Modélisation multi-échelles des défauts d'irradiation dans les métaux cubiques centrés

Clovis Lapointe

CENTRALE LILLE

THESE

Présentée en vue d'obtenir le grade de

DOCTEUR

En

Spécialité : Chimie des matériaux

Par

Clovis Lapointe

DOCTORAT DELIVRE PAR CENTRALE LILLE

Titre de la thèse :
Modélisation multi-échelles des défauts d'irradiation dans le métaux cubiques centrés

Soutenue le 1er Février 2022 devant le jury d'examen :

Président	Alexandre LEGRIS, Professeur à Polytech Lille (UMET)
Rapporteure	Emilie GAUDRY, Professeure à l'Université de Lorraine (IJL)
Rapporteur	Gabriel Stoltz, Professeur à l'Ecole des Ponts ParisTech (CERMICS)
Examinatrice	Céline VARVENNE, Chargée de recherche à l'Université de Aix-Marseille (CINAM)
Examinateur	Christophe DOMAIN, Docteur à EDF R&D
Encadrant	Thomas D. SWINBURNE, Chargé de recherche à l'Université de Aix-Marseille (CINAM)
Encadrant	Mihai-Cosmin MARINICA, Ingénieur de recherche au CEA Saclay (SRMP)
Directeur de thèse	Laurent PROVILLE, Ingénieur de recherche au CEA Saclay (SRMP)
Directrice de thèse	Charlotte S. BECQUART, Professeur à CentraleLille Institut (UMET)

Thèse préparée dans le Laboratoire : Service de Recherche de Métallurgie Physique (SRMP) - DES/ISAS/DMN - CEA Saclay
Ecole doctorale n° 104 : Science de la Matière, du Rayonnement et de l'Environnement (SRME)

Manches sollte, manches nicht
Wir sehen, doch sind wir blind
Wir werfen Schatten ohne Licht

Nach uns wird es vorher geben
Aus der Jugend wird schon Not
Wir sterben weiter, bis wir leben
Sterben lebend in den Tod
Dem Ende treiben wir entgegen
Keine Rast, nur vorwärts streben
Am Ufer winkt Unendlichkeit

Zeit
Bitte bleib stehen, bleib stehen
Zeit
Das soll immer so weitergehen

Warmer Körper ist bald kalt
Zukunft kann man nicht beschwören
Duldet keinen Aufenthalt
Erschaffen und sogleich zerstören
Ich liege hier in deinen Armen
Ach, könnt es doch für immer sein
Doch die Zeit kennt kein Erbarmen
Schon ist der Moment vorbei

Zeit
Bitte bleib stehen, bleib stehen
Zeit
Das soll immer so weitergehen
Zeit
Es ist so schön, so schön
Ein jeder kennt
Den perfekten Moment

— *Zeit, Rammestein*

Remerciements

On parle généralement de la difficulté de l'exercice d'écriture d'un manuscrit de thèse. Ce que l'on omet d'évoquer, plus généralement, c'est une section dont la difficulté est paroxystique et qui doit être écrite quand l'on croit avoir franchi le col de cette longue randonnée qu'est la thèse : les remerciements. Un exercice difficile, non formalisé, non normalisé bref la partie la plus ardue à rédiger, vous avez compris où je voulais en venir... Face à cette "épreuve", je vous livre une version imparfaite, incomplète - j'ai essayé de n'oublier personne mais c'est peine perdue - de ces petites pensées post-soutenance qui ont le mérite de me ressembler.

Mes remerciements vont d'abord à Émilie Gaudry et Gabriel Stoltz, pour le temps et l'énergie qu'ils ont consacrés afin de rapporter - dans les moindres détails - mon manuscrit de thèse. Leur rigueur et leur intérêt pour mon travail ont transparus dans l'ensemble de leurs remarques - toujours pertinentes - et ont permis d'améliorer grandement la qualité de mon écrit.
Je tiens à remercier les membres de mon jury d'avoir accepté d'examiner mes travaux de doctorat : mes rapporteurs - encore une fois - Émilie Gaudry et Gabriel Stoltz, mes examinateurs, Céline Varvenne et Christophe Domain et Alexandre Legris qui m'a fait l'honneur de présider mon jury de soutenance. L'ensemble de leurs questions, de leurs remarques ainsi que le véritable échange scientifique qu'ils ont su construire pendant ma soutenance de thèse m'ont permis d'ouvrir les yeux sur des points clefs de mon travail. J'ai été frappé par leur précision, leur compréhension profonde des enjeux de notre domaine et je n'aurais pas pu rêver d'un meilleur dialogue pour clore ces trois années de travail. Je les en remercie encore une fois.

Mes remerciements vont ensuite à mon directeur de thèse Laurent Proville au SRMP (CEA Saclay) et à ma directrice de thèse Charlotte Becquart à l'UMET (Université de Lille). Je remercie Laurent pour sa rigueur en physique statistique et son franc parlé. J'ai toujours beaucoup apprécié nos échanges où tu savais trouver le juste milieu entre la mise en valeur de mon travail et ta capacité à me pousser au fond de ma compréhension des choses. Tes conseils et les demis journées endiablées de corrections d'articles ont permis d'améliorer grandement la rigueur, la clarté de mes travaux et je t'en remercie.

Nos débuts avec Charlotte n'ont pas été sans frictions. Heureusement, la vie et la thèse sont la résultante d'une intégrale, les premières impressions sont souvent trompeuses et s'estompent très rapidement. Nous avons appris à travailler ensemble et tu m'as permis d'améliorer grandement ma communication orale et écrite. Je n'oublierai pas non plus nos discussions scientifiques dont la qualité n'a fait qu'augmenter avec le temps. Je t'en remercie.

Une thèse c'est avant tout des encadrants avec lesquels on travaille au quotidien : de ce point de vue, j'ai - encore - été fort chanceux. Je voudrais ici remercier les trois personnes - venues de tous horizons - qui m'ont encadré pendant ces trois ans : Alexandra (Sasha) Goryaeva, Tom Swinburne et Cosmin Marinica.
Bien qu'elle n'ait pas participé "formellement" à mon encadrement, Sasha a été "ma grande soeur" pendant ces trois ans. Tu m'as transmis ton amour pour la science et tu as toujours été là dans les moments de creux avec, comme tu l'appelles, ton "soutien psychologique". Je n'oublierai pas toutes nos sorties au bar, au restaurant ou chez toi qui m'ont toujours fait revenir chez moi le sourire aux lèvres.
Tom est la deuxième rencontre "inattendue" de mon encadrement. Tu venais juste d'être embauché au CNRS quand je suis arrivé en thèse et j'ai été ton "crash test" de doctorant. Je me rends compte de la chance que j'ai eue de t'avoir comme encadrant, toujours disponible pour répondre à mes questions, mettant toujours en avant les qualités de mon travail. Tu possèdes des qualités humaines remarquables. J'ai toujours adoré nos discussions scientifiques à travers leur pronfondeur et leur richesse rendues possibles par ta culture au combien large de la physique et des mathématiques. Je n'oublie pas non plus nos sorties quand tu venais au CEA Saclay et toutes tes anecdotes croustillantes dont je ris encore. J'ai été honoré d'être "ton premier thésard".
Je me souviens encore de mon premier entretien téléphonique avec Cosmin et l'histoire quelque peu rocambolesque de mon arrivée en thèse. Je dois avouer que, encore aujourd'hui, je ne comprends pas vraiment pourquoi tu as choisi ma candidature - qui était d'ailleurs tombée dans ta boîte de spam - pour cette thèse que tu proposais alors que ma formation initiale n'était pas en adéquation avec la physique du solide. Heureusement, tu possèdes des capacités de clairvoyance bien meilleures que les miennes. Notre collaboration a tout de suite été évidente et je me demande comment j'aurais pu trouver un encadrant avec lequel j'aurais été aussi bien accordé - *Sad but True* -. Pendant ces trois ans, tu as toujours été là pour m'aider dans mon travail, me donner de nouveaux angles d'attaque ou de nouvelles idées. Tu as aussi toujours été présent personnellement dans les situations difficiles. Ton amour des Sciences, ta culture, ta curiosité et ta générosité ont grandement influé sur ma vision du monde et de la vie en général. C'est un honneur d'avoir été "un membre de ta famille d'adoption au travail" - car je crois que c'est comme ça que tu vois les étudiants que tu prends sous ton aile -

pendant trois ans. Je ne pourrai jamais oublier tous nos moments de rigolades que ce soit au bureau, au restaurant, au bar ou chez toi. De mon point de vue, notre relation dépasse maintenant le cadre professionnel et je sais que nous ne perdrons pas le contact.

Un doctorat c'est aussi un laboratoire d'accueil et donc des chercheurs et d'autres étudiants. Je tiens avant tout à remercier Jean-Luc Béchade d'avoir accepté de m'accueillir dans son laboratoire - le SRMP - pour effectuer ma thèse. Ta gentillesse, ta sympathie, ta bonne humeur et ton implication pour l'ensemble des chercheurs et des étudiants resteront gravées dans ma mémoire. Je ne pourrai jamais assez te remercier d'avoir sauvé ma soutenance de thèse malgré la situation ubuesque qui s'était présentée. Je tiens aussi à remercier notre secrétaire, Rosabelle Berger, pour ta gentillesse, ton implication pour ton aide dans les démarches administratives du laboratoire - elles sont souvent si compliquées et nombreuses ! - que j'avais tendance à faire au dernier moment. Merci pour ta patience et ta bonne humeur. Passer dans ton bureau a réellement été un plaisir pendant ces deux ans.
Mes remerciements vont ensuite à l'ensemble des chercheurs du SRMP. Je sais la chance que j'ai eue de pouvoir effectuer ma thèse dans un laboratoire de haut niveau scientifique et dans une ambiance bienveillante. Le SRMP est un lieu d'émulation et de convivialité pour ses étudiants. Je tiens plus particulièrement à remercier Manuel Athènes, Thomas Jourdan, Estelle Meslin et Jean-Paul Crocombette pour l'ensemble de nos discussions scientifiques toujours prenantes, riches et pour toutes nos discussions informelles lors des "pauses cafés" et des repas où j'entends encore nos rires. J'espère sincèrement que nous pourrons rester en contact.
Je tiens aussi à remercier Fabien Bruneval. J'ai adoré discuter avec toi que ce soit de Sciences - je comprends un peu mieux l'approximation GW maintenant ! -, de jeux vidéos, de films ou de musiques. De même, j'entends encore nos rires dans la salle café ou pendant les repas. Je te remercie aussi de m'avoir impliqué dans ton travail de recherche que je trouve riche, complexe et passionnant. J'espère aussi garder le contact avec toi.
Je remercie sincèrement Maylise Nastar et Emmanuel Clouet d'avoir accepté d'assister à une de mes répétitions de soutenance de thèse. Vos retours, vos commentaires et vos questions plus que pertinentes ont éclairé des points importants de mon travail que j'ai pu mettre en avant lors de ma soutenance. J'ai un grand respect pour vos qualités scientifiques et j'ai été honoré de pouvoir interagir avec vous pendant ces trois ans.
Durant ces trois années, j'ai aussi eu la chance de pouvoir collaborer avec d'autres chercheurs venus d'horizons différents et ayant accepté de coopérer et de partager une partie de leur savoir avec moi. Je tiens à remercier Louis Thiry et Stéphane Mallat du département informatique de l'ENS de la rue d'Ulm ainsi que Simon Gelin et Normand Mousseau du département de physique de l'Université de Montréal pour le temps qu'ils m'ont consacré, nos discussions ainsi que pour la "matière" de grande qualité qu'ils

m'ont fourni et qui a permis d'aboutir à des publications scientifiques.

Évidemment, un laboratoire de recherche c'est aussi d'autres étudiants. La situation inédite qui s'est présentée pendant une bonne partie de ma thèse - la Covid 19 - a réduit grandement la possibilité d'échanger, de sortir, de partager avec les autres étudiants du SRMP. Je regrette cette restriction sociale qui nous a été imposée car je sais à quel point les étudiants de ce laboratoire sont hors normes que ce soit scientifiquement et socialement. Je voudrais d'abord remercier quelques "anciens" partis avant la fin de mon doctorat : Camille, Elric, Anthon, Yvan, Liangzhao, Thomas, Mauricio, Achraf et Guillaume pour toutes nos conversations et nos rires dans la salle café ou nos bureaux. Marie, tu as clairement été ma première partenaire de "pauses cafés" - cette fameuse salle café qu'on m'avait "assignée" comme second bureau - pendant ma première année et demi de thèse. Je pense qu'il ne serait pas raisonnable de faire un décompte des heures que nous avons passées à parler de Sciences, de musiques, de voyages et de tout et de rien. Tu as été un exemple de travail et de gentillesse pour tous les étudiants qui t'ont croisée au SRMP et je sais que nous ne perdrons pas le contact.
Mika, historiquement mon deuxième partenaire de pause café - *the last but not the least* comme on dit - et la personne qui m'a fait découvrir le grand air des arrières du 520 quand j'étais accompagnateur de la pause clope. De même que pour Marie, il n'est pas raisonnable de compter les heures que nous avons passées pendant ces moments "informels" à parler aussi de Sciences, de tout de rien, de la vie. Tu penses souvent que j'ai fait le psychologue amateur pendant nos pauses mais rassure-toi la réciproque est vraie. Tu es vraiment une personne formidable et je suis très heureux que nous gardions le contact et de te compter comme ami. Je ne peux pas m'empêcher d'avoir une pensée émue pour Claire, ta partenaire de bureau, partie trop vite pendant ma thèse...
Je veux aussi remercier l'ensemble des étudiants qui sont arrivés pendant ma thèse pour tous nos moments de rigolades ou de discussions plus sérieuses que ce soit au laboratoire ou durant nos sorties. Je vais essayer de n'oublier personne. Dans un ordre purement arbitraire : Daphnée - je t'ai un peu converti aux pauses "thé" de l'aprèm -, Émile - mon corrupteur de l'aprèm pour me rajouter une pause -, Xixi - nos petites conversations vacances vont me manquer -, Pamela & Charbel - notre couple libanais préféré et adorable -, Anhruo - j'ai eu le plaisir d'encadrer un peu le début de ta thèse et le plaisir de faire connaissance avec toi -, Maxime - avec qui je vais parler de Spiritbox maitenant ? -, Baptiste - mon petit coin de Marne au sein de ce laboratoire -, Jacopo - discret et toujours prêt à aider -, Quentin - j'aurais beaucoup à dire mais ces remerciements sont déja très longs, une rencontre qu'on n'oublie pas et qui va s'inscrire dans la durée - et Orane - je ne suis pas prêt d'oublier toutes nos discussions enflammées. Tu es une personne entière, ne change pas et j'espère qu'on gardera le contact -.
Une thèse c'est un marathon et pendant cette course être accompagné n'a pas de prix. Nous sommes arrivés en même temps au SRMP avec Océane - on ne va pas chipoter

pour un mois - et nous nous sommes réellement trouvés comme compagnons pour ce long voyage. Mes souvenirs sont encore assez vifs pour conter cette épopée mais comme certains disent *cette marge est trop étroite pour la contenir*. Un seul mot me vient pour résumer cette aventure : **thèse**, tu vas me manquer.

Une thèse c'est aussi un cadre hors professionnel et le résultat d'une autre intégrale de plus longue durée. Je tiens à remercier en premier lieu mes parents ainsi que mon oncle Gérard, pour votre soutien tout au long de mon doctorat et lors de mes - longues - études en général. Vous m'avez toujours soutenu et avez cru en moi malgré mes nombreux virages entre les disciplines. Je voudrais aussi remercier Odile, Bruno, Marc et Soonie, vous êtes clairement le prolongement de ma famille, le sang ne fait pas tout, et les années passées ensemble dominent largement cette donnée initiale.
Comme je l'ai dit, *le sang ne fait pas tout*, et pour paraphraser quelqu'un qui sera bientôt cité dans ces remerciements : ***les amis c'est la famille qu'on choisit***[1]. Quand je parle de mes amis, deux noms me viennent directement à l'esprit : Anaïs et Romain. Pour continuer à paraphraser notre cher Pierre de Fermat, je pourrais faire la liste des raisons de vous remercier mais *cette marge est trop étroite pour la contenir*. Deux personnes précedemment citées, deux groupes de personnes : un de Nancy, un de Reims et là encore une extension famiale - 10 ans, et pas loin pour les autres, que l'on se connait ! -. Je vous remercie tous pour tous les moments de rires, d'émotions, de vie que nous avons - et que nous allons - partager. Je ne peux qu'être ému par votre soutien indéfectible dans toutes les situations pendant toutes ces années. La chance que j'ai de vous avoir n'est pas mesurable. Je vais maintenant essayer de n'oublier personne dans cette joyeuse troupe - dont l'ordre d'apparition est là encore aléatoire - et clore ces remerciements avant que leur taille dépasse celle de ce manuscrit - il parait que c'est la coutume pour une thèse -. Alexandre[2], Antoine[2], Arthur, Guillaume, Mélanie, Loïck[2], Léa, Yann, Romain, Brendan, Corentin, Élodie, Émilien, Gabriel, Maxime, Solène, Ophély, Renaud, Rémi, Anaïs, Vincent, Capucine, Agathe, Marie et Lucie, juste un grand merci !

1. Oui, je te vole tes citations Romain

Résumé

La modélisation des métaux sous conditions extrêmes nécessite une approche de type multi-échelles. En effet, pour des raisons de compléxité numérique, il n'est pas possible d'utiliser un formalisme unique, précis et transférable pour toutes les échelles de simulation. Il correspond, en général, une ou plusieurs méthodes utilisables pour une échelle spatiale et temporelle donnée. Ces méthodes se basent sur des approximations - physiques et/ou numériques - dont le nombre croît lorsque les échelles d'espace et de temps simulées augmentent. La modélisation multi-échelle peut donc se résumer comme un - utopique - équilibre entre échelles "spatio-temporelles" simulées et représentatitivité des phénomènes mis en jeu lors des transformations du système étudié. Les méthodes multi-échelles appliquées à la science des matériaux doivent aussi prendre en compte les effets de températures finies afin de simuler des structures dans leurs conditions nominales d'utilisation industrielles. Ces dernières années, les effets de température ont été traités dans le cadre d'approximations locales : harmonique et/ou quasiharmonique. La prise en compte des effets anharmoniques reste difficile - mais nécessaire pour rendre compte de certains phénomènes physiques - et est un sujet de recherche à part entière pour l'amélioration des modèles multi-échelles.

Les objectifs de cette thèse sont de (i) développer de nouveaux outils de simulations afin d'étendre les domaines d'applicabilité des méthodes multi-échelles (ii) et d'estimer des grandeurs de températures finies. Nous nous basons sur un ensemble de méthodes en plein essor dans le domaine de la science des matériaux : le Machine Learning. Ces méthodes permettent de développer des outils statistiques systématiques et d'étudier plus facilement des corrélations. Dans un premier temps, nous développons des méthodes d'estimations rapides de quantités dérivées de propriétés vibrationnelles harmoniques : l'entropie de formation de défauts et les fréquences d'attaque. Le formalisme développé est précis, transférable et permet de réduire grandement le coût numérique (évoluant traditionnellement comme $\mathcal{O}(N^3)$ où N est le nombre de particules dans le système) dont la complexité numérique évolue $\mathcal{O}(N)$. Dans un deuxième temps, nous nous intéressons à quantifier l'influence des effets anharmoniques pour des systèmes métalliques. Nous développons une approche de calcul direct (couplant Machine Learning et méthodes d'*énergie libre*) permettant de calculer le

coefficient d'auto-diffusion, avec une précision *ab initio*, des métaux cubiques centrés. Nous confrontons directement nos résultats avec l'expérience et nous donnons une explication, générale pour les métaux cubiques centrés, du comportement anormal du coefficient d'auto-diffusion à hautes températures.

Table des matières

Liste des Abréviations

ABF	Adaptative Biaising Force
ABP	Adaptative Biaising Potential
ADN	Acide DésoxyriboNucléique
AFS	Angular Fourier Series
AM	Fonctionnelle GGA développée par Armiento et Mattsson
ARN	Acide RiboNucléique
ARTn	Activation Relaxation Technique nouveau
bSO(4)	Bi-spectrum SO(4)
DD	Dislocation Dynamic
DFT	Density Functional Theory
EAM	Embedded Atom Model
EQML	Extended Quadratic Machine Learning
`FEAR`	Free Energy using bayesiAn Reasoning
GGA	Generalized Gradient Approximation
GW	Approximation de l'interaction Coulombienne au premier ordre par utilisation d'un propagateur sous forme de fonction de Green
HOMO	Highest Occupied Molecular Orbital
HTST	Harmonic Transition State Theory
K.L.	Kullback-Leibler
kMC	kinetic Monte Carlo
`LAMMPS`	Large-scale Atomic/Molecular Massively Parallel Simulator
LDA	Local Density Approximation
LML	Linear Machine Learning
MAE	Mean Absolute Error
MCD	Minimum Covariant Determinant
MEAM	Modified Embedded Atom Model
méta-GGA	méta - Generalized Gradient Approxmiation
`MILADY`	MachIne LeArning DYnamics

ML . . . . . Machine Learning

`MOLGW` . . . . . code de DFT utilisant la théorie des perturbations và N-corps dans le cadre de l'approximation GW

NEB . . . . . Nudget Elastic Band

NVE . . . . . Ensemble statistique pour un système contenant un nombre fixé de particules (N), un volume fixé (V) et une énergie fixée (E)

NVT . . . . . Ensemble statistique pour un système contenant un nombre fixé de particules (N), un volume fixé (V) et une température fixée (T)

`PAFI` . . . . . Projected Average Force Integrator

PBE . . . . . Fonctionnelle GGA développée par Perdew, Burke et Ernzerhof

`PHONDY` . . . . . PHONons DYnamics

QMC . . . . . Quantum Monte Carlo

QNML . . . . . Quadratic Noise Machine Learning

RMSE . . . . . Root Mean Square Error

SVD . . . . . Singular Value Decomposition

TB . . . . . Tight Binding

TST . . . . . Transition State Theory

On a fait ce qu'on a fait comme on l'a fait
Mais on l'a fait
Tout se tranforme, rien ne se perd
Ombre et lumière.

— Shonen, Orelsan

Introduction

Dans les matériaux cubiques centrés sous irradiation, on observe la formation de défauts ponctuels (interstitiels et/ou lacunes) liée à l'interaction du rayonnement avec le matériau. Ces défauts possèdent une certaine mobilité et peuvent, au bout d'une certaine durée, former des défauts de grandes tailles (amas, boucles de dislocation...). Ces matériaux sont utilisés dans les centrales nucléaires et notamment pour les cuves pressurisées du circuit primaire, où les amas de défauts peuvent affecter les propriétés mécaniques du matériau. Le suivi et la compréhension des modifications des propriétés mécaniques des matériaux de structures sous conditions extrêmes sont des sujets de premier ordre, notamment pour l'industrie nucléaire. Cet enjeu justifie l'étude - en partant des plus basses échelles - par simulation du comportement des défauts se formant dans les métaux cubiques centrés en conditions extrêmes.

La modélisation des métaux sous conditions extrêmes nécessite une approche de type **multi-échelles**. En effet, pour des raisons de complexité numérique qui seront discutées dans le chapitre 1, il n'est pas possible d'utiliser un formalisme **unique, précis et transférable** pour toutes les échelles de simulation. Il correspond, en général, une ou plusieurs méthodes utilisables pour une échelle spatiale et temporelle donnée. Ces méthodes se basent sur des approximations - physiques et/ou numériques - dont le nombre croît lorsque les échelles d'espace et de temps simulées augmentent. **La modélisation multi-échelle peut donc se résumer comme un - utopique - équilibre entre échelles "spatio-temporelles" simulées et représentativité des phénomènes mis en jeu lors des transformations du système étudié**.

Les études multi-échelles appliquées à la science de matériaux sous irradiation sont en plein essor depuis les années 2000 [1-5] et ont connu des avancées importantes ces dernières années. Nous citerons, par exemple, les études concernant la stabilité relative et le réarrangement des défauts d'irradiation vers des objets étendus [1, 6-11], la prise en compte des effets élastiques pour l'analyse du comportement des dislocations [2, 12-17] et l'analyse de la diffusion des solutés sous irradiation [18, 19]. Nous présentons, dans la figure 1 qui sera reprise et détaillée dans le chapitre 1, le diagramme "espace-temps" accessible à la simulation numérique grâce aux différentes méthodes multi-échelles de la littérature. Ce diagramme est **lacunaire** dans le sens où il existe des domaines "espace-temps" encore inaccessibles aux méthodes de simulations numériques en science des matériaux. Les méthodes multi-échelles appliquées à la science des matériaux doivent également prendre en compte les effets de températures afin de simuler des structures dans leurs conditions nominales d'utilisation industrielles. Ces dernières

années, les effets de température ont été traités dans le **cadre d'approximations locales : *harmonique* et/ou *quasiharmonique*. La prise en compte des effets anharmoniques reste difficile - mais nécessaire pour rendre compte de certains phénomènes physiques - et est un sujet de recherche à part entière pour l'amélioration des modèles multi-échelles.**

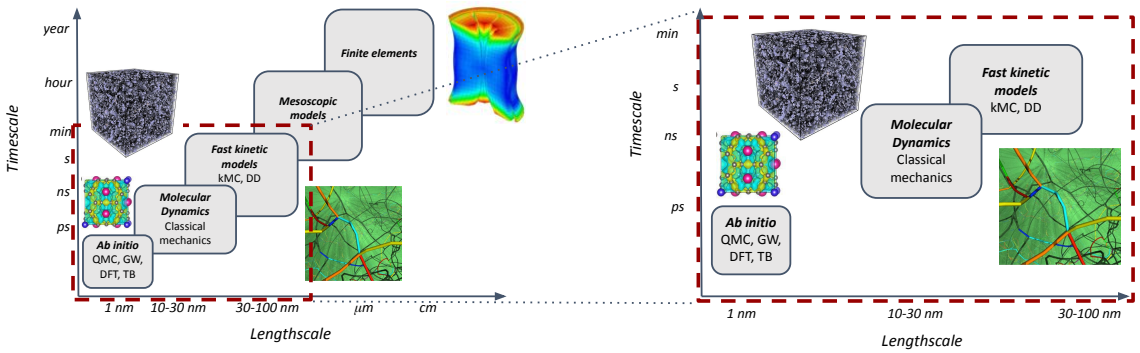

Figure 1: Illustration des différentes méthodes de simulations pour l'analyse multi-échelles. Les échelles d'espace et de temps caractéristiques des différentes méthodes de simulation sont données par l'axe des abscisses et des ordonnées. Les méthodes *ab initio* sont par exemple : le Monte Carlo quantique (QMC), l'approximation GW, la théorie de la fonctionnelle de la densité (DFT) et les modèles de type liaisons fortes (TB). Les modèles cinétiques rapides sont par exemple : le Monte Carlo cinétique (kMC) et la dynamique des dislocations (DD).

Dans ce manuscrit, nous développons des outils permettant de prendre en compte les deux points cités précedemment : **(i) rendre accessible de nouveaux domaines du diagramme "temps-espace" de la simulation multi-échelles et (ii) estimer des grandeurs de températures finies dans le cadre de l'approximation *harmonique* et le cadre *anharmonique***. Nous nous basons sur une nouvelle classe d'approche utilisée en science des matériaux et née dans les années 2010 : les méthodes Machine Learning. Ces approches se basent sur une nouvelle représentation de l'environnement atomique local. Elles permettent d'estimer des observables thermodynamiques avec une plus grande flexibilité que les approches traditionnellement utilisées [20-23]. **Le nouvel espace de représentation des configurations atomiques permet aussi de développer des outils statistiques et d'étudier plus facilement des corrélations**. Nous nous sommes spécifiquement intéressé au développement et à l'utilisation de ces outils pour les cas des métaux cubiques centrés. Ce manuscrit est organisé selon le plan suivant :

Le premier chapitre 1 est dédié à la présentation succincte des méthodes de modélisation atomistique des matériaux. Le principe général de ces méthodes ainsi que leurs domaines - spatiaux et temporels - d'applicabilité sont décrits. Nous introduisons ensuite les méthodes Machine Learning appliquées à la science des matériaux et les perspectives que ces méthodes permettent d'envisager en termes de simulations atomistiques.

Le deuxième chapitre 2 introduit les concepts clefs nécessaires à la mise en place des méthodes Machine Learning pour la science des matériaux. Nous décrivons la notion de descripteurs atomiques locaux et nous faisons une revue des différents descripteurs présents dans la littérature. Nous présentons les méthodes de régressions et de régularisations communément utilisées dans le cas de problèmes en hautes dimensions.

Le troisième chapitre 3 porte sur la construction d'un modèle de régression de l'entropie vibrationnelle harmonique pour les défauts ponctuels dans le fer cubique centré. Ce modèle se base sur les descripteurs atomiques locaux. Dans un premier temps, nous définissons l'entropie vibrationnelle harmonique et ses liens avec l'*énergie libre* et la *densité d'état de modes normaux*. Nous proposons ensuite, en nous basant sur la décomposition locale de la *densité d'état de vibration* [24], un modèle linéaire (en descripteurs atomiques locaux) de l'entropie vibrationnelle harmonique. Nous présentons notre méthode de génération de base de données et appliquons ce modèle au cas des défauts ponctuels dans le fer cubique centré. Ce modèle se montre très précis et transférable. Finalement, nous développons un modèle de régression - linéaire dans l'espace des descripteurs - direct de l'entropie vibrationnelle harmonique locale.

Le quatrième chapitre 4 est consacré à une extension du modèle de régression de l'entropie vibrationnelle, toujours dans le cadre harmonique. Nous commençons par étendre notre modèle linéaire à un modèle quadratique qui se montre plus précis dans le cas des déformations dans le fer cubique centré. Nous développons ensuite une approche corrective afin d'ajuster la différence d'entropie vibrationnelle entre une configuration non-déformée et cette même configuration soumise à une petite déformation. Nous montrons, à travers cette approche théorique, qu'introduire un terme quadratique en descripteur permet d'obtenir une meilleure précision dans le cadre de petites déformations. Ainsi la "physique" des déformations est mieux reproduite par des modèles quadratiques. Nous construisons ensuite des modèles linéaires en descripteurs pour les fréquences d'attaque et les barrières énergétiques dans le silicium amorphe. Ces deux modèles de régression se montrent précis et transférables et ouvrent la possibilité d'un couplage direct avec des méthodes de grandes échelles "spatio-temporelles" tel que le Monte Carlo cinétique. Finalement, nous mettons en oeuvre une analyse statistique et une reformulation, dans l'espace des descripteurs, de la loi empirique de Meyer-Neldel. Nous montrons que cette loi empirique peut être envisagée comme une relation géométrique dans l'espace des descripteurs et nous développons un critère quantitatif de sa validité. Ce type de critère quantitatif pourra être généralisé pour mettre au jour de nouvelles lois de corrélation.

Le cinquième chapitre 5 introduit l'importance des effets anharmoniques dans les matériaux cristallins. Nous donnons une revue des méthodes d'*énergie libre* à biais adaptatifs présentes dans la littérature. Nous précisons l'implémentation des principales méthodes d'*énergie libre* que nous avons utilisées. Nous présentons un cas pratique

de l'utilisation de ces méthodes pour le calcul du module élastique isostatique et du paramètre de maille d'équilibre en fonction de la température pour un potentiel EAM [25, 26] dans le tungstène cubique centré. Cet exemple servira d'étalon pour assurer la convergence de nos calculs d'*énergie libre* dans le chapitre suivant.

Le sixième chapitre 6 s'intéresse à un cas pratique de couplage entre les méthodes de régressions en hautes dimensions et les méthodes d'*énergie libre*. Nous étudions les coefficients d'auto-diffusion dans les métaux cubiques centrés et notamment leur comportement à haute température. Nous présentons une méthode de génération de base de données *ab initio* standards représentative des propriétés des lacunes dans différents métaux cubiques centrés. Nous couplons ensuite les méthodes Machine Learning et les *méthodes d'énergie libre* afin d'estimer, avec une précision *ab initio*, l'*énergie libre d'activation* de la mono-lacune dans différents métaux cubiques centrés. Cette nouvelle approche permet de faire un lien direct entre nos résultats de simulation et les données expérimentales. À l'aide de cette méthode, nous montrons que dans le cas du tungstène et du molybdène, la prise en compte de l'anharmonicité de la mono-lacune permet d'expliquer de façon qualitative le comportement - non-Arrhenius - à hautes températures du coefficient d'auto-diffusion sans introduire d'autres populations de défauts (notamment la di-lacune souvent introduite dans les modèles de la littérature). De plus, nous montrons que les fonctionnelles classiquement utilisées pour les calculs de *théorie de la fonctionnelle de la densité* échouent à reproduire quantitativement le comportement du coefficient d'auto-diffusion.

Dans le septième chapitre (7) nous présentons différentes collaborations, portant sur les méthodes Machine Learning appliquées à la science des matériaux, auxquelles j'ai participé au cours de mon doctorat.

Par soucis de clarté, l'ensemble des tenseurs liés à des quantités physiques seront notés en gras : $\boldsymbol{\mathit{physique}}$. Les tenseurs liés à des quantités statistiques ou d'apprentissage automatique seront soulignés par des barres en fonction de leur ordre de tensorialité : $\underline{\mathit{statistique}}$.

Are you on the square ?
Are you on the level ?
Are you ready to swear right here, right now.

— Square Hammer, Ghost

1

Modélisation multi-échelles : méthodes et échelles caractéristiques

Sommaire

1.1 Vision d'ensemble

La modélisation des matériaux structurels et/ou fonctionnels est un enjeu industriel majeur. En effet, la modélisation de l'évolution des propriétés mécaniques d'un matériau au cours du temps et dans ses conditions d'utilisation est nécessaire afin de dimensionner une structure. On peut notamment penser aux matériaux soumis à des conditions extrêmes ou sous irradiation dans le domaine du nucléaire ou de l'aérospatial. L'objectif concret de la modélisation des matériaux est de prédire les propriétés mécaniques macroscopiques d'un système réel. Néanmoins, cet objectif - relativement simple en apparence - se revèle d'une grande complexité car les phénomènes observés macroscopiquement trouvent leur origine à l'échelle microscopique. À titre d'exemple, la plasticité macroscopique trouve son origine dans l'apparition et le déplacement de dislocations [27-30] et les phénomènes de ségrégation induite sous irradiation ne peuvent être compris sans prendre en compte la diffusion microscopique [18, 31, 32].

La notion de simulation *multi-échelles* permet de prendre en compte l'ensemble des phénomènes physiques allant de la mécanique quantique jusqu'à la mécanique des milieux continus. Nous présentons dans la figure 1.1 une illustration de la notion de simulation *multi-échelle*. Dans ce diagramme, on constate qu'il existe un grand nombre de méthodes numériques permettant de simuler un domaine de l'échelle spatiale et temporelle d'un système. Il serait illusoire de penser pouvoir développer une méthode "générale" prenant directement en compte toutes les échelles de temps et d'espace sans changer de formalisme [33].

On peut classer les méthodes de simulation en science des matériaux en trois grandes classes. Les méthodes dites **atomistiques** (i) dont l'échelle de temps est comprise entre la ps et la μs et l'échelle d'espace entre le nm et le 10^2 nm. Ces méthodes permettent de simuler avec une grande précision les propriétés des systèmes de faibles tailles $\leq 10^7$ atomes en prenant en compte la **nature physique des interactions entre les particules du système**.
Les méthodes dites **mésoscopiques** (ii) dont l'échelle de temps est comprise entre la μs et la min et l'échelle d'espace est comprise entre le $10^1 nm$ et μm. Ces méthodes incluent un nombre suffisant d'atomes (entre 10^6 et 10^8) pour calculer des **moyennes d'ensembles statistiques** et prédire des **observables thermodynamiques** avec précision.
Les méthodes dites **continues** (iii) dont l'échelle de temps est supérieure à la s et l'échelle d'espace est supérieure au μm. Les méthodes continues **ne prennent pas en compte la nature particulaire des systèmes** mais leurs points de départ sont des grandeurs mésoscopiques c'est-à-dire issues de moyennes d'ensembles statistiques.

Dans l'industrie, à quelques exceptions près, la majorité des méthodes de simulations utilisées sont des méthodes dites **continues**. Les méthodes *continues* nécessitent

la connaissance préalable de grandeurs et de modèles issus des échelles inférieures (mésoscopiques et atomistiques). Le paramétrage des méthodes *continues* introduit une interdépendance entre les différentes méthodes citées précedemment afin d'atteindre l'échelle du continu et de rendre compte des **propriétés macroscopiques du matériau**. Dans ce manuscrit, nous nous sommes limités à l'utilisation de méthodes **atomistiques** et **mésoscopiques**. Dans les sections suivantes, nous décrivons les méthodes que nous avons utilisées tout en rappelant les limites de temps et d'espace associées à chaque méthode.

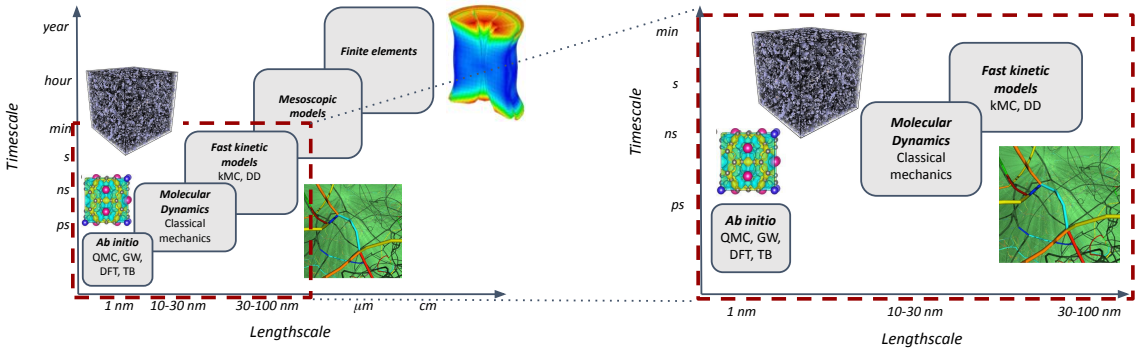

Figure 1.1: Illustration des différentes méthodes de simulation pour l'analyse multi-échelles. Les échelles d'espace et de temps caractéristiques des différentes méthodes de simulation sont données par l'axe des abscisses et des ordonnées. Les méthodes *ab initio* sont par exemple : le Monte Carlo quantique (QMC), l'approximation GW, la théorie de la fonctionnelle de la densité (DFT) et les modèles de type liaisons fortes (TB). Les modèles cinétiques rapides sont par exemple : le Monte Carlo cinétique (kMC) et la dynamique des dislocations (DD).

Nous décrivons ici les trois types de méthodes simulations - à l'échelle microscopique et mésoscopique - que nous avons utilisées dans la suite de ce manuscrit. Nous procédons par échelles d'espace et de temps croissantes. Nous partons donc de l'échelle de l'atome décrit par la mécanique quantique jusqu'aux systèmes de tailles mésoscopiques permettant de calculer des observables thermodynamiques et/ou des propriétés d'objets étendus de l'ordre du *nm*.

