daff, G. Csányi, and N. Marzari, "Achieving dft accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron," Phys. Rev. Mater., vol. 2, p. 013808, 2018.

[30] W. J. Szlachta, A. P. Bartók, and G. Csányi, "Accuracy and transferability of gaussian approximation potential models for tungsten," Phys. Rev. B, vol. 90, p. 104108, 2014.

[31] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys., vol. 134, no. 7, p. 074106, 2011.

[32] J. Behler, "Perspective: Machine learning potentials for atomistic simulations," J. Chem. Phys., vol. 145, no. 17, p. 170901, 2016.

[33] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett., vol. 104, p. 136403, 2010.

[34] M. A. Wood and A. P. Thompson, "Quantum-accurate molecular dynamics potential for tungsten," arXiv preprint arXiv:1702.07042, 2017.

[35] M. A. Wood and A. P. Thompson, "Extending the accuracy of the snap interatomic potential form," J. Chem. Phys., vol. 148, no. 24, 2018.

[36] R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Phys. Rev. B, vol. 99, no. 1, p. 014104, 2019.

[37] R. Drautz, "Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer," Phys. Rev. B, vol. 102, no. 2, p. 024104, 2020.

[38] A. M. Goryaeva, J.-B. Maillet, and M.-C. Marinica, "Towards better efficiency of interatomic linear machine learning potentials," Comp. Mater. Sci., vol. 166, pp. 200 – 209, 2019.

[39] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun., vol. 8, pp. 1–8, 2017.

[40] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, "Machine learning unifies the modeling of materials and molecules," Sci. Adv., vol. 3, no. 12, p. e1701816, 2017.

[41] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, "Towards exact molecular dynamics simulations with machine-learned force fields," Nat. Commun., vol. 9, pp. 1–10, 2018.

[42] B. Grabowski, Y. Ikeda, P. Srinivasan, F. Körmann, C. Freysoldt, A. I. Duff, A. Shapeev, and J. Neugebauer, "*Ab initio* vibrational free energies including anharmonicity for multicomponent alloys," npj Computat. Mater., vol. 5, no. 1, pp. 1–6, 2019.

[43] A. Forslund and A. Ruban, "*Ab initio* surface free energies of tungsten with full account of thermal excitations," Phys. Rev. B, vol. 105, no. 4, p. 045403, 2022.

[44] J. H. Jung, P. Srinivasan, A. Forslund, and B. Grabowski, "High-accuracy thermodynamic properties to the melting point from *ab initio* calculations aided by machine-learning potentials," Npj Comput. Mater., vol. 9, no. 1, pp. 1–12, 2023.

[45] A. M. Goryaeva, C. Domain, A. Chartier, A. Dézaphie, T. D. Swinburne, K. Ma, M. Loyer-Prost, J. Creuze, and M.-C. Marinica, "Compact a15 frank-kasper nano-phases at the origin of dislocation loops in face-centred cubic metals," Nat. Commun., vol. 14, no. 1, p. 3003, 2023.

[46] T. Jourdan, A. Goryaeva, and M.-C. Marinica, "Preferential nucleation of dislocation loops under stress explained by a15 frank-kasper nanophases in aluminum," Phys. Rev. Lett., vol. 132, no. 22, p. 226101, 2024.

[47] P. Hauseux, T.-T. Nguyen, A. Ambrosetti, K. S. Ruiz, S. P. A. Bordas, and A. Tkatchenko, "From quantum to continuum mechanics in the delamination of atomically-thin layers from substrates," Nat. Commun., vol. 11, pp. 1–8, 2020.

[48] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," Proc. Natl. Acad. Sci. U. S. A., vol. 106, no. 3, pp. 697–702, 2009.

[49] A. M. Goryaeva, C. Lapointe, C. Dai, J. Dérès, J.-B. Maillet, and M.-C. Marinica, "Reinforcing materials modelling by encoding the structures of defects in crystalline solids into distortion scores," Nat. Commun., vol. 11, p. 4691, 2020.

[50] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B, vol. 87, p. 184115, 2013.

[51] H. Huo and M. Rupp, "Unified representation of molecules and crystals for machine learning," Mach. Learn.: Sci. Technol., vol. 3, no. 4, p. 045017, 2022.

[52] B. Onat, C. Ortner, and J. R. Kermode, "Sensitivity and dimensionality of atomic environment representations used for machine learning interatomic potentials," J. Chem. Phys., vol. 153, no. 14, p. 144106, 2020.

[53] A. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," Multiscale Model. Simul., vol. 14, no. 3, pp. 1153–1173, 2016.

[54] E. V. Podryabinkin and A. V. Shapeev, "Active learning of linearly parametrized interatomic potentials," Comput. Mater. Sci., vol. 140, pp. 171–180, Dec. 2017.

[55] M. Hirn, S. Mallat, and N. Poilvert, "Wavelet scattering regression of quantum chemical energies," Multiscale Model. Simul., vol. 15, p. 827, 2017.

[56] M. Eickenberg, G. Exarchakis, M. Hirn, and S. Mallat, "Solid harmonic wavelet scattering: Predicting quantum molecular energy from invariant descriptors of 3d electronic densities," Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[57] F. Bruneval, I. Maliyov, C. Lapointe, and M.-C. Marinica, "Extrapolating unconverged GW energies up to the complete basis set limit with linear regression," J. Chem. Theory Comput., vol. 16, no. 7, pp. 4399–4407, 2020.

[58] B. C. Yeo, D. Kim, C. Kim, and S. S. Han, "Pattern Learning Electronic Density of States," Sci. Rep., vol. 9, no. 1, p. 5879, 2019.

[59] C. van der Oord, G. Dusson, G. Cányi, and C. Ortner, "Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials," Mach. Learn.: Sci. Technol., vol. 1, no. 1, p. 015004, 2020.

[60] A. E. A. Allen, G. Dusson, C. Ortner, and G. Csányi, "Atomic permutationally invariant polynomials for fitting molecular force fields," Mach. Learn.: Sci. Technol., vol. 2, no. 2, p. 025017, 2021.

[61] G. Ferré, J.-B. Maillet, and G. Stoltz, "Permutation-invariant distance between atomic configurations," J. Chem. Phys., vol. 143, no. 10, p. 104114, 2015.

[62] G. Ferré, T. Haut, and K. Barros, "Learning molecular energies using localized graph kernels," J. Chem. Phys., vol. 146, no. 11, p. 114107, 2017.

[63] F. Noé and C. Clementi, "Kinetic distance and kinetic maps from molecular dynamics simulation," J. Chem. Theory Comp., vol. 11, no. 10, pp. 5002–5011, 2015.

[64] A. Ziletti, D. Kumar, M. Scheffler, and L. M. Ghiringhelli, "Insightful classification of crystal structures using deep learning," Nat. Commun., vol. 9, 2018.

[65] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, "Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions," Nat. Commun., vol. 10, no. 1, p. 5024, 2019.

[66] F. Noé, S. Olsson, J. Köhler, and H. Wu, "Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning," Science, vol. 365, no. 6457, 2019.

[67] A. Leitherer, A. Ziletti, and L. M. Ghiringhelli, "Robust recognition and exploratory analysis of crystal structures via bayesian deep learning," Nat. Commun., vol. 12, 2021.

[68] H. W. Chung, R. Freitas, G. Cheon, and E. J. Reed, "Data-centric framework for crystal structure identification in atomistic simulations using machine learning," Phys. Rev. Mater., vol. 6, p. 043801, 2022.