1.2 Mécanique quantique et méthodes *ab initio*

La mécanique quantique est régie par l'équation de *Schrödinger* qui peut se traduire de la façon suivante en régime permanent (indépendant du temps) :

$$\hat{\boldsymbol{\mathcal{H}}}|\psi\rangle = \epsilon|\psi\rangle \tag{1.1}$$

Ici, $\hat{\boldsymbol{\mathcal{H}}}$ et $|\psi\rangle$ sont respectivement l'opérateur Hamiltonien et la fonction d'onde du système. L'équation de *Schrödinger* est un problème aux valeurs propres. Les valeurs propres ϵ sont les énergies possibles du système et les vecteurs propres associés

$|\psi\rangle$ sont les fonctions d'onde du système. Nous rappelons ici, que dans le cadre de la mécanique quantique, la fonction d'onde est un objet "sans réalité physique". Choisissons un état possible du sytème $|e_i\rangle$, la réalité physique de la fonction d'onde réside dans le produit scalaire suivant :

$$p(e_i) = |\langle e_i|\psi\rangle|^2 \tag{1.2}$$

Ici, $p(e_i)$ est la probabilité de trouver le système décrit par la fonction d'onde $|\psi\rangle$ dans l'état $|e_i\rangle$. La difficulté majeure de l'équation de *Schrödinger* (1.1) est l'estimation du spectre de l'opérateur Hamiltonien. En effet, dans le cas d'un sytème comportant N noyaux et M électrons, on peut écrire cet opérateur de la façon suivante :

$$\hat{\boldsymbol{\mathcal{H}}} = \underbrace{\sum_{i=1}^{N} \frac{\hat{\boldsymbol{P}}_i^2}{2M_i}}_{\hat{\boldsymbol{T}}_{\boldsymbol{n}}} + \underbrace{\sum_{i=1}^{M} \frac{\hat{\boldsymbol{p}}_i^2}{2m_e}}_{\hat{\boldsymbol{T}}_{\boldsymbol{e}}} - \underbrace{\sum_{i=1}^{N}\sum_{j=1}^{M} \frac{eZ_i}{|\hat{\boldsymbol{R}}_i - \hat{\boldsymbol{r}}_j|}}_{\hat{\boldsymbol{V}}_{\boldsymbol{en}}} + \underbrace{\sum_{1\leq i<j\leq N}^{N,N} \frac{Z_i Z_j}{|\hat{\boldsymbol{R}}_i - \hat{\boldsymbol{R}}_j|}}_{\hat{\boldsymbol{V}}_{\boldsymbol{nn}}} + \underbrace{\sum_{1\leq i<j\leq M}^{M,M} \frac{e^2}{|\hat{\boldsymbol{r}}_i - \hat{\boldsymbol{r}}_j|}}_{\hat{\boldsymbol{V}}_{\boldsymbol{ee}}} \tag{1.3}$$

Ici, $\hat{\boldsymbol{P}}_i$, $\hat{\boldsymbol{R}}_i$, M_i et Z_i sont respectivement l'opérateur quantité de mouvement, l'opérateur position, la masse et la charge du noyau i. De même, $\hat{\boldsymbol{p}}_i$, $\hat{\boldsymbol{r}}_i$, m_e et e sont respectivement l'opérateur quantité de mouvement, l'opérateur position, la masse et la charge de l'électron i. On peut décomposer cet Hamiltonien en cinq contributions principales : (i) l'énergie cinétique des noyaux $\hat{\boldsymbol{T}}_{\boldsymbol{n}}$, (ii) l'énergie cinétique des électrons $\hat{\boldsymbol{T}}_{\boldsymbol{e}}$, (iii) l'énergie d'interaction électrons-noyaux $\hat{\boldsymbol{V}}_{\boldsymbol{en}}$, (iv) l'énergie d'interaction noyaux-noyaux $\hat{\boldsymbol{V}}_{\boldsymbol{nn}}$ et enfin (v) l'énergie d'interaction électrons-électrons $\hat{\boldsymbol{V}}_{\boldsymbol{ee}}$. L'ensemble de ces termes énergétiques est plus on moins complexe à estimer et nécessite certaines approximations que nous allons décrire dans les paragraphes suivants.

1.2.1 Approximation de *Born-Oppenheimer*

Dans la majorité des simulations nécessitant la résolution de l'équation de *Schrödinger*, on fait l'hypothèse de *Born-Oppenheimer*. Cette hypothèse consiste à considérer que le temps de relaxation des électrons est très faible devant le temps de relaxation des noyaux. Dans le cadre de l'approximation de *Born-Oppenheimer*, on peut donc décorréler la dynamique des électrons et des noyaux et considérer que l'état du système va être déterminé par la dynamique la plus rapide, c'est-à-dire celle des électrons. Dans ce cas, les termes énergétiques liés aux noyaux peuvent être considérés comme un potentiel extérieur $\hat{\boldsymbol{V}}_{\boldsymbol{ext}}^{\boldsymbol{n}}$ et le nouvel Hamiltonien s'écrit alors :

$$\hat{\boldsymbol{\mathcal{H}}}_{\mathcal{BO}} = \hat{\boldsymbol{T}}_{\boldsymbol{e}} + \hat{\boldsymbol{V}}_{\boldsymbol{en}} + \hat{\boldsymbol{V}}_{\boldsymbol{ee}} + \hat{\boldsymbol{V}}_{\boldsymbol{ext}}^{\boldsymbol{n}} \tag{1.4}$$

L'ensemble des termes de ce nouvel Hamiltonien ne dépend que de la fonction d'onde des électrons $\psi(\boldsymbol{r})$, où $\boldsymbol{r}$ est le vecteur des positions des électrons. Un terme de l'Hamiltonien de *Born-Oppenheimer*, décrit par l'équation (1.4) reste très ardu à évaluer. Il s'agit du terme d'interaction électrons-électrons $\hat{\boldsymbol{V}}_{\boldsymbol{ee}}$. Il est essentiel de noter ici que l'équation de *Schrödinger* décrite par l'Hamiltonien (1.4) ne peut être résolue de façon exacte que pour **l'atome d'hydrogène**. Suivant la coutume, nous parlerons de méthodes *ab initio*, pour qualifier les méthodes permettant de résoudre de façon approchée l'équation de *Schrödinger* pour un système à N noyaux et M électrons.

1.2.2 Méthode de *Hartree-Fock*

Afin de résoudre l'équation de *Schrödinger* pour le problème à M-corps posé par l'Hamiltonien de l'équation (1.1) sous l'hypothèse de *Born-Oppenheimer* (1.4), il faut partir d'une fonction *a priori*. Les électrons étant des fermions, on peut chercher la fonction d'onde totale des électrons du système sous la forme d'un déterminant de *Slater*. En effet, un déterminant de *Slater* respecte les propriétés d'anti-symétrie et/ou le principe d'*exclusion de Pauli*. On note $\psi_{\mathcal{HF}}(\boldsymbol{r})$ la fonction d'onde solution du problème posé par l'équation de *Schrödinger*. On note alors $\phi_i(\boldsymbol{r_i})$ la base de fonction d'onde de l'électron i et vérifiant :

$$\psi_{\mathcal{HF}}(\boldsymbol{r}) \in \text{Slater}\left(\{\phi_i(\boldsymbol{r_i})\}_{i\leq M}\right) \equiv \mathcal{S}\left\{\phi_i\right\}_{i\leq M} \tag{1.5}$$

Ici, Slater$(\cdot)$ représente le déterminant de *Slater*. En utilisant cette formulation de la fonction d'onde de *Hatree-Fock*, on suppose qu'elle minimise l'énergie du système d'Hamiltonien (1.4). Afin de vérifier l'orthonormalité des fonctions d'onde de chaque électron, on introduit des *multiplicateurs de Lagrange* ϵ_i pour chaque fonction d'onde $\phi_i(\boldsymbol{r_i})$. On aboutit alors au problème de minimisation suivant :

$$\psi_{\mathcal{HF}}(\boldsymbol{r}) = \arg \min_{\psi(\boldsymbol{r})\in\mathcal{S}\{\phi_i\}_{i\leq M}} \left\{ \langle\psi|\hat{\boldsymbol{\mathcal{H}}}_{\mathcal{BO}}|\psi\rangle + \sum_{i=1}^{M} \epsilon_i \left(\langle\phi_i|\phi_i\rangle - 1\right) \right\} \tag{1.6}$$

On peut alors utiliser une méthode variationnelle sur la fonctionnelle donnée par le membre de droite de l'équation (1.6). On aboutit à un système de M équations, pour les fonctions d'onde de chaque électron $\phi_i(\boldsymbol{r_i})$. La méthode de *Hartree-Fock* est **"auto-cohérente"**, c'est-à-dire que l'on donne une fonction d'onde *a priori* et si celle-ci est solution du problème variationnel posé par l'équation (1.6), alors elle est solution de l'équation de *Schrödinger* à M électrons. En effet, le problème variationnel (1.6) implique l'équation de *Schrödinger*, les *multiplicateurs de Lagrange* ϵ_i étant les énergies possibles du système. La méthode d'*Hatree-Fock* ne nécessite pas d'hypothèse supplémentaire mais se révèle très difficile à mettre en pratique à cause des propriétés importantes d'anti-symétrie et du *principe d'exclusion de Pauli*. D'un point de vue numérique, la méthode d'*Hartree-Fock* évolue comme $\mathcal{O}(N^5)$, où N est le nombre d'atomes dans le système. La méthode d'*Hartree-Fock* est très utilisée en chimie quantique [34], mais devient difficile à mettre en place pour les systèmes métalliques. Il existe une autre méthode permettant de résoudre de façon approchée l'équation de *Schrödinger* - Eq. (1.1) - en introduisant le concept de densité électronique notée $n(\boldsymbol{r})$.

1.2.3 Théorie de la fonctionnelle de la densité

La théorie de la fonctionnelle de la densité est née dans les années 1920 et reçoit deux théorèmes importants dans les années 1960 grâce à l'apport de Hohenberg, Kohn et Sham. Ces deux théorèmes sont à l'origine de l'essor de la théorie de la fonctionnelle de la densité (DFT) pour les calculs *ab initio* en sciences des matériaux. Cette théorie permet de calculer la fonction d'onde électronique, solution de l'équation de *Schrödinger* (1.4), sans faire d'hypothèse, *a priori*, sur la forme de la fonction

d'onde. Cela rend cette méthode beaucoup plus simple à mettre en pratique que la méthode de *Hartree-Fock*. La théorie de la fonctionnelle de la densité est une théorie à champ moyen dont le concept clef est la **densité électronique** notée $n(\boldsymbol{r})$, qui vérifie les propriétés suivantes pour un système à M électrons :

$$\int_{\mathbb{R}^3} n(\boldsymbol{r}) d\boldsymbol{r} = M \tag{1.7}$$

$$n(\boldsymbol{r}) \xrightarrow[|\boldsymbol{r}|\to+\infty]{} 0 \tag{1.8}$$

La théorie de la fonctionnelle de la densité tend à traiter tous les électrons du système comme un seul champ moyen défini par la densité électronique $n(\boldsymbol{r})$. La théorie de la fonctionnelle de la densité vise à reformuler l'équation de *Schrödinger* en substituant la fonction d'onde électronique par la densité électronique. Dans ce cadre, la résolution de l'équation de *Schrödinger* admet des densités électroniques comme solution. Nous allons montrer comment effectuer le passage de la fonction d'onde électronique à la densité dans les paragraphes suivants.

Cette transformation se base sur deux théorèmes majeurs : (i) le *premier théorème de Kohn-Hohenberg* Th. (1.1) et (ii) le *deuxième théorème de Kohn-Hohenberg* Th. (1.2) [35].

Théorème 1.1. *La densité électronique définie par l'équation (1.8) permet de reformuler l'Hamiltonien de Born-Oppenheimer. La formulation de la densité électronique est équivalente à la formulation de la fonction d'onde électronique pour l'Hamiltonien Eq. (1.4), ainsi on a :*

$$\hat{\boldsymbol{\mathcal{H}}}_{\mathcal{BO}}\{\boldsymbol{r}\} \longleftrightarrow \hat{\boldsymbol{\mathcal{H}}}_{\mathcal{BO}}\{n(\boldsymbol{r})\} \equiv \hat{\boldsymbol{T}}_{\boldsymbol{e}}\{n(\boldsymbol{r})\} + \hat{\boldsymbol{V}}_{\boldsymbol{ee}}\{n(\boldsymbol{r})\} + \hat{\boldsymbol{V}}_{\boldsymbol{ext}}\{n(\boldsymbol{r})\} \tag{1.9}$$

De plus, l'opérateur $\hat{\boldsymbol{V}}_{\boldsymbol{ext}}\{n(\boldsymbol{r})\}$ est unique.

Théorème 1.2. *L'énergie de l'état fondamental du système composé de N noyaux et M électrons peut être reformulée de la façon suivante :*

$$E[n(\boldsymbol{r})] = \int_{\mathbb{R}^3} \hat{\boldsymbol{\mathcal{H}}}_{\mathcal{BO}}\{n(\boldsymbol{r})\}\, n(\boldsymbol{r}) d\boldsymbol{r} \tag{1.10}$$

Cette formulation est unique et la densité $n_0(\boldsymbol{r})$ minimisant la fonctionnelle $E[n(\boldsymbol{r})]$ est associée à l'état fondamental du système.

L'équation de *Schrödinger* est donc reformulée à l'aide de la densité électronique. Cette reformulation est **unique** et aboutit à un problème de minimisation qui peut être résolu par une méthode auto-cohérente comme dans le cas de la méthode de *Hartree-Fock*. Il reste néanmoins une difficulté résidant dans le terme d'interaction électrons-électrons $\hat{\boldsymbol{V}}_{\boldsymbol{ee}}\{n(\boldsymbol{r})\}$.

1.2.4 Méthode de Kohn-Sham

L'évaluation de l'opérateur d'interaction électrons-électrons dans le cadre de la densité électronique reste très difficile. Afin de pallier ce problème, on utilise les deux théorèmes de Kohn et Hohenberg Th. (1.1) et Th. (1.2). Le potentiel externe au système étant défini de façon unique par le théorème Th. (1.1), on peut supposer un système fictif d'électrons n'interagissant pas entre-eux mais avec un potentiel externe incluant l'interaction électrons-électrons. Le théorème Th. (1.2) reste valide et le problème variationnel Eq. (1.10) peut être réécrit avec un nouvel Hamiltonien n'incluant plus directement l'interaction électrons-électrons. Dans ce cas, on introduit un nouvel opérateur de la densité électronique : $\hat{\boldsymbol{V}}_{\boldsymbol{xc}}\{n(\boldsymbol{r})\}$. Cet opérateur est appelé **opérateur d'échange-corrélation**. Le problème de la densité électronique s'écrit alors :

$$E[n(\boldsymbol{r})] = \hat{\boldsymbol{T}}'_{\boldsymbol{e}}\{n(\boldsymbol{r})\} + \int_{\mathbb{R}^3} \left(\hat{\boldsymbol{V}}_{\boldsymbol{xc}}\{n(\boldsymbol{r})\} + \hat{\boldsymbol{V}}_{\boldsymbol{ext}}\{n(\boldsymbol{r})\}\right) n(\boldsymbol{r}) d\boldsymbol{r} \tag{1.11}$$

Ici, $\hat{\boldsymbol{T}}'_{\boldsymbol{e}}\{n(\boldsymbol{r})\}$ est l'opérateur d'énergie cinétique des électrons non-interagissant. L'opérateur $\hat{\boldsymbol{V}}_{\boldsymbol{ext}}$ contient l'ensemble des interactions du système fictif. Le théorème Th. (1.2) assure que la densité $n_0(\boldsymbol{r})$ solution de ce problème variationnel décrit le système dans son état fondamental. L'objectif est alors de donner une expression la plus précise possible de l'opérateur d'échange-corrélation. Cet opérateur peut être calculé analytiquement dans le cadre d'un gaz d'électrons dilué [36] et abouti au modèle de Thomas-Fermi. Ce modèle exact reste peu applicable aux systèmes plus complexes que les gaz et nécessite des approximations supplémentaires. Ainsi, un grand nombre de fonctionnelles d'échange-corrélation ont été développées :

— on peut considérer que l'opérateur d'échange-corrélation ne dépend que de la densité électronique $n(\boldsymbol{r})$, dans ce cas on parle de LDA (Local Density Approximation). L'idée originale de cette fonctionnelle est donnée par Slater *et al.* [37]. Plusieures paramétrisations LDA ont alors été développées [38, 39].

— on peut aussi généraliser la fonctionnelle d'échange-corrélation en y ajoutant une dépendance en terme de gradient de la densité électronique $\nabla n(\boldsymbol{r})$. L'ajout du terme de gradient permet de prendre en compte l'inhomogénéité de la densité électronique dans les matériaux. Ce nouveau type de fonctionnelle prenant en compte la densité électronique et le gradient de la densité électronique est appelé GGA (Generalised Gradient Approximation). De même que pour la LDA, un grand nombre de fonctionnelles GGA sont présentes dans la littérature [40-46].

— le formalisme GGA peut lui aussi être étendu en introduisant une dépendance en $\Delta n(\boldsymbol{r})$, où Δ est l'opérateur Laplacien. Le Laplacien de la densité peut aussi être remplacé par l'énergie cinétique électronique pour des raisons de stabilité numérique. Ce type d'approche - appelée méta-GGA - a été introduite par Tao *et al.* [47]. Les fonctionnelles méta-GGA permettent de mieux reproduire certains phénomènes que les fonctionnelles GGA [48, 49] pour un coût numérique similaire. Néanmoins leur implémentation est encore peu répandue.

— enfin, il est possible de coupler la *théorie de la fonctionnelle de la densité* et la méthode d'*Hatree-Fock* grâce aux fonctionnelles hybrides [50]. Ces fonctionnelles permettent d'obtenir des résultats très précis mais dont le coût numérique est très élevé par rapport aux autre fonctionnelles citées précedemment.

Le choix de la fonctionnelle d'échange-corrélation résulte d'un équilibre entre : (i) la précision et la représentatitivé requise d'une observable calculée par la DFT et (ii) le coût numérique nécessaire au calcul. Dans certains cas pathologiques, les fonctionnelles les plus simples échouent à reproduire des propriétés essentielles du matériau. C'est le cas pour le fer, où la fonctionnelle LDA ne permet pas de reproduire l'état fondamental férromagnétique.

1.2.5 Domaine d'applicabilité des méthodes dites *ab initio*

Les méthodes dites *ab initio* se limitent à des systèmes de petites tailles et pour de faibles durées (dans le cadre de la dynamique moléculaire *ab initio*) [51]. Les échelles caractéristiques sont de l'ordre de la *ps* et de l'ordre du *nm*. En effet, la complexité numérique de ce type de méthode évolue au moins en $\mathcal{O}(N^3)$ (avec N le nombre d'atomes) dans les systèmes métalliques. Les systèmes simulés peuvent être, au maximum, de l'ordre de 10^3 atomes et sont en général de l'ordre de 10^2 atomes. Il est possible de simuler l'évolution d'un système quantique en utilisant le *théorème de Ehrenfest*. On peut alors effectuer de la *dynamique moléculaire ab initio*. La durée de ce type de simulation se limite à l'ordre de la *ps*. Les méthodes *ab initio* permettent de calculer des observables dont les valeurs sont très proches de l'expérience [52]. De plus, ce type de calcul prend en compte la nature physique de l'échelle atomique des matériaux en considérant directement les interactions entre les noyaux et les électrons. Néanmoins, la complexité numérique de ces méthodes devient très vite limitante vis-à-vis de la taille et de la durée pendant laquelle on peut simuler des systèmes. Nous allons décrire dans la sous-section suivante le passage de l'échelle de 10^3 d'atomes à l'échelle de 10^7 atomes.

1.3 Dynamique moléculaire *classique* et les potentiels *semi-empiriques*

Afin d'effectuer une première transition d'échelle entre des systèmes contenant de l'ordre de 10^2 à 10^7 atomes, il nous faut abandonner le formalisme de la mécanique quantique dont la complexité numérique est trop grande (celle-ci évolue comme au plus $\mathcal{O}(N^3)$, où N est le nombre d'atomes du système). De nouveaux types de formalismes ont été developpés à partir des années 1920 afin de rendre compte des propriétés des systèmes étudiés sans traiter explicitement les électrons. Ce type d'approche est appelé *semi-empirique*. Dans les années 1970-1980 apparait une nouvelle classe de méthodes *semi-empiriques* [26, 53, 54] permettant de décrire correctement **les interactions métalliques**. Nous allons d'abord décrire quelques potentiels *semi-empiriques modèles* avant de nous concentrer sur les potentiels *semi-empiriques* plus précis se basant sur le caractère local ou diffus des électrons dans le matériau.

1.3.1 Une vision sans électrons

Dans la section précédente Sec. 1.2, nous avons vu que les propriétés d'un système peuvent être déterminées à l'aide de la densité électronique $n(\boldsymbol{r})$. La densité électronique est déterminée par le potentiel extérieur appliqué au système $\hat{\boldsymbol{V}}_{\boldsymbol{ext}}$ via le Th. 1.1. Ce potentiel extérieur contient les interactions entre les électrons mais surtout les interactions entre les électrons et les noyaux. La connaissance d'un ensemble de grandeurs relatives aux noyaux tels le vecteur de positions $\boldsymbol{R}$, la charge des noyaux $\boldsymbol{Q}$ (etc.) est suffisante pour décrire un système **si les interactions électrons-électrons sont bien représentées** par $\hat{\boldsymbol{V}}_{\boldsymbol{ext}}$. Les méthodes *semi-empiriques* naissent de ce constat. Ce type de méthodes cherche à décrire les interactions entre les atomes du système par des fonctions simples dont la forme est determinée à partir de considérations qualitatives sur le comportement des électrons. Un potentiel *semi-empirique* sera donc exprimé sous la forme $V(\boldsymbol{R}, \boldsymbol{Q})$ et dépendra notamment du vecteur de positions des noyaux. Dans le cas des systèmes que nous avons étudiés au cours de cette thèse - les métaux cubiques centrés - la connaissance du vecteur $\boldsymbol{R}$ est suffisante pour décrire de façon quantitative les propriétés d'intêret pour ces systèmes.

Des modèles de potentiels semi-empiriques simples ont été développés pour les gaz dilués dans les années 1920. On peut ainsi citer le potentiel de type Lennard-Jones [55] qui est un **potentiel de paires**. Ce potentiel prend en compte la répulsion à faible portée entre deux atomes liée au recouvrement des orbitales atomiques et l'attraction à longue portée liée au potentiel Coulombien. Le potentiel Lennard-Jones peut être décrit de la façon suivante :

$$V_{LJ}(\boldsymbol{R}) = 4\epsilon \sum_{i=1}^{N} \sum_{1 \leq i < j \leq N} \left[\left(\frac{\sigma}{|R_i - R_j|} \right)^{12} - \left(\frac{\sigma}{|R_i - R_j|} \right)^{6} \right] \tag{1.12}$$

Ici, ϵ et σ sont les paramètres du potentiel Lennard-Jones. Les potentiels de Lennard-Jones ont été grandement utilisés dans la littérature pour décrire le comportement des gaz dilués, d'amas modèles ou le comportement de certains verres [56-59]. Malheureusement, **les potentiels de paires sont incapables de reproduire la physique de la liaison métallique** de par leur simplicité. Il est par exemple impossible de prédire correctement les constantes élastiques d'un métal cubique avec un potentiel de paires. De nouveaux formalismes ont alors été developpés, afin de rendre compte de l'interaction à N-corps nécessaire à la description des interactions métalliques.

Les potentiels d'interaction de paires sont une sous-classe de potentiel à N-corps :

$$\mathcal{O}_{i,2} = \sum_{j=1}^{N} f_{i,2}(\boldsymbol{R}^j, \boldsymbol{R}^i) \tag{1.13}$$

Les potentiels d'interaction à N-corps décrivent **l'expression d'une grandeur $\mathcal{O}_i$ d'un atome i comme étant une fonction de l'ensemble des positions de tous les autres atomes du système**.

$$\mathcal{O}_{i,N} = f_{i,N}(\boldsymbol{R}^1, \boldsymbol{R}^2, \cdots, \boldsymbol{R}^{N-1}, \boldsymbol{R}^N) \tag{1.14}$$

La partie potentielle de l'équation de *Schrödinger* Eq. (1.1) est une interaction à N-corps. L'expression de la fonction f_N est en général extrêmement complexe car elle implique tous les couplages possibles entre les différentes composantes $\boldsymbol{R}$. Les travaux de Friedel [60] et la découverte du théorème des moments par Ducastelle *et al.* [53] introduisent la nécessité d'un terme supplémentaire dans l'observable d'énergie de paires afin de rendre compte de la liaison métallique. Nous allons décrire rapidement le formalisme permettant d'obtenir le terme de dépendance à N-corps via la théorie des *liaisons fortes*.

Le modèle des liaisons fortes se base sur une approximation de l'équation de *Schrödinger* (1.1). Dans le cadre de cette approximation, on considère que l'interaction entre un noyau et ses électrons domine toutes les autres interactions. Dans ce cas, ses électrons restent localisés autour de leur noyau d'origine. On peut donc traiter l'équation de *Schrödinger* pour chaque noyau et ses électrons de façon isolée. La résolution de l'équation de *Schrödinger* Eq. (1.1) isolée permet de trouver une fonction d'onde uni-électronique solution pour chaque atome i (dont les coordonnées du noyau sont $\boldsymbol{R}^i$) que l'on note $\psi^i(\boldsymbol{r}-\boldsymbol{R}^i)$. Le modèle des liaisions fortes consiste à considérer que la fonction d'onde totale du sytème à N atomes est une combinaison linéaire des fonctions d'onde pour les atomes isolés. Ainsi la fonction d'onde totale s'écrit sous la forme :

$$\Psi\left(\boldsymbol{r}, \left\{\boldsymbol{R}^i\right\}_{1\leq i\leq n}\right) = \sum_{i=1}^{n} \gamma_i(\boldsymbol{R}^i)\ \psi^i(\boldsymbol{r}-\boldsymbol{R}^i) \tag{1.15}$$

Les $\gamma_i(\boldsymbol{R}^i)$ sont des scalaires reliés à la projection de la fonction d'onde $|\Psi\rangle$ sur les orbitales $\psi^i(\boldsymbol{r}-\boldsymbol{R}^i)$. Il est alors possible de calculer des observables du système, telle que l'énergie totale par l'opération $E_{TB}(\boldsymbol{R}) = \langle\Psi|\hat{\boldsymbol{\mathcal{H}}}|\Psi\rangle$. L'expression de cette énergie totale $E_{TB}(\boldsymbol{R})$ ne dépend plus que du vecteur position des noyaux $\boldsymbol{R}$. Ce type de modèle a d'abord été développé pour les métaux de transition - où le remplissage de la bande d est important - mais a ensuite été étendu aux autres métaux.

Dans le formalisme des liaisons fortes, les électrons apparaissent encore de façon explicite. Si on considère un nombre fini d'orbitales pour construire les fonctions d'onde $\psi^i(\boldsymbol{r}-\boldsymbol{R}^i)$ on peut écrire un modèle paramétrique du potentiel d'interaction $V_{TB}(\boldsymbol{R})$ **ne faisant plus directement intervenir les électrons**. Ce potentiel ne dépend que des positions $\boldsymbol{R}$ des noyaux et d'une base tronquée d'orbitales localisées pour chaque atome. Une telle construction permet d'élaborer des potentiels d'interaction ne traitant plus directement la résolution de l'équation de *Schrödinger*, mais étant représentatifs des propriétés quantiques des électrons (de par la base tronquée d'orbitales localisées). Cette approche implique une nouvelle formulation de l'observable d'énergie locale pour l'atome de coordonnées $\boldsymbol{R}^i$ du système sous la forme :

$$E(\boldsymbol{R}^i) = -\Theta \sum_{i=1}^{n} \sqrt{\sum_{1\leq j\neq i\leq n} h(\boldsymbol{R}^{j,i})} + H(\boldsymbol{R}^i) \tag{1.16}$$

Ici, Θ est une constante relative à l'occupation de la bande d'électrons d, $h(\boldsymbol{R}^{j,i})$ est reliée à l'intégrale de sauts pour les orbitales de l'atome i et l'atome j et $H(\boldsymbol{R}^i)$ est

une énergie locale. On constate que l'expression donnée par la théorie des liaisons fortes Eq. (1.16) exclut l'utilisation d'un formalisme d'interaction de paires pour les métaux de transition. L'expression analytique de ce terme d'interaction à N-corps dans l'expression d'énergie de paires permet de passer outre le formalisme de la *théorie de la fonctionnelle de la densité* et de réduire l'estimation de l'observable d'énergie du système à l'expression d'une fonction analytique. On peut distinguer deux grandes écoles nées dans les années 1970 et prenant en compte l'interaction à N-corps pour décrire la liaison métallique : (i) l'école Française autour de Friedel *et al.* [60] et de Ducastelle *et al.* [53] via la théorie des *liaisons fortes* et des *moments* décrite plus haut qui a ensuite été mise en oeuvre par Finnis, Sinclair et Sutton [61, 62] et (ii) l'école Américaine de Daw et Baskes [25, 26] via le formalisme *Embededded Atom Method* (EAM) qui sera décrit dans la sous-section suivante.

Nous retiendrons ici les principales caractéristiques des approches *semi-empiriques* à travers les points suivants : elles possèdent des (i) formes analytiques simples, (ii) sont représentatives de certaines propriétés d'interactions du matériau et (iii) ne dépendent que du vecteur $\boldsymbol{R}$ de positions des noyaux du système. Dans la suite de ce chapitre, nous allons décrire l'approche EAM développée par Daw et Baskes [25, 26] et qui est aujourd'hui la plus largement répandue pour construire des potentiels *semi-empiriques* pour les métaux de transition.

1.3.2 Embedded Atom Method

Le formalisme Embedded Atom Method a été developpé dans les années 1980 par Daw et Baskes [25, 26]. On veut exprimer l'énergie totale d'un système de N atomes en utilisant seulement les positions des noyaux $\boldsymbol{R}$ tout en rendant compte de la non-localité des électrons autour de leur noyau. Dans le cadre du formalisme Embedded Atom Method, on considère l'observable d'énergie suivante :

$$E\left[n(\boldsymbol{r}), \{\boldsymbol{R}_i\}_{0\leq i\leq n}\right] = E_e[n(\boldsymbol{r})] + \int_{\mathbb{R}^3} \hat{\boldsymbol{V}}_{\boldsymbol{en}}\left\{n(\boldsymbol{r}), \{\boldsymbol{R}_i\}_{0\leq i\leq n}\right\} d\boldsymbol{r} + \sum_{0\leq i<j\leq n} \frac{Z_i Z_j}{|\boldsymbol{R}_i - \boldsymbol{R}_j|} \tag{1.17}$$

Le premier terme de cette observable correspond aux interactions électrons-électrons, le second terme aux interactions électrons-noyaux et le dernier terme aux interactions noyaux-noyaux. Sous l'hypothèse que la densité électronique $n(\boldsymbol{r})$ peut être décomposée en une combinaison linéaire de densités locales centrées, notées n^i pour l'atome i, l'observable d'énergie peut être ré-écrite comme une somme de potentiels de paires :

$$V_{EAM}(\boldsymbol{R}) = \sum_{0\leq i<j\leq n} f(|\boldsymbol{R}_i - \boldsymbol{R}_j|) + \sum_{i=0}^{n} g(n^i) \tag{1.18}$$

Ici, g et f sont des fonctions analytiques définissant le modèle EAM utilisé. Dans les faits, la densité locale n^i centrée sur l'atome i s'écrit $n^i = \sum_{j=0}^{n} \alpha_j(\boldsymbol{R}_j)\, n(|\boldsymbol{R}_i - \boldsymbol{R}_j|)$, où les $\alpha_j(\boldsymbol{R}_j)$ sont des scalaires reliés aux projections de la densité électronique sur l'atome i et ne dépend plus que du vecteur de positions des noyaux $\boldsymbol{R}$. De même

que pour les liaisons fortes, les potentiels EAM ne dépendent plus que du vecteur de coordonnées des noyaux $\boldsymbol{R}$ et se décomposent à l'aide de fonctions simples ne nécessitant plus la résolution de l'équation de *Schrödinger* à M électrons.

Le formalisme **EAM** est un formalisme ne prenant en compte que les distances entre atomes. Il est important de noter que ce formalisme peut être étendu aux angles entre les atomes. Ce formalisme est appelé Modify Embedded Atom Method (MEAM) et a été developpé par Baskes *et al.* [63].

1.3.3 Dynamique moléculaire *classique*

L'utilisation des potentiels *semi-empiriques* permet de réduire la complexité numérique d'évaluation de l'énergie d'un système. En introduisant une fonction de *cut-off* pour les interactions énergétiques, la complexité numérique des potentiels *semi-empiriques* évolue comme $\mathcal{O}(N)$ où N est le nombre d'atomes dans le système, au lieu de au plus $\mathcal{O}(N^3)$ pour les méthodes *ab initio*. Les potentiels *semi-empiriques* permettent donc d'effectuer des calculs rapides d'énergie pour des systèmes relativement grands. De plus, l'expression simple des potentiels *semi-empiriques* permet de calculer les forces induites par le potentiel sur chaque atome. **La dynamique moléculaire consiste à laisser le système évoluer selon le *principe fondamental de la dynamique***. On peut alors suivre l'évolution de la quantité de mouvement et des positions du système - évoluant dans le potentiel $V(\boldsymbol{R})$ - au cours du temps :

$$\begin{cases} d\boldsymbol{R} &= \boldsymbol{M}^{-1} \cdot \boldsymbol{P}\, dt \\ d\boldsymbol{P} &= -\boldsymbol{\nabla}_{\boldsymbol{R}} V(\boldsymbol{R})\, dt \end{cases} \tag{1.19}$$

Ici, $\boldsymbol{P} \in \mathbb{R}^{3N\times 1}$ et $\boldsymbol{M} \in \mathbb{R}^{3N\times 3N}$ sont respectivement le vecteur quantité de mouvement et la matrice de masses du système. $\boldsymbol{\nabla}_{\boldsymbol{R}} \in \mathbb{R}^{3N\times 3N}$ est l'opérateur gradient par rapport au vecteur $\boldsymbol{R}$. L'équation (1.19) décrit la dynamique d'un système de coordonnées $(\boldsymbol{R}_0, \boldsymbol{P}_0)$ initiales à l'instant t_0 dans l'espace des phases. On peut alors calculer les positions et la quantité de mouvement du système $(\boldsymbol{R}_t, \boldsymbol{P}_t)$ à n'importe quel instant t en intégrant le système d'équations (1.19) entre t_0 et t. La dynamique décrite par le système d'équations Eq. (1.19) est valable dans un système dont le nombre d'atomes N, le volume V et l'énergie E sont constants au cours du temps (ensemble NVE).

L'ensemble NVE n'est pas très commode pour modéliser des systèmes physiques dans des conditions de température finie. Dans le cadre de notre étude, nous voudrions pouvoir accéder à l'ensemble NVT - c'est-à-dire pour les systèmes dont le nombre d'atomes N, le volume V et la température T sont fixés pendant la simulation - via la dynamique moléculaire. Il est possible de changer d'ensemble statistique en introduisant de nouvelles variables dans l'expression du *principe fondamental de la dynamique* Eq. (1.19). On peut ainsi ajouter la variable de tempéraure et/ou ajouter des forces non-conservatives. L'équation

de Langevin permet de travailler dans l'ensemble NVT en ajoutant la variable de température et des forces non-conservatives :

$$\begin{cases} d\boldsymbol{R} &= \boldsymbol{M}^{-1} \cdot \boldsymbol{P}\, dt \\ d\boldsymbol{P} &= -\boldsymbol{\nabla}_{\boldsymbol{R}} V(\boldsymbol{R})\, dt - \gamma \boldsymbol{M}^{-1} \cdot \boldsymbol{P}\, dt + \sqrt{2\gamma\beta^{-1}}\, d\boldsymbol{W} \end{cases} \tag{1.20}$$

Ici, $\gamma > 0$ est le coefficient de viscosité et $\beta^{-1} = k_B T$. $\boldsymbol{W} \in \mathbb{R}^{3N}$ est un processus de Wiener tel que $\forall t_n, t_{n+1}$ indépendants, $\int_{t_n}^{t_{n+1}} d\boldsymbol{W} \overset{\text{loi}}{\sim} \boldsymbol{\mathcal{N}}(0, \sqrt{t_{n+1} - t_n})$ où $\boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{\Sigma}_{t_n}^{t_{n+1}})$ est une loi normale multi-dimensionnelle de moyenne nulle et dont la matrice de covariance vérifie $\boldsymbol{\Sigma}_{t_n}^{t_{n+1}} = (t_{n+1} - t_n)\mathbf{1}_{3N}$. L'intérêt de l'équation de Langevin Eq. (1.20) dans les calculs de moyennes d'ensembles statistiques sera developpé dans le chapitre 5.

L'utilisation de potentiels *semi-empiriques* permet de calculer de **longues trajectoires dynamiques dans l'espace des phases** pour des systèmes allant jusqu'à 10^7 atomes.

1.3.4 Domaine d'applicabilité des méthodes *semi-empiriques*

Les potentiels *semi-empiriques* ne prenant pas explicitement en compte les électrons du système, leur complexité numérique est beaucoup plus faible que celle des méthodes *ab initio*. La complexité des méthodes *semi-empiriques* évolue, en général, comme $\mathcal{O}(N)$ ce qui permet de simuler des systèmes de tailles allant de 10^4 à 10^7 atomes (comprises entre le nm et le $10^2 nm$). Grâce à la dynamique moléculaire *classique*, les échelles de temps simulées peuvent aller de la ns à la μs. Les méthodes *semi-empiriques* prennent en compte certaines propriétés physiques de l'échelle nanoscopique mais ne reproduisent pas certaines propriétés prédites par les méthodes de type *ab initio* et restent numériquement coûteuses pour des système de plus de 10^7 atomes. **Il est important de noter ici que le pouvoir prédictif des méthodes *semi-empiriques* est souvent limité. En effet, leur mode d'ajustement peu flexible basé sur des "fonctions de bases physiques" peut conduire à des comportements non-physiques pour certains systèmes contenant des structures de taille nanoscopique [64]. Ce pouvoir prédictif - cette transférabilité - limité pour des objets étendus de l'ordre du nm (boucle de dislocation) fixe leur limite d'utilisation en termes d'espace**.

Néanmoins, les méthodes *semi-empiriques* permettent d'effecteur des calculs de longues *trajectoires dynamiques* pour des systèmes comportant des défauts étendus telles que les dislocations. Enfin, ces méthodes permettent d'ouvrir une nouvelle transition d'échelle vers les calculs de moyennes d'ensembles statistiques.

1.4 Méthodes statistiques : Monte Carlo thermodynamiques et cinétiques

Les méthodes *semi-empiriques* décrites dans la partie précédente permettent de construire des trajectoires longues dans l'espace des phases par rapport aux méthodes

ab initio. Néanmoins, la durée de ces trajectoires et la taille des systèmes simulés restent relativement modestes et ne permettent pas de simuler précisement l'évolution ou les propriétés cinétiques de systèmes réels. Les méthodes statistiques permettent d'effectuer un **échantillonnage efficace de l'espace des phases afin de réaliser : (i) des calculs de propriétés thermodynamiques de systèmes de plus grandes tailles que par la *dynamique moléculaire classique* et (ii) simuler des trajectoires cinétiques plus longues que la *dynamique moléculaire classique*.** Dans les méthodes statistiques, nous nous intéressons à l'ensemble des trajectoires possibles du système dans l'espace des phases.