[69] A. Allera, A. M. Goryaeva, P. Lafourcade, J.-B. Maillet, and M.-C. Marinica, "Neighbors map: An efficient atomic descriptor for structural analysis," Comput. Mater. Sci., vol. 231, p. 112535, 2024.

[70] T. Fukuya and Y. Shibuta, "Machine learning approach to automated analysis of atomic configuration of molecular dynamics simulation," Comp. Mater. Sci., vol. 184, p. 109880, 2020.

[71] V. Fung, J. Zhang, E. Juarez, and B. G. Sumpter, "Benchmarking graph neural networks for materials chemistry," Npj Comput. Mater., vol. 7, 12 2021.

[72] Z. Fan and E. Ma, "Predicting orientation-dependent plastic susceptibility from static structure in amorphous solids via deep learning," Nat. Commun., vol. 12, 2021.

[73] P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer, and P. Friederich, "Graph neural networks for materials science and chemistry," Commun. Mater., vol. 3, pp. 1–18, 2022.

[74] S. Banik, D. Dhabal, H. Chan, S. Manna, M. Cherukara, V. Molinero, and S. K. Sankaranarayanan, "CEGANN: Crystal edge graph attention neural network for multiscale classification of materials environment," npj Comput. Mater., vol. 9, pp. 1–12, 2023.

[75] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials," J. Comp. Phys., vol. 285, p. 316, 2015.

[76] A. Seko, A. Takahashi, and I. Tanaka, "Sparse representation for a potential energy surface," Phys. Rev. B, vol. 90, p. 024101, 2014.

[77] A. Seko, A. Takahashi, and I. Tanaka, "First-principles interatomic potentials for ten elemental metals via compressed sensing," Phys. Rev. B, vol. 92, p. 054113, 2015.

[78] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, and S. P. Ong, "Accurate force field for molybdenum by machine learning large materials data," Phys. Rev. Mater., vol. 1, p. 043603, 2017.

[79] H. Yanxon, D. Zagaceta, B. C. Wood, and Q. Zhu, "Neural network potential from bispectrum components: A case study on crystalline silicon," J. Chem. Phys., vol. 153, no. 5, p. 054118, 2020.

[80] J. Zeng, D. Zhang, D. Lu, P. Mo, Z. Li, Y. Chen, M. Rynik, L. Huang, Z. Li, S. Shi, Y. Wang, H. Ye, P. Tuo, J. Yang, Y. Ding, Y. Li, D. Tisi, Q. Zeng, H. Bao, Y. Xia, J. Huang, K. Muraoka, Y. Wang, J. Chang, F. Yuan, S. L. Bore, C. Cai, Y. Lin, B. Wang, J. Xu, J.-X. Zhu, C. Luo, Y. Zhang, R. E. A. Goodall, W. Liang, A. K. Singh, S. Yao, J. Zhang, R. Wentzcovitch, J. Han, J. Liu, W. Jia, D. M. York, W. E, R. Car, L. Zhang, and H. Wang, "DeePMD-kit v2: A software package for deep potential models," J. Chem. Phys., vol. 159, p. 054801, 2023.

[81] H. Wang, L. Zhang, J. Han, and W. E, "DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics," Comput. Phys. Comm., vol. 228, p. 178, 2018.

[82] K. Schütt, O. Unke, and M. Gastegger, "Equivariant message passing for the prediction of tensorial properties and molecular spectra," in International Conference on Machine Learning, pp. 9377–9388, PMLR, 2021.

[83] V. G. Satorras, E. Hoogeboom, and M. Welling, "E(n) equivariant graph neural networks," in International conference on machine learning, pp. 9323–9332, PMLR, 2021.

[84] J. Brandstetter, R. D. Hesselink, E. van der Pol, E. J. Bekkers, and M. Welling, "Geometric and physical quantities improve E(3) equivariant message passing," ICLR, 2021. arXiv:2110.02905.

[85] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth, and B. Kozinsky, "Learning local equivariant representations for large-scale atomistic dynamics," Nat. Commun., vol. 14, no. 1, p. 579, 2023.

[86] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," Nat. Commun., vol. 13, no. 1, p. 2453, 2022.

[87] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csanyi, "Mace: Higher order equivariant message passing neural networks for fast and accurate force fields," Adv. Neural Inf. Process. Syst., vol. 35, pp. 11423–11436, 2022.

[88] I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. Simm, R. Drautz, C. Ortner, B. Kozinsky, and G. Csányi, "The design space of E(3)-equivariant atom-centered interatomic potentials," arXiv preprint arXiv:2205.06643, 2022.

[89] J. Gasteiger, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," ICLR, 2020.

[90] J. Gasteiger, S. Giri, J. T. Margraf, and S. Gunnemann, "Fast and uncertainty-aware directional message passing for non-equilibrium molecules," NeurIPS, 2020.

[91] A. P. Bartók and G. Csányi, "Gaussian approximation potentials: A brief tutorial introduction," Int. J. Quantum Chem., vol. 115, pp. 1051–1057, Aug. 2015.

[92] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," Ann. Statist., vol. 36, no. 3, pp. 1171–1220, 2008.

[93] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, "Machine Learning Force Fields: Construction, Validation, and Outlook," J. Phys. Chem. C, vol. 121, pp. 511–522, 2017.

[94] V. Botu and R. Ramprasad, "Adaptive machine learning framework to accelerate *ab initio* molecular dynamics," Int. J. Quantum Chem., vol. 115, no. 16, pp. 1074–1083, 2015.

[95] V. Botu and R. Ramprasad, "Learning scheme to predict atomic forces and accelerate materials simulations," Phys. Rev. B, vol. 92, p. 094306, 2015.

[96] Z. Li, J. R. Kermode, and A. De Vita, "Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces," Phys. Rev. Lett., vol. 114, p. 096405, Mar. 2015.

[97] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big Data of Materials Science: Critical Role of the Descriptor," Phys. Rev. Lett., vol. 114, p. 105503, 2015.

[98] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine Learning Force Fields," Chem. Rev., 2021.

[99] C. C. Aggarwal, Neural Networks and Deep Learning. Springer, Berlin, Heidelberg, 2018.

[100] C. E. Rasmussen, Gaussian Processes in Machine Learning. Springer, Berlin, Heidelberg, 2004.

[101] N. Bernstein, G. Csáyi, and V. L. Deringer, "De novo exploration and self-guided learning of potential-energy surfaces," Npj Comput. Mater., vol. 5, pp. 1–9, Oct. 2019.

[102] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, "Machine learning a general-purpose interatomic potential for silicon," Phys. Rev. X, vol. 8, p. 041048, Dec 2018.

[103] V. L. Deringer, M. A. Caro, and G. Csányi, "Machine learning interatomic potentials as emerging tools for materials science," Adv. Mater., vol. 31, no. 46, p. 1902765, 2019.

[104] N. W. Ashcroft and N. D. Mermin, Solid state physics. Science: Physics, Philadelphia: Saunders College, 1976.

[105] J. Rickman and R. LeSar, "Free-energy calculations in materials research," Annu. Rev. Mater. Res., vol. 32, p. 195, 2002.

[106] E. B. Tadmor and R. E. Miller, Modeling materials: continuum, atomistic and multiscale techniques. Cambridge University Press, 2011.