L'ensemble des trajectoires de durée t détermine les propriétés cinétiques d'un système et, dans la limite $t \to \infty$, détermine l'état d'équilibre thermodynamique du système. Seules ces propriétés cinétiques et thermodynamiques "macroscopiques" (a minima mésoscopiques) peuvent être déterminées expérimentalement et peuvent donc être comparées avec des simulations. Ces grandeurs sont le résultat de moyennes statistiques sur un grand nombre d'atomes. C'est par exemple le cas de l'énergie interne d'un système ou du coefficient de diffusion d'une impureté ou d'un défaut. Nous décrivons brièvement deux grandes méthodes permettant de calculer des grandeurs dérivées de moyennes d'ensembles statistiques : (i) les méthodes Monte Carlo thermodynamiques et (ii) les méthodes Monte Carlo cinétiques.

1.4.1 Méthodes Monte Carlo thermodynamiques

Les méthodes dites de Monte Carlo thermodynamiques visent à calculer la valeur d'équilibre d'une observable thermodynamique $\mathcal{O}$ d'un système dans un ensemble statistique donné. Dans la suite de ce manuscrit, nous allons nous intéresser plus précisément à l'ensemble NVT.

Dans le cadre de l'ensemble NVT, on peut montrer que le calcul de la valeur d'équilibre d'une observable $\langle \mathcal{O} \rangle_\pi$ est directement lié à la mesure de probabilité $\pi(\boldsymbol{q}, \boldsymbol{p})$ ($\boldsymbol{q}$ et $\boldsymbol{p}$ sont respectivement les coordonnées géneralisées et les quantités de mouvement du système). Cette mesure de probabilité se définit de la façon suivante :

$$\pi(\boldsymbol{q}, \boldsymbol{p}) = \frac{e^{-\beta \mathcal{H}(\boldsymbol{q}, \boldsymbol{p})}}{\int_{\mathcal{Q} \times \mathcal{P}} e^{-\beta \mathcal{H}(\boldsymbol{q'}, \boldsymbol{p'})} d\boldsymbol{q'} d\boldsymbol{p'}} \tag{1.21}$$

Ici, $\beta = (k_B T)^{-1}$ et k_B est la constante de Boltzmann. Cette mesure de probabilité est plus communément appelée mesure de Boltzmann. Dans ces conditions, la valeur moyenne $\langle \mathcal{O} \rangle_\pi$ de l'ensemble NVT est donnée par la formulation suivante :

$$\langle \mathcal{O} \rangle_\pi = \int_{\mathcal{Q} \times \mathcal{P}} \mathcal{O}(\boldsymbol{q}, \boldsymbol{p}) \pi(\boldsymbol{q}, \boldsymbol{p}) d\boldsymbol{q} d\boldsymbol{p} \tag{1.22}$$

où $\mathcal{Q} \times \mathcal{P}$ est l'espace des phases et $\mathcal{H}(\boldsymbol{q}, \boldsymbol{p})$ est l'Hamiltonien du système considéré. L'intégrale nécessaire au calcul doit se faire sur l'intégralité de l'espace des phases et se revèle donc très difficile à évaluer sans un **échantillonnage efficace**. Cet échantillonnage

s'effectue grâce à l'application du *théorème ergodique* [65]. Ce théorème montre qu'il est équivalent d'effectuer une moyenne empirique de $\mathcal{O}$ le long d'une trajectoire determinée par une dynamique $\phi_{0,\tau}$ permettant d'échantillonner la mesure Boltzmannienne que de calculer $\langle\mathcal{O}\rangle_\pi$. On définit la dynamique $\phi_{0,\tau}$ comme vérifiant la conditions suivante : $\phi_{0,0} = \mathbf{1}$. En partant des coordonnées initiales $(\boldsymbol{q_0}, \boldsymbol{p_0})$ la dynamique $\phi_{0,\tau}$ génère une trajectoire dans l'espace des phases selon la relation suivante :

$$\phi_{0,\tau}(\boldsymbol{q_0}, \boldsymbol{p_0}) = (\boldsymbol{q_\tau}, \boldsymbol{p_\tau}) \tag{1.23}$$

où, $(\boldsymbol{q_\tau}, \boldsymbol{p_\tau})$ sont les coordonnées généralisées au temps τ générées par ϕ. Dans la limite $\tau \to \infty$, on obtient alors l'égalité suivante :

$$\langle\mathcal{O}\rangle_\pi = \lim_{\tau\to+\infty} \frac{1}{\tau} \sum_{\tau'\leq\tau} \mathcal{O}(\phi_{0,\tau'}(\boldsymbol{q_0}, \boldsymbol{p_0})) \tag{1.24}$$

Le *théorème ergodique* et la notion de dynamique $\phi_{0,\tau}$ seront abordés en détail dans le chapitre 5. L'équation de *Langevin* [66, 67] décrite par l'équation (1.20) est un exemple de dynamique vérifiant les conditions du *théorème ergodique*. Nous citerons aussi l'algorithme fondateur des méthodes stochastiques pour l'évaluation d'observables de mécaniques statistiques : l'algorithme de **Metropolis-Hasting** [68, 69] est utilisé dans la majorité des codes Monte Carlo afin d'échantillonner la mesure canonique. L'équation (1.24) donne une façon simple d'évaluer une valeur moyenne d'ensemble statistique d'une observable $\mathcal{O}$ du système, notamment son énergie interne U. L'énergie interne a d'ailleurs été mainte fois étudiée dans un grand nombre de systèmes en sciences des matériaux [70]. Il est néanmoins utile de noter que la trajectoire décrite par la dynamique $\phi_{0,\tau}$ n'est pas une image d'une transformation "physique" du système.

Supposons un système initial $\mathcal{S}_0$ composé de deux éléments A et B. Le système $\mathcal{S}_0$ est une solution homogène composée d'autant d'atomes A que d'atomes B. Supposons maintenant que A et B ont tendance à la démixtion, on applique alors la dynamique ϕ^τ à $\mathcal{S}_0$. Le système $\mathcal{S}_0$ va évoluer vers son équilibre thermodynamique $\mathcal{S}_1$ où les composants A et B vont se séparer. Une question se pose alors : la trajectoire décrite entre $\mathcal{S}_0$ et $\mathcal{S}_1$ induite par $\phi_{0,\tau}$ est-elle représentative de l'évolution cinétique réelle du système ? La réponse est non : la dynamique $\phi_{0,\tau}$ est un **chemin thermodynamique possible** mais pas nécessairement un **chemin cinétique possible** du système. Les méthodes Monte Carlo thermodynamiques ne décrivent pas la cinétique réelle d'un système. Il est alors nécessaire d'introduire une nouvelle classe de méthode appelée méthodes Monte Carlo cinétiques.

1.4.2 Méthodes Monte Carlo cinétiques

Les méthodes Monte Carlo cinétiques permettent d'obtenir le temps associé à une transformation d'un système. Elles nécessitent la connaissance de ce que l'on nomme les taux de transitions pour une collection d'événements $\{\mathcal{E}_i\}$. On appelle événement $\mathcal{E}$ l'association des deux instances : (i) un état du système d'énergie minimum noté $\mathcal{E}, m$ et un état de point de selle du système connecté à $\mathcal{E}, m$ et noté $\mathcal{E}, s$.

On peut alors définir le taux de transition entre un état $\mathcal{E}, m$ et un état $\mathcal{E}, s$, $R_{\mathcal{E},m\to s}$. Considérons un bassin noté i et l'ensemble des transitions reliant i vers d'autres bassins k ($\Gamma_{i\to k}$). On note $\{R_{\mathcal{E},i\to k}\}_{k\leq N_k}$ l'ensemble des N_k taux de transitions de i vers les k. Ces taux de transition sont dimensionnellement homogènes à un $temps^{-1}$. Il est alors possible de décrire un temps "physique" de sortie du bassin i. Ainsi, le temps de sortie du bassin i, $T_{i\to}$. suit la loi de probablité suivante :

$$T_{i\to\cdot} \overset{\text{loi}}{\sim} \mathcal{P}\left(\sum_{k=1}^{N_i} R_{\mathcal{E},i\to k}\right) \tag{1.25}$$

où $\mathcal{P}\left(\sum_{k=1}^{N_i} R_{\mathcal{E},i\to k}\right)$ est une loi de Poisson de paramètre $\sum_{k}^{N_k} R_{\mathcal{E},i\to k}$. En considérant l'ensemble des transitions possibles en partant du bassin initial i vers un bassin k, les méthode Monte Carlo permettent d'estimer le temps nécessaire pour effectuer cette transition. L'objectif des méthodes de Monte Carlo cinétiques est de donner une description correcte, et dont le coût numérique est acceptable, des taux de transitions qui peuvent être étendus à tous les bassins de l'espace des phases $R(\boldsymbol{q}, \boldsymbol{p})$. Dans la majorité des méthodes Monte Carlo cinétiques utilisées dans la littérature, on ne connait pas l'expression $R(\boldsymbol{q}, \boldsymbol{p})$ sur l'ensemble de l'espace des phases mais seulement sur une restriction de l'espace des phases. Cet espace des phases réduit définit le type de la méthode Monte Carlo cinétique utilisée.

Nous allons alors distinguer plusieurs types de méthodes Monte Carlo cinétiques :

— (i) Les méthodes Monte Carlo cinétiques atomistiques [71] où l'on suit directement les transitions possibles des atomes. On peut alors distinguer les méthodes de Monte Carlo sur réseau et les méthodes hors-réseaux [72, 73]. Ces méthodes nécessitent la connaissance de l'ensemble des transitions possibles du système pour chaque configuration atomique. Dans le cas des méthodes Monte Carlo cinétiques hors-réseau, les taux de transitions $R(\boldsymbol{q}, \boldsymbol{p})$ sont calculables sur l'ensemble de l'espace des phases. Dans le cas des Monte Carlo cinétiques sur réseau, la connaissance et le calcul des $R(\boldsymbol{q}, \boldsymbol{p})$ se réduisent à un espace des phases discrétisé sur l'ensemble des sites du réseau que l'on peut noter $(\mathbb{Q}, \mathbb{P})$.

— (ii) Les méthodes Monte Carlo cinétiques objets [10, 17, 74-78] où l'on suit l'évolution d'objets tels que des défauts ponctuels ou étendus telle des boucles de dislocation. Les objets peuvent interagir entre eux afin de créer de nouveaux objets. Le but de ces méthodes est de suivre certaines populations d'objets d'interêt au cours de la simulation. Dans les méthodes Monte Carlo cinétiques objets, l'espace des phases accessibles se réduit aux coordonnées des différents objets présents dans le système et est noté $(\mathfrak{q}, \mathfrak{p}) \subset (\boldsymbol{q}, \boldsymbol{p})$. Cet espace des phases est aussi discret mais de dimension encore plus faible que pour les méthodes Monte Carlo atomistiques sur réseau.

— (iii) Les méthodes Monte Carlo cinétiques basées sur les événements consistent à décrire de façon la plus exhaustive possible la collection d'événements $\{\mathcal{E}_i\}$ du système. Ce Monte Carlo suit la réalisation d'événements parmi la collection $\{\mathcal{E}_i\}$ au cours du temps. À la différence des autres méthodes où l'on a le choix ou non d'accepter la transition entre t et $t + \delta t$, on choisit une transition dans la collection d'événements $\{\mathcal{E}_i\}$ et on incrémente le temps de la durée caractéristique de la transition choisie. Les méthodes Monte Carlo cinétiques basées sur les événements font abstraction de la dépendance des taux R vis-à-vis de l'espace des phases. Dans ces méthodes, les taux de transition sont simplement rassemblés dans la collection $\{\mathcal{E}_i\}$.

In fine, on a les relations suivantes vis-à-vis de la description de l'espace des phases :

$$(\mathfrak{q}, \mathfrak{p}) \subset (\mathbb{Q}, \mathbb{P}) \subset (\boldsymbol{q}, \boldsymbol{p}) \tag{1.26}$$

La description des phénomènes pris en compte par les méthodes Monte Carlo cinétiques est d'autant plus fine que l'espace des phases est décrit de façon précise. Néanmoins, cette description fine de l'espace des phases augmente le temps numérique nécessaire à la convergence de ces méthodes. L'utilisation des méthodes Monte Carlo cinétiques se traduit par un compromis entre le temps numérique et la précision de la description de l'espace des phases.

1.4.3 Domaine d'applicabilité des méthodes Monte Carlo

Les méthodes type Monte Carlo permettent le calcul des grandeurs mésoscopiques ou macroscopiques grâce aux moyennes d'ensembles statistiques. Ces méthodes permettent aussi bien de calculer des grandeurs thermodynamiques, c'est-à-dire représentatives de l'état d'équilibre du système en temps infini, que des grandeurs cinétiques, c'est-à-dire décrivant l'évolution temporelle du système au cours d'une transformation. Ces méthodes peuvent être appliquées à des systèmes atomistiques ou à des systèmes "continus" contenant des représentations atomistiques [17, 74]. Ces méthodes permettent de simuler des systèmes de tailles allant de 10^6 à 10^8 atomes (comprise entre $10^1 nm$ et μm) et sur des échelles de temps pouvant aller de la μs à la min.

La principale limitation en termes de taille et de temps d'évolution d'un système dépend de la représentativité des interactions énergétiques entre les atomes du système. C'est également le cas lors de l'utilisation de la *dynamique moléculaire* ou de méthodes à champs moyens [17, 74] qui peuvent se révéler peu transférables pour des systèmes de grandes tailles et pour de longues évolutions temporelles. L'autre limitation majeure est le choix, *a priori*, d'une collection d'événements $\{\mathcal{E}_i\}$. En effet, une collection trop grande nécessite un grand nombre de simulations en aval ou *à la volée* dans l'algorithme Monte Carlo et une collection trop petite risque de négliger des transitions d'intérêt pour l'évolution cinétique du système.

1.5 Approches Machine Learning pour la simulation multi-échelles des matériaux

Les approches Machines Learning en sciences des matériaux visent à étendre les échelles de temps et d'espace accessibles à la simulation dans le diagramme présenté dans la figure 1.1. Pour cela, on peut définir différents outils d'apprentissage automatique permettant de couvrir différents domaines de l'étude multi-échelles. On va distinguer trois classes de méthodes différentes : (i) les **métamodèles**, (ii) les **potentiels Machine Learning**, (iii) les **méthodes couplantes non-supervisées**. Nous proposons de mettre à jour le diagramme donné par la figure 1.1 en figurant les domaines accessibles par les trois méthodes citées. Ce nouveau diagramme multi-échelles est présenté dans la figure 1.2.

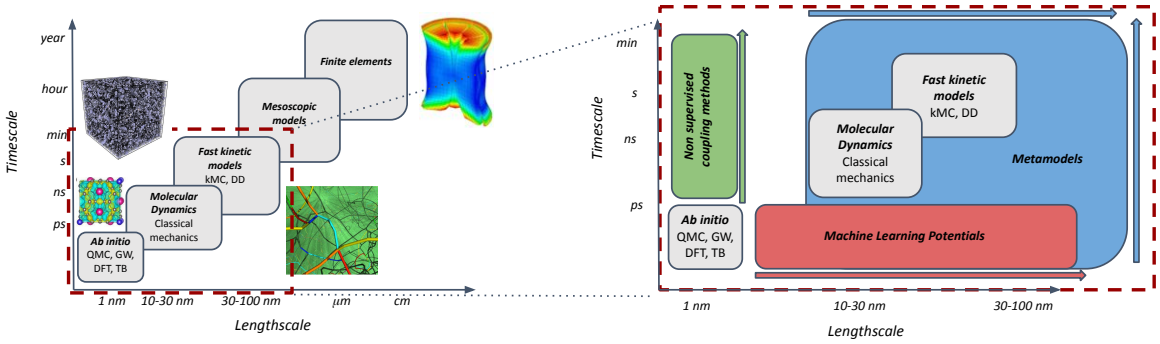

Figure 1.2: Illustration des différentes méthodes de simulations pour l'analyse multi-échelles. Les échelles de temps et d'espace caractéristiques des différentes méthodes de simulation sont données par l'axe des abscisses et des ordonnées. Ce diagramme est mis à jour en incluant de nouveaux domaines temps-espace accessibles par les méthodes Machine Learning.

1.5.1 Métamodèles

On appelle métamodèle un modèle de régression de hautes dimensions permettant de calculer la valeur d'une observable en partant d'une représentation des coordonnées $\boldsymbol{q}$ d'un système. Des métamodèles ont été developpés dans la littérature avant l'essor des méthodes Machine Learning. Des exemples simples sont les modèles de régression des barrières de migration dans les alliages métalliques en se basant sur la concentration locale en éléments d'alliage [79]. Dans ce cas, l'observable étudiée est la barrière de migration et la représentation du système consiste en la concentration locale.

Les métamodèles, dans le cadre des méthodes Machine Learning, se basent sur des représentations plus systématiques appelées **descripteurs** dont les principales caractéristiques seront décrites dans le chapitre 2. L'utilisation de métamodèles permet de calculer la valeur d'une observable d'intérêt pour un coût numérique plus faible

que son évaluation directe. Dans le chapitre 3, nous développons un métamodèle de régression de l'entropie vibrationnelle dont la complexité numérique évolue comme $\mathcal{O}(N)$ alors que la méthode "traditionnelle" de calcul de cette grandeur nécessite une procédure dont la complexité évolue comme $\mathcal{O}(N^3)$ (où N est le nombre d'atomes du système).

Un métamodèle se construit autour d'une représentation des coordonnées atomiques $\underline{D}(\boldsymbol{q})$. On choisit ensuite un certain nombre de systèmes représentatifs du problème que l'on veut étudier et que l'on rassemble sous la forme d'une base de données $\{\underline{D}(\boldsymbol{q_k})\}_{k\leq m}$ composée de m systèmes. Enfin, on introduit une fonction f permettant de faire le lien entre l'observable et la représentation. Le métamodèle de l'observable $\mathcal{O}^{\mathcal{M}}$ s'écrit alors de la façon suivante :

$$\{\underline{D}(\boldsymbol{q_k})\}_{1\leq k\leq m} \rightarrow \{\mathcal{O}_k^{\mathcal{M}}\}_{1\leq k\leq m} = \{f(\boldsymbol{q_k})\}_{1\leq k\leq m} \tag{1.27}$$

La base de données $\{\underline{D}(\boldsymbol{q_k})\}_{1\leq k\leq m}$ est appelée base d'**entraînement**. Cette base de données est un des points centraux des métamodèles qui seront décrits dans le chapitre 3. Il est important de noter que les métamodèles permettent de décrire l'observable $\mathcal{O}$ seulement dans certaines conditions et ne permettent pas forcément de calculer d'observables secondaires issues de $\mathcal{O}$ telles que ses dérivées (par exemple, les forces ou les contraintes à partir de l'observable d'énergie). Dans le chapitre 3, notre modèle de régression de l'entropie vibrationnelle n'est valable que pour des minima du paysage énergétique considéré et n'a pas de raison d'être correctement défini sur un chemin de migration. Ces conditions d'application les distinguent des **potentiels Machine Learning**.

Ce type de procédure permet de simuler des systèmes dont la taille est beaucoup plus grande que le système initial. Si le métamodèle est bien pensé, celui-ci sera **transférable** et permettra de construire une estimation de l'observable d'une qualité proche de l'observable de référence utilisée pour l'entraînement. En d'autres termes, un ajustement d'un métamodèle sur des données issues et contenant les "informations physiques" de la théorie de la fonctionnelle de la densité permettra d'obtenir et de prédire des valeurs de l'observable d'intérêt pour des systèmes plus grands tout en conservant "la physique" de la théorie de la fonctionnelle de la densité. Les nouveaux domaines du diagramme temps-espace accessibles par l'utilisation de métamodèles sont décrits par les flèches et les bulles bleues en Fig. 1.2.

1.5.2 Potentiels Machine Learning

Les potentiels Machine Learning sont une classe particulière de métamodèles de l'observable d'**énergie potentielle** d'un système. Ces métamodèles doivent être valables sur l'ensemble de l'espace des phases. De plus, les dérivées partielles par rapport aux positions du système doivent évidemment être représentatives des **forces exercées dans le système**. La base de données d'entraînement doit donc faire l'objet d'une construction minutieuse. La construction de ce type de potentiel sera décrite

plus précisement dans le chapitre 6.

Les potentiels Machine Learning visent à construire une nouvelle classe de potentiels *semi-empiriques* plus précis et plus transférables que les potentiels décrits dans la section 1.3. Les potentiels Machine Learning pour la simulation atomistique ont commencé à apparaître dans les années 2000. Il en existe un grand nombre basés sur différents formalismes. Les détails techniques des formalismes cités ici seront décrits dans le chapitre 2.

L'approche de régression la plus simple est l'approche linéaire entre l'observable d'énergie locale d'un atome du système et la représentation de son environnement local. Les **potentiels Machine Learning linéaires** [20, 80-83] sont plus précis et plus transférables que les potentiels *semi-empiriques classiques*. Ils possèdent un comportement "physique" sur l'ensemble de leur base de données d'entraînement mais leur erreur vis-à-vis des quantités thermodynamiques prédites (telle que l'énergie) est toujours relativement grande (de l'ordre de 10 meV/atome) par rapport à la base de données d'entraînement. Afin d'obtenir des résultats plus précis, vis-à-vis de la base de données d'entraînement, il est nécessaire d'introduire une non-linéarité de la fonction f Eq. (1.27). On distingue deux grandes classes de méthodes non-linéaires : (i) les réseaux de neurones et (ii) les méthodes à noyaux. Ces classes de méthodes seront décrites plus précisément dans le chapitre 2.

Les méthodes à **réseaux de neurones** décrivent l'observable d'énergie du système comme une fonction non-linéaire des représentations locales des atomes. L'expression analytique de cette fonction n'est pas connue et est contenue dans le réseau de neurones. Les méthodes à réseaux de neurones [84-88] permettent d'obtenir de grandes précisions vis-à-vis des données d'entraînement mais sont moins transférables que les approches linéaires - leur erreur sur des données qui ne sont pas contenues dans la base d'entraînement est plus grande que les modèles linéaires -. Les **méthodes à noyaux**, décrivent l'observable d'énergie du système comme une fonction non-linéaire des représentations locales des atomes. Cependant, contrairement aux réseaux de neurones, l'expression analytique de la fonction f est connue. Les méthodes à noyaux [89-96] permettent d'obtenir de grandes précisions vis-à-vis des données d'entraînement mais sont moins transférables que les approches linéaires. On trouve un grand nombre de potentiels de type Machine Learning dans la littérature. Afin de comparer les potentiels entre-eux la littérature utilise une référence commune développée en 2010 : GAP (Gaussian Approximation Potential) [97, 98]. GAP est un potentiel Machine Learning se basant sur le formalisme à noyau.

Les potentiels Machine Learning visent à étendre la transférabilité des potentiels *semi-empiriques* classiques en termes de temps et d'espace, grâce à leur meilleure précision. En effet, les *potentiels semi-empiriques* classiques peuvent conduire à des comportements non-physiques pour certains systèmes contenant des structures étendues de l'ordre du *nm* [64]. **Les potentiels Machine Learning permettent d'obtenir une**

"transition d'échelle" tout en conservant la "physique" de la base de données d'entraînement. Nous définissons la transition d'échelle comme étant la capacité d'une méthode, pour une échelle donnée, à décrire une grandeur physique avec la précision de l'échelle inférieure. Dans le chapitre 6, nous allons ajuster l'énergie d'activation de la mono-lacune dans différents métaux cubiques centrés afin d'effectuer des calculs d'*énergie libre* tout en conservant une précision *ab initio*, le tout dans un temps de calcul acceptable. Les nouveaux domaines du diagramme temps-espace accessibles par l'utilisation des potentiels Machine Learning sont donnés par les flèches et les bulles rouges en Fig. 1.2.

1.5.3 Méthodes couplantes non-supervisées

Nous appelons **méthodes couplantes non-supervisées**, une classe de méthodes utilisant déja des potentiels Machine Learning comme champ de force et permettant un échantillonnage optimisé de l'espace des phases. **Ces méthodes permettent de calculer les propriétés de températures finies du système**. On va distinguer deux types de grandeurs à température finie : (i) les grandeurs thermodynamiques et (ii) les grandeurs cinétiques. Cette classe de méthode permet de calculer avec précision la valeur d'un **potentiel thermodynamique** pour un ensemble statistique donné. Dans des conditions réelles, on se focalisera généralement sur l'*énergie libre* $F = U - TS$ (où U est l'énergie interne, T la température et S est l'entropie du système) ou sur l'*enthalpie libre* $G = U + PV - TS$ (où P est la pression et V est le volume du système). L'*énergie libre* est le potentiel thermodynamique de l'ensemble NVT et l'*enthalpie libre* est le potentiel thermodynamique de l'ensemble NPT. Dans la suite de ce manuscrit, nous nous intéressons plus précisément au cas de l'ensemble NVT et donc à l'*énergie libre*.

La grandeur thermodynamique essentielle de l'ensemble NVT est l'*énergie libre*. L'utilisation d'un métamodèle de l'entropie vibrationnelle [22] ainsi que l'évaluation d'un potentiel Machine Learning afin d'estimer la valeur d'*énergie libre* d'un système pour une température donnée est un exemple de méthode couplante non-supervisée. Dans le cas plus général, la précision des potentiels Machine Learning peut être utilisée de façon couplée avec une méthode d'échantillonnage d'énergie libre telle que celle développée par Wang *et al.* [99], des méthodes à potentiels biaisants adaptatifs [66, 100-102] ou des méthodes à forces biaisantes adaptatives [66, 103-106]. Lors d'une migration, la grandeur essentielle de l'ensemble NVT est l'*énergie libre de migration*. Le calcul de cette grandeur nécessite l'utilisation d'une variable globale du système appelée **coordonnée de réaction**. La construction de la *coordonnée de réaction* pour une migration donnée est, aujourd'hui encore, un exercice très ardu. Seule une poignée de méthodes permet à l'heure actuelle de construire une *coordonnée de réaction*. La méthode des *strings* [107-110] permet de construire la *coordonnée de réaction* grâce aux chemins d'énergies minimales. On peut aussi se baser sur une *coordonnée NEB*. Ce type d'approche a notamment été développé par Swinburne *et al.* [111] et permet de construire, *à la volée*, une *coordonnée de réaction* en température. Des approches de type auto-encodeur permettent de construire des *coordonnées de réaction* sans aucun *a*

priori [112]. Certaines méthodes se basent sur un échantillonnage direct dans l'espace de représentation des données où la distribution Boltzmannienne présente une forme plus simple [112-115].

Dans le chapitre 6, nous allons utiliser ce type de méthodes couplantes afin de calculer le coefficient d'auto-diffusion dans différents métaux cubiques centrés. Ce type d'approche permet d'obtenir des observables de **grande échelle de temps pour des systèmes de tailles faibles**. Les nouveaux domaines du diagramme temps-espace accessibles par l'utilisation des méthodes couplantes non-supervisées sont donnés par les fléches et les bulles vertes Fig. 1.2.

1.6 Conclusions de chapitre

La figure 1 montre les différents domaines d'applicabilité des méthodes de simulation numériques en termes de temps et d'espace. Le changement d'échelle nécessite de changer régulièrement de formalisme et de méthodes afin de conserver un coût numérique acceptable. On note que ce diagramme multi-échelle est **"lacunaire"**. En effet un grand nombre de domaines sont inaccessibles aux méthodes de simulations "classiques" telles que la DFT, la *dynamique moléculaire classique* ou les méthodes Monte Carlo. Les méthodes Machine Learning, appliquées aux simulations atomistiques, ont pour but d'étendre les domaines accessibles de ce diagramme. On note alors deux grandes directions possibles : (i) augmenter l'échelle en espace ou (ii) augmenter l'échelle en temps.

L'augmentation du domaine d'échelles d'espace est possible grâce à l'utilisation de métamodèles. Nous décrirons dans le chapitre 3 l'utilisation d'un métamodèle de l'entropie vibrationnelle harmonique de défauts ponctuels. Son coût numérique est plus faible que les méthodes "classiques" de calcul de cette grandeur. Ce type de métamodèle permet d'espérer estimer des grandeurs complexes, telle que l'entropie vibrationnelle harmonique, pour des objets étendus - comme des boucles de dislocation - où les méthodes "classiques" sont impossibles à mettre en place.

L'augmentation du domaine d'échelles en temps est possible grâce aux méthodes couplantes non-supervisées. Nous décrirons dans le chapitre 6 un schéma de calcul complet permettant d'obtenir la dépendance en température du coefficient d'auto-diffusion pour différents métaux cubiques centrés. Cette méthode dépasse le cadre de l'approximation harmonique et permet de reproduire des propriétés importantes de ces métaux en température, comme leur expansion thermique.

Please remain calm
The end has arrived
We cannot save you
Enjoy the ride.

— Parasite Eve, Bring Me The Horizon

2

Représentation des environnements atomiques locaux et régressions

Sommaire

Ce chapitre va présenter plus en détails la notion de **descripteurs atomiques** et de **modèles de régression** décrits dans le chapitre 1. Ces deux éléments sont les coeurs des métamodèles et des potentiels de type Machine Learning définis par l'équation (1.27). Les métamodèles et les potentiels de type Machine Learning mettent en jeu deux éléments importants : (i) la représentation de l'environnement atomique $\underline{D}(\boldsymbol{q})$ introduite dans la section 1.5.1 et (ii) la fonction permettant la régression de l'observable à partir de $\underline{D}(\boldsymbol{q})$, f. Dans ce chapitre, nous allons d'abord décrire de façon quantitative la notion de descripteurs atomiques en sciences des matériaux ainsi que les descripteurs que nous avons utilisés dans la suite de ce manuscrit. Puis nous décrirons les différentes méthodes de régression pour les problèmes en hautes dimensions.

2.1 Descripteurs atomiques en sciences des matériaux

Nous allons ici définir la notion de descripteurs atomiques locaux (voir Section 2.1.2) introduite dans le chapitre précédent 1. Les descripteurs atomiques sont au centre des méthodes Machine Learning appliquées aux matériaux. Dans un premier temps, nous rappelons l'introduction et l'intérêt de l'utilisation des descripteurs locaux dans les méthodes de simulations multi-échelles Sec. 2.1.1. Dans un deuxième temps, nous décrivons (de façon non exhaustive) les différents descripteurs atomiques locaux développés dans la littérature. Nous nous focalisons particulièrement sur ceux que nous allons utiliser dans les chapitres suivants.

2.1.1 Descripteurs atomiques locaux : définition et exemple simple

Comme nous l'avons expliqué dans le chapitre 1, les méthodes Machine Learning appliquées à la science des matériaux permettent d'accéder à de nouveaux domaines du diagramme espace-temps (Fig. 1.2).Afin d'être efficaces, elles nécessitent de ne pas prendre directement les coordonnées du système $\boldsymbol{q}$ comme entrée, mais une représentation de ces coordonnées que l'on appelle descripteur. Les descripteurs sont les fonctions faisant le lien entre l'espace des configurations et ce que l'on appelle l'**espace de représentation**.

La construction de l'espace de représentation est donnée par le choix du descripteur utilisé. On définit un descripteur local centré sur l'atome i de la façon suivante :

$$\begin{aligned} \underline{D}_i : \mathbb{R}^{\mathcal{V}(i)} &\rightarrow \mathbb{R}^{\mathcal{D}} \\ \boldsymbol{q}_{\mathcal{V}(i)} &\rightarrow \underline{D}_i(\boldsymbol{q}) \end{aligned} \tag{2.1}$$

Ici $\boldsymbol{q}_{\mathcal{V}(i)} \in \mathbb{R}^{\mathcal{V}(i)}$ est le vecteur de coordonnées atomiques du système $\boldsymbol{q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_N)$ restreint au voisinage de l'atome i et $\underline{D}_i$ est une fonction permettant la projection de l'environnement local de l'atome i dans l'espace de représentation. Dans le cas d'une quantité physique extensive - et qui se décompose localement de façon exacte -, on

peut alors construire un descripteur **global** du système par sommation $\underline{D} = \sum_{i=1}^{N} \underline{D}_i$. Ce descripteur prend bien en compte l'ensemble de l'espace des configurations $\mathbb{R}^{3N}$. Le vecteur associé à $\underline{D}$ est de dimension $\mathcal{D}$. Il existe un descripteur local par atome dans le système traduisant le changement d'espace suivant : $\mathbb{R}^{3N} \rightarrow \mathbb{R}^{\mathcal{D}\times N}$. Dans le cas général $\mathcal{D} \gg 3$, le problème posé dans l'espace des descripteurs possède donc une dimensionalité intrinsèque plus grande que le problème originalement posé dans l'espace des phases. L'idée principale derrière l'espace de représentation est la **linéarisation du problème original**. En effet, si on projette le problème initial dans un espace de suffisamment grande dimension, celui-ci pourra être linéarisé. Ce problème est analogue à celui de la cartographie : dans le but de représenter une carte, on utilise un système de projection pour passer d'un problème à trois dimensions à deux dimensions et on "aplanit" le problème. L'espace des descripteurs a le même rôle : transformer un problème de topologie complexe (ces notions de topologie seront développées plus précisément dans le chapitre (4)) en un problème suffisamment "plat" pour qu'il soit linéarisable. La dimensionalité du problème devient alors $\mathcal{D} \ll 3N$. Comme pour la cartographie, le choix du système de projection ne doit pas être fait au hasard et dépend du problème que nous voulons étudier. Les descripteurs doivent respecter des **propriétés de symétries** importantes et leurs composantes doivent être suffisamment non-colinéaires pour décrire de façon quasi-unique [1] une configuration.
Nous allons maintenant présenter les principales propriétés de symétries que doit vérifier un descripteur. On nomme $\mathfrak{G}$ le groupe de symétrie du système que nous étudions. Soit $\boldsymbol{g} \in \mathfrak{G}$ et $\boldsymbol{q}$ le vecteur de coordonnées du système, on a :

$$\forall \boldsymbol{g} \in \mathfrak{G}, \quad \boldsymbol{g}(\boldsymbol{q}) = \boldsymbol{q} \tag{2.2}$$

L'équation (2.2) définit rigoureusement le groupe de symétrie du système comme étant l'ensemble des transformations laissant invariantes les coordonnées du système. Les groupes de symétries les plus importants pour la science des matériaux sont : (i) le groupe des **permutations**, qui assure l'équivalence pour les échanges de deux particules identiques ; (ii) le groupe des **translations** ; (iii) le groupe des **transformations orthogonales** comme les rotations et les réflexions pour les systèmes cristallins. L'équivalence entre deux configurations différentes par la composition d'un élément d'un groupe de symétrie du système doit aussi se traduire dans l'espace des descripteurs c'est-à-dire pour le descripteur local centré sur l'atome i :

$$\forall \boldsymbol{g} \in \mathfrak{G}, \quad \underline{D}_i \circ \boldsymbol{g} = \underline{D}_i \tag{2.3}$$

Ici, $\circ$ représente l'opérateur de composition de $\boldsymbol{q}$ par $\boldsymbol{g}$. Cette condition peut être exprimée de façon plus élégante par la proposition suivante : *les descripteurs doivent posséder au moins les symétries du système*. La notion de descripteur que nous décrivons jusqu'ici reste abstraite. Nous allons donc donner deux exemples simples de "descripteurs atomiques" respectant les conditions présentées plus haut. Le premier exemple simple de descripteur atomique local est la **coordinance de l'atome** i. La

1. Dans le cas d'un passage $\mathbb{R}^{3N} \rightarrow \mathbb{R}^{\mathcal{D}}$ l'injectivité ne peut pas être assurée et donc la bijectivité non plus.

coordinance vérifie toutes les propriétés de symétries et d'invariance du système et permet de donner des informations qualitatives sur l'environnement local de l'atome i. L'analyse de la coordinance permet de détecter, par exemple, la présence d'une lacune ou d'un atome interstitiel dans l'environnement local de l'atome i. Néanmoins, cette information n'est pas quantitative, au sens où la coordinance seule ne pourra pas permettre de construire un modèle d'énergie locale de l'atome i qui sera précis. Afin de construire un modèle plus quantitatif, il faut choisir un descripteur contenant plus d'informations que la coordinance. On peut citer un autre exemple simple de descripteur local permettant d'effectuer des analyses quantitatives : il s'agit de la **matrice de Couloumb** [116-118]. La matrice de Coulomb est généralement utilisée pour des systèmes de petites tailles et où les interactions électrostatiques sont importantes, par exemple les molécules organiques. Dans ce cas, les coordonnées du système sont remplacées par la matrice suivante :

$$C_{ij} = \begin{cases} Z_i^{2.4} & \text{si } i = j \\ \frac{Z_i Z_j}{|\boldsymbol{q_i} - \boldsymbol{q_j}|} & \text{sinon} \end{cases} \tag{2.4}$$

Ici, Z_i et $\boldsymbol{q_i}$ représentent respectivement la charge électrique et la position de l'atome i. La matrice de Coulomb représente un deuxième exemple de descripteur simple. Celle-ci respecte les propriétés de symétries et d'invariance du système. Mais, contrairement à la coordinance, la matrice de Couloumb contient suffisamment d'informations pour prédire de façon quantitative des observables du système, telle que son énergie [116-118]. Un grand nombre de descripteurs atomiques plus complexes et plus ou moins sophistiqués ont été developpés dans la littérature. Nous choisissons ici de décrire les "grandes classes" de descripteurs : (i) les descripteurs **radiaux et angulaires**, (ii) les descripteurs basés sur des **réseaux de neurones**, (iii) les descripteurs **spectraux** et (iv) les descripteurs **tensoriels**.

Descripteurs radiaux et angulaires

La construction analytique de descripteurs a été initiée par Behler et Parrinello [84, 119, 120] et Bartók *et al.* [97, 98, 121] et forme la base actuelle des descripteurs locaux utilisés en sciences des matériaux. Les descripteurs radiaux construisent des invariants en ne prenant en compte que les distances entre paires d'atomes. Le premier descripteur radial utilisé est $\boldsymbol{G_3}$ créé par Behler et Parrinello [84]. D'autres descripteurs, prenant en compte les distances de paires et les angles de triplets ont alors été developpés. On peut citer le descripteur $\boldsymbol{G_2}$ [84, 119, 120] et le descripteur Angular Fourier Series (AFS) issu de Bartók *et al.* [97, 98, 121]. Ces descripteurs permettent de construire des modèles de régression ou des potentiels Machine Learning dont la précision est plus grande que le formalisme EAM ou MEAM. Ces descripteurs possèdent en général entre 10 et 100 composantes.

Descripteurs basés sur des réseaux de neurones

La construction peut aussi être spécifique pour un système donné [91, 122-124] où les descripteurs peuvent être directement construits à l'aide de méthodes de type *deep learning* [113, 125-127]. Dans ce cas, les descripteurs sont construits de façon automatique par le réseau de neurones, qui construit les symétries et les invariances. Il est important de noter que la **forme analytique des descripteurs n'est pas connue** ici. Celle-ci est "cachée" dans les poids du réseau de neurones. Ce type d'approche ne permet pas de garantir l'invariance par certaines opérations de symétrie, si celles-ci ne figuraient pas dans la base de données d'entraînement.

Descripteurs spectraux

Les descripteurs spectraux se basent sur la décomposition de la densité atomique d'un système sur une base de fonctions hypersphériques. Les propriétés des fonctions hypersphériques permettent alors de construire des invariants. La forme et la qualité des descripteurs vont dépendre de la fonction de distribution choisie. On peut choisir la densité atomique comme une somme de **fonctions deltas** centrées sur chaque atome du système : on aboutit alors au descripteur bi-spectrum SO(4) [121]. On peut aussi choisir de travailler avec de Gaussiennes de largeurs finies et on aboutit alors au descripteur SOAP [97, 98, 121]. L'utilisation de Gaussiennes plutôt que de fonctions deltas permet d'assurer la construction d'un descripteur plus lisse. Les deux descripteurs cités précédemment sont des descripteurs locaux, mais il est aussi possible de construire des descripteurs spectraux multi-échelles. C'est le cas des *coefficients de Scattering* développés par Mallat *et al.* [128-131]. Dans le cas des *coefficients de Scattering*, la construction de l'invariance est plus générale et ne se base plus sur les propriétés spécifiques des fonctions hypersphériques. L'utilisation de la convolution avec des ondelettes de différentes échelles permet de construire un descripteur dépassant l'information locale. On peut distinguer deux familles principales pour ces descripteurs : (i) les descripteurs compacts, tel que le bi-spectrum SO(4) [121] et possédant généralement entre 10 et 50 composantes et (ii) les descripteurs non-compacts tels que SOAP [97, 98, 121] et les *coefficients de Scattering* [128-131] possédant entre 100 et 4000 composantes. Le caractère compact ou non de ce type de descripteur dépend notamment de la façon dont sont couplées les informations radiales et angulaires dans le descripteur (par exemple entre le bi-spectrum SO(4) et SOAP).

Descripteurs tensoriels

Les descripteurs tensoriels représentent une nouvelle classe de méthodes permettant de construire de façon systématique des invariants à un groupe de symétrie. L'approche de construction de l'invariant est proche de celles des *coefficients de Scattering*. Les méthodes tensorielles utilisent cette même "astuce" mais ne se limitent pas à une base de fonctions hypersphériques. On peut alors contruire des descripteurs invariants par rotation et par permutations à l'aide de fonctions de bases polynomiales [132, 133]. La construction du descripteur peut faire intervenir des interactions à N atomes. C'est le cas des descripteurs construits par Oord et Allen [134, 135]. Des cas particuliers de ces

méthodes aboutissent aux descripteurs cités précedemment. Ainsi, si on utilise comme fonction de base les fonctions hypersphériques et qu'on se limite à un développement tensoriel d'ordre 2, les descripteurs tensoriels aboutissent à SOAP [97]. On citera aussi le descripteur ACE (atomic cluster expansion) [136, 137], lui aussi basé sur les propriétés d'invariance des harmoniques sphériques et faisant intervenir explicitement des interactions à J-corps pour une expansion à l'ordre J. Les descripteurs tensoriels possèdent un nombre extrêmenent élevé de composantes (entre 1000 et 10000) ce qui les rend généralement difficilement utilisables directement dans la pratique.

Autres types de descripteurs et descripteurs utilisés

On peut enfin constuire des descripteurs plus "exotiques" composés de la concaténation de plusieurs descripteurs. C'est ce que l'on appelle des descripteurs hybrides [20]. Il est aussi possible de construire des descripteurs plus "physiques", par exemple basés sur la projection de la fonction d'onde sur des orbitales atomiques données [21].

Dans la suite, nous allons porter une attention particulière aux descripteurs analytiques locaux implémentés dans le package `MILADY` [20, 138] et sur les *coefficients de Scattering*. La sous-section suivante (2.1.2) sera dédiée à la description quantitative des descripteurs suivants : (i) le descripteur $\boldsymbol{G_2}$ premier descripteur développé par Behler et Parrinello [84, 119, 120], (ii) les descripteurs Angular Fourier Series (AFS) [121] ; (iii) le descripteur bi-spectrum SO(4) [121] et (iv) les *coefficients de Scattering* [128-131].

2.1.2 Principaux descripteurs atomiques utilisés

Historiquement, le premier descripteur atomique quantitatif a été développé par Behler et Parrinello [84, 119, 120] et consiste en un descripteur radial. Le descripteur $\boldsymbol{G_2}(\eta_{max}, R^s_{max})$ consiste seulement en une norme de paires dans l'espace des Gaussiennes et est défini par l'expression suivante :

$$\boldsymbol{G^i_2}(\eta, R^s) = \sum_{k \in \mathcal{V}(i)} e^{-\eta(r_{ik}-R^s)^2} f_c(r_{ik}) \tag{2.5}$$

On note ici $\mathcal{V}(i)$ l'ensemble des atomes dans le voisinage de l'atome i, r_{ik} est la distance euclidienne entre l'atome i et l'atome k, enfin f_c est appelée fonction de *cut-off*. C'est une fonction lisse telle que pour toute distance $r \geq r_{cut}$ on a $f_c(r) \equiv 0$. Le scalaire $0 \leq \eta \leq \eta_{max}$ permet de définir l'étalement de la Gaussienne et donc la vitesse à laquelle l'interaction de paires décroît. Le paramètre R^s sert à ajuster le "centre" de la distribution de valeur du descripteur. Le choix de la valeur de ces paramètres est discuté par Behler et Parrinello [84, 119, 120]. Le descripteur $\boldsymbol{G_2}$ vérifie les propriétés d'invariance par permutations, translations et rotations de façon évidente. Ce premier descripteur ne contient aucune information angulaire ce qui restreint son utilisation à des problèmes simples. Le formalisme de ce descripteur est très fortement inspiré de celui des potentiels EAM [25, 26] mais constitue la première introduction de la notion de descripteur atomique local. Le descripteur résultant pour les collections $\{\eta_k\}_{0\leq k\leq d_1}$ et $\{R^S_l\}_{0\leq l\leq d_2}$ s'écrira $\underline{D_i} = (\boldsymbol{G^i_2}(\eta_1, R^s_1), \boldsymbol{G^i_2}(\eta_1, R^s_2), \cdots, \boldsymbol{G^i_2}(\eta_i, R^s_j), \cdots, \boldsymbol{G^i_2}(\eta_{d_1}, R^s_{d_2}))$

et la dimension de ce descripteur sera $D = d_1 d_2$.

Le descripteur Angular Fourier Series a été initialement introduit par Bartók *et al.* [121] et consiste en un descripteur produit entre les informations radiales et les informations angulaires de l'environnement atomique local. Considérons le descripteur $\mathcal{A}^i_{n,l}$ possédant n canaux radiaux, l canaux angulaires et centré sur l'atome i. On a alors l'expression analytique suivante :

$$\mathcal{A}^i_{n,l} = \sum_{k,k' \in \mathcal{V}(i)} g_n(r_{ik}) g_n(r_{ik'}) \cos\left(l\theta_{ik,ik'}\right) f_i(r_{ik}) f_i(r_{ik'}) \tag{2.6}$$

On note ici $\mathcal{V}(i)$ l'ensemble des atomes dans le voisinage de l'atome i, r_{ik} est la distance euclidienne entre l'atome i et l'atome k et $\theta_{ik,ik'}$ représente l'angle formé par le triplet d'atomes i, k, k' centré sur i. Le descripteur contient l'ensemble des informations de paires et de triplets d'atomes dans le voisinage $\mathcal{V}(i)$ de l'atome i. Concernant les fonctions quantitatives impliquées dans l'expression (2.6) : (i) f_i est une fonction lisse de *cut-off*, telle que pour toute distance $r \geq r_{cut}$ on a $f_i(r) \equiv 0$. (ii) Les fonctions radiales $g_n(r)$ sont des fonctions polynomiales décroissantes avec la distance r et de degrés $\alpha + 2$ avec $0 \leq \alpha \leq n$. Une description plus précise de ces fonctions est donnée par Bartók *et al.* [121]. (iii) Les fonctions angulaires $\cos(l\theta)$ sont les polynômes de Tchebyshev avec $0 \leq l \leq l_{max}$. En tant que descripteur produit $\mathcal{A}^i_{n_{max},l_{max}}$ possède $n_{max}(l_{max}+1)$ canaux et vérifie toutes propriétés d'invariance par permutations, translations et rotations. D'un point de vue qualitatif, le descripteur AFS propose une grille radiale et une grille angulaire indépendantes et dont la finesse peut être choisie de façon très flexible. La limitation principale de l'AFS est sa nature produit. En effet, dans le cas d'un problème nécessitant une grille radiale et une grille angulaire fine, son nombre de canaux peut vite devenir très grand. On pourra noter un autre inconvénient lié à la construction même de ce descripteur : les fonctions g_n sont construites par un processus d'orthonormalisation de fonctions polynômiales. Dans le cas de fonctions se recouvrant fortement (ce qui est le cas des fonctions utilisées par Bartók *et al.* [121]) cette procédure d'orthonormalisation peut créer des instabilités numériques pour les grandes valeurs de α. Dans le reste de ce manuscrit, nous continuerons à noter le descripteur AFS sous la forme $\mathcal{A}_{n_{max},l_{max}}$.

Le descripteur bi-spectrum SO(4), $bSO(4)_{j_{max}}$, est un descripteur spectral basé sur la décomposition de la densité atomique locale sur la base des fonctions *hyper-sphériques* en 4 dimensions [97, 121]. Il existe en effet une bijection entre l'espace réel $\mathbb{R}^3$ et l'hyper-sphère unité $\mathcal{S}^4 \in \mathbb{R}^4$. L'environnement atomique de l'atome i est décrit par sa densité $\rho_i(\boldsymbol{r})$ et se décompose de la façon suivante sur les fonctions *hyper-sphériques* :

$$\rho_i(\boldsymbol{r}) = \sum_{k \in \mathcal{V}(i)} w_k \delta(\boldsymbol{r} - \boldsymbol{r_k}) \tag{2.7}$$

$$= \sum_{k \in \mathcal{V}(i)} \sum_{j=0}^{\infty} \sum_{m,m'=-j}^{j} \boldsymbol{c}^{m,m'}_{i,j} U^{m,m'}_j \tag{2.8}$$

Comme pour les descripteurs AFS, le voisinage $\mathcal{V}(i)$ est défini comme étant la sphère de rayon r_{cut} centrée sur l'atome i. La densité est identiquement nulle en dehors de ce voisinage. Ici w_k est une scalaire dépendant de l'espèce chimique, $\boldsymbol{c}_{i,j}^{m,m'}$ est le résultat du produit scalaire entre la fonction de densité centrée sur l'atome i et l'*hyper harmonique sphérique* $U_j^{m,m'}$. En utilisant l'équation (2.8) et les coefficients $\boldsymbol{c}_{i,j}^{m,m'}$ on peut déduire la décomposition en puissance spectrale et le bi-spectrum de la densité atomique centrée sur i. On définit alors les composantes du bi-spectrum SO(4) de la façon suivante avec $j \leq j_{max}$ et $|j_1 - j_2| \leq j \leq j_1 + j_2$:

$$B^i_{jj_1j_2} = (\boldsymbol{c}_{i,j}^{m,m'})^\dagger \boldsymbol{H}^{\boldsymbol{j_1 j_2}} (\boldsymbol{c}_{i,j_1}^{m_1,m'_1} \otimes \boldsymbol{c}_{i,j_2}^{m_2,m'_2}) \tag{2.9}$$

Ici, $\boldsymbol{H}^{\boldsymbol{j_1 j_2}}$ sont reliés aux coefficients de Clebsch-Gordan pour la représentation du groupe SO(4). Une description détaillée de ces coefficients est donnée par Bartók *et al.* [97, 121]. Le bi-spectrum SO(4) respecte les propriétés d'invariance par translations et par permutations de façon évidente. La propriété d'invariance par rotation est plus subtile et rendue possible par les propriétés des fonctions hypersphériques et la construction même du descripteur [97, 121]. Une partie des composantes du bi-spectrum SO(4) est nulle par propriété des coefficients de Clebsch-Gordan ; on peut alors utiliser les composantes non nulles ou seulement les composantes diagonales c'est-à-dire celle correspondant à $j_1 = j_2$ [97, 121, 139]. Le bi-spectrum SO(4) donne une description très sensible de l'environnement atomique. En effet une faible différence entre des configurations $\boldsymbol{q}_1$ et $\boldsymbol{q}_2$ en termes de norme euclidienne peut induire une forte différence en termes de représentations irréductibles et donc une forte différence entre $\underline{D}(\boldsymbol{q}_1)$ et $\underline{D}(\boldsymbol{q}_2)$ (en termes de norme euclidienne). Bien que le bi-spectrum SO(4) soit très sensible au changement de symétries, celui-ci possède un inconvénient majeur venant de sa construction. En effet, il est impossible de découpler l'information radiale et angulaire dans les composantes du bi-spectrum SO(4) en raison de leur projection sur la sphère $\mathcal{S}^4$. D'un point de vue qualitatif, le $bSO(4)$ donne une description angulaire très fine mais nécessite un temps de calcul conséquent en raison de l'expression récurrente des coefficients de Clebsch-Gordan [139] ce qui limite la valeur maximale de j. Dans le reste de ce manuscrit, nous continuerons à noter le bi-spectrum SO(4) sous la forme $bSO(4)_{j_{max}}$.

La *transformée en ondelettes* proposée par Mallat *et al.* [128, 129] permet la construction d'un descripteur respectant les propriétés d'invariance par translations et par rotations avec un formalisme explicitement multi-échelles. Pour cela, on exprime la densité du système $\mathcal{C}$ comme une somme de Gaussiennes centrées sur les positions des atomes :

$$\rho(\boldsymbol{r}) = \sum_{k \in \mathcal{C}} g(\boldsymbol{r} - \boldsymbol{r_k}). \tag{2.10}$$

Les *coefficients de Scattering* $S^{\mathcal{J},L}\rho[j,\ell], j \in \mathcal{J}, 0 \leq l \leq L$ sont calculés par convolutions successives de la densité ρ avec les ondelettes $\psi_{j,\ell}^m$ de différentes échelles $j \in \mathcal{J}$.

L'ensemble est ensuite rendu invariant par rotations et par translations par intégration sur l'ensemble des translations et des rotations :

$$S^{\mathcal{J},L}\rho[j,\ell] = \int_{\mathbb{R}^3} \left(\sum_{m=-\ell}^{\ell} |\rho * \psi_{j,\ell}^m(\boldsymbol{r})|^2 \right)^{1/2} \mathrm{d}\boldsymbol{r} \tag{2.11}$$

$$\psi_{j,\ell}^m(\boldsymbol{r}) = \frac{1}{(\sqrt{2\pi})^3}\, e^{-\frac{1}{2}\left|\frac{\boldsymbol{r}}{2^j}\right|^2} \left|\frac{\boldsymbol{r}}{2^j}\right|^\ell\, Y_\ell^m\left(\frac{\boldsymbol{r}}{|\boldsymbol{r}|}\right) \tag{2.12}$$

Ici, Y_ℓ^m sont les *harmoniques sphériques* sur la sphère $\mathcal{S}^3$. Le formalisme développé par Mallat *et al.* [128, 129] permet la prise en compte explicite des interactions entre les différentes échelles, ce qui n'est pas le cas des descripteurs locaux. Néanmoins, l'interaction des différentes échelles via les convolutions ne permet pas de distinguer l'origine des contributions dans les *coefficients de Scattering*. Le descripteur et l'information qu'il contient résultent d'une "moyenne" entre les différentes échelles, ce qui rend l'analyse difficile. Dans le reste de ce manuscrit, nous continuerons à noter les *coefficients de Scattering* sous la forme $S^{\mathcal{J},L}$.

Les descripteurs atomiques décrits précédemment possèdent les qualités pour décrire de façon systématique des environnements atomiques tout en respectant des propriétés fortes d'invariances par permutations, translations et rotations. Néanmoins les descripteurs seuls ne permettent pas de remonter à des quantités thermodynamiques du système : il est nécessaire d'utiliser un modèle de régression entre l'espace des descripteurs et la quantité thermodynamique. Le choix de ce modèle est tout aussi important que le choix du descripteur lui même si on veut obtenir une régression robuste et transférable. Les principaux modèles de régressions présents dans la littérature vont être décrits dans la section suivante Sec. 2.2.

2.2 Modèles de régressions pour des quantités physiques

L'essor des méthodes de type *Machine Learning* a permis de nombreuses avancées en science des matériaux [140]. Tous ces progrès ont un point commun : l'utilisation de méthodes numériques de régressions et de classifications efficaces en grande dimension. Nous allons décrire ces méthodes dans cette section en commençant par (i) les réseaux de neurones artificiels [141] dans la sous-section 2.2.1 ; (ii) les méthodes à noyau [142] dans la sous-section 2.2.2 et nous terminerons avec (iii) des méthodes plus simples telles que les régressions linéaires ordinaires et Bayésiennes [142] dans la sous-section 2.2.3. Nous nous attacherons ici à décrire succinctement le principe de chaque méthode et dans quelles situations elles sont les plus efficaces.

Nous définissons d'abord ce que nous qualifions être un "problème de régression". Soit, $\underline{y} \in \mathbb{R}^m$ un vecteur de données que l'on veut ajuster, $\underline{\underline{x}} \in \mathbb{R}^{m\times n}$ une matrice de

données d'entrée (ou matrice de design). Le problème de régression se traduit par le choix d'une famille de fonction $f_w : \mathbb{R}^{m\times n} \to \mathbb{R}^m$ telle que :

$$\underline{\hat{y}} = f_w(\underline{\underline{x}})$$

La fonction f_w dans l'équation (2.2) définit la forme du modèle de régression. On introduit une *fonction de coût* $L\left\{\underline{y}, f_w(\underline{\underline{x}})\right\}$ qui va définir quantitativement la qualité du modèle de régression. La *fonction de coût* peut, par exemple, être l'erreur quadratique moyenne entre les quantités à ajuster et les quantités prédites par le modèle : $\|\underline{y} - f_w(\underline{\underline{x}})\|^2$. On cherche alors f^* solution du problème d'optimisation suivant :

$$f^* = \arg\min_{f_w} L\left\{\underline{y}, f_w(\underline{\underline{x}})\right\} \tag{2.13}$$

Le problème de régression se traduit donc par deux quantités : (i) la fonction f qui fait le lien entre l'espace d'entrée et la quantité à ajuster et (ii) la *fonction de coût* associée au problème. Les méthodes que nous allons présenter dans les sous-sections suivantes sont entièrement définies et se distiguent par le choix de ces deux quantités.

2.2.1 Modèles non-linéaires : réseaux de neurones artificiels

Les réseaux de neurones artificiels sont nés à la fin des années 1950 avec l'article fondateur de Rosenblatt [143] introduisant le concept de *perceptron* dont l'architecture s'inspire des cortex cérébraux et des connexions entre les neurones biologiques. Les méthodes de type *perceptron* vont, petit à petit, se complexifier par augmentation du nombre de couches et d'interconnexions du réseau grâce à l'amélioration des algorithmes d'optimisation de poids [144] et l'introduction de l'algorithme de rétro-propagation des erreurs par Werbos [145]. Au fil des années, les réseaux de neurones sont devenus la référence en terme de problème de classification d'images grâce à la mise en place de bases de données telles que `ImageNet` [146], `MNIST` [147] etc. Un nouveau pas en termes de classification d'image a été franchi par l'introduction des réseaux de neurones convolutionnels par Le Cun *et al.* [148] qui sont aujourd'hui les structures obtenant les meilleurs scores de classification sur la base de données [146, 147].
Un réseau de neurones artificiels est constitué de n couches contenant n_i neurones dans la couche i. Le passage d'entrée à la sortie du réseau de neurones peut être vu comme une cascade "d'opérations élémentaires". Chacune de ces opérations se décompose en deux temps. Á titre d'exemple plaçons-nous sur la couche i du réseau : (i) une phase d'aggrégation linéaire des sorties de tous les neurones de la couche $i-1$; (ii) l'application d'une fonction non-linéaire $\sigma(\cdot)$ sur l'entrée de la couche $i-1$ et l'ajout d'un biais b_i. Ces deux opérations élémentaires sont représentées schématiquement par la figure 2.1. L'opération élémentaire (illustrée dans la figure 2.1) entre la couche $i-1$ et i se traduit de la façon suivante :

$$\underline{\mathcal{N}}_i = \sigma_{i-1}(\underline{\underline{W}}^{i-1} \cdot \underline{\mathcal{N}}_{i-1} + \underline{b}_{i-1}) \tag{2.14}$$

où $\underline{\mathcal{N}}_i$ et $\underline{\mathcal{N}}_{i-1}$ sont respectivement les vecteurs de sorties de la couche i et $i-1$. Les paramètres $\underline{\underline{W}}^i \in \mathbb{R}^{n_{l_i}\times n_{l_{i-1}}}$ et $\underline{b}_i \in \mathbb{R}^{n_{l_i}}$ sont des matrices et vecteurs de poids qui

définissent les propriétés du réseau. L'ensemble du réseau à n couches contenant n_{l_i} neurones dans la couche l_i peut être décrit par la fonction suivante entre le vecteur d'entrée $\underline{I} \in \mathbb{R}^i$ et le vecteur de sortie $\underline{O} \in \mathbb{R}^o$:

$$\mathcal{NN}(\underline{I}) = \sigma_n(\underline{\underline{W}}^n(\sigma_{n-1}(\underline{\underline{W}}^{n-1}(\underbrace{\ldots}_{n-2 \text{ fois}} \sigma_1(\underline{\underline{W}}^1 \cdot \underline{I} + \underline{b}_1)\ldots) + \underline{b}_{n-1}) + \underline{b}_n) \tag{2.15}$$

Les paramètres $\underline{\underline{W}}^i \in \mathbb{R}^{n_{l_i} \times n_{l_{i-1}}}$ et $\underline{b}_i \in \mathbb{R}^{n_{l_i}}$ sont des matrices et vecteurs de poids qui définissent les propriétés du réseau. L'ensemble de ces poids doit être optimisé lors de la procédure d'entraînement afin de construire un modèle de régression.

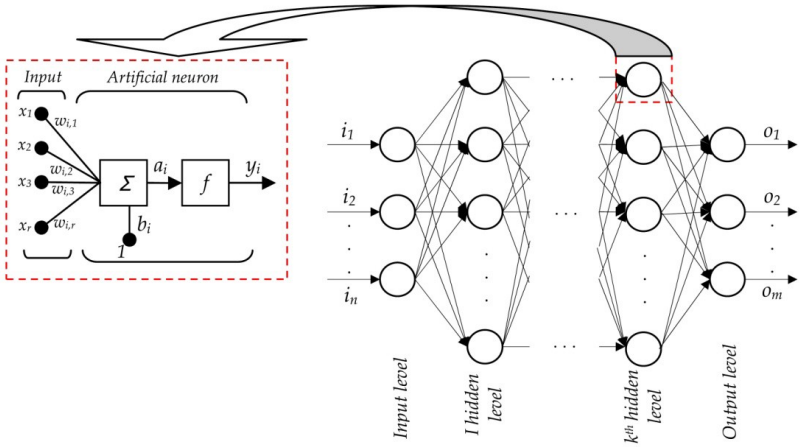

Figure 2.1: Illustration schématique de l'architecture d'un réseau de neurones artificiels. Les différents neurones du réseau sont représentés par les cercles blancs. Ces neurones sont organisés en couches successives dont les connexions sont schématisées par les traits noirs entre les neurones. Les informations d'entrées d'un neurone de la couche i sont constituées d'une combinaison linéaire des informations de sortie des neurones de la couche précédente $i-1$. On applique ensuite une linéarité à cette combinaison linéaire afin de créer la sortie du neurone de la couche i.

Cette architecture simple développée en premier lieu par Rosenblatt [143] consiste donc en une suite d'opérations linéaires suivie de l'application d'une non-linéarité. En se basant sur cette architecture, on peut démontrer deux théorèmes importants concernant les réseaux de neurones artificiels.

Le *théorème de classification* (i) qui peut être énoncé de la façon suivante :

Théorème 2.1. *Un réseau de neurones artificiels contenant une seule couche est un classificateur linéaire pour un problème linéairement séparable.*

Ce théorème se démontre facilement en remarquant que dans le cas d'un réseau monocouche, le vecteur de paramètres $\underline{W}$ peut être vu comme la normale à l'hyperplan séparant les classes de données. Le vecteur $\underline{b}$ est un *offset* permettant de placer l'hyperplan "au bon endroit". Enfin, la non linéarité permet de discriminer de quel côté de l'hyperplan se situent les données. Une illustration de la méthode de démonstration de ce théorème est donnée par la figure 2.2.

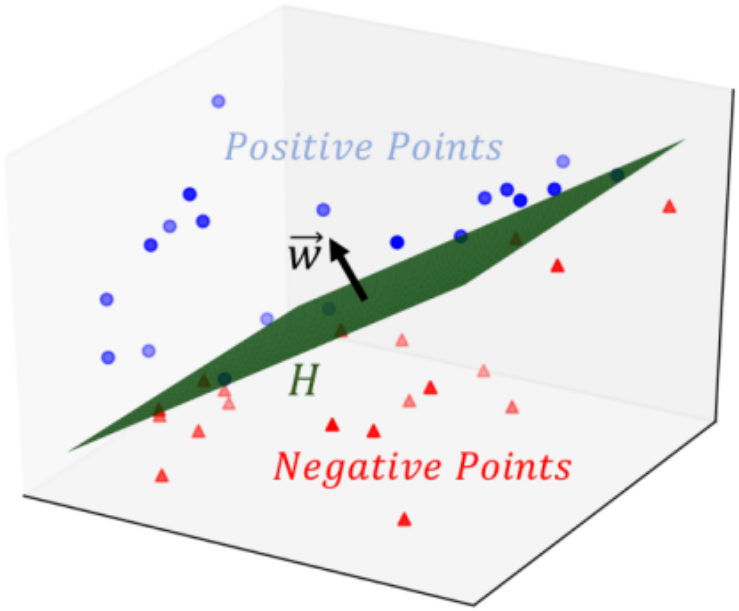

Figure 2.2: Illustration schématique de la démonstration du *théorème de classification* qui consiste seulement à placer l'hyperplan adéquat pour séparer et discriminer les données. Ici $\overline{w}$ correspond au vecteur de paramètres $\underline{W}$ et H est l'hyperplan de séparation des données.

Le *théorème d'approximation universelle* dû à Cybenko [149] pour le cas de fonction sigmoïde utilisée pour la non-linéarité peut être énoncé de la façon suivante :

Théorème 2.2. *L'ensemble des réseaux de neurones* $\mathcal{NN} : \mathbb{R}^i \to \mathbb{R}$ *est dense dans l'ensemble des fonctions* $f : \mathbb{R}^i \to \mathbb{R}$.

En d'autres termes, un réseau de neurones artificiels peut approximer n'importe quelle fonction continue et peut donc être utilisé pour construire des modèles de régression en grande dimension. Ces deux théorèmes importants expliquent l'utilisation grandissante des réseaux de neurones en *science des données*. En effet, les réseaux de neurones artificiels possèdent un grand nombre de paramètres à optimiser ce qui nécessite d'importantes bases de données. La phase d'optimisation est assurée de la façon suivante : considérons m vecteurs de données d'entrée $\underline{X}_k \in \mathbb{R}^i$ et les vecteurs de données de sortie $\underline{y}_k \in \mathbb{R}^o$. On définit alors une *fonction de coût* L :

$$L\left\{\underline{\underline{X}}, \underline{\underline{y}}\right\} = \frac{1}{m}\sum_{k=1}^{m} \ell\left\{\mathcal{NN}(\underline{X}_k), \underline{y}_k\right\} \tag{2.16}$$

Ici, $\underline{\underline{X}}$ et $\underline{\underline{y}}$ sont respectivement la concaténation des vecteurs $\boldsymbol{X}_k$ et $\boldsymbol{y}_k$. Cette *fonction de coût* doit être minimale afin de minimiser l'erreur entre les données réelles $\underline{\underline{y}}$ et les prédictions du réseau de neurones $\mathcal{NN}(\boldsymbol{y})$. La *fonction de coût* peut prendre différentes formes en fonction du type de problèmes. A titre d'exemple pour un problème de régression, cette fonction peut être l'écart quadratique $\|\underline{\underline{y}} - \mathcal{NN}(\underline{\underline{y}})\|^2$. La notion

de *fonction de coût* a un rôle central dans les problèmes de régression et sera aussi utilisée dans les sous-sections 2.2.2 et 2.2.3. Dans le cadre des réseaux de neurones artificiels, la *fonction de coût* est minimisée grâce à l'algorithme de rétro-propagation des erreurs [145]. Au cours de cette procédure, qu'on appelle entraînement, la valeur de la fonction de coût sur le jeu de données d'entraînement va diminuer et peut même atteindre la valeur de zéro. On obtient alors un modèle d'**interpolation** très efficace mais qui peut être très mauvais pour l'**extrapolation**. Ce phénomène est appelé *hyper-ajustement*, si on sépare nos données d'entrée en deux catégories : (i) des données d'entraînement, (ii) des données de vérification ; l'*hyper-ajustement* se traduit par une réduction de l'erreur sur les données d'entraînement et une augmentation de l'erreur sur la base de données de vérification. Les phénomènes d'*hyper-ajustement* limitent grandement la transférabilité des modèles construits, c'est-à-dire la capacité d'un modèle à avoir une erreur faible de prédiction sur des données qui n'ont pas été utilisées pour l'entraînement. Différentes procédures peuvent être utilisées afin d'augmenter la transférabilité d'un modèle telle que : (i) la régularisation qui sera décrite plus précisément dans la sous-section 2.2.2 qui agit directement sur la *fonction de coût* ; (ii) la *k-cross validation* [150, 151] qui permet de réduire le *sur-ajustement* en trouvant le meilleur échantillonnage possible de données pour entraîner le modèle. Dans le cas des réseaux de neurones artificiels, il peut exister une phase de généralisation [152, 153], c'est-à-dire que l'erreur sur la base de données de vérification sera du même ordre de grandeur que l'erreur sur la base de données d'entraînement. Dans ce cas, le modèle construit possède une très bonne capacité d'**interpolation** et d'**extrapolation**.

Les réseaux de neurones artificiels présentent d'excellentes performances pour la reconnaissance d'images notamment grâce à l'architecture convolutive [154]. Néanmoins le niveau de compréhension théorique de ce qui fait fonctionner ces outils reste très flou. En effet, leur structure de plus en plus complexe (augmentation du nombre de couches) et leur caractère très non-linéaire ne permettent pas de construire une théorie simple. Dans les faits ceux-ci restent des "boîtes noires" notamment en ce qui concerne les modèles de régressions, ce qui rend obscure la compréhension d'un "modèle physique" derrière les millions (voir des dizaines de millions) de paramètres optimisés dans l'architecture du réseau.

2.2.2 Modèles non-linéaires : méthodes à noyau

Les méthodes à noyau offrent un formalisme pour les problèmes de régression et de classification. Le cadre théorique des méthodes à noyaux s'appuie sur deux théorèmes importants démontrés par Mercer [155] et Schölkopf *et al.* [156] et grâce à l'apport non négligeable de Aizerman *et al.* [157]. Nous allons détailler ces deux théorèmes mais nous devons d'abord définir ce qu'est un noyau. Considérons une fonction $\phi : \mathbb{R}^n \to \mathbb{R}^n$. On définit un noyau K par son opérateur intégral T_K de la façon suivante :

$$T_K\left\{\phi\right\}(\underline{x}) = \int_{\mathbb{R}^n} K(\underline{x}, \underline{x}')\phi(\underline{x}')\underline{dx}' \tag{2.17}$$

Il découle de cette définition deux théorèmes importants. Le *théorème de Mercer* [155, 157] qui se formule de la façon suivante :

Théorème 2.3. *Soit un noyau $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ vérifiant les propriétés suivantes :*

$$\forall \underline{x}, \underline{x}' \in \mathbb{R}^n \quad K(\underline{x}, \underline{x}') = K(\underline{x}', \underline{x})$$
$$\forall \underline{x}, \underline{x}', \underline{\alpha}, \underline{\beta} \in \mathbb{R}^n \quad \int_{\mathbb{R}^n \times \mathbb{R}^n} K(\underline{x}, \underline{x}') \underline{\alpha} \underline{\beta} d\underline{x} d\underline{x}' \geq 0$$

Alors, il existe un espace de Hilbert $\mathcal{H}$ où on peut définir un produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ tel que :

$$\forall \underline{x}, \underline{x}' \in \mathbb{R}^n \quad K(\underline{x}, \underline{x}') = \langle \underline{x}, \underline{x}' \rangle_{\mathcal{H}}$$

Le *théorème de Mercer* permet de définir la notion de mesure de similarité entre deux vecteurs $\underline{x}$ et $\underline{x}'$ grâce à l'utilisation d'un noyau K et de définir la norme $K(\underline{x} - \underline{x}', \underline{x} - \underline{x}') = \|\underline{x} - \underline{x}'\|_{\mathcal{H}}$. Nous reviendrons un peu plus tard sur l'intérêt de la construction de cette norme.

Le deuxième théorème découlant de l'utilisation des noyaux est le *théorème du représentant* démontré par Schölkopf *et al.* [156]. Il peut être énoncé de la façon suivante :

Théorème 2.4. *Considérons une fonction de coût L entre des données de sortie $\underline{\underline{y}} \in \mathbb{R}^{m \times o}$ et un modèle f utilisant des données d'entrée $\underline{\underline{x}} \in \mathbb{R}^{m \times i}$. On cherche à résoudre le problème d'optimisation suivant dans l'espace de Hilbert engendré par le noyau K :*

$$f^* = \arg\min_{f \in \mathcal{H}} L\left\{\underline{\underline{y}}, f(\underline{\underline{x}})\right\} + g(\|f\|) \tag{2.18}$$

Ici g est une fonction positive et strictement croissante et est appelée régularisation. Sous les hypothèses du théorème de Mercer Th. 2.3 il existe $\underline{w} \in \mathbb{R}^m$ telle que la solution du problème d'optimisation Eq. (2.18) peut s'écrire de la façon suivante :

$$f^*(\cdot) = \sum_{k=1}^{m} w_k K(\cdot, \underline{x}_k) \tag{2.19}$$

On appelle *régularisation* le terme de pénalisation ajouté dans la *fonction de coût* (cf. Sec 2.2.4) portant sur la norme de f. Le contrôle de la norme de f implique un contrôle sur la norme du vecteur de poids $\underline{w}$. Cette méthode (générale utilisée aussi pour les réseaux de neurones et les méthodes simples de régression) permet d'éviter le *sur-ajustement*. Cet effet se comprend très bien pour les méthodes à noyau. En effet, considérons une base de données $(\underline{\underline{x}}, \underline{y})$ et supposons que la donnée $\tilde{y}_w$ associée vecteur $\underline{x}_w$ soit abbérante. Par exemple, $\tilde{y}_w = 100 y_w$ avec y_w la valeur cohérente associée au vecteur $\underline{x}_w$. Lors d'ajustement du modèle (afin de minimiser la *fonction de coût*), la norme du poids $\|w_w\|$ associée à $K(\cdot, \underline{x}_w)$ va être élevée. Or, si on veut faire une prédiction sur un nouveau vecteur $\underline{x}$ sur lequel le modèle n'a pas été ajusté, la prédiction sera fortement faussée par le poids fort de la donnée $\tilde{y}_w$. Le terme de *régularisation* permet de pallier le problème en discriminant les modèles contenant des normes de poids élevées et en privéligiant les modèles ayant un bon comportement (au

sens de l'erreur) sur l'ensemble des données. Ce type de procédure permet d'améliorer grandement la transférabilité et la robustesse des modèles construits.

Le *théorème du représentant* Th. 2.4 assure que n'importe quelle fonction peut être approximée par une décomposition sur un noyau K vérifiant les hypothèses du *théorème de Mercer* Th. 2.3. L'intérêt principal des méthodes à noyau par rapport aux réseaux de neurones artificiels est la notion de mesure de similarité donnée par la norme associée $\|\cdot\|_{\mathcal{H}}$ dans l'espace de *Hilbert* $\mathcal{H}$. L'espace $\mathcal{H}$ est appelé espace de représentation des données et présente l'avantage de posséder une norme associée. On peut donc savoir "où se situe" une nouvelle donnée par rapport aux autres grâce à $\|\cdot\|_{\mathcal{H}}$ ce qui n'est pas le cas d'un réseau de neurones artificiels car la norme associée à l'espace de représentation n'est pas directement accessible. Le formalisme à noyau et sa norme associée permettent de **donner une estimation de l'erreur de prédiction d'une nouvelle donnée sachant les données sur lesquelles le modèle a déjà été entraîné**. Schématiquement, si on considère le "barycentre" de la base de données d'entraînement $\langle \underline{x} \rangle$, et un nouveau vecteur $\underline{s}$ alors plus la distance $\|\langle \underline{x} \rangle - \underline{s}\|_{\mathcal{H}}$ est grande, plus l'erreur attendue sur la prédiction du vecteur $\underline{s}$ sera grande. Ce contrôle systématique de l'estimation de l'erreur est l'avantage majeur de l'utilisation des méthodes à noyau. Les méthodes à noyau ont déja été utilisées en science des matériaux notamment par Bartók *et al.* [89, 95, 98, 158] afin d'ajuster des *potentiels machine learning* comme GAP [89]. Les *potentiels machine learning* développés par Bartók *et al.* [89, 95, 98, 158] sont de très bons interpolants sur la base de données sur laquelle ils ont été entraînés. Ceux-ci présentent néanmoins un pouvoir d'extrapolation assez limité [159]. L'un des inconvénients majeurs des méthodes à noyau est la complexité numérique directe du modèle. En effet, pour un système à N particules et une base de données d'entrainement de taille m, celle-ci évolue comme $\mathcal{O}(mN)$: l'évaluation du modèle est donc d'autant plus lente que la base de données d'entraînement est grande. Il est possible, dans le cas des grandes bases de données, d'utiliser des méthodes de sparcification ou d'analyse de rang afin d'éliminer les configurations redondantes. On peut alors passer d'une complexité évoluant comme $\mathcal{O}(mN)$ à $\mathcal{O}(m_1 N)$ avec $m_1 \ll m$ le nombre de configurations non redondantes de la base de données. Or, pour avoir un bon pouvoir de prédiction, la base de données doit être en général grande (ou très grande c'est-à-dire de l'ordre de 10^5 configurations). Dans le cas de GAP [89] l'évaluation de l'énergie d'un système de 10^3 atomes de fer est environ 10^5 plus longue qu'un potentiel MEAM [63]. Nous voudrions nous tourner vers des modèles de régression simples, plus robustes et transférables que les méthodes précédement citées afin de garder une "compréhension physique" du modèle et avoir une complexité numérique faible. Ce type de modèle va être décrit dans la sous-section 2.2.3.