[107] A. Glensk, B. Grabowski, T. Hickel, and J. Neugebauer, "Understanding anharmonicity in fcc materials: From its origin to *ab initio* strategies beyond the quasiharmonic approximation," Phys. Rev. Lett., vol. 114, p. 195901, 2015.

[108] J. G. Kirkwood, "Statistical mechanics of fluid mixtures," J. Chem. Phys., vol. 3, no. 5, pp. 300–313, 1935.

[109] D. Frenkel and B. Smit, Understanding molecular simulation: from algorithms to applications. Elsevier, 2023.

[110] D. Alfè, G. Price, and M. Gillan, "Thermodynamics of hexagonal-close-packed iron under earth's core conditions," Phys. Rev. B, vol. 64, no. 4, p. 045123, 2001.

[111] L. Vocadlo and D. Alfè, "*Ab initio* melting curve of the fcc phase of aluminum," Phys. Rev. B, vol. 65, no. 21, 2002.

[112] B. Grabowski, L. Ismer, T. Hickel, and J. Neugebauer, "*Ab initio* up to the melting point: Anharmonicity and vacancies in aluminum," Phys. Rev. B, vol. 79, no. 13, p. 134106, 2009.

[113] A. I. Duff, T. Davey, D. Korbmacher, A. Glensk, B. Grabowski, J. Neugebauer, and M. W. Finnis, "Improved method of calculating *ab initio* high-temperature thermodynamic properties with application to zrc," Phys. Rev. B, vol. 91, no. 21, p. 214311, 2015.

[114] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," Multiscale Model. & Simul., vol. 14, no. 3, pp. 1153–1173, 2016.

[115] C. Chipot and A. Pohorille, Free energy calculations. Springer-Verlag Berlin Heidelberg, 2007.

[116] C. H. Bennett, "Efficient estimation of free energy differences from monte carlo data," J. Comput. Phys., vol. 22, no. 2, pp. 245 – 268, 1976.

[117] C. Abrams and G. Bussi, "Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration," Entropy, vol. 16, no. 1, pp. 163–199, 2013.

[118] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling," J. Comput. Phys., vol. 23, no. 2, pp. 187–199, 1977.

[119] J. Kästner, "Umbrella sampling," Wiley Interdiscip. Rev. Comput. Mol. Sci., vol. 1, no. 6, pp. 932–942, 2011.

[120] F. Wang and D. Landau, "Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram," Phys. Rev. E, vol. 64, no. 5, p. 056101, 2001.

[121] F. Wang and D. P. Landau, "Efficient, multiple-range random walk algorithm to calculate the density of states," Phys. Rev. Lett., vol. 86, no. 10, p. 2050, 2001.

[122] T. Huber, A. E. Torda, and W. F. Van Gunsteren, "Local elevation: a method for improving the searching properties of molecular dynamics simulation," J. Comput. Aided Mol. Des., vol. 8, no. 6, pp. 695–708, 1994.

[123] A. Laio and M. Parrinello, "Escaping free-energy minima," Proc. Natl. Acad. Sci. U. S. A., vol. 99, no. 20, pp. 12562–12566, 2002.

[124] A. Barducci, G. Bussi, and M. Parrinello, "Well-tempered metadynamics: a smoothly converging and tunable free-energy method," Phys. Rev. Lett., vol. 100, no. 2, p. 020603, 2008.

[125] A. Barducci, M. Bonomi, and M. Parrinello, "Metadynamics," Wiley Interdiscip. Rev. Comput. Mol. Sci., vol. 1, no. 5, pp. 826–843, 2011.

[126] L. Maragliano and E. Vanden-Eijnden, "A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations," Chem. Phys. Lett., vol. 426, no. 1-3, pp. 168–175, 2006.

[127] J. B. Abrams and M. E. Tuckerman, "Efficient and direct generation of multidimensional free energy surfaces via adiabatic dynamics without coordinate transformations," J. Phys. Chem. B, vol. 112, no. 49, pp. 15742–15757, 2008.

[128] T. D. Swinburne, J. Janssen, M. Todorova, G. Simpson, P. Plechac, M. Luskin, and J. Neugebauer, "Anharmonic free energy of lattice vibrations in fcc crystals from a mean-field bond," Phys. Rev. B, vol. 102, no. 10, p. 100101, 2020.

[129] E. Darve and A. Pohorille, "Calculating free energies using average force," J. Chem. Phys., vol. 115, no. 20, p. 9169, 2001.

[130] E. Darve, M. A. Wilson, and A. Pohorille, "Calculating free energies using a scaled-force molecular dynamics algorithm," Mol. Simul., vol. 28, no. 1-2, p. 113, 2002.

[131] J. Hénin and C. Chipot, "Overcoming free energy barriers using unconstrained molecular dynamics simulations," J. Chem. Phys., vol. 121, no. 7, pp. 2904–2914, 2004.

[132] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, "Adaptive biasing force method for scalar and vector free energy calculations," J. Chem. Phys., vol. 128, no. 14, 2008.

[133] J. Comer, J. C. Gumbart, J. Henin, T. Lelièvre, A. Pohorille, and C. Chipot, "The adaptive biasing force method: Everything you always wanted to know but were afraid to ask," J. Phys. Chem. B, vol. 119, no. 3, p. 1129, 2015. PMID: 25247823.

[134] T. Lelièvre, M. Rousset, and G. Stoltz, "Long-time convergence of an adaptive biasing force method," Nonlinearity, vol. 21, no. 6, p. 1155, 2008.

[135] T. Lelièvre, M. Rousset, and G. Stoltz, "Computation of free energy profiles with parallel adaptive dynamics," J. Chem. Phys., vol. 126, no. 13, p. 134111, 2007.

[136] L. Zheng, M. Chen, and W. Yang, "Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems," Proc. Natl. Acad. Sci. U. S. A., vol. 105, no. 51, pp. 20227–20232, 2008.

[137] H. Fu, X. Shao, C. Chipot, and W. Cai, "Extended adaptive biasing force algorithm. an on-the-fly implementation for accurate free-energy calculations," J. Chem. Theory Comput., vol. 12, no. 8, pp. 3506–3513, 2016.

[138] A. Lesage, T. Lelièvre, G. Stoltz, and J. Hénin, "Smoothed biasing forces yield unbiased free energies with the extended-system adaptive biasing force method," J. Phys. Chem. B, vol. 121, no. 15, pp. 3676–3685, 2017.

[139] A. Lesage, T. Lelievre, G. Stoltz, and J. Hénin, "Smoothed biasing forces yield unbiased free energies with the extended-system adaptive biasing force method," J. Phys. Chem. B, vol. 121, no. 15, pp. 3676–3685, 2017.

[140] L. Mones, N. Bernstein, and G. Csányi, "Exploration, sampling, and reconstruction of free energy surfaces with gaussian process regression," J. Chem. Theory Comput., vol. 12, no. 10, pp. 5100–5110, 2016.

[141] L. Bonati, G. Piccini, and M. Parrinello, "Deep learning the slow modes for rare events sampling," Proc. Natl. Acad. Sci. U. S. A., vol. 118, no. 44, p. e2113533118, 2021.

[142] H. Fu, H. Liu, J. Xing, T. Zhao, X. Shao, and W. Cai, "Deep-learning-assisted enhanced sampling for exploring molecular conformational changes," J. Phys. Chem. B, vol. 127, no. 46, pp. 9926–9935, 2023.