2.2.3 Autres modèles simples de régression : modèle linéaire

Les réseaux de neurones artificiels et les méthodes à noyau sont de puissants outils d'interpolation qui ne nécessitent pas de connaître la fonction à ajuster de façon a priori. En physique, le choix d'une fonction a priori est parfois intéressant car il reflète une réalité théorique. Si l'on veut ajuster un modèle d'énergie cinétique E_c en fonction de la vitesse d'un projectile v, on va directement penser à utiliser un modèle de la forme $E_c \propto v^2$ car celui-ci reflète une "réalité physique" sous-jacente. Ce type d'approche est d'ailleurs utilisé dans le formalisme des potentiels EAM [25, 26] et MEAM [63] où la forme du modèle est inspirée par la forme attendue des orbitales atomiques. Nous allons, dans cette sous-section, nous intéresser au modèle de régression particulièrement simple qu'est la régression linéaire. Commençons par le théorème suivant :

Théorème 2.5. *Soit une fonction f bijective sur $\mathbb{R}$, $\underline{y} \in \mathbb{R}^m$ un vecteur de données de sortie, $\underline{\underline{x}} \in \mathbb{R}^{m\times n}$ une matrice de données d'entrée, un vecteur $\underline{\alpha} \in \mathbb{R}^n$ et un bruit Gaussien de moyenne nulle et de matrice de covariance $\underline{\underline{\Sigma}} = \sigma^2\underline{\underline{1}} \in \mathbb{R}^{m\times m}$ tel que :*

$$\underline{y} = f(\underline{\underline{x}} \cdot \underline{\alpha} + \underline{\epsilon})$$

Considérons la fonction de coût $L\left\{\underline{y}, \underline{\underline{x}}\right\} = \|f^{-1}(\underline{y}) - \underline{\underline{x}} \cdot \underline{w}\|^2$. Il existe alors une solution au problème d'optimisation suivant :

$$\underline{w}^* = \arg\min_{\underline{w}} L\left\{\underline{y}, \underline{\underline{x}}\right\} = \left[\underline{\underline{x}}^T \cdot \underline{\underline{x}}\right]^{-1} \cdot \underline{\underline{x}}^T \cdot f^{-1}(\underline{y})$$

Ce *théorème de Gauss-Markov* [160] démontré à quelques années d'intervalle par Gauss et Markov pose les bases théoriques de la régression linéaire au sens des moindres carrés ordinaires. La régression linéaire peut être étendue aux cas de *fonctions de coûts* régularisées. Il existe notamment des solutions analytiques du problème des moindres carrés ordinaires avec régularisation au sens des normes L_1 et L_2 [142]. Les solutions fournies par la résolution du problème des moindres carrés ordinaires sont généralement peu stables même en utilisant une *fonction de coût* régularisée (cf. Sec (2.2.4)). Nous allons maintenant décrire la méthode de *régression linéaire Bayésienne* qui présente de meilleures qualités de robustesse et de transférabilité.

Considérons le problème de régression suivant $\underline{y} = \underline{\underline{x}} \cdot \underline{w} + \underline{\epsilon}$ où les notations sont les mêmes que celles utilisées dans le théorème Th. (2.5). Dans le cadre de l'approche Bayésienne, on peut définir la probabilité conditionnelle suivante $p(\underline{y}|\underline{w}, \underline{\underline{x}}, \sigma)$, qui représente la probabilité d'obtenir le vecteur $\underline{y}$ sachant $\underline{w}$, $\underline{\underline{x}}$ et σ. Cette probablité est appelée vraisemblance de $\underline{y}$ sachant $\underline{w}$, $\underline{\underline{x}}$ et σ et on vérifie facilement que $-\log(p(\underline{y}|\underline{w}, \underline{\underline{x}}, \sigma))$ vérifie les propriétés d'une *fonction de coût*. En effet, dans le cas d'un bruit Gaussien on montre que :

$$p(\underline{y}|\underline{w}, \underline{\underline{x}}, \sigma) \propto \exp\left(-\|\underline{y} - \underline{\underline{x}} \cdot \underline{w}\|^2/2\sigma^2\right) \tag{2.20}$$

Dans ce cas $-\log(p(\underline{y}|\underline{w},\underline{\underline{x}},\sigma))$ est exactement la *fonction de coût* associée au problème des moindres carrés ordinaires. L'approche Bayésienne permet aussi d'ajouter une distribution a priori pour les valeurs de $\underline{w}$ que l'on prend généralement Gaussien. On a alors $p_0(\underline{w}|\sigma_w) \propto \exp(-\|\underline{w}\|^2/2\sigma_w^2)$ où σ_w est la variance a priori du paramètre $\underline{w}$. On peut alors ré-écrire la vraisemblance de $\underline{y}$ sachant w, $\underline{\underline{x}}$ et σ comme étant $p(\underline{y}|\underline{w},\underline{\underline{x}},\sigma,\sigma_w) = p(\underline{y}|\underline{w},\underline{\underline{x}},\sigma)p_0(\underline{w}|\sigma_w)$. On peut aisement vérifier que la *fonction de coût* associée à $-\log(p(\underline{y}|\underline{w},\underline{\underline{x}},\sigma,\sigma_w))$ correspond à la même *fonction de coût* que dans l'équation (2.20) mais en ajoutant un terme de régularisation de norme L_2. Dans la majorité des cas, on choisit des distributions Gamma pour $p_0(\sigma), p_0(\sigma_w)$ afin de mener les calculs analytiquement [142]. Jusqu'ici le formalisme Bayesien permet seulement d'écrire le problème des moindres carrés ordinaires dans un formalisme probabiliste. Dans les faits, la probabilité $p(\underline{y}|\underline{w},\underline{\underline{x}})$ n'a que peu d'intérêt car elle porte sur la probabilité de retrouver une donnée qu'on a de façon sûre à partir d'une paramétrisation que l'on considère comme connue. Ainsi, la probabilité $p(\underline{w}|\underline{y},\underline{\underline{x}})$ est beaucoup plus intéressante. En effet, on voudrait estimer la probabilité d'une paramétrisation $\underline{w}$ en fonction de nos données d'entrées c'est-à-dire $\underline{y}$ et $\underline{\underline{x}}$. Le formalisme Bayésien permet de calculer assez aisément cette probabilité en définissant l'intrégrale suivante sur les distributions a priori $p_0(\sigma)p_0(\sigma_w)$. La probabilité $p(\underline{w}|\underline{y},\underline{\underline{x}})$ est alors donnée par la formule suivante :

$$p(\underline{w}|\underline{y},\underline{\underline{x}}) = \mathcal{N}^{-1}\int_{\sigma,\sigma_w} p(\underline{y}|\underline{w},\underline{\underline{x}},\sigma)p_0(\underline{w}|\sigma_w)d\sigma d\sigma_w \tag{2.21}$$

où $\mathcal{N}^{-1} = \int d^{\mathcal{D}}\underline{w}\int_{\sigma,\sigma_w} L(\underline{y}|\underline{w},\underline{y},\sigma)p_0(\underline{w}|\sigma_w)d\sigma d\sigma_w$ permet d'assurer la normalisation. On voudrait que cette probabilité soit maximisée pour la paramétrisation optimale $\underline{w}^*$. Par passage au logarithme (permettant d'éviter le calcul de la constante $\mathcal{N}$), on obtient le problème variationnel suivant :

$$\underline{w}^* = \arg\max_{\underline{w}\in\mathbb{R}^n} \log\int_{\sigma,\sigma_w} p(\underline{y}|\underline{w},\underline{\underline{x}},\sigma)p_0(\underline{w}|\sigma_w)d\sigma d\sigma_w \tag{2.22}$$

La régression linéaire Bayésienne présente une meilleure robustesse et une meilleure transférabilité que la méthode des moindres carrés ordinaires grâce à l'inverse de la probabilité conditionnelle donnée ci-dessus. Dans la suite, et notamment dans les chapitres (3) et (4), nous allons utiliser la régression linéaire Bayésienne afin de construire des modèles de régression transférables pour des quantités thermodynamiques.

2.2.4 Les méthodes de régularisations

Les méthodes de régularisation sont absolument essentielles dans le cadre des méthodes de régression en hautes dimensions. Comme nous l'avons décrit plus-haut, elle permettent d'éviter le sur-ajustement de certaines données et donc d'augmenter la transférabilité des modèles. Il existe un grand nombre de méthodes de régularisation qui dépendent du type de problème étudié. Une méthode de régularisation est définie par la forme de la fonction g donnée dans l'équation (2.18). La méthode de régularisation

la plus utilisée est la régularisation L_2 due à Tikhonov et Phillips [161]. Le terme de régularisation L_2 s'écrit alors sous la forme :

$$g(\underline{w}) = \lambda \|\underline{w}\|^2 \tag{2.23}$$

où $\underline{w}$ est le vecteur de poids associé au modèle et λ est le paramètre de la régularisation. La régularisation de type L_2 permet d'obtenir un bon comportement moyen de la régression. En d'autres termes, la régression est moins précise pour les données d'entraînement que la régression non-régularisée mais le modèle ne prédit pas de valeurs "aberrantes" sur de nouvelles données. Nous montrerons que la régularisation de type L_2 équivaut, dans le formalisme Bayésien, à choisir une distribution *a priori* Gaussienne sur le vecteur de poids $\underline{w}$.

L'autre type de régularisation classiquement utilisé est la régularisation L_1 ou Lasso introduite par Tibshirani [162]. Cette régularisation peut s'exprimer de la façon suivante :

$$g(\underline{w}) = \lambda |\underline{w}| \tag{2.24}$$

où $\underline{w}$ est le vecteur de poids associé au modèle et λ est le paramètre de la régularisation. La régularisation L_1 permet de fixer des poids exactement à 0 dans la régression. Cette méthode permet d'éliminer la dépendance de certaines composantes de la régression. Ce type de régularisation est particulièrement utile pour les modèles possédant un grand nombre de paramètres. On peut aussi coupler la régularisation L_1 et L_2 : c'est ce qu'on appelle la régularisation par Elastic Net [163].

Dans le cadre des réseaux de neurones, on peut aussi mettre en place des méthodes de régularisation L_1 ou L_2 directement introduites dans les fonctions de coût. Une autre méthode de régularisation pour les réseaux de neurones est le *dropout* [164]. Cette méthode consiste à fixer certains poids du réseau de neurones à 0 après l'entraînement. Cette méthode est équivalente à une approche Bayésienne de l'ajustement des poids. Le *dropout* permet de limiter grandement les phénomènes de sur-ajustement dans les réseaux de neurones.

Dans la littérature, il existe d'autres types de régularisations. Celles-ci sont basées sur des fonctions g prenant en compte le caractère non-homogène de la répartition de données ou bien sur des décompositions SVD (Singular Value Decomposion) plus robustes que la simple utilisation de la norme L_1 ou L_2 [134, 135]. Notons ici que les méthodes de sparcification telle que la SVD ne sont applicables que pour des matrices de design de dimensions raisonnables ($\max(m, n) = 10^4$). Si la dimension de cette matrice d'entrée de données est significativement plus grande, seule une approche numérique de minimisation de la *fonction de coût* est possible.

2.3 Conclusions de chapitre

Dans la sous-section 2.1.1, nous avons montré l'intérêt croissant des méthodes de type *machine learning* en science des matériaux notamment afin d'effectuer la transition d'échelle entre la DFT et la dynamique moléculaire. Nous avons introduit les descripteurs atomiques locaux 2.1.2 qui permettent de décrire de façon systématique les environnements atomiques locaux tout en conservant les propriétés de symétrie du système étudié. Nous avons aussi donné une liste non-exhaustive de descripteurs locaux présents dans la littérature et que nous allons utiliser par la suite.

Nous avons décrit les grands types de méthodes utilisables pour les problèmes de régressions et de classifications. L'ensemble de ces méthodes sont des solutions de problèmes d'approximation : les réseaux de neurones et les méthodes à noyau permettent de construire des approximations sans postuler de forme a priori. Ce n'est pas le cas pour les méthodes plus simples telle que la régression linéaire où il est nécessaire de donner une expression analytique du modèle à ajuster. Les réseaux de neurones artificiels et les méthodes à noyaux sont hautement non-linéaires et permettent de répondre avec grande précision à des problèmes d'interpolation. Néanmoins, leur transférabilité reste relativement faible ce qui engendre des problèmes lors de l'extrapolation du modèle construit sur des données extérieures à la base de données d'entraînement. On notera qu'une estimation de l'erreur sur une nouvelle donnée vis-à-vis du modèle déja construit est fournie par les méthodes à noyau. La notion de régularisation joue un rôle central dans la réduction de l'*hyper-ajustement* et est devenue une méthode standard pour tous les modèles de régressions précédemment cités. L'ensemble de ces méthodes est aujourd'hui très largement implémenté dans des langages tels que `python` grâce aux bibliothèques `scikit-learn` [165] (pour les régression linéaires et les méthodes à noyau) et `tensorflow` [166] ou `pytorch` [167] (pour les réseaux de neurones artificiels).

Dans les faits, les méthodes sophistiquées que constituent les réseaux de neurones artificiels et les méthodes à noyaux sont relativement opaques vis-à-vis du modèle construit et ne permettent pas (en général) de comprendre "la physique" cachée derrière les modèles de régression construits. Les méthodes plus simples, telles que les régressions linéaires, nécessitent un a priori "physique" sur la forme du modèle à ajuster et permettent d'obtenir de grandes capacités de transférabilité notamment grâce à l'approche Bayésienne. Dans la suite de notre propos, nous allons nous concentrer sur ces modèles simples et montrer que leur utilisation couplée à celle des descripteurs permet d'obtenir des résultats très encourageants sur des quantités thermodynamiques telle que l'entropie vibrationnelle Chap. 3 ou les fréquences d'attaque Chap. 4 dans le cadre de l'approximation harmonique.

Oh, I'll never kill myself to save my soul
I was gone, but how was I to know ?

— Unsainted, Slipknot

3

Modèles de régression de l'entropie vibrationnelle dans le cadre de l'approximation harmonique

Sommaire

3.1 Rappels et définitions

Nous débutons ce chapitre en rappelant des définitions essentielles sur l'entropie thermodynamique dans sa formulation moderne (Boltzmannienne) et en explicitant ses différentes contributions en Sec. 3.1.1. Nous travaillons dans ce chapitre dans le cadre de l'approximation harmonique dont la définition est rappelée dans la sous-section 3.1.2. Enfin nous nous focalisons sur la contribution vibrationnelle de l'entropie dont nous rappelons son lien étroit avec les modes de vibrations et les transitions de phases en Sec. 3.1.3.

3.1.1 Entropie microcanonique : définition, contributions

Considérons un système physique constitué de N particules de coordonnées $(\boldsymbol{q}, \boldsymbol{p}) = \mathbb{R}^{3N\times 3N}$ dans l'espace des phases noté $\mathcal{Q}\times\mathcal{P} \in \mathbb{R}^{3N\times 3N}$. Ce système possède une énergie $\mathcal{E}(\boldsymbol{q}, \boldsymbol{p})$. L'entropie microcanonique pour une énergie donnée E, $S(E)$ est alors définie par :

$$S(E) = k_B \ln\left(\frac{1}{h(\delta E)}\int_{\mathcal{Q}\times\mathcal{P}} \mu_{\delta E}(|\mathcal{E}(\boldsymbol{q}, \boldsymbol{p}) - E|)d\boldsymbol{q}d\boldsymbol{p}\right). \tag{3.1}$$

Dans cette définition, $h(\delta E)$ est l'élément d'action minimale associé à la fenêtre δE et vérifiant $h(\delta E) \propto \delta E$, k_B est la constante de Boltzmann et $\mu_{\delta E}(|\mathcal{E}(\boldsymbol{q}, \boldsymbol{p}) - E|)$ est la mesure de comptage suivante :

$$\mu_{\delta E}(|\mathcal{E}(\boldsymbol{q}, \boldsymbol{p}) - E|) = \begin{cases} 0 \text{ si } |\mathcal{E}(\boldsymbol{q}, \boldsymbol{p}) - E| \geq \frac{1}{2}\delta E \\ 1 \text{ si } |\mathcal{E}(\boldsymbol{q}, \boldsymbol{p}) - E| \leq \frac{1}{2}\delta E \end{cases} \tag{3.2}$$

L'entropie microcanonique est donc proportionnelle au logarithme du nombre de micro-états d'énergies $\mathcal{E}(\boldsymbol{q}, \boldsymbol{p})$ comprises entre $E - \frac{1}{2}\delta E$ et $E + \frac{1}{2}\delta E$. Cette définition introduite par Boltzmann permet de faire le lien entre l'entropie macroscopique S et une grandeur statistique portant sur les états possibles du système.

L'entropie, au sens de Boltzmann, peut être décomposée en trois contributions distinctes que nous allons détailler ici en nous plaçant (sans perte de généralité) à l'énergie E.
La première contribution est d'ordre **configurationnelle**. Elle est proportionelle au logarithme du nombre de configurations géométriques d'énergies $\mathcal{E}_c$ compris entre $E - \frac{1}{2}\delta E$ et $E + \frac{1}{2}\delta E$ toute chose étant égale par ailleurs. Deux configurations géométriques différentes peuvent avoir la même énergie E si elles sont images l'une de l'autre par application d'une transformation de symétrie laissant invariant le système. Ce nombre de configurations correspond aux nombres d'arrangements possibles du système permettant de conserver son énergie entre $E - \frac{1}{2}\delta E$ et $E + \frac{1}{2}\delta E$.
La deuxième contribution de l'entropie est d'ordre **électronique**. Elle est proportionnelle au logarithme du nombre d'états électroniques d'énergies $\mathcal{E}_e$ compris entre $E - \frac{1}{2}\delta E$ et $E + \frac{1}{2}\delta E$ toute chose étant égale par ailleurs.
Enfin, la troisième contribution est d'ordre **vibrationnelle**. Elle est proportionnelle au logarithme du nombre d'états de phonons d'énergies $\mathcal{E}_{vib}$ compris entre E et $E \pm \frac{1}{2}\delta E$

toute chose étant égale par ailleurs. Les phonons sont des pseudo-particules associées aux vibrations du réseau du système. C'est cette contribution qui va nous intéresser et que nous allons quantifier plus finement dans le cadre de l'approximation harmonique en Sec. 3.1.3.
Ces trois contributions peuvent être vues comme une image du désordre du système que ce soit au niveau de son organisation globale (**configurationnelle**), de son organisation **électronique** ou de son organisation **phononique**. Plus un système est désordonné, plus il possède un nombre grand d'états d'énergie $\mathcal{E}$ compris entre $E - \frac{1}{2}\delta E$ et $E + \frac{1}{2}\delta E$.

L'entropie est une variable d'état extensive. Considérons un système $\mathcal{C}$ possédant une entropie S. L'extensivité se traduit par la propriété suivante si le système $\mathcal{C}$ est répliqué k fois ($k\mathcal{C}$) :

$$\mathcal{C} \to k\mathcal{C} \Longrightarrow S \to kS \tag{3.3}$$

L'extensivité est une propriété importante et devra être un moteur de vérification des modèles de régression. Les modèles de régression ne vérifiant pas l'extensivité thermodynamique de l'entropie devront être abandonnés.

En thermodynamique, on définit la notion de *potentiel thermodynamique* comme étant une fonction d'état $\mathfrak{P}$. Cette grandeur est minimale à l'équilibre thermodynamique. L'état d'équilibre prévu par la thermodynamique peut donc être reformulé comme la solution d'un problème d'optimisation dans l'espace des phases et dont la fonction à minimiser est $\mathfrak{V}$. Pour un système isolé ne pouvant échanger ni matière ni chaleur avec le milieu extérieur le 2^{nd} *principe de la thermodynamique* indique que le *potentiel thermodynamique* est $\mathfrak{P} = -S$. L'utilisation de $-S$ comme *potentiel thermodynamique* est restrictive car la majorité des systèmes *réels* échange de l'énergie avec le milieu extérieur. Ainsi, pour un système ne pouvant échanger de matière avec l'extérieur mais pouvant échanger de l'énergie avec un thermostat à température T le *potentiel thermodynamique* devient $\mathfrak{P} = U - TS$. Cette nouvelle variable d'état, $F \equiv U - TS$, est appelée *énergie libre*. U est définie comme étant l'énergie interne du système (et qui sera définie rigoureusement dans la sous-section suivante 3.1.2). L'*énergie libre* est la grandeur qui définit l'état d'équilibre thermodynamique d'un système *réel* à nombre de particules fixé N, à température fixée T et à volume fixé V, ensemble que l'on appelle NVT. Le calcul de l'*énergie libre* d'un système nécessite de calculer son entropie S via le quantification des trois contributions citées plus haut.

3.1.2 Énergie potentielle dans le cadre harmonique

Le calcul de l'*énergie libre* nécessite aussi la connaissance de l'énergie interne du système que l'on peut décomposer en deux contributions :

$$U = \langle E_{cin} \rangle_\pi + \langle V \rangle_\pi \tag{3.4}$$

Le 1^{er} terme du membre de droite représente la moyenne d'ensemble vis à vis de la mesure canonique $\pi(\boldsymbol{q}, \boldsymbol{p})$ - dont la définition est donnée en Sec. 1.4 et rappelée en

annexe (A) - de l'énergie cinétique microscopique $E_{cin}(\boldsymbol{q},\boldsymbol{p})$ du système. Le $2^{ème}$ terme du membre de droite représente la moyenne d'ensemble vis à vis de la mesure canonique $\pi(\boldsymbol{q},\boldsymbol{p})$ de l'énergie potentielle microscopique $V(\boldsymbol{q},\boldsymbol{p})$ du système. Dans le cadre de l'ensemble canonique, il est aisé de montrer que les grandeurs thermodynamiques telle que l'énergie interne sont directement reliées à la fonction de partition Z du système. Ainsi pour un *Hamiltonien* $\mathcal{H}(\boldsymbol{q},\boldsymbol{p}) = E_{cin}(\boldsymbol{q},\boldsymbol{p}) + V(\boldsymbol{q},\boldsymbol{p})$ l'expression de la fonction de partition canonique Z pour une température donnée T est :

$$Z = \frac{1}{h^{3N}} \int_{\mathcal{Q}\times\mathcal{P}} e^{-\beta\mathcal{H}(\boldsymbol{q},\boldsymbol{p})} d\boldsymbol{q}d\boldsymbol{p} \tag{3.5}$$

où on rapelle que $\beta = (k_B T)^{-1}$ et h est la constante de Planck. Les variables thermodynamiques peuvent alors être directement déduites de l'expression de la fonction de partition. On a par exemple $U = -\frac{\partial}{\partial\beta}(\ln Z)$. La seule difficulté restante est de pouvoir calculer la fonction de partition via les expressions de $E_{cin}(\boldsymbol{q},\boldsymbol{p})$ et $V(\boldsymbol{q},\boldsymbol{p})$ de l'*Hamiltonien* du système. L'expression de l'énergie cinétique microscopique d'un système à N particules se réduit à l'expression simple suivante :

$$E_{cin}(\boldsymbol{q},\boldsymbol{p}) = \sum_{i=1}^{N}\sum_{\alpha=1}^{3} \frac{1}{2m_i} \boldsymbol{p}_{i\alpha}\cdot\boldsymbol{p}_{i\alpha} \tag{3.6}$$

Le vecteur $\boldsymbol{p}_{i\alpha}$ représente la quantité de mouvement de la particule i dans la direction α de l'espace, m_i la masse de la particule i et $\cdot$ le produit scalaire usuel. Dans le cadre de l'approximation harmonique, nous nous plaçons proche d'un minimum de l'énergie potentielle pour les coordonnées $\boldsymbol{q}_0$. On suppose, sans perte de généralité, que le potentiel ne dépend que des coordonnées $\boldsymbol{q}$. On a alors $V_0 = V(\boldsymbol{q}_0)$ la valeur du potentiel au minimum. L'approximation harmonique consiste à effectuer un développement de Taylor (en coordonnées) en se limitant à l'ordre 2 autour de la position d'équilibre $\boldsymbol{q}_0$. On a alors l'expression du potentiel $V(\boldsymbol{q})$ pour les coordonnées $\boldsymbol{q}$ proches de $\boldsymbol{q}_0$ (pour un système à N particules) :

$$V(\boldsymbol{q}) = V_0 + \boldsymbol{\nabla}V(\boldsymbol{q}_0)\cdot(\boldsymbol{q}-\boldsymbol{q}_0) + \frac{1}{2}(\boldsymbol{q}-\boldsymbol{q}_0)\cdot\boldsymbol{H}\left\{\boldsymbol{q}_0\right\}\cdot(\boldsymbol{q}-\boldsymbol{q}_0) + o(\gamma\|\boldsymbol{q}-\boldsymbol{q}_0\|^3) \tag{3.7}$$

où $\boldsymbol{\nabla}V(\boldsymbol{q}_0) \in \mathbb{R}^{3N}$ est le gradient du potentiel V par rapport aux coordonnées $\boldsymbol{q}$ évalué en $\boldsymbol{q}_0$, $\|\cdot\|$ est la norme Euclidienne et γ est une constante assurant l'homogénéité. Enfin, $\boldsymbol{H}\left\{V(\boldsymbol{q}_0)\right\} \in \mathbb{R}^{3N\times 3N}$, est la matrice Hessienne de l'énergie potentielle par rapport aux coordonnées $\boldsymbol{q}$ évaluée en $\boldsymbol{q}_0$ et définie par :

$$\boldsymbol{H}\left\{\boldsymbol{q}_0\right\} = \sum_{i,\alpha=1}^{N,3}\sum_{j,\beta=1}^{N,3} \frac{\partial^2}{\partial q_{i\alpha}\partial q_{j\beta}}\left\{V(\boldsymbol{q}_0)\right\} \boldsymbol{q}_{i\alpha}\otimes\boldsymbol{q}_{j\beta} \tag{3.8}$$

L'approximation harmonique permet de donner une expression analytique de la fonction de partition des phonons d'un système à N particules. Cette expression va être détaillée dans la sous-section suivante 3.1.3.

On définit l'*approximation harmonique* par le développement de Taylor donné par

l'équation (3.7) où on se limite à l'ordre 2 dans le développement $\boldsymbol{q} - \boldsymbol{q}_0$. Ce développement décrit l'énergie potentielle, autour d'un minimum du paysage énergétique, comme une forme quadratique de $\boldsymbol{q} - \boldsymbol{q}_0$. L'*approximation harmonique* est une approximation locale autour d'un minimum $\boldsymbol{q}_0$ d'un paysage énergétique. La forme quadratique associée est alors toujours positive et traduit le fait que le bassin est convexe quelque soit $\boldsymbol{q} - \boldsymbol{q}_0$ et que le système subit toujours une force (proportionnelle à $\boldsymbol{q} - \boldsymbol{q}_0$) qui tend à faire revenir la configuration $\boldsymbol{q}_0$. Dans un cas uni-dimensionnel, l'approximation harmonique se réduit à une équation parabolique pour l'énergie potentielle $V(q)$:

$$V(q) = \frac{d^2}{dq^2}\left\{V(q_0)\right\}(q - q_0)^2 \tag{3.9}$$

Le terme $\frac{d^2}{dq^2}\left\{V(q_0)\right\}$ décrit la courbure du potentiel au point q_0 définissant le minimum. Dans le cas plus général, l'ensemble de valeur propre de la forme quadratique associée à l'équation (3.8) décrit la courbure du bassin en fonction du déplacement $\boldsymbol{q} - \boldsymbol{q}_0$. Plus la courbure est élévée, plus la force de rappel associée au déplacement $\boldsymbol{q} - \boldsymbol{q}_0$ est élevée. Le cadre *harmonique* est une approximation locale du paysage énergétique autour d'un minimum donné. Dans le cas de déplacement $\boldsymbol{q} - \boldsymbol{q}_0$ faible, on peut toujours correctement décrire un mimimum à l'aide de cette approximation. Néanmoins, pour des déplacements $\boldsymbol{q} - \boldsymbol{q}_0$ (par exemple dans le cas de système à température finie) les hypothèses de cette approximation ne sont plus forcément valides. Nous discuterons de ces situations dans le chapitre 5, mais pour le reste de ce chapitre nous nous limitons au cadre de l'*approximation harmonique*.

3.1.3 Entropie vibrationnelle et modes de vibrations

L'énergie potentielle étant définie à une constante près, nous posons $V_0 = 0$. Dans la suite, nous considérons les déplacements $\boldsymbol{q}' = \boldsymbol{q} - \boldsymbol{q}_0$. Nous pouvons maintenant écrire l'*Hamiltonien* d'un système à N particules dans le **cadre harmonique** autour de la coordonnée d'équilibre $\boldsymbol{q}_0$:

$$\mathcal{H}_{ha}(\boldsymbol{q}', \boldsymbol{p}) = \sum_{i,\alpha=1}^{N,3} \frac{1}{2m_i}\boldsymbol{p}_{i\alpha} \cdot \boldsymbol{p}_{i\alpha} + \sum_{i,\alpha=1}^{N,3}\sum_{j,\beta=1}^{N,3} \boldsymbol{q}'_{i\alpha} \cdot \boldsymbol{H}\left\{V(\boldsymbol{q}_0)\right\} \cdot \boldsymbol{q}'_{j\beta} \tag{3.10}$$

La matrice Hessienne du système est symétrique par construction et par conséquent il existe une base orthogonale de vecteurs $\{\mathbb{U}_\nu\}_{1\leq\nu\leq 3N}$ qui **diagonalise la matrice dynamique** et une matrice de rotation $\boldsymbol{L}$ telles qu'on a les relations suivantes :

$$\boldsymbol{q}'_{i\alpha} = \frac{1}{\sqrt{m_i}}\sum_{\nu=1}^{3N} L_{i\alpha,\nu}\mathbb{U}_\nu \tag{3.11}$$

$$\boldsymbol{p}'_{i\alpha} = \frac{1}{\sqrt{m_i}}\sum_{\nu=1}^{3N} L_{i\alpha,\nu}\mathbb{V}_\nu \tag{3.12}$$

$$\boldsymbol{L} \cdot \boldsymbol{L}^T = \mathbf{1} \tag{3.13}$$

Cette base de vecteurs $\{\mathbb{U}_\nu\}_{1\leq\nu\leq 3N}$ est appelée base des *modes normaux*. Ecrivons maintenant l'*Hamiltonien* du système (3.10) dans la nouvelle base $\{(\mathbb{U}_\nu, \mathbb{V}_\nu)\}_{1\leq\nu\leq 3N}$:

$$\mathcal{H}_{ha}(\mathbb{U}_\nu, \mathbb{V}_\nu) = \frac{1}{2}\sum_{\nu=1}^{3N} \mathbb{V}_\nu \cdot \mathbb{V}_\nu + \frac{1}{2}\sum_{\nu,\nu'=1}^{3N} \mathbb{U}_\nu \cdot \left(\boldsymbol{L}^T \cdot \mathfrak{D}\left\{V(\boldsymbol{q}_0)\right\} \cdot \boldsymbol{L}\right) \cdot \mathbb{U}_{\nu'} \tag{3.14}$$

$$\mathfrak{D}\left\{\boldsymbol{q}_0\right\} = \sum_{i,\alpha=1}^{N,3}\sum_{j,\beta=1}^{N,3} \frac{1}{\sqrt{m_i m_j}} \frac{\partial^2}{\partial q_{i\alpha}\partial q_{j\beta}} \left\{V(\boldsymbol{q}_0)\right\} \boldsymbol{q}'_{i\alpha} \otimes \boldsymbol{q}'_{j\beta} \tag{3.15}$$

Le tenseur $\mathfrak{D}\left\{\boldsymbol{q}_0\right\}$ est appelé *matrice dynamique* du système au point de coordonnées $\boldsymbol{q}_0$. Cette matrice est symétrique et a pour vecteurs propres la famille $\mathbb{U}_\nu$. Ses valeurs propres sont notées ω_ν^2. Nous obtenons finalement une expression simple de l'*Hamiltonien* harmonique en nous plaçant dans la base $\left\{(\mathbb{U}_\nu, \dot{\mathbb{U}}_\nu)\right\}_{1\leq\nu\leq 3N}$ et considérant un système composé de N particules :

$$\mathcal{H}_{vib}(\mathbb{U}_\nu, \mathbb{V}_\nu) = \frac{1}{2}\sum_{\nu=1}^{3N} \mathbb{V}_\nu \cdot \mathbb{V}_\nu + \frac{1}{2}\sum_{\nu=1}^{3N} \omega_\nu^2 \mathbb{U}_\nu \cdot \mathbb{U}_\nu \tag{3.16}$$

Cet *Hamiltonien* (3.16) définit la dynamique des vibrations du réseau cristallin. La dynamique des vibrations du réseau correspond à un cas particulier - au point Γ - de dynamique des pseudo-particules appelées phonons. Les phonons de vecteur d'onde $\boldsymbol{k}$ correspondent aux *modes normaux* de la matrice dynamique suivante :

$$\mathfrak{D}\left\{\boldsymbol{q}_0, \boldsymbol{k}\right\} = \sum_{p\in\mathcal{R}}\sum_{i_o,\alpha=1}^{N,3}\sum_{j_p,\beta=1}^{N,3} \frac{1}{\sqrt{m_{i_o} m_{j_p}}} \frac{\partial^2}{\partial q_{i_o\alpha}\partial q_{j_p\beta}} \left\{V(\boldsymbol{q}_0)\right\} e^{i\boldsymbol{k}\cdot\boldsymbol{R}_{o,p}} \boldsymbol{q}'_{i_o\alpha} \otimes \boldsymbol{q}'_{j_p\beta} \tag{3.17}$$

Ici, $\mathcal{R}$ représente l'ensemble des images périodiques générées par le groupe de translation du réseau et $\boldsymbol{R}_{o,p}$ correspond au vecteur reliant l'atome de coordonnées $\boldsymbol{q}_o$ et sa $p^{ième}$ image périodique. Les phonons sont des bosons et obéissent à la statistique de *Bose-Einstein*. Comme pour les électrons, les phonons obéissent à l'équation de *Schrödinger* et au *principe d'incertitude*. Leurs niveaux d'énergies sont quantifiés et on peut montrer qu'un phonon de pulsation ω_ν peut accéder aux niveaux d'énergies $\hbar\omega_\nu\left(n+\frac{1}{2}\right)$ pour $n \in \mathbb{N}$. Il nous est maintenant possible de calculer la fonction de partition $Z_{vib}(N,T)$ associée aux modes de vibration - des phonons au point Γ - et en déduire l'expression de l'entropie vibrationnelle S_{vib}, pour une température donnée T et pour un système de N particules, grâce à la relation thermodynamique suivante :

$$S_{vib}(N,T) = \frac{\partial}{\partial T}\left(\beta^{-1}\ln Z_{vib}(N,T)\right) \tag{3.18}$$

Un calcul détaillé en annexe (A) permet de montrer que sous l'hypothèse suivante, $\forall\omega_\nu / \frac{k_B T}{\hbar\omega_\nu} \gg 1$ (ce qui correspond à une température supérieure à celle de Debye), on obtient une expression simple de l'entropie vibrationnelle dans le cadre de l'approximation harmonique [168] :

$$S_{vib}(T,N) = k_B \sum_{\nu=1}^{3N}\left[\ln\left(\frac{k_B T}{\hbar\omega_\nu}\right) + 1\right] \tag{3.19}$$

Pour une température donnée, la valeur de l'entropie vibrationnelle d'un système est directement reliée aux pulsations des modes de vibration ω_ν. L'entropie vibrationnelle est d'autant plus grande qu'il existe de basses fréquences pour le système. Ainsi, dans le fer cubique centré le mode de vibration basse fréquence, appelé *mode mou*, dans la direction $\langle 111 \rangle$ est initiateur de la transition de phase entre la phase cubique centrée et la phase cubique à faces centrées [169-171]. Ce mode basse fréquence et de grande longueur d'onde dans la direction $\langle 111 \rangle$ joue un rôle important dans la nucléation des *paires de décrochement* de la dislocation $\frac{1}{2}\langle 111 \rangle$ et peut être délocalisé sur des distances supérieures à 10 Å [14]. Dans le fer, le *mode mou* dans la direction $\langle 111 \rangle$ est aussi initiateur de la transition entre la phase cubique centrée et la phase de lave *C15* sous irradiation créant des structures relativement exotiques [7] révélatrices de la cinétique et de la thermodynamique de créations des défauts d'irradiation dans le fer [11].

3.2 Formalisme de *Green* appliqué à l'entropie vibrationnelle

Le cadre de l'approximation harmonique nous a permis de donner une expression analytique de l'entropie vibrationnelle $S_{vib}(T, N)$ pour un système de N particules à la température T par l'équation (3.19). Par analogie directe avec la structure électronique - notamment grâce aux travaux de Friedel [172, 173] - nous allons utiliser le formalisme des fonctions de *Green* pour traiter le problème aux valeurs propres associé aux modes de vibrations voir Eq. (3.20). Comme dans le cas de la structure électronique, l'utilisation des fonctions de *Green* permet de construire la *densité d'état* de particules solution du problème aux valeurs propres associé (les électrons et l'équation de *Schrödinger* dans le cas de la structure électronique et les modes de vibration et l'équation (3.20) dans le cas vibrationnel). Grâce au travaux de Friedel [172, 173], on sait que la *densité d'état* - ici de modes de vibration - peut alors être décomposée localement. Dans le cas des *modes normaux*, nous allons étudier le lien étroit entre les solutions de *Green* pour un problème aux valeurs propres et la *densité d'état* de modes de vibration d'un système Sec. 3.2.1. Nous détaillerons ensuite le passage de la *densité d'état* portant sur les *modes propres* d'un système et la *densité d'état locale* qui permet de réduire l'entropie vibrationnelle harmonique à un problème purement local Sec. 3.2.1.

3.2.1 *Fonction de Green* et problèmes aux valeurs propres

Dans le cadre de l'approximation harmonique, un développement de Taylor du 2^{nd} ordre autour des coordonnées $\boldsymbol{q}_0$ de l'énergie potentielle permet d'obtenir une formulation simple du *principe fondamental de la dynamique* donnée par l'équation (3.8). Nous traduisons cette équation sous la forme d'un problème aux valeurs propres suivant, que nous formulons pour plus de simplicité, dans le domaine de *Fourier* en temps :

$$\left(\mathfrak{D}\left\{\boldsymbol{q}_0\right\} - \omega^2 \mathbf{1}\right) \cdot \hat{\mathbf{e}}_\nu(\omega) = \mathbf{0} \tag{3.20}$$

Ici, on suppose travailler avec un système à N particules. Cette équation aux valeurs propres fait intervenir la *matrice dynamique* du système $\mathfrak{D}\{\boldsymbol{q}_0\}$ au point de coordonnées $\boldsymbol{q}_0$, $\mathbf{1}$ est la matrice identité et $\hat{\mathbf{e}}_\nu$ la fonction propre associée au mode propre ν. L'expression de la *fonction de Green* associée au problème des *modes normaux* (3.20) peut être calculée analytiquement. En se basant sur le traitement par fonctions de *Green* de l'équation de *Schrödinger* [172, 173] développé dès les années 1950 - Sec. 1.3 -, P. H. Dederichs *et al.* [24] ont obtenu un résultat totalement analogue pour l'équation séculaire Eq. (3.20). L'expression de la *fonction de Green* $\boldsymbol{\mathcal{G}} \in \mathbb{C}^{3N\times 3N}$ associée au problème Eq. (3.20) est la suivante :

$$\boldsymbol{\mathcal{G}}(\omega) = \sum_{\nu=1}^{3N} \frac{\hat{\mathbf{e}}_\nu^T(\omega) \otimes \hat{\mathbf{e}}_\nu(\omega)}{\omega_\nu^2 - \omega^2}. \tag{3.21}$$

Les pôles du dénominateur sont les pulsations ω_ν des modes de vibration du système. Nous allons maintenant nous intéresser au lien entre cette *fonction de Green* et l'expression de l'entropie vibrationnelle classique dans le cadre harmonique donné par l'équation (3.19) et passant par la *densité d'état* de modes de vibration.