[143] L. Cao, G. Stoltz, T. Lelièvre, M.-C. Marinica, and M. Athènes, "Free energy calculations from adaptive molecular dynamics simulations with adiabatic reweighting," J. Chem. Phys., vol. 140, no. 10, 2014.

[144] P. Terrier, M.-C. Marinica, and M. Athènes, "Using bayes formula to estimate rates of rare events in transition path sampling simulations," J. Chem. Phys., vol. 143, no. 13, 2015.

[145] M. Athènes and P. Terrier, "Estimating thermodynamic expectations and free energies in expanded ensemble simulations: Systematic variance reduction through conditioning," J. Chem. Phys., vol. 146, no. 19, p. 194101, 2017.

[146] F. Cyrot-Lackmann, "On the electronic structure of liquid transitional metals," Advances in Physics, vol. 16, no. 63, pp. 393–400, 1967.

[147] C. Freysoldt, B. Grabowski, T. Hickel, J. Neugebauer, G. Kresse, A. Janotti, and C. G. Van de Walle, "First-principles calculations for point defects in solids," Rev. Mod. Phys., vol. 86, p. 253, 2014.

[148] A. M. Goryaeva, J. Dérés, C. Lapointe, P. Grigorev, T. D. Swinburne, J. R. Kermode, L. Ventelon, J. Baima, and M.-C. Marinica, "Efficient and transferable machine learning potentials for the simulation of crystal defects in bcc Fe and W," Phys. Rev. Mater., vol. 5, p. 103803, Oct. 2021.

[149] T. Lelièvre, G. Stoltz, and M. Rousset, Free energy computations: A mathematical perspective. Imperial College Press, London, 2010.

[150] A. Laio and M. Parrinello, "Escaping free-energy minima," Proc. Natl. Acad. Sci. U. S. A., vol. 99, no. 20, pp. 12562–12566, 2002.

[151] S. Chen and B. N. Brockhouse, "Lattice vibrations of tungsten," Solid State Commun., vol. 2, no. 3, pp. 73–77, 1964.

[152] A. Larose and B. Brockhouse, "Lattice vibrations in tungsten at 22 c studied by neutron scattering," Can. J. Phys., vol. 54, no. 17, pp. 1819–1823, 1976.

[153] G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," Phys. Rev. B, vol. 54, pp. 11169–11186, 1996.

[154] G. Henkelman, G. Jóhannesson, and H. Jónsson, Methods for Finding Saddle Points and Minimum Energy Paths, pp. 269–302. Dordrecht: Springer Netherlands, 2002.

[155] G. Henkelman, B. P. Uberuaga, and H. Jónsson, "A climbing image nudged elastic band method for finding saddle points and minimum energy paths," J. Chem. Phys., vol. 113, no. 22, pp. 9901–9904, 2000.

[156] G. Henkelman, "Atomistic simulations of activated processes in materials," Annu. Rev. Mater. Res., vol. 47, pp. 199–216, 2017.

[157] S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," J. Comput. Phys., vol. 117, no. 1, pp. 1–19, 1995.

[158] G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," Phys. Rev. B, vol. 54, no. 16, p. 11169, 1996.

[159] J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, and B. Kozinsky, "On-the-fly active learning of interpretable bayesian force fields for atomistic rare events," Npj Comput. Mater., vol. 6, no. 1, p. 20, 2020.

[160] Y. Xie, J. Vandermause, L. Sun, A. Cepellotti, and B. Kozinsky, "Bayesian force fields from active learning for simulation of inter-dimensional transformation of stanene," Npj Comput. Mater., vol. 7, no. 1, p. 40, 2021.

[161] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," Nat. Commun., vol. 13, no. 1, p. 2453, 2022.

[162] O. Kaba and S. Ravanbakhsh, "Equivariant networks for crystal structures," Adv. Neural Inf. Process. Syst., vol. 35, pp. 4150–4164, 2022.

[163] V. H. A. Nguyen and A. Lunghi, "Predicting tensorial molecular properties with equivariant machine learning models," Phys. Rev. B, vol. 105, no. 16, p. 165131, 2022.

[164] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett., vol. 108, p. 058301, Jan 2012.

[165] M. Rupp, "Machine learning for quantum mechanics in a nutshell," Int. J. Quantum Chem., vol. 115, no. 16, pp. 1058–1073, 2015.

[166] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. Von Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," New J. Phys., vol. 15, p. 095003, sep 2013.

[167] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, "Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials," J. Chem. Phys., vol. 148, no. 24, p. 241730, 2018.

[168] J. Behler, "Perspective: Machine learning potentials for atomistic simulations," J. Chem. Phys., vol. 145, no. 17, p. 170901, 2016.

[169] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B, vol. 87, no. 18, p. 184115, 2013.

[170] J. Byggmästar, K. Nordlund, and F. Djurabekova, "Simple machine-learned interatomic potentials for complex alloys," Phys. Rev. Mater., vol. 6, no. 8, pp. 1–11, 2022.

[171] H. Zong, G. Pilania, X. Ding, G. J. Ackland, and T. Lookman, "Developing an interatomic potential for martensitic phase transformations in zirconium by machine learning," Npj Comput. Mater., vol. 4, no. 1, pp. 1–8, 2018.

[172] J. R. Kermode, A. Gleizer, G. Kovel, L. Pastewka, G. Csányi, D. Sherman, and A. De Vita, "Low speed crack propagation via kink formation and advance on the silicon (110) cleavage plane," Phys. Rev. Lett., vol. 115, no. 13, p. 135501, 2015.

[173] K. T. Schutt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet A deep learning architecture for molecules and materials," J. Chem. Phys., vol. 148, no. 24, p. 241722, 2018.

[174] A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, "Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach," Chem. Phys. Lett., vol. 509, no. 1, pp. 1–11, 2011.

[175] W. F. Reinhart, A. W. Long, M. P. Howard, A. L. Ferguson, and A. Z. Panagiotopoulos, "Machine learning for autonomous crystal structure identification," Soft Matter, vol. 13, no. 27, pp. 4733–4745, 2017.

[176] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[177] K. Schütt, P. Kessel, M. Gastegger, K. Nicoli, A. Tkatchenko, and K.-R. Müller, "Schnetpack: A deep learning toolbox for atomistic systems," J. Chem. Theory Comput., vol. 15, no. 1, pp. 448–455, 2018.

[178] K. Choudhary and B. DeCost, "Atomistic Line Graph Neural Network for improved materials property predictions," npj Comput. Mater., vol. 7, no. 1, pp. 1–8, 2021.

[179] C. van der Oord, M. Sachs, D. P. Kovács, C. Ortner, and G. Csányi, "Hyperactive learning for data-driven interatomic potentials," Npj Comput. Mater., vol. 9, no. 1, p. 168, 2023.

[180] M. Eickenberg, G. Exarchakis, M. Hirn, S. Mallat, and L. Thiry, "Solid harmonic wavelet scattering for predictions of molecule properties," J. Chem. Phys., vol. 148, no. 24, p. 241732, 2018.

[181] E. Homer, D. Hensley, C. Rosenbrock, A. Nguyen, and G. Hart, "Machine-learning informed representations for grain boundary structures," Front. Mater., vol. 6, 2019.

[182] R. Kakarala, The bispectrum as a source of phase-sensitive invariants for Fourier descriptors: a group-theoretic approach. PhD thesis, Irvine University, 1992.

[183] D. J. C. MacKay, Information Theory, Inference, and Learning Algorithms. Cambridge: Cmbridge University, 2005.

[184] D. L. Phillips, "A technique for the numerical solution of certain integral equations of the first kind," J. ACM, vol. 9, no. 1, pp. 84–97, 1962.

[185] C. Rasmussen and C. Williams, Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning series, MIT Press, 2005.

[186] N. Aronszajn, "Theory of reproducing kernels," Trans. Am. Math. Soc., vol. 68, no. 3, pp. 337–404, 1950.

[187] C. Nyshadham, M. Rupp, B. Bekker, A. V. Shapeev, T. Mueller, C. W. Rosenbrock, G. Csányi, D. W. Wingate, and G. L. Hart, "Machine-learned multi-system surrogate models for materials prediction," Npj Comput. Mater., vol. 5, no. 1, p. 51, 2019.

[188] P. Rowe, G. Csányi, D. Alfè, and A. Michaelides, "Development of a machine learning potential for graphene," Phys. Rev. B, vol. 97, no. 5, p. 054303, 2018.

[189] V. L. Deringer and G. Csányi, "Machine learning based interatomic potential for amorphous carbon," Phys. Rev. B, vol. 95, no. 9, p. 094203, 2017.

[190] C. W. Rosenbrock, K. Gubaev, A. V. Shapeev, L. B. Pártay, N. Bernstein, G. Csányi, and G. L. Hart, "Machine-learned interatomic potentials for alloys and alloy phase diagrams," Npj Comput. Mater., vol. 7, no. 1, p. 24, 2021.

[191] G. Dhaliwal, P. B. Nair, and C. V. Singh, "Machine learned interatomic potentials using random features," Npj Comput. Mater., vol. 8, no. 1, p. 7, 2022.

[192] J. P. Darby, J. R. Kermode, and G. Csányi, "Compressing local atomic neighbourhood descriptors," Npj Comput. Mater., vol. 8, no. 1, p. 166, 2022.

[193] G. Wahba, D. R. Johnson, F. Gao, and J. Gong, "Adaptive tuning of numerical weather prediction models: Randomized GCV in three and four dimensional data assimilation.," Mon. Weather Rev., vol. 123, p. 3358, 1995.

[194] P. J. Rousseeuw and K. van Driessen, "A fast algorithm for the minimum covariance determinant estimator," Technometrics, vol. 41, no. 3, pp. 212–223, 1999.

[195] M. Hubert, M. Debruyne, and P. J. Rousseeuw, "Minimum covariance determinant and extensions," WIRES Comp. Stat., vol. 10, p. e1421, 2018.

[196] M. Hubert and M. Debruyne, "Minimum covariance determinant," WIRES Comp. Stat., vol. 2, pp. 36–43, 2010.

[197] A. Glielmo, P. Sollich, and A. De Vita, "Accurate interatomic force fields via machine learning with covariant kernels," Phys. Rev. B, vol. 95, p. 214302, Jun 2017.

[198] A. Glielmo, C. Zeni, and A. De Vita, "Efficient nonparametric n-body force fields from machine learning," Phys. Rev. B, vol. 97, p. 184307, May 2018.

[199] A. Glielmo, C. Zeni, Á. Fekete, and A. De Vita, Building nonparametric n-body force fields using gaussian process regression, pp. 67–98. Cham: Springer International Publishing, 2020.

[200] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," Adv. Neural Inf. Process. Syst., vol. 20, 2007.

[201] "Multidimensional b-spline interpolation of data on a regular grid." https://github.com/jacobwilliams/bspline-fortran.

[202] A. Glielmo, P. Sollich, and A. De Vita, "Accurate interatomic force fields via machine learning with covariant kernels," Phys. Rev. B, vol. 95, p. 214302, Jun 2017.

[203] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Incompleteness of atomic structure representations," Phys. Rev. Lett., vol. 125, p. 166001, Oct 2020.

[204] S. N. Pozdnyakov and M. Ceriotti, "Incompleteness of graph neural networks for points clouds in three dimensions," Mach. Learn.: Sci. Technol., vol. 3, p. 045020, Nov. 2022.

[205] J. F. Ziegler and J. P. Biersack, The Stopping and Range of Ions in Matter, p. 93. Boston, MA: Springer US, 1985.

[206] J. Byggmästar, A. Hamedani, K. Nordlund, and F. Djurabekova, "Machine-learning interatomic potential for radiation damage and defects in tungsten," Phys. Rev. B, vol. 100, p. 144105, Oct 2019.

[207] K. Nordlund, N. Runeberg, and D. Sundholm, "Repulsive interatomic potentials calculated using hartree-fock and density-functional theory methods," Nucl. Instrum. Methods Phys. Res. B, vol. 132, no. 1, pp. 45–54, 1997.

[208] N. W. Ashcroft and N. D. Mermin, Solid state physics. Holt-Saunders International Editions: Science : Physics, Holt, Rinehart and Winston, 1976.

[209] M. T. Dove, Introduction to lattice dynamics. No. 4, Cambridge university press, 1993.

[210] G. Bussi and D. Branduardi, "Free-energy calculations with metadynamics: Theory and practice," Reviews in Computational Chemistry Volume 28, pp. 1–49, 2015.

[211] A. Z. Guo, E. Sevgen, H. Sidky, J. K. Whitmer, J. A. Hubbell, and J. J. de Pablo, "Adaptive enhanced sampling by force-biasing using neural networks," J. Chem. Phys., vol. 148, no. 13, p. 134108, 2018.

[212] J. Baima, A. M. Goryaeva, T. D. Swinburne, J.-B. Maillet, M. Nastar, and M.-C. Marinica, "Capabilities and limits of autoencoders for extracting collective variables in atomistic materials science," Phys. Chem. Chem. Phys., vol. 24, pp. 23152–23163, 2022.

[213] A. Zhong, C. Lapointe, A. M. Goryaeva, J. Baima, M. Athènes, and M.-C. Marinica, "Anharmonic thermo-elasticity of tungsten from accelerated Bayesian adaptive biasing force calculations with data-driven force fields," Phys. Rev. Mater., vol. 7, p. 023802, Feb. 2023.

[214] M. Watanabe and W. P. Reinhardt, "Direct dynamical calculation of entropy and free energy by adiabatic switching," Phys. Rev. Lett., vol. 65, no. 26, p. 3301, 1990.

[215] W. G. Hoover and F. H. Ree, "Use of computer experiments to locate the melting transition and calculate the entropy in the solid phase," J. Chem. Phys., vol. 47, no. 12, p. 4873, 1967.

[216] P. A. Monson and D. A. Kofke, Solid-Fluid Equilibrium: Insights from Simple Molecular Models, p. 113. John Wiley & Sons, Inc., 2007.

[217] W. G. Hoover, M. Ross, K. W. Johnson, D. Henderson, J. A. Barker, and B. C. Brown, "Soft-sphere equation of state," J. Chem. Phys., vol. 52, no. 10, p. 4931, 1970.

[218] W. G. Hoover, S. G. Gray, and K. W. Johnson, "Thermodynamic properties of the fluid and solid phases for inverse power potentials," J. Chem. Phys., vol. 55, no. 3, p. 1128, 1971.

[219] D. Frenkel and A. J. C. Ladd, "New monte carlo method to compute the free energy of arbitrary solids. application to the fcc and hcp phases of hard spheres," J. Chem. Phys., vol. 81, no. 7, p. 3188, 1984.