Considérons un système constitué de N particules. On appelle *densité d'état* de modes de vibration $\Omega(\omega)$ la grandeur vérifiant les deux relations suivantes :

$$n_{[\omega,\omega+\delta\omega[} = \int_\omega^{\omega+\delta\omega} \Omega(\omega')d\omega' \tag{3.22}$$

$$3N = \int_0^\infty \Omega(\omega)d\omega \tag{3.23}$$

Dans cette expression $n_{[\omega,\omega+\delta\omega[}$ représente le nombre de *modes normaux* du système ayant leur pulsation ω_ν comprise entre ω et $\omega + \delta\omega$.

Il est possible d'effectuer une décomposition de *fonction de Green* Eq. (3.21) en posant $\omega = \Omega + i\eta$ dans la limite $\eta \to 0$:

$$\frac{1}{\omega_\nu^2 - (\omega + i\eta)^2} = P\left(\frac{1}{\omega_\nu^2 - \omega^2}\right) + i\pi\delta(\omega_\nu^2 - \omega^2) \tag{3.24}$$

Ici, $P(\cdot)$ est la partie principale et $\delta(\cdot)$ est la distribution de Dirac. Cette décomposition de la *fonction de Green* Eq. (3.21) effectuée par P. H. Dederichs *et al.* [24] (Section 2) permet de montrer que :

$$\frac{2\omega}{\pi}\Im\left(\mathrm{Tr}\left\{\boldsymbol{\mathcal{G}}(\omega)\right\}\right) = \sum_{\nu=1}^{3N} \delta(\omega - \omega_\nu) \tag{3.25}$$

Ici, $\Im(\cdot)$ représente la partie imaginaire, $\mathrm{Tr}(\cdot)$ l'opérateur trace et δ la distribution de *Dirac*. On reconnait alors l'expression de la *densité d'état* de modes de vibration. Il existe donc un lien direct entre Ω et la partie imaginaire de la trace de $\boldsymbol{\mathcal{G}}$ que nous allons

pouvoir exploiter pour donner une formulation continue de l'entropie vibrationnelle. En effet, on peut écrire l'expression $S_{vib}(N,T)$ à l'aide de la *densité d'état* :

$$S_{vib}(N,T) = -k_B \int_0^\infty \left[\ln\left(\frac{\hbar\omega}{k_B T}\right) - 1\right] \Omega(\omega) d\omega \tag{3.26}$$

Cette forme continue porte sur des entités "globales" que sont les *modes normaux* (en tant que solution d'un problème aux valeurs propres). Nous allons voir qu'il existe une relation simple entre la *base propre* des *modes normaux* et la *base locale* centrée sur les particules du système.

3.2.2 Des *modes normaux* à un formalisme local

Considérons la base *propre* des *modes normaux* $\{\hat{\mathbf{e}}_\nu\}_{1\leq\nu\leq 3N}$ associée au problème Eq. (3.20) et la base *locale* $\{\hat{\mathbf{e}}_{i\alpha}\}_{1\leq i\alpha\leq 3N}$ centrée sur la particule i et suivant la direction α. Ces deux bases étant orthonormées, il existe une matrice de rotation $\boldsymbol{\xi}$ [24, 172, 173] telle qu'on a :

$$\hat{\mathbf{e}}_\nu = \sum_{i,\alpha=1}^{N,3} \xi^{i\alpha}(\nu)\, \hat{\mathbf{e}}_{i\alpha}, \tag{3.27}$$

La norme au carré d'un élément $|\xi^{i\alpha}(\nu)|^2$ de cette matrice est égale à la probabilité de trouver le mode de vibration $\hat{\mathbf{e}}_{i\alpha}$ projeté sur la particule i et dans la direction α. Cette décomposition locale des modes de vibration permet d'exprimer la notion de *densité d'état locale* définie par les relations suivantes :

$$\varrho^{i\alpha}(\omega) = \frac{2\omega}{\pi} \Im\left(\mathcal{G}_{i\alpha}(\omega)\right). \tag{3.28}$$

$$\varrho^{i\alpha}(\omega) = \sum_{\nu=1}^{3N} |\xi^{i\alpha}(\nu)|^2 \delta(\omega - \omega_\nu), \tag{3.29}$$

La *densité d'état locale* sur la particule i et dans la direction α est notée $\varrho^{i\alpha}(\omega)$. Une simple sommation sur les particules du système et les directions de l'espace permet d'obtenir la *densité d'état* de *modes normaux*. Cette nouvelle expression *locale* nous permet de reformuler la forme continue de l'entropie vibrationnelle $S_{vib}(N,T)$ Eq. (3.26) sous la forme :

$$\begin{aligned} S_{vib}(N,T) &= \sum_{i=1}^{N} \Bigg[\underbrace{-k_B \sum_{\alpha=1}^{3} \sum_{\nu=1}^{3N} \left[\ln\left(\frac{\hbar\omega_\nu}{k_B T}\right) - 1\right] |\xi^{i\alpha}(\nu)|^2}_{S_i(T),\ \textit{information locale}} \Bigg] \\ &= \sum_{i=1}^{N} \left[\sum_{\alpha=1}^{3} s^{i\alpha}(T)\right], \end{aligned} \tag{3.30}$$

Ici, $s^{i\alpha}$ représente l'entropie de la particule i suivant la direction α et $S_i(T) = \sum_{\alpha=1}^{3} s^{i\alpha}(T)$ est simplement l'entropie locale de la particule i. Le formalisme de *Green* nous a donc permis de reformuler le problème de l'entropie vibrationnelle comme un problème linéaire en termes de source *locale* d'entropie. Cette description, en termes locaux, va nous permettre de choisir le type de *noyaux* (voir Sec. 2.2.2) qui sera le plus adapté pour construire un modèle de régression de l'entropie vibrationnelle.

3.2.3 Régression linéaire dans l'espace des descripteurs

Nous allons maintenant faire le lien entre le formalisme local décrit ici et l'espace de représentation des descripteurs présenté dans le chapitre 2. Nous allons montrer que l'utilisation de l'espace des descripteurs permet d'obtenir un modèle de régression simple et robuste pour le cas de l'entropie vibrationnelle. Considérons la matrice de descripteurs du sytème composé de N particules à une température T, $\underline{\underline{D}}(T) \in \mathbb{R}^{\mathcal{D}\times N}$ où $\mathcal{D}$ est la dimension du descripteur utilisé. Nous faisons ici l'hypothèse qu'il existe une relation linéaire entre le vecteur d'entropie locale $S_i(T)$ et la colonne i de la matrice de descripteur $\underline{D}_i(T)$, c'est-à-dire qu'il existe un vecteur de poids $\underline{w}_i \in \mathbb{R}^{\mathcal{D}}$ tel que :

$$S_i(T) = \underline{w}_i \cdot \underline{D}_i(T) \tag{3.31}$$

Cette relation linéaire permet d'assurer l'*extensivité* de l'entropie. Nous supposons, afin de construire un modèle général de régression, que $\forall i, \quad \underline{w}_i \equiv \underline{w}$. Nous pouvons alors donner une expression de l'entropie vibrationnelle $S_{vib}(N,T)$ en termes de vecteurs de descripteurs du système :

$$S_{vib}(N,T) = \sum_{i=1}^{N}\left[\sum_{\alpha=1}^{3} s^{i\alpha}(T)\right] = \underline{w}\cdot\left(\sum_{i=1}^{N}\underline{D}^{i}(T)\right) = N\underline{w}\cdot\langle\underline{D}(T)\rangle \tag{3.32}$$

On note que l'équation (3.32) vérifie le caractère extensif de l'entropie vibrationnelle. L'opérateur $\langle\underline{D}(T)\rangle$ représente la moyenne sur les colonnes de la matrice de descripteurs du système $\underline{\underline{D}}(T)$. L'expression obtenue dans l'équation (3.32) suggère qu'une régression linéaire entre $S_{vib}(N,T)$ et le vecteur moyen $\langle\underline{D}(T)\rangle$ sera le type de régression suffisant. L'utilisation d'un modèle linéaire dans l'espace des descripteurs permet de réduire la dimensionnalité du problème de l'espace des phases de dimension $3N$ à un problème de dimension $\mathcal{D}$. De plus, la dimension de l'espace des descripteurs est fixée quelque soit la taille du système ce qui rend consistant le problème d'un modèle de régression. Il est en effet impossible de construire un modèle de régression dont le nombre de paramètres dépend de la taille du système. Un modèle entraîné sur un système à N particules ne pourrait être utilisé que pour des systèmes à N particules, ce qui rendrait impossible tout changement d'échelle. L'utilisation d'un modèle linéaire (plutôt qu'une approche à noyau cf. Chap (2)) basé sur la décomposition locale formelle de l'entropie vibrationnelle dans le cadre de l'approximation harmonique nous permet d'obtenir des résultats robustes et d'une grande transférabilité qui seront décrits dans la section (3.4).

3.3 Génération et détails de la base de données

Dans cette section, nous allons dans un premier temps expliciter le problème posé par la complexité numérique du calcul de l'entropie vibrationnelle et définir l'entropie vibrationnelle de formation qui est la grandeur d'intérêt que nous avons choisi d'inférer l'entropie vibrationnelle de formation Eq. (3.33). Nous avons généré deux bases de données afin d'entraîner notre modèle.

— Une base de donnée initiale générée pour le fer cubique centré grâce à la méthode *Activation Relaxation Technique* décrite en Sec. 3.3.2 (*ARTn*)

— Cette base de données a ensuite été étendue par modification du volume et/ou de l'état de déformation de la base de données *ARTn* et par forte pertubation des configurations *ARTn* donnant suite à des configurations aléatoires contenant un très grand nombre de défauts ainsi que discuté en Sec. 3.3.3.

3.3.1 Positionnement du problème et grandeur d'intérêt

L'évaluation de l'entropie vibrationnelle harmonique par simulation numérique représente un défi de complexité numérique. En effet, la méthode "standard" d'évaluation consiste en la construction de la matrice dynamique du système $\mathfrak{D}$, opération évoluant comme $\mathcal{O}(N^2)$, puis à la diagonalisation directe de celle-ci afin d'obtenir les valeurs propres ω_ν^2 et les vecteurs propres $\hat{\mathbf{e}}_\nu$. Cette diagonalisation correspond à une complexité en $\mathcal{O}(N^3)$ pour les algorithmes standards de diagonalisation. La complexité globale de la méthode standard évoluant comme $\mathcal{O}(N^3)$, les systèmes pouvant être étudiés sont limités à une taille maximum d'environ 10^5 atomes avec l'aide des clusters de calculs. Cette limitation est aussi bien d'ordre temporelle que d'ordre mémoire. Ainsi le calcul complet de l'ensemble des valeurs propres de la matrice *Hessienne* d'un système de 2×10^5 atomes a nécessité 20 TB de mémoire et environ 7 h de calcul sur près de 3000 CPU récents [14]. La voie principale d'amélioration de la méthode standard est l'évolution des algorithmes de diagonalisation. Néanmoins, on pourra citer quelques méthodes élégantes telle que celle proposée par Huang *et al.* [174] permettant d'éviter la diagonalisation directe de la matrice dynamique du système et évoluant comme $\mathcal{O}(N)$. Cette méthode se base sur une reconstruction directe de la *densité d'état de vibration* du système à l'aide d'une base de polynômes. Malgré l'apport conceptuel de cette méthode, celle-ci reste difficilement compétitive face à la méthode "standard" pour les systèmes de grande taille (>10 000 atomes). En effet, la méthode développée par Huang *et al.* est itérative. Sa convergence - au sens du nombre de fonctions de base et d'évaluations de force - n'est donc pas connue pour un système quelconque et nécessite donc une démarche heuristique (coûteuse en temps) à chaque nouveau système étudié.

Nous proposons ici d'utiliser un modèle type *Machine Learning* basé sur les descripteurs atomiques locaux décrits dans le chapitre 2. Pour ce genre de méthodes, on calcule d'abord la matrice de descripteur associée au système $\underline{\underline{D}} \in \mathbb{R}^{\mathcal{D}\times N}$. Cette opération évolue comme $\mathcal{O}(N)$ pour les descripteurs décrits dans le chapitre précédent (2). L'évaluation du modèle possède une complexité qui dépend du type de méthode de régression utilisé. Dans le cas des méthodes à noyaux, la complexité numérique évolue comme $\mathcal{O}(\mathcal{D}^n M^p)$ où M représente le nombre de configurations présentes dans la base de données d'entraînement. Dans le cas général, n et p peuvent prendre des valeurs entre 0 et 2, mais dans le cas de la régression linéaire $n = 1$ et $p = 0$ ce qui aboutit à une complexité numérique en $\mathcal{O}(N)$. Le modèle linéaire basé sur les descripteurs atomiques locaux permet donc d'améliorer grandement la complexité

numérique du calcul de l'entropie vibrationnelle harmonique et permet donc de simuler des systèmes plus grands. La comparaison schématique de deux méthodes de calcul est présentée dans la figure 3.1.

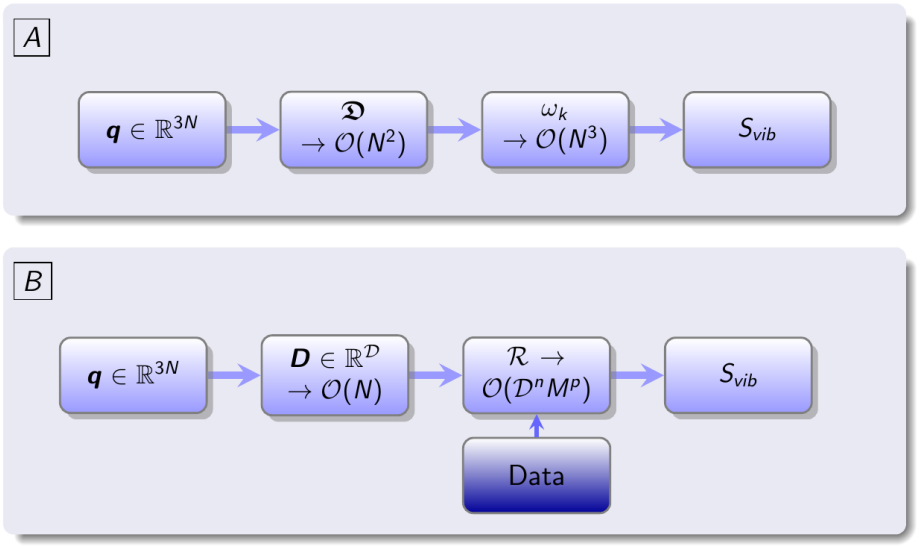

Figure 3.1: Comparaison de deux stratégies de calcul de l'entropie vibrationnelle pour un système de N particules : (A) le calcul complet du spectre de *modes normaux* par construction et diagonalisation directe de la matrice dynamique du système $\mathfrak{D}$ dont la complexité numérique évolue en $\mathcal{O}(N^3)$; (B) la méthode type *Machine Learning* par utilisation d'une base de données de taille M. Les coefficients de la régression n et p peuvent prendre des valeurs entre 0 et 2 selon le type de méthode de régression choisi. Dans le cas de la régression linéaire $n = 1$ et $p = 0$, ce qui aboutit à une complexité numérique en $\mathcal{O}(N)$.

Nous avons choisi de nous intéresser à l'entropie vibrationnelle (harmonique) de formation de défauts ponctuels. Cette grandeur est dérivée de l'entropie vibrationnelle harmonique et se définit comme suit. Considérons un système composé de N_b atomes de cristal parfait et de N_d défauts ponctuels ; on définit pour une température donnée T l'entropie de formation $S_f(T, N_d)$ comme étant :

$$S_f(T, N_d) = S_d(T, N_b \pm N_d) - \frac{N_b \pm N_d}{N_b} S_b(T, N_b) \tag{3.33}$$

Ici, $S_b(T, N_b)$ et $S_d(T, N_b \pm N_d)$ sont respectivement l'entropie vibrationnelle harmonique du cristal parfait contenant N_b atomes et l'entropie vibrationnelle harmonique du système contenant N_d défauts. Nous définissons cette relation à volume constant (V) pour le système avec défauts et le cristal parfait. L'entropie vibrationnelle de formation est une grandeur intensive et elle peut être directement reliée aux pulsations des modes de vibration ω_ν en utilisant l'équation (3.19) :

$$S_f(T, N_d) = k_B \ln \left(\frac{\prod_{\nu_b=1}^{3N_b} (\hbar\omega_{\nu_b})^{\frac{N_b \pm N_d}{N_b}}}{\prod_{\nu_d=1}^{3(N_b \pm N_d)} \hbar\omega_{\nu_d}} \right) \tag{3.34}$$

L'entropie vibrationnelle de formation est donc d'autant plus grande que les fréquences propres du système avec défauts sont petites par rapport aux fréquences propres du cristal parfait. L'entropie vibrationnelle de formation est la grandeur d'intérêt pour étudier la thermodynamique et la cinétique de la création de défauts pour des systèmes réels (c'est-à-dire pour $T > 0$). Dans le cadre d'un modèle linéaire basé sur les descripteurs locaux, l'équation (3.33) peut être reformulée sous la même forme que l'équation (3.32) :

$$S_f(T, N_d) = (N_b \pm N_d)\ \underline{w} \cdot [\langle \underline{D}(T) \rangle_d - \langle \underline{D}(T) \rangle_b] \tag{3.35}$$

où $\langle \underline{D}(T) \rangle_d$ et $\langle \underline{D}(T) \rangle_b$ représentent la moyenne sur les colonnes des matrices de descripteurs de la configuration avec défauts et de cristal parfait. Il nous faut maintenant pouvoir entraîner ce modèle sur une base de données représentative du problème, c'est-à-dire représentative des défauts ponctuels dans le fer.

3.3.2 Génération de la base de données par la méthode *ARTn*

La base d'entraînement est le véritable *nerf de la guerre* en science des données. La base de données doit être la plus complète et la plus représentative du problème de régression étudié. Dans le cas contraire, elle va contenir des biais importants qui vont compromettre le modèle de régression ainsi que sa transférabilité. En effet, dans les applications en physique, les modèles de régression ne doivent pas se limiter à des solutions interpolantes mais aussi à des solutions ayant **un bon pouvoir prédictif**. La base de données d'entraînement conditionne grandement le pouvoir prédictif des modèles de régression.

Afin de construire une base de données la moins biaisée possible, nous avons utilisé l'*Activation Relaxation Technique nouveau* (*ARTn*) [175-178]. L'*Activation Relaxation Technique* est une méthode systèmatique d'exploration du paysage énergétique d'un système à température nulle et a été developpée par Barkema *et al.* [175] et ensuite modifiée par Malek et Mousseau [176] (*ARTn*). Cette méthode a ensuite été améliorée par Marinica *et al.* [178, 179]. Nous décrivons ici brièvement cette méthodologie itérative qui consiste à trouver les points-selle associés au minimum initial. On part d'un état d'énergie minimale arbitraire de coordonnées $\boldsymbol{q}^0$ dans le paysage. L'algorithme *ARTn* va appliquer un déplacement aléatoire $\boldsymbol{\delta q}^0$ et partiellement évaluer le spectre de la matrice Hessienne du système $\mathbb{H}\{\boldsymbol{q}^0\}$ dans l'hyperplan orthogonal à $\boldsymbol{q}^0$. L'évaluation se fait sur la valeur propre la plus petite λ^0_{min} et son vecteur propre associé $\boldsymbol{\phi}(\boldsymbol{q}^0)$. Les coordonnées du système vont alors être mises à jour afin de faire "remonter" le système du bassin dans lequel il se trouve. Ainsi l'évolution des coordonnées du système pour la $i^{ième}$ étape de *ARTn* suit l'équation suivante :

$$\boldsymbol{q}^{i+1} = \begin{cases} \boldsymbol{q}^i + \boldsymbol{\delta q}^i \text{ si } \lambda^i_{min} > 0 \\ \boldsymbol{q}^i - \frac{\|\boldsymbol{\delta q}^i\|}{\|\boldsymbol{\phi}(\boldsymbol{q}^i)\|}\boldsymbol{\phi}(\boldsymbol{q}^i) \text{ si } \lambda^i_{min} < 0 \end{cases} \tag{3.36}$$

Si la valeur propre λ^i_{min} est positive, cela signifie que le système évolue encore dans le "bas" du bassin. Si la valeur propre λ^i_{min} est négative, cela signifie que le système a

dépassé un point d'inflexion du bassin. La méthode *ARTn* consiste alors à trouver les coordonnées du point de selle associé à cette inflexion. On "remonte" alors le bassin en suivant la direction $-\boldsymbol{\phi}(\boldsymbol{q}^i)$. La convergence de la méthode est assurée par les conditions décrites par Marininca *et al.* [179] et permet donc de trouver les points-selle associés au mimimum initial. Une fois au point de selle, il suffit d'appliquer une perturbation orthogonale à la direction d'arrivée puis une relaxation pour trouver un autre état d'énergie minimum dans le paysage énergétique. La méthode des *graphs de connectivité* [179] permet d'assurer l'unicité des minima découverts et de les classer hiérarchiquement en fonction de leur énergie. La méthode *ARTn* est un outil idéal pour construire une base de données de configurations pour notre modèle de régression. Néanmoins, cette méthode reste coûteuse en temps de calcul à cause de l'algorithme de diagonalisation partiel (méthode de *Lanczos*) nécessaire à la convergence pour le point-selle[1]. Pour cette raison, nous nous sommes restreint à des études utilisant la *statique moléculaire semi-empirique* et non des méthodes *ab initio*.

Nous nous sommes servis d'une partie de la base de données générée avec la méthode *ARTn* par Marinica *et al.* [179] pour le fer cubique centré. Cette base de données a été générée en utilisant le *potentiel semi-empirique* de type Embedded Atom Model (EAM) développé par Ackland *et al.* [180]. La base de données comporte des configurations de I_n auto-interstitiel avec $n = 2, 3, 4$ et de V_n lacunes avec $n = 4$. Deux configurations sont considérées comme non équivalentes si elles respectent les conditions suivantes : (i) leur énergies diffèrent de plus de 10^{-2} eV ; (ii) dans le cas des auto-interstitiels, la somme des carrés des valeurs propres du tenseur d'inertie du défaut sont différents, ces défauts étant identifiés par la méthode de Wigner-Seitz [181]. Les informations sur la base de données *ARTn* sont résumées dans la première ligne de la Table 3.1.

Nous allons ici décrire la méthodologie employée pour calculer les entropies vibrationnelles et les descripteurs atomiques associés à chaque configuration. L'entropie de formation est calculée à l'aide du package `PHONDY` [170, 182-184] couplé avec `LAMMPS` [185]. La configuration est d'abord relaxée avec une tolérance de 10^{-6}eV/Å pour la norme des forces. L'évaluation numérique de l'entropie vibrationnelle harmonique se fait par échantillonnage direct de la matrice Hessienne par différence finie à 2 points. On impose un déplacement symétrique de 10^{-3} Å sur un atome et on évalue les $3N$ forces associées à ce déplacement. On vérifie si la configuration relaxée est bien un minimum du paysage énergétique en s'assurant que toutes les fréquences propres des *modes normaux* sont réelles (hormis les fréquences nulles imposées par l'invariance par translation). Si c'est le cas, le spectre de *modes normaux* ainsi que l'entropie vibrationnelle de la configuration à 1000 K sont calculés. Sinon la configuration est à nouveau relaxée après avoir appliqué un déplacement aléatoire suivant une loi uniforme d'amplitude $[-10^{-4}\text{Å}, 10^{-4}\text{Å}]$ sur tous les atomes. Nous avons ensuite calculé la matrice de descripteurs atomiques locaux associée à la configuration relaxée grâce au package `MILADY` [20, 138]. Un schéma de la méthode de calcul est présenté figure 3.2. Afin de mener à bien les calculs de statique moléculaire, nous avons choisi d'utiliser

1. Environ 100 évaluations de forces sont nécéssaires pour estimer λ^i_{min}.

deux *potentiels semi-empiriques* différents pour le fer cubique centré : (i) le potentiel EAM développé par Ackland *et al.* [180] (AM04) ; (ii) le potentiel de type Modified Embedded Atom Model (MEAM) developpé par Alireza et Asadi [186]. Le potentiel développé par Ackland *et al.* est utilisé de façon standard pour des comparaisons de stabilité dans le fer cubique centré [7, 9]. Nous voulions aussi étudier l'influence de la prise en compte des triplets pour le potentiel *semi-empirique* et nous avons choisi le potentiel développé par Alireza et Asadi [186]. La base de données *ARTn* est composée de systèmes contenus dans les boites cubiques de fer cubique centré de volume $(8a_0)^3$ avec $a_0 = 2,8553$ Å, ce qui correspond à 1024 atomes pour le cristal parfait. Nous avons effectué deux types de simulations de *statique moléculaire* pour le potentiel AM04 : (i) des simulations à volume constant V_{ct} ; (ii) des simulations à volume atomique constant V_{at}. Une discussion sur la différence entre ces deux types de calculs sera abordée dans la sous-section 3.4.1.

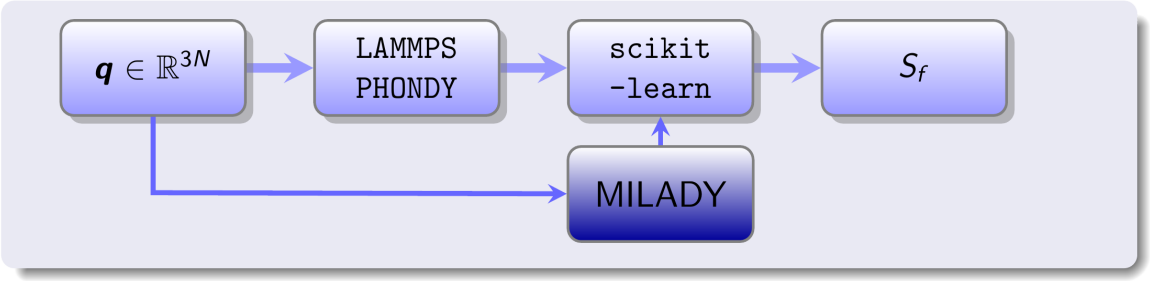

Figure 3.2: Schéma de la méthode de calcul de l'entropie de formations de configurations par utilisation du package `PHONDY` [170, 182-184] couplé avec `LAMMPS` [185]. Parallèlement, les descripteurs atomiques sont calculés à l'aide du package `MILADY` [20, 138]. Le modèle de régression est construit à l'aide du package `scikit-learn` [165].

3.3.3 Extension de la base de données : changement de volume, déformations et configurations aléatoires

La base de données *ARTn* est représentative des défauts d'irradiation possibles dans une boîte de simulations de $(8a_0)^3$ n'ayant pas subi de déformations. Afin de rendre compte de volumes différents et de taux des déformations différents sur le système, nous avons choisi d'étendre la base de données *ARTn* pour le potentiel MEAM seulement. Une discussion sur ce choix sera présentée de façon exhaustive dans la sous-section (3.4.2). Nous avons étendu la base de données *ARTn* de deux façons différentes.

Dans un premier temps, nous avons appliqué le tenseur de déformations suivant à toutes les configurations de la base de données *ARTn* :

$$\boldsymbol{\epsilon} = \sum_{i=1}^{3} \epsilon \, \boldsymbol{x}_i \otimes \boldsymbol{x}_i \tag{3.37}$$

Ici, ϵ est le taux de déformation et $\boldsymbol{x}_i \in \mathbb{R}^3$ sont les vecteurs engendrant la boîte de simulation. Ce tenseur correspond à une déformation homogène et isotrope de taux de déformation ϵ. Nous avons choisi d'appliquer les valeurs de taux de déformations suivants ; $\epsilon = -1\%, 1\%, 2\%, 3\%$. La méthodologie décrite dans la sous-section 3.3.2 a ensuite été appliquée afin de calculer les entropies de formation et la matrice de descripteurs locaux des configurations. Cette nouvelle base de données est appelée base de données *ARTn déformée*. Les données concernant cette base *ARTn déformée* sont rassemblées dans les lignes 2 à 6 de la Table 3.1. **Cette extension permet de tester la robustesse de notre modèle pour de petites perturbations géométriques des systèmes**.

Dans un second temps, nous avons changé le volume et le nombre de défauts présents dans les boîtes de simulations. Pour cela, nous avons étendu à $(10a_0)^3$ et $(12a_0)^3$ le volume de la base de données *ARTn*. Des déplacements aléatoires ont ensuite été appliqués sur les configurations afin de créer des paires de *Frenkel*. La méthodologie décrite dans la sous-section (3.3.2) a ensuite été appliquée afin de calculer les entropies de formation et la matrice de descripteurs locaux des configurations. Malgré la relaxation effectuée en *statique moléculaire*, le nombre de défauts ponctuels présent dans la boîte de simulation reste élevé comme le montre la figure 3.3, allant jusqu'a 22 lacunes et 26 auto-interstitiels dans le même système. Cette nouvelle base de données est appelée base de données *aléatoire*. **Il est important de noter qu'il existe une forte corrélation positive entre le nombre d'atomes auto-insterstitiels et le nombre de lacunes dans un même système car le nombre d'atomes est resté constant dans la boîte lors de la création des paires de *Frenkel***. Nous avons aussi procédé à une analyse de la distribution des entropies de formations pour la base de données *aléatoire*. Celle-ci est présentée dans la figure 3.4(a). On constate que la base de données *aléatoire* contient des configurations possédant une entropie de formation allant jusqu'à 250 k_B, ce qui correspond à des configurations se situant très loin de l'équilibre thermodynamique. Ces conditions hors-équilibres sont fréquemment recontrées sous irradiation [11] . Nous avons aussi procédé à une analyse de la distribution de distance entre les défauts d'une même configuration. Cette analyse a été menée par utilisation de la méthode de Wigner-Seitz [181] implémentée dans le package `Ovito` [187] et est présentée par la figure 3.4(b). On constate que les défauts au sein d'un même système peuvent se situer à des distances inférieures à 10 Å, c'est-à-dire à $2r_{cut}$ du *potentiel semi-empirique*. **On ne peut donc pas faire l'hypothèse que les défauts sont indépendants les uns des autres. Cette base de données sera utilisée pour tester la transférabilité des modèles de régression de l'entropie vibrationnelle de formation. Les informations sur la base de données *aléatoire* sont données par les deux dernières lignes de la Table 3.1**.

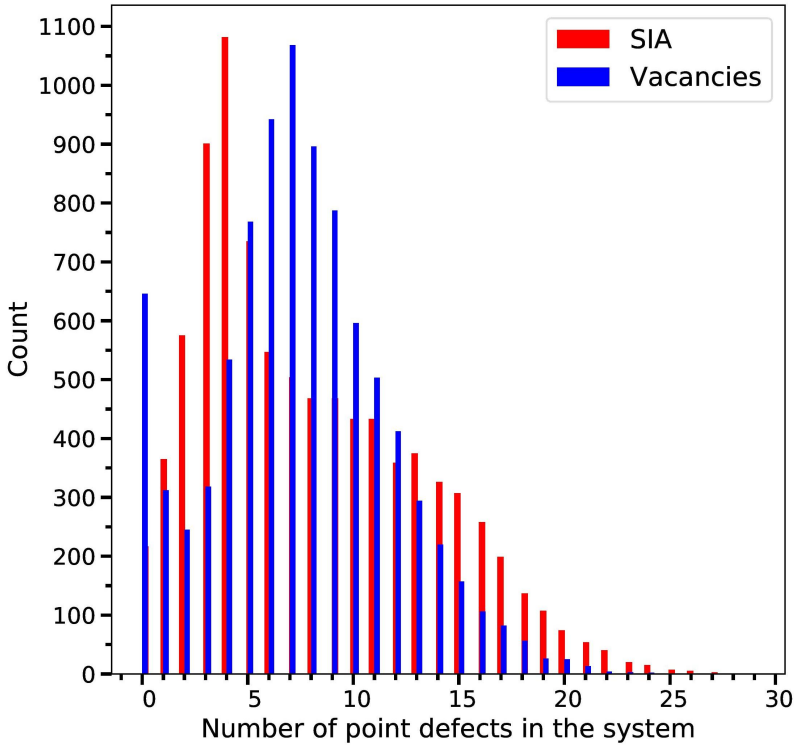

Figure 3.3: Analyse de la distribution de défauts ponctuels présents dans la base de données *aléatoire*. Cette base de données est construite à partir de la base de données *ARTn* par augmentation du volume du système à $(10a_0)^3$ ou $(12a_0)^3$ et par création de paires de *Frenkel*.

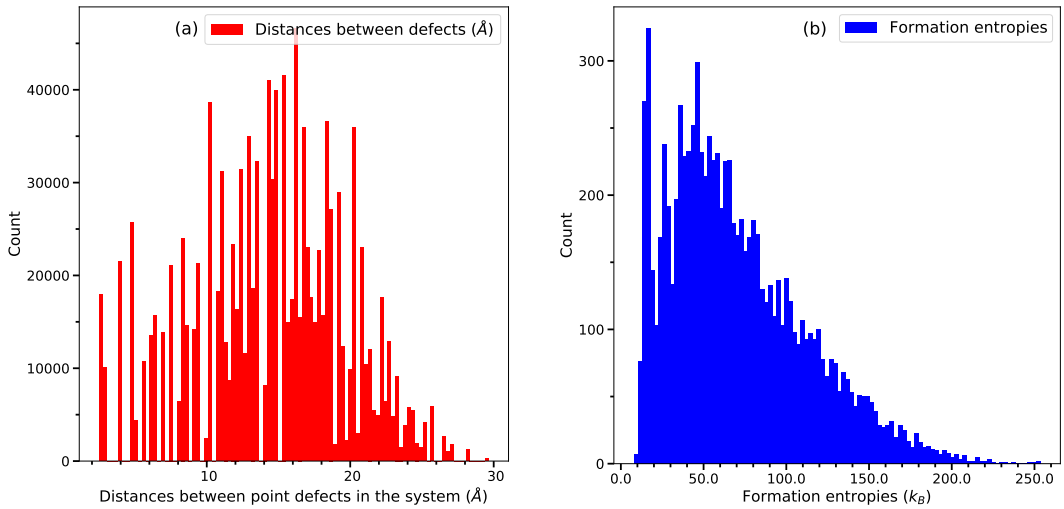

Figure 3.4: Analyse de la distribution des entropies de formation (b) et des distances entre les défauts ponctuels (a) dans la base de données *aléatoire*. Les entropies de formation suivent la même distribution que le nombre de défauts ponctuels dans les boites présenté par la Fig. 3.3. On constate que environ 1/3 des défauts ponctuels sont séparés de moins de 10 Å $= 2r_{cut}$.

Table 3.1: Base de données utilisée pour l'entraînement du modèle de régression. N est le nombre d'atomes pour le système parfait, N_{cf} est le nombre de configurations différentes pour chaque classe de défaut. I_{2-4} et V_4 représentent les clusters de défauts contenant respectivement 2 à 4 auto-interstiels et des 4 lacunes. La taille des systèmes contenant des défauts sont respectivement $N+(2\dots4)$ et $N-4$ pour I_{2-4} et V_4. ϵ est le taux de déformation homogène et isotrope appliqué sur les configurations

Système (N, ϵ)	Type de défaut ponctuel (N_{cf})				Total	Base
	I_2	I_3	I_4	V_4		
1024, $\epsilon = +0\%$	434	1105	1280	1701	4520	*ARTn*
1024, $\epsilon = -1\%$	434	1105	1280	1701	4520	
1024, $\epsilon = +1\%$	434	1105	1280	1701	4520	Base
1024, $\epsilon = +2\%$	434	1105	1280	1701	4520	*déformée*
1024, $\epsilon = +3\%$	434	1105	1280	1701	4520	
2000, $\epsilon = +0\%$	434	1105	1280	1701	4520	Base
3456, $\epsilon = +0\%$	434	1105	1280	1701	4520	*aléatoire*
Total	3038	7735	8960	11907	31640	

3.4 Modèle linéaire de régression de l'entropie vibrationnelle

Dans cette section, nous allons présenter les résultats obtenus en utilisant le modèle linéaire de régression détaillé dans les sous-sections 3.3.1 et 3.2.3. Dans la sous-section 3.4.1, nous présentons un modèle très simple de régression entre l'entropie de formation à volume constant et à volume atomique constant basé sur la théorie *élastique isotrope*. Nous allons voir que cette théorie n'est pas suffisante pour décrire correctement le problème de l'entropie vibrationnelle ce qui est l'une des motivations de l'approche par utilisation des descripteurs locaux. Dans la sous-section 3.4.2, nous présentons les résultats obtenus en utilisant la régression linéaire *Bayésienne* (voir chapitre 2) dans l'espace des descripteurs locaux pour ajuster l'entropie vibrationnelle de formation des différentes bases de données présentées dans la section 3.3. Nous étudierons avec un soin particulier la base de données *aléatoire* qui va montrer la transférabilité et la robustesse du modèle linéaire.