[220] S. Fukushima, E. Ushijima, H. Kumazoe, A. Koura, F. Shimojo, K. Shimamura, M. Misawa, R. K. Kalia, A. Nakano, and P. Vashishta, "Thermodynamic integration by neural network potentials based on first-principles dynamic calculations," Phys. Rev. B, vol. 100, no. 21, p. 214108, 2019.

[221] B. Grabowski, T. Hickel, and J. Neugebauer, "Formation energies of point defects at finite temperatures," Phys. Status Solidi B, vol. 248, no. 6, p. 1295, 2011.

[222] B. Cheng and M. Ceriotti, "Computing the absolute gibbs free energy in atomistic simulations: Applications to defects in solids," Phys. Rev. B, vol. 97, no. 5, p. 054102, 2018.

[223] L. Wang, A. van de Walle, and D. Alfè, "Melting temperature of tungsten from two *ab initio* approaches," Phys. Rev. B, vol. 84, no. 9, p. 092102, 2011.

[224] J. Q. Broughton and G. H. Gilmer, "Molecular dynamics investigation of the crystal–fluid interface. i. bulk properties," J. Chem. Phys., vol. 79, no. 10, p. 5095, 1983.

[225] J. P. Ryckaert and G. Ciccotti, "Introduction of andersen's demon in the molecular dynamics of systems with constraints," J. Chem. Phys., vol. 78, no. 12, p. 7368, 1983.

[226] G. Clavier, N. Desbiens, E. Bourasseau, V. Lachet, N. Brusselle-Dupend, and B. Rousseau, "Computation of elastic constants of solids using molecular simulation: comparison of constant volume and constant pressure ensemble methods," Mol. Simul., vol. 43, no. 17, pp. 1413–1422, 2017.

[227] S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," J. Comp. Phys., vol. 117, no. 1, pp. 1 – 19, 1995.

[228] C. Kittel, Introduction to solid state physics. Wiley, 2004.

[229] M.-C. Marinica, L. Ventelon, M. Gilbert, L. Proville, S. Dudarev, J. Marian, G. Bencteux, and F. Willaime, "Interatomic potentials for modelling radiation defects and dislocations in tungsten," J. Phys. Condens. Matter, vol. 25, no. 39, p. 395502, 2013.

[230] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett., vol. 104, no. 13, p. 136403, 2010.

[231] A. M. Miksch, T. Morawietz, J. Kästner, A. Urban, and N. Artrith, "Strategies for the construction of machine-learning potentials for accurate and efficient atomic-scale simulations," Mach. Learn.: Sci. Technol., vol. 2, no. 3, p. 031001, 2021.

[232] K. Rasch, R. Siegel, and H. Schultz, "Quenching and recovery investigations of vacancies in tungsten," Philos. Mag. A, vol. 41, no. 1, pp. 91–117, 1980.

[233] C. S. Becquart and C. Domain, "*Ab initio* calculations about intrinsic point defects and He in W," Nucl. Instrum. Methods Phys. Res. B, vol. 255, no. 1, p. 23, 2007.

[234] C. S. Becquart, C. Domain, U. Sarkar, A. Debacker, and M. Hou, "Microstructural evolution of irradiated tungsten: *Ab initio* parameterisation of an okmc model," J. Nucl. Mater., vol. 403, no. 1-3, pp. 75–88, 2010.

[235] M. Muzyk, D. Nguyen-Manh, K. Kurzydłowski, N. Baluc, and S. Dudarev, "Phase stability, point defects, and elastic properties of wv and w-ta alloys," Phys. Rev. B, vol. 84, no. 10, p. 104115, 2011.

[236] D. Kato, H. Iwakiri, and K. Morishita, "Formation of vacancy clusters in tungsten crystals under hydrogen-rich condition," J. Nucl. Mater., vol. 417, no. 1-3, pp. 1115–1118, 2011.

[237] L. Ventelon, F. Willaime, C.-C. Fu, M. Heran, and I. Ginoux, "*Ab initio* investigation of radiation defects in tungsten: Structure of self-interstitials and specificity of di-vacancies compared to other bcc transition metals," J. Nucl. Mater., vol. 425, no. 1-3, pp. 16–21, 2012.

[238] K. Heinola, F. Djurabekova, and T. Ahlgren, "On the stability and mobility of di-vacancies in tungsten," Nucl. Fusion, vol. 58, no. 2, p. 026004, 2017.

[239] D. R. Mason, D. Nguyen-Manh, and C. S. Becquart, "An empirical potential for simulating vacancy clusters in tungsten," J. Phys.: Cond. Matter, vol. 29, no. 50, p. 505501, 2017.

[240] A. M. Goryaeva, J. Dérès, C. Lapointe, P. Grigorev, T. D. Swinburne, J. R. Kermode, L. Ventelon, J. Baima, and M.-C. Marinica, "Efficient and transferable machine learning potentials for the simulation of crystal defects in bcc fe and w," Phys. Rev. Mater., vol. 5, no. 10, p. 103803, 2021.

[241] R. Armiento and A. E. Mattsson, "Functional designed to include surface effects in self-consistent density functional theory," Phys. Rev. B, vol. 72, no. 8, p. 085108, 2005.

[242] A. E. Mattsson, R. Armiento, J. Paier, G. Kresse, J. M. Wills, and T. R. Mattsson, "The AM05 density functional applied to solids," J. Chem. Phys., vol. 128, no. 8, p. 084714, 2008.

[243] A. E. Mattsson and R. Armiento, "Implementing and testing the AM05 spin density functional," Phys. Rev. B, vol. 79, p. 155101, 2009.

[244] M. Methfessel and A. Paxton, "High-precision sampling for brillouin-zone integration in metals," Phys. Rev. B, vol. 40, no. 6, p. 3616, 1989.

[245] S. H. Vosko, L. Wilk, and M. Nusair, "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis," Can. J. Phys, vol. 59, p. 1200, 1980.

[246] J. P. Perdew and A. Zunger, "Self-interaction correction to density-functional approximations for many-electron systems," Phys. Rev. B, vol. 23, pp. 5048–5079, 1981.

[247] J. P. Perdew and W. Yue, "Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation," Phys. Rev. B, vol. 33, pp. 8800–8802, 1986.

[248] J. P. Perdew, "Density-functional approximation for the correlation energy of the inhomogeneous electron gas," Phys. Rev. B, vol. 33, pp. 8822–8824, 1986.

[249] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett., vol. 77, no. 18, p. 3865, 1996.

[250] T. Lelièvre, M. Rousset, and G. Stoltz, "Computation of free energy profiles with parallel adaptive dynamics," J. Chem. Phys., vol. 126, p. 134111, Apr. 2007.

[251] M. C. Abramo, C. Caccamo, D. Costa, P. V. Giaquinta, G. Malescio, G. Munaò, and S. Prestipino, "On the determination of phase boundaries via thermodynamic integration across coexistence regions," J. Chem. Phys., vol. 142, no. 21, 2015.

[252] R. Paula Leite, R. Freitas, R. Azevedo, and M. de Koning, "The uhlenbeck-ford model: Exact virial coefficients and application as a reference system in fluid-phase free-energy calculations," J. Chem. Phys., vol. 145, no. 19, 2016.

[253] S. Menon, Y. Lysogorskiy, J. Rogal, and R. Drautz, "Automated free-energy calculation from atomistic simulations," Phys. Rev. Mater., vol. 5, no. 10, p. 103801, 2021.

[254] R. A. Causey and T. J. Venhaus, "The use of tungsten in fusion reactors: a review of the hydrogen retention and migration properties," Physica scripta, vol. 2001, no. T94, p. 9, 2001.

[255] J. V. Vas, J. Pan, N. Wang, J. Xu, R. Medwal, M. Mishra, J. Y. Pae, M. V. Matham, L. Paul, and R. S. Rawat, "Plasma processed tungsten for fusion reactor first-wall material," J. Mater. Sci., vol. 56, no. 17, pp. 10494–10509, 2021.

[256] J. You, E. Visca, C. Bachmann, T. Barrett, F. Crescenzi, M. Fursdon, H. Greuner, D. Guilhem, P. Languille, M. Li, S. McIntosh, A. Müller, J. Reiser, M. Richou, and M. Rieth, "European DEMO divertor target: Operational requirements and material-design interface," Nucl. Mater. Energy, vol. 9, pp. 171–176, 2016.

[257] M. Merola, D. Loesser, A. Martin, P. Chappuis, R. Mitteau, V. Komarov, R. Pitts, S. Gicquel, V. Barabash, L. Giancarli, J. Palmer, M. Nakahira, A. Loarte, D. Campbell, R. Eaton, A. Kukushkin, M. Sugihara, F. Zhang, C. Kim, R. Raffray, L. Ferrand, D. Yao, S. Sadakov, A. Furmanek, V. Rozov, T. Hirai, F. Escourbiac, T. Jokinen, B. Calcagno, and S. Mori, "ITER plasma-facing components," Fusion Eng. Des., vol. 85, no. 10, pp. 2312–2322, 2010. Proceedings of the Ninth International Symposium on Fusion Nuclear Technology.

[258] J. You, E. Visca, T. Barrett, B. Böswirth, F. Crescenzi, F. Domptail, M. Fursdon, F. Gallay, B.-E. Ghidersa, H. Greuner, M. Li, A. Müller, J. Reiser, M. Richou, S. Roccella, and C. Vorpahl, "European divertor target concepts for DEMO: Design rationales and high heat flux performance," Nucl. Mater. Energy, vol. 16, pp. 1–11, 2018.

[259] J. H. Stathis and D. Bolef, "Elastic constants of tungsten between 4.2 and 77 K," J. Appl. Phys., vol. 51, no. 9, pp. 4770–4773, 1980.

[260] D. I. Bolef and J. De Klerk, "Elastic constants of single-crystal Mo and W between 77° and 500°K," J. Appl. Phys., vol. 33, no. 7, pp. 2311–2314, 1962.

[261] F. H. Featherston and J. R. Neighbours, "Elastic constants of tantalum, tungsten, and molybdenum," Phys. Rev., vol. 130, p. 1324, 1963.

[262] R. Lowrie and A. M. Gonas, "Single-crystal elastic properties of tungsten from 24° to 1800°C," J. Appl. Phys., vol. 38, no. 11, p. 4505, 1967.

[263] K. Wang and R. R. Reeber, "High temperature bulk moduli and self-diffusion for tantalum and tungsten," High Temp. and Mater. Sci., vol. 36, no. 2-3, 1996.

[264] S. Saxena and J. Zhang, "Thermochemical and pressure-volume-temperature systematics of data on solids, examples: tungsten and mgo," Phys. Chem. Miner., vol. 17, no. 1, pp. 45–51, 1990.

[265] Y. S. Touloukian and E. H. Buyco, Specific Heat Metallic Elements and Alloys, vol. 4. West Lafayette: CINDAS/Purdue Research Foundation, 1971.

[266] Y. S. Touloukian, R. K. Kirby, R. E. Taylor, and P. D. Desai, Thermal Expansion Metallic Elements and Alloys, vol. 12. West Lafayette: IFI/Plenum, Purdue Research Foundation, 1975.

[267] P. Varotsos and K. Alexopoulos, "Calculation of diffusion coefficients at any temperature and pressure from a single measurement. i. self diffusion," Phys. Rev. B, vol. 22, no. 6, p. 3130, 1980.

[268] C. Falter and W. Zierau, "Determination of the high-temperature bulk modulus from self-diffusion experiments demonstrated for tungsten," J. Appl. Phys., vol. 51, no. 4, pp. 2070–2073, 1980.

[269] P. Gustafson, "Evaluation of the thermodynamic properties of tungsten," Int. J. Thermophys., vol. 6, no. 4, pp. 395–409, 1985.

[270] N. Juslin and B. D. Wirth, "Interatomic potentials for simulation of He bubble formation in W," J. Nucl. Mater., vol. 432, no. 1-3, p. 61, 2013.

[271] D. R. Mason, D. Nguyen-Manh, and C. S. Becquart, "An empirical potential for simulating vacancy clusters in tungsten," J. Phys.: Condens. Matter, vol. 29, p. 505501, Nov. 2017.

[272] L. Ventelon, F. Willaime, C.-C. Fu, M. Heran, and I. Ginoux, "*Ab initio* investigation of radiation defects in tungsten: Structure of self-interstitials and specificity of di-vacancies compared to other bcc transition metals," J. Nucl. Mater., vol. 425, no. 1-3, p. 16, 2012.

[273] V. L. Deringer, M. A. Caro, and G. Csányi, "Machine Learning Interatomic Potentials as Emerging Tools for Materials Science," Adv. Mater., vol. 31, no. 46, p. 1902765, 2019.

[274] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian Process Regression for Materials and Molecules," Chem. Rev., vol. 121, pp. 10073–10141, Aug. 2021.

[275] X. Wang, Y. Wang, L. Zhang, F. Dai, and H. Wang, "A tungsten deep neural-network potential for simulating mechanical property degradation under fusion service environment," Nucl. Fusion, 2022.

[276] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," Phys. Rev. Lett., vol. 77, p. 3865, 1997.

[277] L. Proville, D. Rodney, and M.-C. Marinica, "Quantum effect on thermally activated glide of dislocations," Nat. Mater., vol. 11, no. 10, p. 845, 2012.

[278] T. D. Swinburne, P.-W. Ma, and S. L. Dudarev, "Low temperature diffusivity of self-interstitial defects in tungsten," New J. Phys., vol. 19, no. 7, p. 073024, 2017.

[279] K. Arakawa, M.-C. Marinica, S. Fitzgerald, L. Proville, D. Nguyen-Manh, S. L. Dudarev, P.-W. Ma, T. D. Swinburne, A. M. Goryaeva, T. Yamada, T. Amino, S. Arai, Y. Yamamoto, K. Higuchi, N. Tanaka, H. Yasuda, T. Yasuda, and H. Mori, "Quantum de-trapping and transport of heavy defects in tungsten," Nat. Mater., vol. 19, no. 5, pp. 508–511, 2020.

[280] J. Park, H. Huang, R. Siegel, and R. Balluffi, "A quantitative study of vacancy defects in quenched tungsten by combined field-ion microscopy and electrical resistometry," Philos. Mag. A, vol. 48, no. 3, pp. 397–419, 1983.

[281] J. Mundy, S. Rothman, N. Lam, H. Hoff, and L. Nowicki, "Self-diffusion in tungsten," Phys. Rev. B, vol. 18, no. 12, p. 6566, 1978.

[282] J. Mundy, "Electrical resistivity-temperature scale of tungsten," Philos. Mag. A, vol. 46, no. 2, pp. 345–349, 1982.