3.4.1 Insuffisance de la théorie *élastique isotrope*

Nous présentons ici un modèle très simple de régression de la différence d'entropie de formation entre une simulation à volume constant et une simulation à volume atomique constant, visant à reproduire les propriétés d'extensivité de l'entropie vibrationnelle. Dans toute cette sous-section, les données utilisées pour ajuster le modèle sont issues de la base de données *ARTn* pour le *potentiel semi-empirique* développé par Ackland *et al.* [180] pour le fer cubique centré. Nous cherchons à calculer le terme de correction $\Delta S^{P\to V}$ suivant :

$$S_v^{P\neq 0} = S_v^{P=0} + \Delta S^{P\to V} \tag{3.38}$$

Cette correction $\Delta S^{P\to V}$ correspond à la différence entre l'entropie de formation calculée à volume constant $S_v^{P\neq 0}$ et l'entropie de formation calculée à volume atomique constant (c'est-à-dire à pression nulle) $S_v^{P=0}$. L'expression de ce terme de correction peut être obtenue analytiquement dans le cas de défauts à symétrie sphérique grâce à la théorie *élastique isotrope*. Ce calcul a été mené par Mishin *et al.* [188] et aboutit à l'expression suivante pour $\Delta S^{P\to V}$:

$$\Delta S^{P\to V} = \frac{1+\nu}{3(1-\nu)}\beta\,\mathrm{Tr}(\boldsymbol{\sigma})V_0 \tag{3.39}$$

Ici, ν est le coefficient de *Poisson* du matériau, $\boldsymbol{\sigma}$ est le tenseur des contraintes de la configuration, V_0 est le volume du système mesuré pour la simulation à volume constant et β est le coefficient de dilatation thermique du matériau. Dans le cadre de la théorie *élastique isotrope*, le coefficient de *Poisson* ν et le coefficient de dilatation thermique β doivent être uniques pour le matériau quelque soit le type de défauts contenus dans le système simulé. Si nous fixons la valeur de ν, cela implique que la distribution des valeurs de β doit être unimodale pour toutes les configurations de la base de données *ARTn*. Nous présentons cette distribution dans la figure 3.5. On constate immédiatement que la distribution est bimodale : (i) un mode correspond aux auto-interstitiels et possède une faible variance ; (ii) l'autre mode correspond aux lacunes et possède une variance beaucoup plus élevée. La théorie élastique isotrope ne permet donc pas de reproduire la propriété d'extensivité de l'entropie vibrationnelle si le type de défauts contenu dans le système n'est pas connu de façon a priori. **Il est donc nécessaire d'utiliser une approche plus raffinée que l'élasticité isotrope pour le problème de régression des entropies vibrationnelles de formation.**

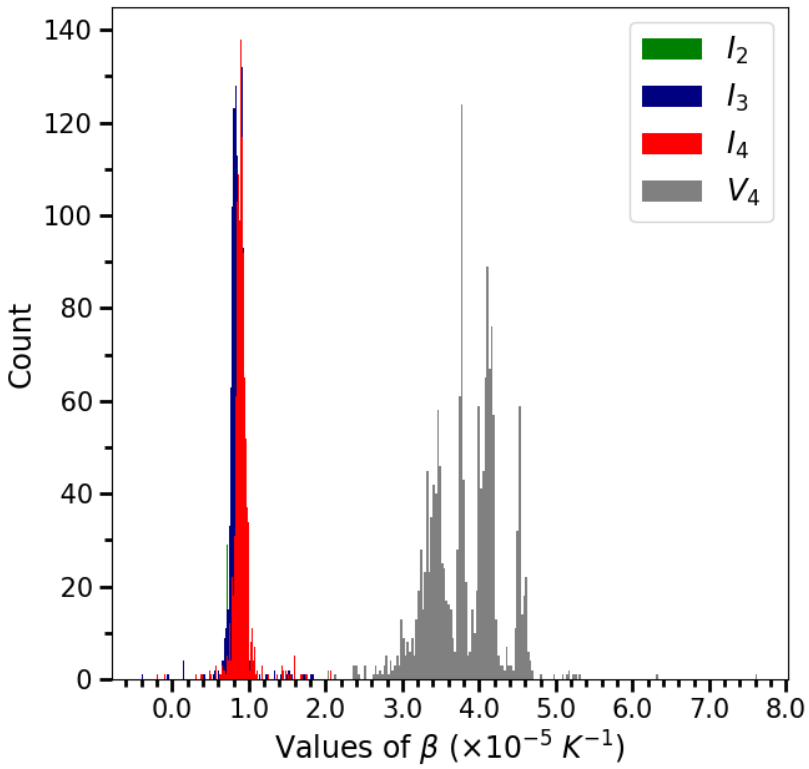

Figure 3.5: Analyse de la distribution du coefficient de dilatation thermique β pour les différentes classes de défauts ponctuels de la base de données *ARTn*. On constate que la distribution n'est pas unimodale ce qui montre que β n'est pas unique pour tous les types de défauts ponctuels dans un matériau donné.

3.4.2 Modèles linéaires dans l'espace des descripteurs

Nous allons ici présenter les résultats des modèles linéaires dans l'espace des descripteurs permettant d'ajuster les différentes bases de données décrites dans la section précédente 3.3. Dans cette sous-section, nous avons utilisé les descripteurs suivants afin de construire le modèle de régression : (i) les Angular Fourier Series (AFS) $\mathcal{A}_{r,\theta}$ possédant r composantes radiales et θ composantes angulaires, (ii) le bi-spectrum SO(4) $bSO(4)_{j_{max}}$ possédant un "moment cinétique" maximum égal à j_{max} et (iii) les coefficients de *scattering* $S^{\mathcal{J},L}$ possédant un "moment cinétique maximun' égal à L et couplant les échelles incluses dans le sous-ensemble $\mathcal{J}$. La définition formelle de ces descripteurs est donnée dans le chapitre 2. Dans l'intégralité de cette sous-section, le rayon de coupure utilisé pour les descripteurs r_{cut} est égal à 5 Å.

Nous débutons par la base de données *ARTn* et nous comparons les deux *potentiels semi-empiriques* (i) AM04 [180] et (ii) MEAM [186] ainsi que les différents descripteurs cités plus haut. Nous cherchons ici à obtenir le modèle le plus précis possible. Afin de quantifier la notion de précision des modèles construits, nous introduisons deux grandeurs statistiques : (i) l'erreur quadratique moyenne (Root Mean Square Error,

RMSE) et (ii) l'erreur moyenne absolue (Mean Absolute Error, MAE) ;

$$\sqrt{M_s^{-1}\|\underline{S}_s - \underline{w}\cdot\underline{\underline{D}}_s\|^2} \quad \text{(RMSE)} \tag{3.40}$$

$$M_s^{-1}\|\underline{S}_s - \underline{w}\cdot\underline{\underline{D}}_s\|^1 \quad \text{(MAE)} \tag{3.41}$$

Ici M_s est le nombre de configurations présentes dans la base de données, $\underline{S}_s \in \mathbb{R}^{M_s}$ est le vecteur d'entropies de formations à ajuster, $\underline{w} \in \mathbb{R}^{\mathcal{D}}$ est le vecteur de poids et $\underline{\underline{D}}_s \in \mathbb{R}^{\mathcal{D}\times M_s}$ est la matrice de descripteurs moyennés (cf. Sec. (3.2.3)) de la base de données. Enfin, $\|\cdot\|^1$ et $\|\cdot\|^2$ réprésentent respectivement la norme L_1 et le carré de la norme L_2. Un modèle de régression est d'autant plus précis que sa RMSE et sa MAE sont proches de zéro. Les différents modèles de régression en fonction du *potentiel semi-empirique* et du type de descripteur sont présentés dans la figure 3.6. Les résultats sont visualisés dans le plan $\left(\underline{w}\cdot\underline{\underline{D}}_s, \underline{S}_s\right)$. Un modèle est d'autant plus précis que les points sont proches de la droite $y = x$, la variance des points dans ce plan est une image directe de la RMSE. On constate que le modèle linéaire Eq. (3.32) motivé par le formalisme de *Green* est un très bon estimateur de l'entropie vibrationnelle de formation. Sur une plage d'entropie de l'ordre de 20 k_B à 25 k_B, la RMSE est inférieure à 1 k_B. Pour les modèles linéaires proposés ici, le nombre de paramètres ajustables est de l'ordre de la centaine et la base de données compte environ 4000 configurations. Le bon ajustement n'est donc pas lié à un phénomène de *sur-ajustement* symptomatique des *méthodes à noyaux* (cf. Sec. 2.2.2). Pour tous les descripteurs utilisés, on constate que les lacunes sont plus facilement ajustées que les auto-interstitiels, ce qui s'explique par la grande variabilité des configurations d'auto-interstitiels. Nous allons maintenant comparer les différents modèles obtenus, d'abord en fonction du *potentiel semi-empirique* utilisé et ensuite en fonction des descripteurs utilisés.

En termes de comparaison des *potentiels semi-empiriques*, on constate que pour un même descripteur les indicateurs statistiques du potentiel MEAM sont meilleurs que ceux du potentiel AM04 (EAM), ce que l'on constate aussi visuellement avec une plus grande dispersion pour AM04 que pour MEAM. Nous allons essayer d'interpréter cette différence significative. L'une des principales différences entre le potentiel AM04 et MEAM est le formalisme de construction. Le potentiel AM04 a été ajusté à l'aide de fonctions *splines cubiques* alors que le potentiel MEAM, lui, est ajusté avec des fonctions présentant une meilleure régularité. L'estimation du Hessien du système nécessite d'évaluer numériquement la dérivée seconde du potentiel pour chacune des coordonnées des atomes présents à l'intérieur du système. Dans le cas des fonctions *splines cubiques*, la dérivée seconde sera linéaire par morceaux (et donc au mieux de classe $\mathcal{C}^0$). Le modèle de régression que nous proposons est équivalent à une estimation à un point de la courbure multi-dimensionnelle d'un bassin dans le paysage énergétique du système. Une bonne estimation à un point nécessite des propriétés importantes de régularité des fonctions estimées. La plus grande précision du modèle de régression pour le potentiel MEAM est donc due à la meilleure régularité de celui-ci. Après cette constatation, nous avons choisi d'utiliser seulement le potentiel MEAM pour les autres modèles de régression de cette sous-section. Une étude quantitative des défauts de régularité de certains *potentiels semi-empiriques* est présentée en Annexe C.

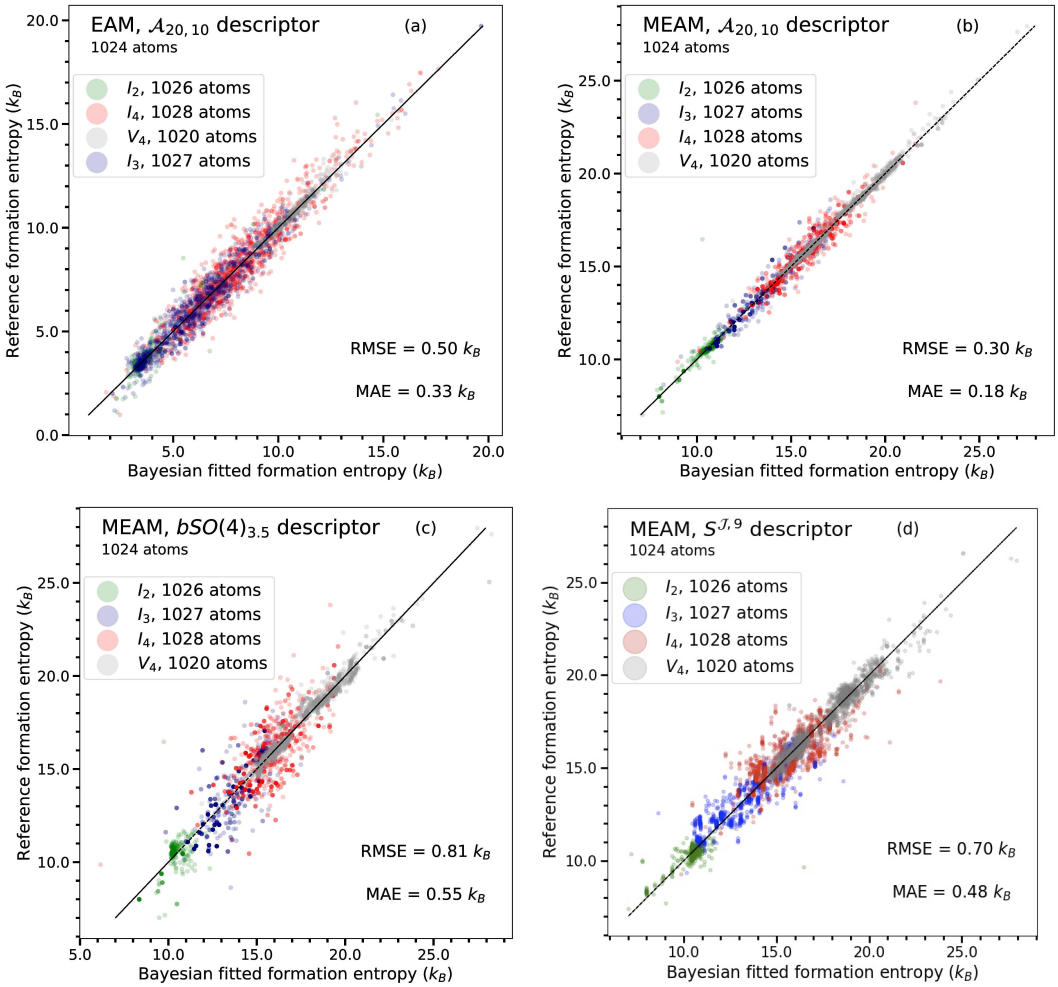

Figure 3.6: Comparaison entre les entropies calculées par diagonalisation directe du Hessien et l'ajustement du modèle linéaire en fonction du potentiel : (a) EAM et (b-d) MEAM pour les 2-4 auto-interstitiels I_{2-4} et les quadri-lacunes V_4 pour des systèmes de volume $(8a_0)^3$. Le nombre de configurations pour chaque classe de défauts est donné par la Tab.3.1. Les descripteurs utilisés sont : (a-b) $\mathcal{A}_{20,10}$, (c) $bSO(4)_{3.5}$ et (d) $S^{\mathcal{J},L}$ avec les échelles suivantes $\mathcal{J} = \{0, 0.25, 0.5, 0.75, 1, 2, 3, 4, 5\}$.

Comparons maintenant les différents descripteurs. On constate qu'il existe de fortes variations des indicateurs statistiques en fonction du type de représentation utilisée. Les descripteurs AFS semblent être les mieux adaptés pour le problème de régression, comparativement aux coefficients de *Scattering* ou le bSO(4). Ce résultat peut être expliqué par les raisons suivantes : (i) les grilles radiales et angulaires du descripteur AFS sont plus facilement adaptables grâce à la définition des AFS en produit direct. Ce n'est pas le cas pour le bSO(4) où les informations radiales et angulaires sont "mixées" par la définition des fonctions *hyper-sphériques* ; (ii) le descripteur AFS est un descripteur local, plus à même de décrire un problème de termes sources locales (cf. Sec. 3.2.2 et Sec. 3.2.3) que les coefficients de *Scattering* qui sont construits pour les analyses multi-échelles et qui sont donc non locaux. La meilleure précision du

descripteur AFS montre que le problème de la régression d'entropie vibrationnelle de formation pour les défauts ponctuels dans le fer cubique centré nécessite une grille radiale fine, ce qui est en adéquation avec la description radiale complexe des atomes formant les auto-interstitiels.
Nous allons ici décrire les résultats obtenus par les modèles de régression linéaires pour la base de données *déformée*. Pour cette base de données, nous avons décidé de ne considérer que le cas du potentiel MEAM et d'utiliser le descripteur AFS $\mathcal{A}_{20,10}$, les résultats du modèle de régression sont présentés dans la figure 3.7. On constate que les indicateurs statistiques conservent des valeurs très satisfaisantes de l'ordre de $0.4\,k_B$ pour la RMSE. La figure 3.7.(a) montre le même comportement que pour la base de données *ARTn*. Le modèle linéaire ajuste mieux l'entropie vibrationnelle des lacunes que des interstitiels. La figure 3.7.(b) présente les résultats du modèle en fonction des taux de déformations appliqués. On observe que les configurations les plus déformées ont tendance à être un peu moins bien ajustées par le modèle.

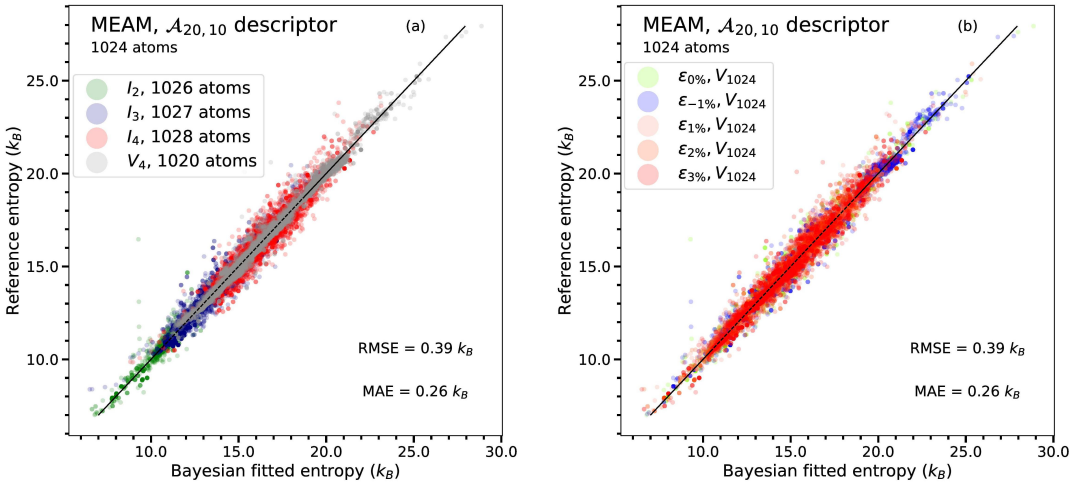

Figure 3.7: Illustration des résultats du modèle linéaire pour la base de données *déformée* pour les I_{2-4}/V_4 amas de défauts (pour le potentiel MEAM) en utilisant le descripteur $\mathcal{A}_{20,10}$. Les systèmes initiaux possèdaient un volume de $(8a_0)^3$ et ont été déformés par appliquation d'une déformation homogène et isotrope. Le taux de déformation varie de -1% à 3%. La figure (a) illustre les résultats du modèle de régression en fonction du type de défauts ponctuels ; la figure (b) illustre les mêmes résultats mais en fonction du taux de déformation.

Les modèles de régression présentés plus haut ont été ajustés sur l'ensemble de leur base de données, ils ne présentent donc pas de preuve directe de transférabilité. Nous allons ici nous servir de la base de données *aléatoire* afin de démontrer la transférabilité et la robustesse de l'approche linéaire pour le problème de régression de l'entropie vibrationnelle de formation. Pour cela, nous utilisons la méthode de *training/testing*. La base de données est alors découpée aléatoirement en deux : une première partie de proportion $(1-p)$ sert à entraîner le modèle ; la deuxième partie de proportion p

sert à vérifier la transférabilité. On calcule ensuite les indicateurs statistiques pour les deux sous-ensembles de la base de données. Afin d'éviter des biais de représentatitivé des ensembles d'entraînement et de vérification, on réitère 100 fois la procédure et on calcule la moyenne des indicateurs statistiques. Les résultats de la procédure de *training/testing* sont présentés par la figure 3.8.(a). On constate que la valeur des indicateurs statistiques est très satisfaisante, de l'ordre de 1 k_B, pour une plage d'entropie vibrationnelle de formation allant jusqu'à 250 k_B. De plus, on observe que ces indicateurs restent stables même pour une très grande proportion du sous-ensemble de vérification. **Le modèle linéaire semble donc prometteur à des fins d'extrapolations pour des estimations d'entropies vibrationnelles de formation**. Afin d'aller encore plus loin dans la vérification de la transférabilité du modèle linéaire, nous avons choisi d'entraîner le modèle sur la base de données *ARTn* et d'utiliser la base de données *aléatoire* comme ensemble d'extrapolation. Les résultats sont présentés dans la figure 3.8.(b). On observe que la RMSE est seulement de $1,5\,k_B$ alors que le modèle n'a été entraîné que sur des configurations de défauts isolés. Le modèle linéaire est transférable à des configurations aléatoires de défauts ponctuels et intéragissant à moins de $2r_{cut}$ du potentiel MEAM utilisé (cf. Sec. 3.3.3). À ce stade, on peut noter les deux conclusions suivantes : (i) le problème de l'entropie vibrationnelle de formation peut effectivement être réduit à un problème local en terme de sources d'entropie vibrationnelle comme le suggérait le formalisme de *Green* pour la *densité d'état* de *modes normaux* ; (ii) l'utilisation d'un modèle linéaire dans l'espace des descripteurs fournit des ajustements extrêmement stables et d'une très grande transférabilité. Le modèle linéaire entraîné sur les bassins énergétiques de la base de données *ARTn* possède un très bon pouvoir prédictif sur des bassins énergétiques disjoints (base de données *aléatoire*) de la base de données d'entraînement. De plus ce modèle est capable de reconstruire les très grandes longueurs d'ondes des modes de vibration associés aux *dumbbells* d'orientation $\langle 111\rangle$ [169-171] jouant un grand rôle dans la transition $\alpha-\gamma$ martensitique du fer et la nucléation de paires de décrochement pour la dislocation $\frac{1}{2}\langle 111\rangle$ [14]. Ces longueurs d'ondes dépassent largement les 10 Å mais sont tout de même reconstruites par la régression. Cette excellente transférabilité pourrait être utilisée pour prédire l'entropie de formation de défauts de grande taille, telles que les boucles de dislocation dont le calcul numérique standard est impraticable à cause de la taille du système (de l'ordre de 10^6 atomes). Il est aussi important de noter que le modèle linéaire (dont la justification vient du formalisme de *Green*) n'est valable que dans le cadre de l'*approximation harmonique*.

Afin de nous assurer de la transférabilité de notre modèle, nous avons confronté ses prédictions avec un calcul direct pour un objet étendu dans le fer cubique centré. Nous avons choisi une **boucle de dislocation de 283 atomes interstitiels** dans une boîte cubique de simulation de 128 283 atomes (soit un volume de $(40a_0)^3$). Cette boucle est présentée dans la figure 3.9. Afin de rendre compte des effets de tailles finies, nous avons modifié la base de données *aléatoire* grâce à la procédure décrite dans la section 3.4.3. Le calcul par diagonalisation directe de la boîte de cristal parfait

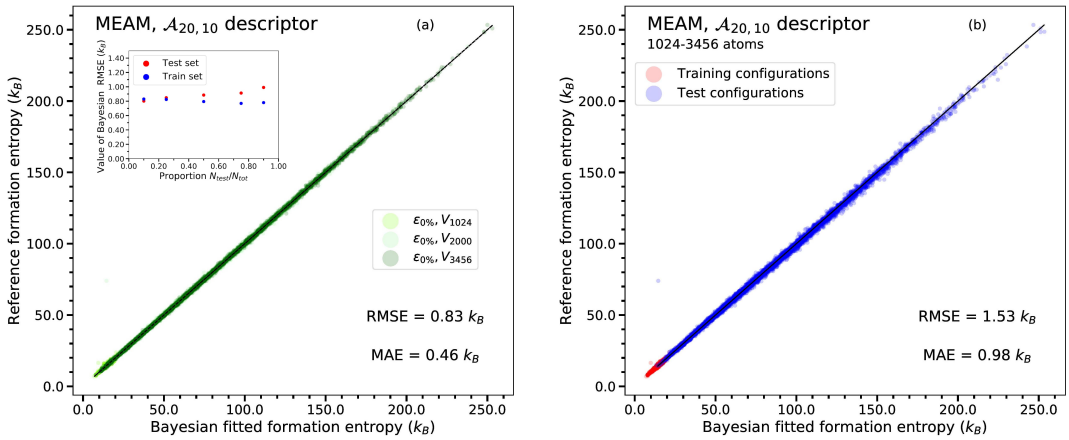

Figure 3.8: La robustesse et la transférabilité du modèle linéaire sont testées par (a) la procédure de *training/testing* présentée dans la section (3.3.3) sur la base de données *ARTn* et *aléatoire* (les entropies sont calculées en utilisant le potentiel MEAM [186] et le descripteur utilisé pour la régression est $\mathcal{A}_{20,10}$). Les résultats de la procédure de *training/testing* sont donnés pour différentes proportions d'ensemble de vérification (voir l'encart). (b) Le pouvoir prédictif du modèle linéaire est testé en l'entraînant sur la base de données *ARTn* et en extrapolant sur la base de données *aléatoire*. On constate que les indicateurs statistiques sont du même ordre de grandeur que pour la figure 3.8.(a) alors que le modèle est en régime d'extrapolation.

de $(40a_0)^3$ et la boîte contenant la boucle de dislocation a nécessité environ 14 heures sur 3000 CPU. **Notre modèle linéaire dans l'espace des descripteurs permet une estimation en environ 10 minutes sur un ordinateur de bureau avec une erreur de seulement 5% par rapport à la diagonalisation directe. Notre modèle linéaire, dans l'espace des descripteurs, se montre donc transférable et rapide pour des objets étendus**.

Nous avons travaillé essentiellement sur les défauts dans le fer cubique centré pour vérifier la validité d'un modèle linéaire dans l'espace des descripteurs pour ajuster l'entropie vibrationnelle de formation. Afin d'aller plus loin dans notre méthodologie de validation, nous avons décidé de tester l'efficacité du modèle linéaire sur des systèmes non cristallins. Pour cela nous nous sommes intéressés à des systèmes très simples : les *clusters Lennard-Jones*. Ces systèmes contenant un faible nombre d'atomes sont un cas d'école de paysage énergétique contenant un grand nombre bassins d'attraction [57]. Ces systèmes sont en général utilisés pour tester l'efficacité des méthodes d'exploration dans les paysages énergétiques complexes [58, 59, 189]. Nous nous sommes ici intéressés aux *clusters Lennard-Jones* LJ_{38} contenant 38 atomes pour lesquels il existe une base de données de grande taille fournie par le Professeur David J. Wales : `http://www-wales.ch.cam.ac.uk/CCD.html`. L'entropie des LJ_{38} est calculée sans aucune difficulté par diagonalisation directe de la matrice Hessienne. Nous avons sélectionné 10 000 configurations différentes de LJ_{38} dont nous avons calculé

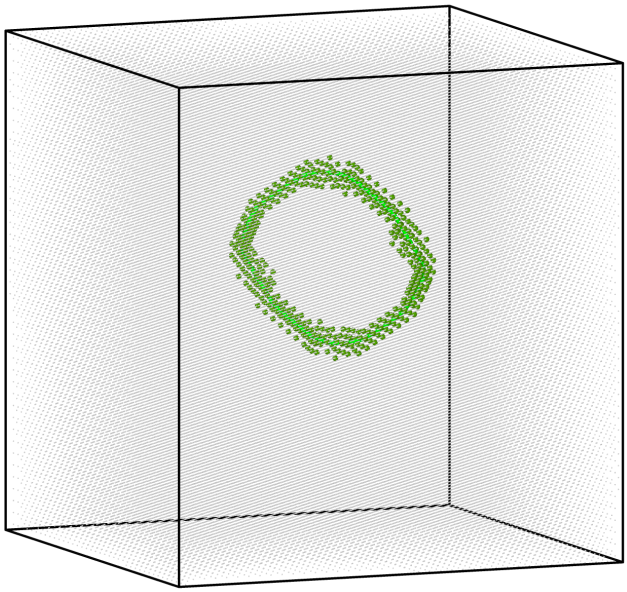

Figure 3.9: Visualisation de la boucle interstitielle dans le fer cubique centré grâce au logiciel `OVITO` [187]. Celle-ci est contenue dans une boîte cubique de volume $(40a_0)^3$.

l'entropie vibrationnelle ainsi que les descripteurs. Ici, nous avons choisi d'utiliser le descripteur AFS $\mathcal{A}_{20,10}$ (cf. Sec. 2.1.2) et un rayon de coupure de 5 Å. Le modèle linéaire a été ajusté et les résultats obtenus sont présentés dans la figure 3.10 avec en encart les résultats de la procédure de *training/testing* décrite précédemment. On constate que comme dans le cas du fer cubique centré, le modèle linéaire possède une très bonne précision et une très bonne transférabilité illustrée par les résultats du *training/testing*.

3.4.3 Entropies vibrationnelles locales

L'hypothèse majeure de notre modèle linéaire de régression est la proportionnalité entre l'entropie locale d'un atome et son environnement atomique. En effet, notamment dans les matériaux cristallins, les modes de vibration peuvent être très délocalisés et leur longueur d'onde peut aisement dépasser la taille de la boite de simulation. C'est le cas du mode de vibration $\langle 111 \rangle$ dans le fer cubique centré qui peut être délocalisé sur plusieurs dizaines d'Ångströms. **Dans cette section, nous éprouvons l'hypothèse de cette proportionnalité locale en entraînant directement le modèle linéaire sur les entropies locales**.

Nos simulations - présentées dans les sections 3.4.2 - ont été effectuées sur des systèmes de tailles finies et possèdent donc un biais sur les modes de vibration de basses fréquences. En effet, la longueur d'onde maximum pouvant résonner dans un système de taille finie est $2L$ avec L la plus grande dimension de la boite de simulation. Les *modes normaux* de très basse fréquence et donc les plus délocalisés peuvent alors

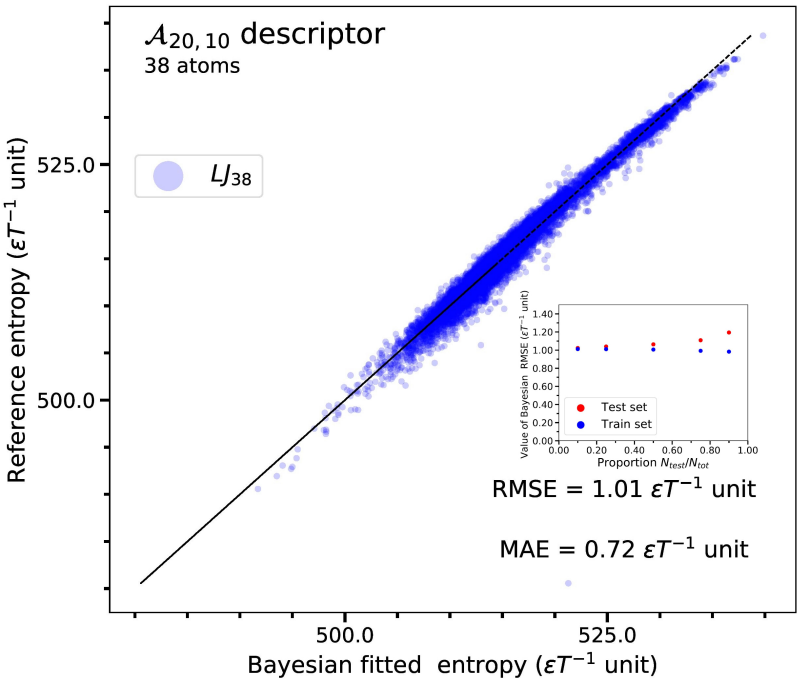

Figure 3.10: Modèle linéaire appliqué aux *clusters Lennard-Jones* contenant 38 atomes LJ_{38} pour ajuster leur entropies vibrationnelles en unités de ϵT^{-1}. Les résultats de la procédure de *training/testing* sont présentés en encart. Les indicateurs statistiques restent stables même pour de grandes proportions d'ensemble de vérification.

ne pas pouvoir exister dans un système de taille finie. Afin de lever ce biais de simulation, nous avons implémenté une méthode de maillage de l'espace réciproque (*points-q*) [168] dans le package `PHONDY` [170, 182-184] couplé avec `LAMMPS` [185] afin d'améliorer l'échantillonnage des vecteurs d'ondes (la nouvelle matrice dynamique échantillonnée est donnée par l'équation (3.17)). **La grille de *points-q* utilisée pour nos simulations est uniforme et permet d'obtenir le même échantillonnage de l'espace réciproque qu'une boite cubique de 127680 atomes (grille $4\times4\times4$ pour une boite cubique de 2000 atomes), ce qui permet d'augmenter significativement le nombre de mode de vibration basses fréquences admissibles**. Le calcul des entropies locales a ensuite été effectué dans les mêmes conditions que celles décrites dans la section 3.3 en utilisant la décomposition locale de la *densité d'état* de *modes normaux* présentée dans la section 3.2 moyennée sur l'ensemble des *points-q* utilisés pour échantillonner l'espace réciproque [168]. Ces calculs sont très coûteux en temps CPU car ils nécessitent autant de diagonalisations que de *points-q* utilisés pour former la grille dans l'espace reciproque. Nous nous sommes donc limités à la base de données *aléatoire* à 2000 atomes. Celle-ci est représentative d'une grande partie des défauts ponctuels pour le fer cubique centré.

Les résultats obtenus sur la base de données *ARTn* pour une température de $T = 1000\ K$ sont présentés dans la figure 3.11 en utilisant le descripteur AFS(20,10) avec un rayon de coupure $r_{cut} = 5$Å. **On constate que le modèle possède une très bonne**

précision avec une RMSE de $0.01\ k_B$**. Ici, la moyenne des entropies locales est de** $6.73 k_B$ **par atome et les entropies les plus élevées correspondent à des systèmes où des modes de basses fréquences sont fortement représentés**. On constate que la variance du modèle reste stable pour de grandes valeurs d'entropies par atomes et commence à dévier pour des entropies dépassant $7.8\ k_B$ par atome. Notre modèle linéaire est donc capable de reconstruire des modes de vibrations dont la délocalisation dépasse le rayon de coupure utilisé par les descripteurs. Il est néanmoins important de constater que le modèle dévie de la réalité pour des entropies locales très élevées, de l'ordre de $7.8\ k_B$ par atome.

Nous précisons, ici, l'importance de la représentativité de la base de données. En effet, l'erreur d'estimation de l'entropie vibrationnelle de formation de la boucle de dislocation décrite par la figure 3.9 était de l'ordre de 200% sans la prise en compte des *points-q* et seulement de 5% quand ceux-ci étaient pris en compte. **La représentativité et la "physique" de la base de donnée sont donc absolument cruciales pour ce type d'approche**.

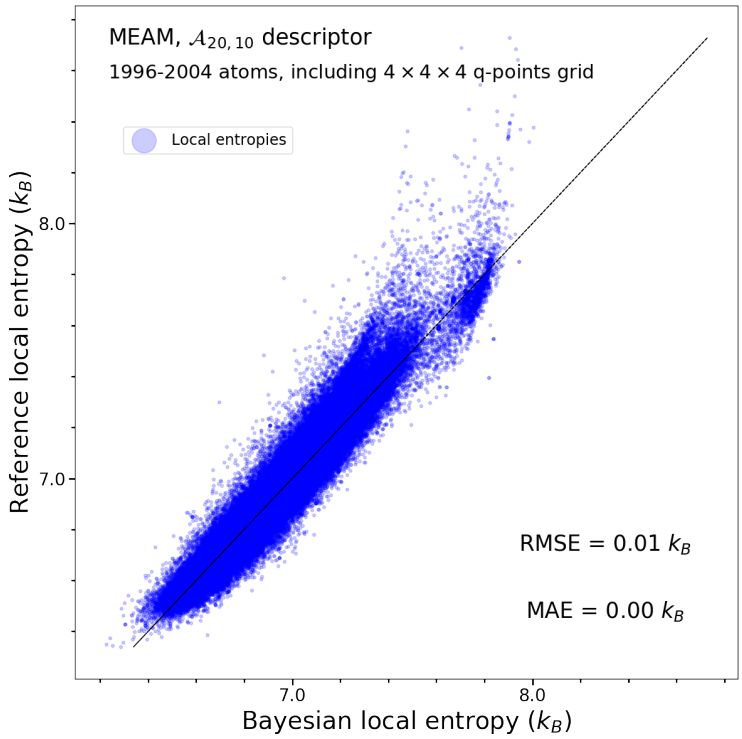

Figure 3.11: Application du modèle linéaire pour la régression des entropies locales de la base de données *aléatoire* à 2000 atomes pour une température de $T = 1000\,K$. Le descripteur utilisé est l'AFS(20,10) avec un rayon de coupure $r_{cut} = 5$ Å. On constate que le modèle possède une très grande précision (RMSE de $0.01\ k_B$) et ne voit sa variance augmenter que pour des valeurs d'entropie locale de $7.8\ k_B$ par atome.

3.5 Conclusions de chapitre

Dans ce chapitre nous avons montré que le problème de l'entropie vibrationnelle harmonique peut être reformulé comme un problème de termes sources locaux [24]. Cette décomposition locale (exacte dans le cadre de l'approximation harmonique) de l'entropie vibrationnelle de formation grâce au formalisme de *Green* pour la *densité d'état* de *modes normaux* permet de décrire les interactions aussi bien à courtes que à longues distances par une approche locale. L'utilisation de l'espace de représentation des descripteurs atomiques locaux associée à une simple régression linéaire permet d'estimer avec une grande précision l'entropie vibrationnelle des bases de données considérées dans les sous-sections 3.3.2 et 3.3.3. De plus, nous avons vérifié la transférabilité du modèle linéaire en entraînant celui-ci sur une réunion de bassins de basses énergies et en extrapolant sur des bassins disjoints de plus hautes énergies. Les grandeurs statistiques relatives à la précision du modèle restent stables même dans le cas de ce régime de pure extrapolation. Ce modèle linéaire a ensuite été testé sur des structures non cristallines (*clusters de Lennard-Jones*) afin d'éprouver encore sa transférabilité. Les conclusions sont les mêmes que pour les bases de données métalliques : un modèle robuste et transférable. **Notre modèle linéaire a aussi prouvé sa transférabilité pour prédire l'entropie vibrationnelle d'un objet étendu (une boucle de dislocation). L'estimation de cette grandeur est de l'ordre de la dizaines de minute sur un ordinateur de bureau alors que la méthode "standard" de diagonalisation nécessite plusieurs heures sur des milliers de CPU. Enfin, nous avons montré l'importance de la base de données utilisée à travers l'utilisation des *points-q* pour cette même boucle**.

Le modèle linéaire developpé dans l'espace des descripteurs locaux présente une complexité numérique évoluant comme $\mathcal{O}(N)$ avec N le nombre d'atomes dans le système alors que la méthode standard de calcul de l'entropie vibrationnelle nécessite l'échantillonnage et la diagonalisation directe de la matrice dynamique du système, opération dont la complexité évolue comme $\mathcal{O}(N^3)$. Le calcul de l'entropie vibrationnelle est donc généralement réduit à des systèmes de petites tailles dépassant rarement 10^4 atomes. Le modèle linéaire dans l'espace des descripteurs, grâce à sa faible complexité et sa grande transférabilité, pourrait donc être utilisé pour prédire l'entropie vibrationnelle de défauts de grandes tailles telles que les boucles de dislocations, et dont il est impossible de calculer aisément l'entropie vibrationnelle par la méthode standard [14]. L'entropie vibrationnelle est aussi un indicateur important des transitions de phases (3.1.3). Une estimation rapide de l'entropie vibrationnelle couplée avec une méthode d'exploration du paysage énergétique pourrait conduire à la découverte de nouveaux chemins de transition de phases et/ou de nouvelles phases telle que la phase de *lave C15* dans le fer prédite théoriquement par Marinica *et al.* [7].

Visionaire and deepest fake
Dirty gold, the coulours change
Hands are frozen, fell no pain
I just wanna hold the flame.

— Circle With Me, Spiritbox

4

Structuration de l'espace des données : déformations et Théorie de l'État de Transition (TST)

Sommaire

4.1 Possibilité d'extension du modèle de régression d'entropie vibrationnelle à des ordres supérieurs

Dans le chapitre précédent Chap. 3, nous avons montré que l'on pouvait construire un modèle de régression d'entropie vibrationnelle harmonique grâce à la décomposition locale de l'entropie. En utilisant des descripteurs locaux, Chap. 2, et en supposant une relation linéaire entre l'entropie locale de l'atome i - notée S_i - et le descripteur $\underline{D}_i$ nous avons construit un modèle linéaire de régression simple présentant une grande précision et une grande transférabilité. De plus, ce modèle préserve les propriétés d'extensivité de l'entropie.