[283] X. Zhang, S. V. Divinski, and B. Grabowski, "*Ab initio* machine-learning unveils strong anharmonicity in non-arrhenius self-diffusion of tungsten," arXiv preprint arXiv:2311.00633, 2023.

[284] F. Maury, M. Biget, P. Vajda, A. Lucasson, and P. Lucasson, "Frenkel pair creation and stage i recovery in w crystals irradiated near threshold," Radiat. Eff., vol. 38, no. 1-2, pp. 53–65, 1978.

[285] T. D. Swinburne and M.-C. Marinica, "Unsupervised calculation of free energy barriers in large crystalline systems," Phys. Rev. Lett., vol. 120, no. 13, p. 135503, 2018.

[286] "Pafi: Evaluation of free energy barriers beyond harmonic tst." https://github.com/tomswinburne/pafi/.

[287] V. N. Staroverov, G. E. Scuseria, J. Tao, and J. P. Perdew, "Tests of a ladder of density functionals for bulk solids and surfaces," Phys. Rev. B, vol. 69, no. 7, p. 075102, 2004.

[288] G. I. Csonka, J. P. Perdew, A. Ruzsinszky, P. H. Philipsen, S. Lebègue, J. Paier, O. A. Vydrov, and J. G. Ángyán, "Assessing the performance of recent density functionals for bulk solids," Phys. Rev. B, vol. 79, no. 15, p. 155107, 2009.

[289] A. Glielmo, P. Sollich, and A. De Vita, "Accurate interatomic force fields via machine learning with covariant kernels," Phys. Rev. B, vol. 95, no. 21, p. 214302, 2017.

[290] B. Cantor, I. T. H. Chang, P. Knight, and A. J. B. Vincent, "Microstructural development in equiatomic multicomponent alloys," Mater. Sci. Eng. A, vol. 375-377, pp. 213–218, jul 2004.

[291] J.-W. Yeh, S.-K. Chen, S.-J. Lin, J.-Y. Gan, T.-S. Chin, T.-T. Shun, C.-H. Tsau, and S.-Y. Chang, "Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes," Adv. Eng. Mater., vol. 6, no. 5, pp. 299–303, 2004.

[292] N. A. P. Kiran Kumar, C. Li, K. J. Leonard, H. Bei, and S. J. Zinkle, "Microstructural stability and mechanical behavior of FeNiMnCr high entropy alloy under ion irradiation," Acta Mater., vol. 113, pp. 230–244, 2016.

[293] C. Lu, L. Niu, N. Chen, K. Jin, T. Yang, P. Xiu, Y. Zhang, F. Gao, H. Bei, S. Shi, M.-r. He, I. M. Robertson, W. J. Weber, and L. Wang, "Enhancing radiation tolerance by controlling defect mobility and migration pathways in multicomponent single-phase alloys," Nat. Commun., vol. 7, no. December, p. 13564, 2016.

[294] K. Jin, C. Lu, L. M. Wang, J. Qu, W. J. Weber, Y. Zhang, and H. Bei, "Effects of compositional complexity on the ion-irradiation induced swelling and hardening in Ni-containing equiatomic alloys," Scr. Mater., vol. 119, no. July, pp. 65–70, 2016.

[295] O. El-Atwani, N. Li, M. Li, A. Devaraj, J. K. Baldwin, M. M. Schneider, D. Sobieraj, J. S. Wróbel, D. Nguyen-Manh, S. A. Maloy, and E. Martinez, "Outstanding radiation resistance of tungsten-based high-entropy alloys," Sci. Adv., vol. 5, no. 3, p. eaav2002, 2019.

[296] O. El Atwani, H. T. Vo, M. A. Tunes, C. Lee, A. Alvarado, N. Krienke, J. D. Poplawsky, A. A. Kohnert, J. Gigax, W. Y. Chen, M. Li, Y. Q. Wang, J. S. Wróbel, D. Nguyen-Manh, J. K. Baldwin, O. U. Tukac, E. Aydogan, S. Fensin, and E. Martinez, "A quinary WTaCrVHf nanocrystalline refractory high-entropy alloy withholding extreme irradiation environments," Nat. Commun., vol. 14, no. 1, pp. 1–12, 2023.

[297] D. Sobieraj, J. S. Wróbel, T. Rygier, K. J. Kurzydłowski, O. El Atwani, A. Devaraj, E. Martinez, and D. Nguyen-Manh, "Chemical short-range order in derivative Cr-Ta-Ti-V-W high entropy alloys from the first-principles thermodynamic study," Phys. Chem. Chem. Phys., 2020.

[298] L. Tan, K. Ali, P. S. Ghosh, A. Arya, Y. Zhou, R. Smith, P. Goddard, D. Patel, H. Shahmir, and A. Gandy, "Design principles of low-activation high entropy alloys," J. Alloys Compd., vol. 907, p. 164526, 2022.

[299] H. J. Monkhorst and J. D. Pack, "Special points for Brillouin-zone integrations," Phys. Rev. B, vol. 13, p. 5188, 1976.

[300] E. Walker and P. Bujard, "Anomalous temperature behaviour of the shear elastic constant c44 in tantalum," Solid State Commun., vol. 34, no. 8, pp. 691–693, 1980.

[301] H. Ogi, S. Kai, H. Ledbetter, R. Tarumi, M. Hirao, and K. Takashima, "Titanium's high-temperature elastic constants through the hcp–bcc phase transformation," Acta Mater., vol. 52, no. 7, pp. 2075–2080, 2004.

[302] E. Walker, "Anomalous temperature behaviour of the shear elastic constant c44 in vanadium," Solid State Commun., vol. 28, no. 7, pp. 587–589, 1978.

[303] R. Lowrie and A. Gonas, "Single-crystal elastic properties of tungsten from 24° to 1800°C," J. Appl. Phys., vol. 38, no. 11, pp. 4505–4509, 1967.

[304] J. S. Wróbel, M. R. Zemła, D. Nguyen-Manh, P. Olsson, L. Messina, C. Domain, T. Wejrzanowski, and S. L. Dudarev, "Elastic dipole tensors and relaxation volumes of point defects in concentrated random magnetic Fe-Cr alloys," Comput. Mater. Sci., vol. 194, p. 110435, 2021.

[305] J. Wróbel, D. Nguyen-Manh, S. Dudarev, and K. Kurzydłowski, "Point defect properties of ternary fcc Fe-Cr-Ni alloys," Nucl. Instr. Meth. Phys. Res. B, vol. 393, pp. 126–129, 2017.

[306] J. B. Piochaud, T. P. C. Klaver, G. Adjanor, P. Olsson, C. Domain, and C. S. Becquart, "First-principles study of point defects in an fcc Fe-10Ni-20Cr model alloy," Phys. Rev. B, vol. 89, no. 2, p. 024101, 2014.

[307] D. Connétable, J. Huez, É. Andrieu, and C. Mijoule, "First-principles study of diffusion and interactions of vacancies and hydrogen in hcp-titanium," J. Phys. Condens. Matter, vol. 23, no. 40, p. 405401, 2011.

[308] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," J. Chem. Phys., vol. 129, no. 12, 2008.

[309] H. Robbins and S. Monro, "A stochastic approximation method," Ann. Math. Stat., pp. 400–407, 1951.

[310] Z. Cournia and C. Chipot, "Applications of free-energy calculations to biomolecular processes. a collection," J. Chem. Inf. Model., vol. 64, no. 7, pp. 2129–2131, 2024. PMID: 38587007.