Dans ce chapitre, nous souhaitons étendre notre modèle de régression, basé sur la localité de l'entropie vibrationnelle harmonique, grâce à une approche d'**ordre plus élevée** dans l'espace des descripteurs. Un modèle d'ordre n dans l'espace des descripteurs peut être formalisé de la façon suivante :

$$S_i = \sum_{1 \leq k \leq n} \left(\underline{W}^{(k)} \cdot^{(k)} \bigotimes^{k} \underline{D}_i \right) + o\left(\left\| \bigotimes^{n+1} \underline{D}_i \right\| \right) \tag{4.1}$$

Où $\underline{W}^{(k)} \in \mathbb{R}^{\mathcal{D}^k}$ est un tenseur de poids d'ordre k, $\cdot^{(k)}$ représente l'opérateur de k contraction défini pour les tenseurs d'ordre k et $\bigotimes^k$ est le k produit tensoriel. Le terme d'ordre $n+1$ est considéré comme étant négligeable. Afin de clarifier cette formulation générale, décrivons le cas $n = 2$. Nous obtenons alors une formulation quadratique simple :

$$S_i = \underline{w} \cdot \underline{D}_i + \underline{\underline{W}} \cdot^{(2)} (\underline{D}_i \otimes \underline{D}_i) + o\left(\|\underline{D}_i \otimes \underline{D}_i \otimes \underline{D}_i\| \right) \tag{4.2}$$

Ici $\underline{w}$ et $\underline{\underline{W}}$ correspondent respectivement à $\underline{W}^{(1)}$ et $\underline{W}^{(2)}$. Cette nouvelle formulation reste compatible avec la décomposition locale de l'entropie vibrationnelle harmonique (i) $S = \sum_i S_i$ et (ii) la proportionnalité entre l'entropie locale et l'environnement local $S_i \sim \underline{D}_i$. Ce modèle quadratique permet d'introduire naturellement un **couplage entre les différentes composantes des descripteurs**. Ce couplage des différentes composantes permet d'enrichir le descripteur et d'obtenir des résultats d'une meilleure précision pour les configurations de la base de données d'entraînement. Néanmoins, comme dans tous les processus non-linéaires, l'erreur va augmenter drastiquement pour des configurations "distantes" (en termes de norme dans l'espace des descripteurs, cf. Sec 2.2.2) de la base de données d'entraînement. Il faut donc trouver un juste équilibre entre la précision voulue, la transférabilité et l'ordre de non-linéarité. Nous allons montrer, dans les sous-sections suivantes : (i) qu'il est possible d'ajuster correctement un **modèle quadratique** grâce à un préconditionnement Sec. 4.1.1, (ii) qu'un tel modèle quadratique permet d'obtenir des résultats plus précis que le simple modèle linéaire notamment dans le cas des déformations Sec. 4.1.2.

4.1.1 Modèle quadratique et pré-conditionnement

L'ajustement du modèle de régression décrit par l'équation (4.2) nécessite quelques précautions. En effet, le nombre de paramètres de ce type de modèle évolue comme $\mathcal{O}(\mathcal{D}^2)$. La dimension de l'espace des paramètres à ajuster est donc $\mathcal{O}(\mathcal{D}^2)$ ce qui correspond à un espace de beaucoup plus grande dimension que dans le cas d'un modèle linéaire ($\mathcal{D}$). Dans un espace de grande dimension, on peut supposer qu'il existe un grand nombre de minima locaux de la fonction de coût. La solution de ce problème d'optimisation va donc dépendre du choix de départ des poids ainsi que de l'algorithme de minimisation choisi. Il sera donc possible d'obtenir un grand nombre de modèles, plus ou moins sensibles et précis en partant de la même base de données. Nous proposons ici une méthode de pré-conditionnement du modèle quadratique. Pour cela, on ajuste dans un premier temps le modèle linéaire solution du problème d'optimisation suivant :

$$\underline{w}^* = \arg\min_{\underline{w}} \left\{ \|\underline{S} - \underline{w} \cdot \underline{\underline{D}}^T\|^2 + \lambda^l \|\underline{w}\|^2 \right\} \tag{4.3}$$

Ici, $\underline{S} \in \mathbb{R}^{1\times M}$ est le vecteur correspondant aux entropies de la base de données choisie. $\underline{\underline{D}} \in \mathbb{R}^{M\times\mathcal{D}}$ est la matrice de design du problème. La ligne k de la matrice est $\sum_i \underline{D}_i^k$, la somme courant sur tous les atomes de la configuration k de la base de données. Enfin λ^l est le paramètre de régularisation de la régression. On va ensuite ajuster la partie quadratique du modèle sur les données $\Delta^l\underline{S} \equiv \underline{S} - \hat{\underline{S}}$, où $\hat{\underline{S}}$ est la prédiction donnée par le modèle linéaire de l'équation (4.3). Ce préconditionnement permet d'assurer deux aspects essentiels : (i) la solution du modèle linéaire est unique et permet de "fixer" une région physique dans l'espace de la *fonction de coût* et (ii) ce pré-conditionnement permet, après un constat numérique, d'assurer que $\Delta^l\underline{S}$ va suivre une distribution normale ce qui va rendre plus robuste la régularisation L_2 de la régression quadratique. L'ajustement quadratique répond alors au problème d'optimisation suivant :

$$\underline{\underline{W}}^* = \arg\min_{\underline{\underline{W}}} \left\{ \|\Delta^l\underline{S} - \underline{\underline{W}} \cdot^{(2)} \underline{D}_{(3)}\|^2 + \lambda^q \|\underline{\underline{W}}\|^2 \right\} \tag{4.4}$$

où $\underline{\underline{W}} \in \mathcal{D}\times\mathcal{D}$ est un tenseur de paramètres ajustables et $\underline{D}_{(3)} \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}\times M}$ est un tenseur d'ordre 3 tel que $D_{ijk} = \sum_n D_n^i(k) D_n^j(k)$ avec $D_n^j(k)$ la $j^{-ième}$ composante du $n^{-ième}$ atome de la configuration k. λ^q est le paramètre de régularisation de la régression. La résolution des problèmes d'optimisation Eq. (4.3) et Eq. (4.4) permet de construire le modèle décrit par l'équation (4.2). Ce modèle d'ordre 2 est appelé modèle Extended Quadratic Machine Learning (que nous appellerons dans la suite : modèle EQML).

Nous allons utiliser le modèle EQML pour les cas "mals décrits" par le modèle linéaire de régression. Dans le chapitre 3, nous avons proposé la base de données *ARTn déformée* et nous avons constaté que le modèle linéaire possèdait une moins bonne précision pour le cas particulier de cette base de données. En effet, les déformations même homogènes et isotropes sont difficiles à quantifier pour les descripteurs à cause des faibles variations d'angles et de distances entre les atomes. L'utilisation du modèle EQML sur cette base de données introduit plus d'informations et permet une meilleure précision du modèle.

4.1.2 Application à la base de données *ARTn déformée*

Dans cette sous-section, nous allons comparer la précision et la transférabilité du modèle linéaire et du modèle EQML pour le cas de la base de données *ARTn déformée*. Nous rappelons que cette base est constituée de systèmes de fer cubique centré de volume $(8a_0)^3$ contenant des défauts ponctuels. Ces systèmes ont ensuite subi une déformation homogène et isotrope par application d'un tenseur de déformation $\boldsymbol{\epsilon}$. L'entropie vibrationnelle (harmonique) de formation de défauts a été calculée avec le package `PHONDY` [170, 182-184] couplé avec `LAMMPS` [185] en utilisant le potentiel MEAM développé par Alireza et Asadi [186] (les détails concernant les simulations numériques sont donnés dans la sous-section 3.3.2). Afin de pouvoir utiliser le modèle EQML pour cette base de données, nous devons choisir un descripteur "riche" et comportant un nombre relativement faible de composantes afin d'éviter des phénomènes de *sur-ajustement* du terme quadratique. Notre choix s'est porté sur le bi-spectrum SO(4) avec $j_{max} = 4.0$ et $r_{cut} = 5$Å . Les résultats obtenus pour la base de données *ARTn déformée* sont donnés dans la figure 4.1. Les résultats des procédures de *train/test* sont présentés en encart de ces figures. La figure 4.1.(a) présente les résultats de la régression avec le modèle linéaire simple tandis que la figure 4.1.(b) présente ceux du modèle EQML. Grâce aux courbes de *training/testing* (décrites dans le chapitre 3), présentées en encart des figures 4.1, on constate que les deux modèles présentés possèdent une très bonne transférabilité même pour une grande proportion de données dans l'ensemble de *test*. Il est important de noter que l'erreur du modèle EQML n'augmente pas drastiquement avec la proportion de données dans l'ensemble de *test* ce qui montre qu'il n'y a pas d'effet de *sur-ajustement*. De plus, l'erreur quadratique moyenne du modèle EQML est deux fois inférieure à celle du modèle linéaire ($0.34k_B$ contre $0.82k_B$). Pour un même calcul de descripteur, le modèle EQML permet donc de décrire de façon plus **complète** le cas difficile des déformations que par le modèle linéaire. Le point *technique* de l'utilisation d'un modèle quadratique réside dans son ajustement. Nous contrôlons cet ajustement par le préconditionnement linéaire, la régularisation et par l'utilisation d'une base de données riche et représentative de notre problème. Notre approche initiale, basée sur l'utilisation d'un modèle linéaire de régression dans l'espace des descripteurs, avait été motivée par des arguments "physiques" portant sur la décomposition locale de l'entropie vibrationnelle harmonique ($S = \sum_{i=1}^{N} S_i$ et $S_i \sim \underline{D}_i$). Le modèle EQML développé vérifie toujours ces propriétés.

Dans les sections suivantes, nous développons une **approche corrective** permettant de calculer la **variation d'entropie vibrationnelle harmonique** entre une configuration initiale et une configuration déformée (issue de la configuration initiale) par application d'un tenseur de déformations $\boldsymbol{\epsilon}$. Pour cela, nous avons développé une approche analytique mettant en évidence la nécessité d'introduire des couplages entre les *différents modes normaux* du système. Ce terme de correction analytique conduit à l'utilisation d'un modèle **purement quadratique dans l'espace des descripteurs** - il n'y a pas d'utilisation d'un préconditionnement linéaire - afin d'ajuster la variation d'entropie vibrationnelle harmonique. Nous comparons l'efficacité de cette méthode corrective avec le modèle EQML et nous développons le lien entre la **structure de l'espace des phases et celle de l'espace des descripteurs**.

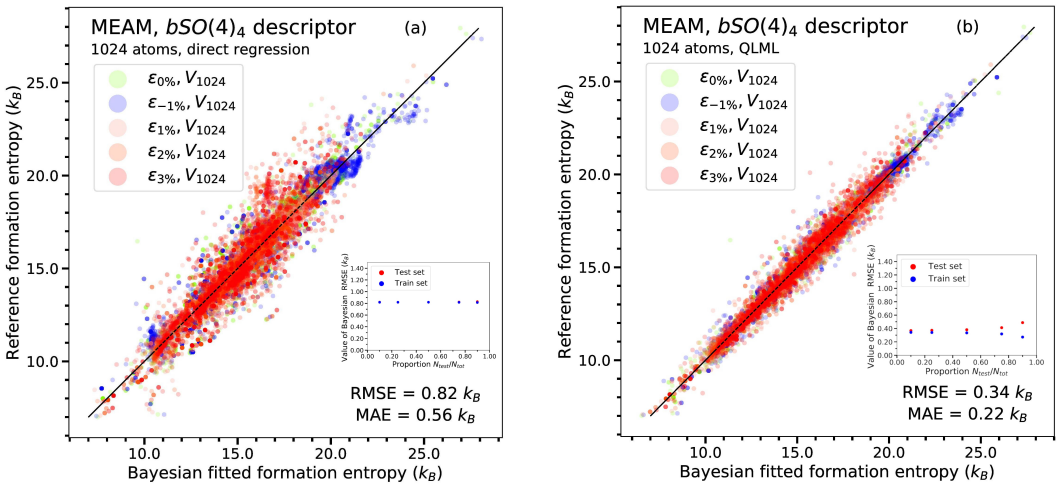

Figure 4.1: Comparaison des résultats du modèle linéaire et du modèle EQML pour la base de données *déformée* pour les I_{2-4}/V_4 amas de défauts (pour le potentiel MEAM) en utilisant le descripteur $bSO(4)_4$. Les systèmes initiaux possèdaient un volume de $(8a_0)^3$ et ont été déformés par application d'une déformation homogène et isotrope. Le taux de déformation varie de -1% à 3%. La figure (a) illustre les résultats du modèle linéaire de régression en fonction du type de défauts ponctuels ; la figure (b) illustre les résultats obtenus avec le modèle EQML. En encart sont présentés les résultats de la procédure de *training/testing* pour les deux modèles. On constate que les deux modèles présentent une bonne transférabilité et que le modèle EQML possède une meilleure précision au sens de la RMSE.

4.2 Structure de l'espace des phases et de l'espace des descripteurs

L'entropie vibrationnelle (même dans le cadre harmonique) a un lien étroit avec la courbure de l'espace des phases $\mathcal{Q} \times \mathcal{P} \in \mathbb{R}^{3N \times 3N}$ (où N est le nombre d'atomes dans le système). Une esquisse intuitive de ce lien sera développée dans les sections suivantes. Cette caractéristique "structurale" de l'entropie vibrationnelle la rend difficile à calculer et à appréhender. L'énergie d'une configuration dans son état fondamental est une caractéristique ponctuelle d'un bassin : elle peut être résumée à l'aide d'un seul vecteur de coordonnées et est définie par une seule quantité scalaire non ambigüe. L'entropie vibrationnelle est une grandeur plus subtile. Elle est diffuse et ne peut aucunement être déduite à l'aide d'un seul vecteur de coordonnées (au moins dans l'espace des phases). **Elle est représentative de l'état de courbure général d'un bassin**. Ainsi, plus un bassin est "plat", plus son entropie vibrationnelle est grande.

Nous proposons, dans ce chapitre, de nous intéresser au lien qui existe entre la structure de l'espace des phases et la structure de l'espace de représentation. Dans la sous-section (4.2.1), nous allons établir un lien direct entre ces deux espaces - dans le cas des déformations - en nous basant sur le formalisme de *Green* pour la *densité d'état* de *modes normaux* décrits dans le chapitre 3. Nous proposerons un terme de correction analytique permettant d'améliorer la précision du modèle de régression de l'entropie de formation. Ce modèle correctif sera appliqué à la base de données *déformée* Sec. 3.3.3

pour le fer cubique centré. Dans un second temps, nous appliquerons une dérivation directe du modèle linéaire developpé dans le chapitre 3 afin d'ajuster les fréquences d'attaque (dans le cadre harmonique). La fréquence d'attaque est une image directe de la variation de courbure entre un état initial et un état final dans l'espace des phases. Ce modèle sera testé dans un système non cristallin, le silicium amorphe, qui présente un paysage énergétique riche et complexe.

Le chapitre précédent, Chap. 3, montre qu'un modèle linéaire dans l'espace des descripteurs, basé sur le formalisme de *Green* pour la *densité d'état* de *modes normaux*, permet d'ajuster l'entropie vibrationnelle de différents systèmes. D'autre part, d'après la définition de l'entropie microcanonique donnée par l'équation (3.1), celle-ci peut toujours être formulée comme une grandeur proportionnelle au logarithme d'un certain volume. Dans le cas de l'entropie configurationnelle calculée à une énergie E (le cas le plus intuitif) ce volume est simplement le nombre de configurations $\mathcal{N}_c$ possibles (renormalisé par δE) du système dont l'énergie est comprise entre $E - \frac{1}{2}\delta E$ et $E + \frac{1}{2}\delta E$. Dans le cas des modes de vibration, le raisonnement n'est pas aussi intuitif. Nous allons le détailler de façon simplifiée en considérant seulement le régime haute température ($T > T_{Debye}$ où tous les modes de vibration sont activés et peuvent être considérés dans la limite classique) et dans le cadre de l'approximation harmonique.

Nous cherchons à exprimer l'entropie vibrationnelle - d'un bassin centré aux coordonnées $\boldsymbol{q}_0$ - sous la formulation suivante :

$$S_{vib} = k_B \ln\left(\mathcal{V}(\boldsymbol{q}_0)/v_u\right) \tag{4.5}$$

Où $\mathcal{V}$ est un certain volume de l'espace des phases et v_u un volume unitaire de l'espace des phases. Intuitivement $\mathcal{V}(\boldsymbol{q}_0)$ est le volume de l'espace des phases accessible par les *modes normaux* du système à une température donnée T et v_u est simplement l'action minimale imposée par le *principe d'incertitude* h^{3N}. Pour plus de clarté, nous nous plaçons dans l'espace des *modes propres* et des quantités de mouvements associée à l'équation (3.20) : $(\hat{\mathbf{e}}_{\nu_1}, \cdots, \hat{\mathbf{e}}_{\nu_{3N}}, \hat{\mathbf{p}}_{\nu_1}, \cdots, \hat{\mathbf{p}}_{\nu_{3N}})$. Nous voulons alors calculer une expression simple de $\mathcal{V}(\hat{\mathbf{e}}_{\nu_1}, \cdots, \hat{\mathbf{e}}_{\nu_{3N}}, \hat{\mathbf{p}}_{\nu_1}, \cdots, \hat{\mathbf{p}}_{\nu_{3N}})$ dont la définition formelle est :

$$\mathcal{V}(\hat{\mathbf{e}}_{\nu_1}, \cdots, \hat{\mathbf{e}}_{\nu_{3N}}, \hat{\mathbf{p}}_{\nu_1}, \cdots, \hat{\mathbf{p}}_{\nu_{3N}}) = \int_{\mathcal{Q}\times\mathcal{P}} \sqrt{|\det(\boldsymbol{g}_{\mathcal{H}})|}\, d\hat{\mathbf{e}}_{\nu_1} \wedge \cdots \wedge d\hat{\mathbf{e}}_{\nu_{3N}} \wedge d\hat{\mathbf{p}}_{\nu_1} \wedge \cdots \wedge d\hat{\mathbf{p}}_{\nu_{3N}}$$

$$\boldsymbol{g}_{\mathcal{H}} = \left(\begin{array}{ccc|ccc} & \vdots & & \frac{\partial \mathcal{H}}{\partial \hat{p}_{\nu_1}}\left(\frac{d\hat{p}_{\nu_1}}{dt}\right)^{-1} & & \\ \cdots & 0 & \cdots & & \ddots & \\ & \vdots & & & & \frac{\partial \mathcal{H}}{\partial \hat{p}_{\nu_{3N}}}\left(\frac{d\hat{p}_{\nu_{3N}}}{dt}\right)^{-1} \\ \hline -\frac{\partial \mathcal{H}}{\partial \hat{e}_{\nu_1}}\left(\frac{d\hat{e}_{\nu_1}}{dt}\right)^{-1} & & & & \vdots & \\ & \ddots & & \cdots & 0 & \cdots \\ & & -\frac{\partial \mathcal{H}}{\partial \hat{e}_{\nu_{3N}}}\left(\frac{d\hat{e}_{\nu_{3N}}}{dt}\right)^{-1} & & \vdots & \end{array}\right)$$

Où $\det(\cdot)$ est l'application déterminant, $\boldsymbol{g}_{\mathcal{H}}$ est le tenseur métrique associé à la *dynamique Hamiltonnienne* du système et $\wedge$ est le produit extérieur. Le calcul du volume

de la famille des modes propres du système est difficile. Plaçons-nous dans la base des modes *normaux* $\{(\mathbb{U}_\nu, \mathbb{V}_\nu)\}_{1\leq\nu\leq 3N}$ et dont l'énergie est décrite par l'équation (3.16). Dans cette nouvelle base, la famille des modes propres se représente comme une matrice diagonale ce qui permet d'exprimer simplement $\mathcal{V}(\boldsymbol{q}, \boldsymbol{p}) = \prod_k \mathcal{V}(\mathbb{U}_k, \mathbb{V}_k)$. Il reste alors à déterminer l'expression de $\mathcal{V}(\mathbb{U}_k, \mathbb{V}_k)$. Pour cela, plaçons-nous dans le plan $(\mathbb{U}_k, \mathbb{V}_k)$ et calculons le volume engendré. L'équation (3.16) permet d'aboutir à une simple ellipse pour le mode k dont l'énergie associée est E_k. Le cas de ce mode est décrit par la figure 4.2 :

$$\frac{1}{2}\omega_k^2\mathbb{U}_k^2 + \frac{1}{2}\mathbb{V}_k^2 = E_k \tag{4.6}$$

Le volume $\mathcal{V}(\mathbb{U}_k, \mathbb{V}_k)$ est la surface de l'ellipse $\frac{2\pi E_k}{\omega_k}$. Il nous suffit alors d'appliquer le *théorème d'équipartition de l'énergie* et comme $T > T_{Debye}$ tous les modes de vibration sont activés. On peut donc écrire que $\forall k,\ E_k = k_B T$. On peut maintenant injecter les résultats précédents dans l'équation (4.5) et obtenir la même expression que l'équation (3.19) :

$$S_{vib} = k_B \sum_{k=1}^{3N} \ln\left(\frac{k_B T}{\hbar\omega_k}\right) \tag{4.7}$$

Ce simple calcul nous permet d'effleurer le lien étroit qui existe entre un volume de l'espace des phases et l'entropie vibrationnelle. Dans la prochaine partie, nous allons montrer qu'il est possible d'établir un lien entre l'espace des descripteurs et l'espace des phases grâce au formalisme de *Green* pour la *densité d'état* de *modes normaux*. Par souci de clarté, nous avons choisi de ne présenter que les principaux résultats de notre démarche. Toutes les démonstrations relatives à cette sous-section sont présentées en annexe (B). La principale difficulté de cette démarche est d'établir un lien formel entre l'espace des phases et l'espace des descripteurs. Nous verrons que ce lien peut être obtenu par l'intermédiaire de la *matrice de covariance des descripteurs*.

4.2.1 Extension du formalisme de *Green* : déformations et structure de l'espace des descripteurs

Dans cette sous-section, nous notons $\mathcal{C}^{(0)}$ une configuration initiale de coordonnées $\boldsymbol{q}_0 \in \mathbb{R}^{3N}$. Nous appliquons un tenseur de déformation $\boldsymbol{\epsilon} \in \mathbb{R}^{3\times 3}$ à la configuration $\mathcal{C}^{(0)}$. Celle-ci subit alors la transformation de coordonnées suivante : $\boldsymbol{q}_0 \to \boldsymbol{q}_0 + \delta\boldsymbol{q}$, avec $\delta\boldsymbol{q} = \boldsymbol{\epsilon}\cdot\boldsymbol{q}_0$. Les nouvelles coordonnées du système sont alors notées $\boldsymbol{q}_\epsilon$ et la configuration est notée $\mathcal{C}^{(\epsilon)}$. Nous donnons aussi un exemple typique de tenseur de déformation :

$$\boldsymbol{\epsilon}(\delta) = \begin{pmatrix} \delta & 0 & 0 \\ 0 & \delta & 0 \\ 0 & 0 & \delta \end{pmatrix} \tag{4.8}$$

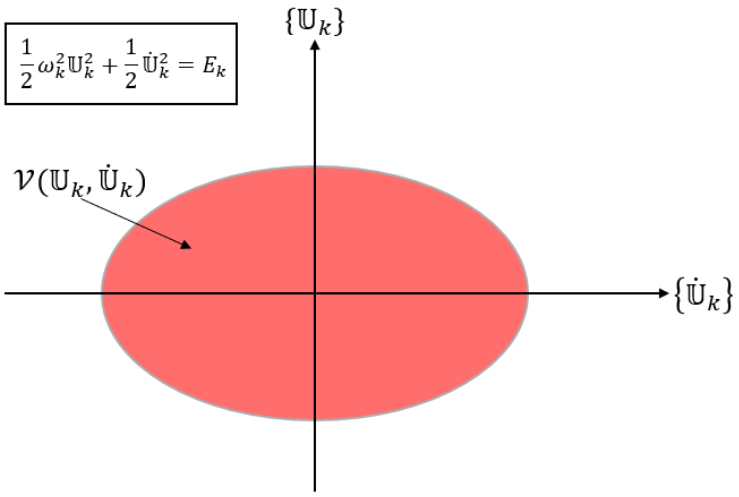

Figure 4.2: Projection du *mode normal* k d'énergie E_k. Le portrait de phase associé est une simple ellipse dont la surface est aisément calculable. L'équation de l'ellipse pour le *mode normal* k est donnée en encart.

On peut directement relier le tenseur des déformations et la variation de volume du système ΔV entre la configuration $\mathcal{C}^{(0)}$ et $\mathcal{C}^{(\epsilon)}$. Ce changement de volume s'exprime de la façon suivante :

$$\Delta V = \det\{\boldsymbol{\epsilon}\} V^{(0)} \tag{4.9}$$

Ici, $\det(\cdot)$ est l'opérateur déterminant et $V^{(0)}$ est le volume de la configuration $\mathcal{C}^{(0)}$.

Nous allons ici calculer un terme analytique de correction entre une configuration initiale $\mathcal{C}^{(0)}$ et l'image de cette même configuration après avoir subi une petite déformation $\boldsymbol{\epsilon}$ noté $\mathcal{C}^{(\epsilon)}$. Dans cette situation, on comprend que le tenseur $\boldsymbol{\epsilon}$ va induire une variation ΔV du volume du système et donc une variation $\Delta\omega_\nu$ des pulsations propres du système. Nous cherchons à quantifier la variation de l'entropie ΔS par rapport aux variations précédentes. Pour cela, on utilise la *densité d'état* de *modes normaux*. L'équation (3.26) permet d'exprimer facilement cette variation :

$$\Delta S(\boldsymbol{\epsilon}) = -k_B \int_0^{+\infty} \left[\ln\left(\frac{\hbar\omega}{k_B T}\right) - 1 \right] \Delta\Omega(\omega, \boldsymbol{\epsilon}) d\omega \tag{4.10}$$

Nous devons maintenant déterminer une expression analytique de la variation de *densité d'état* de mode de vibration $\Delta\Omega(\omega, \boldsymbol{\epsilon})$. Pour cela, nous allons appliquer une approche pertubative car la déformation $\boldsymbol{\epsilon}$ est petite (au sens de la norme $\|\cdot\|_\infty$ décrite dans l'annexe B.1). Cette approche pertubative est présentée de façon schématique dans la figure 4.3. Nous introduisons le tenseur $\boldsymbol{\delta}$ dont la définition formelle est donnée par l'équation (B.11). Ce tenseur quantifie la différence entre les fonctions de *Green* solutions du problème aux valeurs propres des *modes normaux* Eq. (3.20) pour les configurations $\mathcal{C}^{(\epsilon)}$ et $\mathcal{C}^{(0)}$. Le tenseur $\boldsymbol{\delta}$ est petit au sens de la norme $\|\cdot\|_\infty$ et agit sur les *modes propres* du système. L'effet de $\boldsymbol{\delta}$ est schématisé dans la figure 4.3, où

le tenseur va induire un changement de volume de l'ellipse engendrée par le couple $(\mathbb{U}_k, \mathbb{V}_k)$. En s'appliquant sur l'ensemble des *modes propres* du système, $\boldsymbol{\delta}$ va induire un changement du volume $\mathcal{V}(\mathbb{U}_k, \mathbb{V}_k)$ qu'il nous faut quantifier.

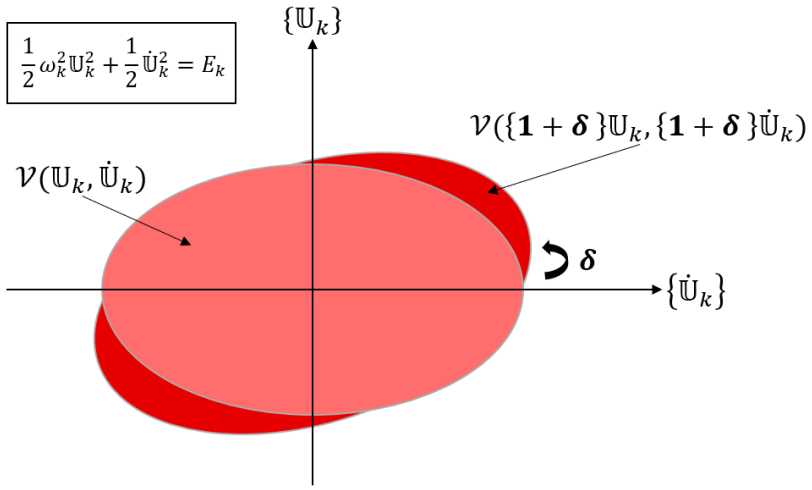

Figure 4.3: Projection du *mode normal* k. Le tenseur $\boldsymbol{\delta}$ déforme le portrait de phase du mode k ce qui implique une variation de volume qu'il nous faut quantifier pour calculer le terme de correction d'entropie vibrationnelle.

Considérons la fonction de *Green* $\boldsymbol{\mathcal{G}}^{(0)}(\omega)$ solution de l'équation (3.20) pour la configuration $\mathcal{C}^{(0)}$ et la fonction de *Green* $\boldsymbol{\mathcal{G}}(\omega, \boldsymbol{\epsilon})$ solution du problème des *modes normaux* pour la configuration $\mathcal{C}^{(\epsilon)}$. On formalise l'approche perturbative sous la forme suivante :

$$\boldsymbol{\mathcal{G}}(\omega, \boldsymbol{\epsilon}) = \boldsymbol{\mathcal{G}}^{(0)}(\omega) + \boldsymbol{\mathcal{G}}^{(1)}(\omega, \boldsymbol{\epsilon}) + \cdots + \boldsymbol{\mathcal{G}}^{(n)}(\omega, \boldsymbol{\epsilon}) \tag{4.11}$$

où $\boldsymbol{\mathcal{G}}^{(k)}(\omega, \boldsymbol{\epsilon})$ est la fonction de *Green* perturbative à l'ordre k. Les fonctions perturbatives vérifient les relations suivantes : $\forall k > 0$, $\|\boldsymbol{\mathcal{G}}^{(k+1)}(\omega, \boldsymbol{\epsilon})\|_\infty \ll \|\boldsymbol{\mathcal{G}}^{(k)}(\omega, \boldsymbol{\epsilon})\|_\infty$. Nous allons ici travailler sur un modèle perturbatif à l'ordre 1. Afin de mener les calculs présentés en annexe B, nous utilisons aussi les propriétés spécifiques des fonctions de *Green* associées au problème des *modes normaux* Eq. (3.20) : (i) $\boldsymbol{\mathcal{G}}^{(0)}(\omega) \cdot \boldsymbol{\mathcal{G}}^{(0)}(\omega) = \partial_{\omega^2}\boldsymbol{\mathcal{G}}^{(0)}(\omega)$; (ii) $\boldsymbol{\mathcal{G}}^{(0)}(\omega) \cdot \boldsymbol{\mathcal{G}}^{-1}(\omega, \boldsymbol{\epsilon}) \simeq \mathbf{1}$ où $\mathbf{1}$ est le tenseur identité ; (iii) $\|\Delta\boldsymbol{\mathcal{G}}(\omega, \boldsymbol{\epsilon}) \cdot \boldsymbol{\mathcal{G}}^{-1}(\omega, \boldsymbol{\epsilon})\|_\infty \ll 1$ où $\|\cdot\|_\infty$ est la norme décrite en annexe B.1. Nous sommes maintenant en mesure de donner une expression analytique de la variation de la *densité d'état* des modes de vibration décrite dans l'équation (4.10) en utilisant la méthode pertubative à l'ordre 1. Le calcul détaillé de ce terme est décrit en annexe B.3 et est grandement basé sur les développements de P. H. Dederichs *et al.* [24] pour les calculs de correction en température. On aboutit finalement à l'expression suivante :

$$\Delta\Omega(\omega, \boldsymbol{\epsilon}) = \frac{1}{\pi}\Im\left(\partial_\omega \left\{ \mathrm{Tr}\left(\frac{\Delta\boldsymbol{\mathcal{G}}(\omega, \boldsymbol{\epsilon})}{\boldsymbol{\mathcal{G}}(\omega, \boldsymbol{\epsilon})}\right)\right\}\right) \tag{4.12}$$

où $\Im(\cdot)$ représente la partie imaginaire et $\Delta\boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon}) = \boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon}) - \boldsymbol{\mathcal{G}}^{(0)}(\omega)$ est la variation de fonction de *Green* entre $\mathcal{C}^{(\epsilon)}$ et $\mathcal{C}^{(0)}$. La notation sous forme de fraction est abusive et représente l'inverse du tenseur $\boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon})$. On constate que l'expression de la variation de *densité d'état* des modes de vibration est simplement reliée aux parties imaginaires des fonctions de *Green* $\Delta\boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon})$ et $\boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon})$. Grâce à l'expression de $\Delta\Omega(\omega,\boldsymbol{\epsilon})$, on peut maintenant calculer ΔS par intégration sur ω et passage au logarithme. Les subtilités du calcul sont détaillées en annexe B.3. On obtient l'expression de $\Delta S(\boldsymbol{\epsilon})$:

$$\Delta S(\boldsymbol{\epsilon}) = -k_B \left\{ \sum_{i,\alpha=1}^{N,3} \left(\sum_{1\leq\nu^\epsilon\neq\nu^0\leq 3N} \frac{\omega_{\nu^\epsilon}^{-2}|\xi_\epsilon^{i\alpha}|^2 - \omega_{\nu^0}^{-2}|\xi_0^{i\alpha}|^2}{\sqrt{\sum_{\nu^0=0}^{3N} f(\omega_{\nu^0},\omega_{\nu^\epsilon})|\xi_0^{i\alpha}|^4}} \right) \right\} \tag{4.13}$$

Les notions utilisées sont les mêmes que dans le chapitre 3 et la fonction f est définie plus rigoureusement dans l'annexe B.3. La variation d'entropie $\Delta S(\boldsymbol{\epsilon})$ peut donc s'écrire comme une différence de quantités entre les modes de vibration de la configuration déformée ν^ϵ et les modes de vibration de la configuration non-déformée ν^0. Cette différence implique un couplage direct entre les modes des configurations $\mathcal{C}^{(\epsilon)}$ et $\mathcal{C}^{(0)}$. Notre modèle linéaire ne peut pas directement rendre compte de ce couplage. Afin de relier la quantité décrite par l'équation (4.13) à l'espace des descripteurs, nous voulons approximer le produit scalaire suivant :

$$\sum_{\nu=1}^{3N}\sum_{\alpha'=1}^{3} \xi^{i\alpha'}(\nu')\,\hat{\boldsymbol{e}}_{i\alpha'}^T \cdot \sum_{\nu=1}^{3N}\sum_{\alpha=1}^{3} \xi^{i\alpha}(\nu)\,\hat{\mathbf{e}}_{i\alpha} = \sum_{\nu=1}^{3N}\sum_{\alpha=1}^{3} |\xi^{i\alpha}(\nu)|^2 \tag{4.14}$$

Dans le chapitre 3, nous avons introduit une relation linéaire entre la base des *modes normaux* $\{\hat{\mathbf{e}}_{i\alpha}\}_{1\leq i\alpha\leq 3N}$ dans l'espace des phases et la base des descripteurs $\{D_d\}_{1\leq d\leq\mathcal{D}}$ dans l'espace de représentation :

$$\sum_{\nu=1}^{3N}\sum_{\alpha=1}^{3} |\xi^{i\alpha}(\nu)|^2 = \underline{w}\cdot\underline{D}^i \tag{4.15}$$

L'expression (4.13) introduit naturellement un couplage entre différents modes de vibration en raison de l'expression de la fonction f présente au dénominateur de celle-ci. Nous proposons ici d'introduire un couplage direct des différentes composantes des descripteurs grâce au *noyau* suivant $K(\underline{D}^i,\underline{D}^{i'})$:

$$K(\underline{D}^i,\underline{D}^{i'}) = \mathrm{Tr}\left\{ \underline{\underline{W}}^{\frac{1}{2}} \cdot \underline{D}^i \cdot \left[\underline{D}^{i'}\right]^T \cdot \left[\underline{\underline{W}}^{\frac{1}{2}}\right]^T \right\} \tag{4.16}$$

Ici, $\underline{\underline{W}} \in \mathbb{C}^{\mathcal{D}\times\mathcal{D}}$ est une matrice de paramètres ajustables, approximant l'espace des phases à partir de l'espace des descripteurs, $\underline{D}^i \in \mathbb{R}^{1\times\mathcal{D}}$ est le vecteur de descripteurs associé à l'atome i. On note que l'expression (4.16) est formellement un *noyau* si la matrice $\underline{\underline{W}}$ est définie positive (d'après la définition du chapitre 2 et les propriétés de l'opérateur trace). Finalement, nous faisons l'approximation suivante :

$$\sum_{\nu=1}^{3N}\sum_{\alpha=1}^{3} \frac{|\xi^{i\alpha}(\nu)|^2}{\sqrt{\sum_{\nu^0=1}^{3N} f(\omega_{\nu^0},\omega_\nu)|\xi^{i\alpha}(\nu^0)|^4}} \propto K(\underline{D}^i,\underline{D}^i) \tag{4.17}$$

On peut noter que le *noyau* $K(\underline{D}^i, \underline{D}^i)$ se réduit au carré d'un simple produit scalaire $|\underline{w} \cdot \underline{D}^i|^2$ si $\underline{\underline{W}} = \text{diag}(\underline{w})^{\frac{1}{2}}$. De plus, ce *noyau* conserve l'extensivité de l'entropie vibrationnelle. Des cas particuliers du lien entre ce *noyau* et un produit scalaire simple sont discutés dans l'annexe B.5. Finalement, en utilisant les propriétés de l'opérateur trace, nous pouvons ajuster $\Delta S(\boldsymbol{\epsilon})$ à l'aide d'une quantité scalaire :

$$\Delta S^{vib}(\boldsymbol{\epsilon}) = -k_B \,\text{Tr}\left\{\underline{\underline{W}}^{\frac{1}{2}}(\boldsymbol{\epsilon}) \cdot \left[\underline{\underline{D}}^T(\boldsymbol{\epsilon}) \cdot \underline{\underline{D}}(\boldsymbol{\epsilon})\right] - \underline{\underline{W}}^{\frac{1}{2}}(\mathbf{0}) \cdot \left[\underline{\underline{D}}^T(\mathbf{0}) \cdot \underline{\underline{D}}(\mathbf{0})\right]\right\} \tag{4.18}$$

où $\left[\underline{\underline{D}}^T(\mathbf{0}) \cdot \underline{\underline{D}}(\mathbf{0})\right] \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ et $\left[\underline{\underline{D}}^T(\boldsymbol{\epsilon}) \cdot \underline{\underline{D}}(\boldsymbol{\epsilon})\right] \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ sont liées aux matrices de covariance des configurations initiale et déformée. $\underline{\underline{W}}^{\frac{1}{2}}(\boldsymbol{\epsilon}) \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ et $\underline{\underline{W}}^{\frac{1}{2}}(\mathbf{0}) \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ sont des matrices de paramètres ajustables. Pour plus de clarté, dans la suite, nous introduisons le tenseur suivant :

$$\underline{\underline{\mathcal{M}}} \equiv \left[\underline{\underline{D}}^T(\boldsymbol{\epsilon}) \cdot \underline{\underline{D}}(\boldsymbol{\epsilon})\right] - \left[\underline{\underline{D}}^T(\mathbf{0}) \cdot \underline{\underline{D}}(\mathbf{0})\right] \tag{4.19}$$

Sous l'hypothèse que $\underline{\underline{W}}^{\frac{1}{2}}(\boldsymbol{\epsilon}) \simeq \underline{\underline{W}}^{\frac{1}{2}}(\mathbf{0})$ et en utilisant les propriétés de la trace, on peut alors écrire simplement l'équation (4.13) sous la forme :

$$\Delta S(\boldsymbol{\epsilon}) = \text{Tr}\left\{\underline{\underline{\mathcal{W}}}_{\mathcal{D}} \cdot \underline{\underline{\mathcal{M}}}\right\} \tag{4.20}$$

Les composantes du tenseur $\underline{\underline{\mathcal{W}}}_{\mathcal{D}} \in \mathbb{C}^{\mathcal{D}\times\mathcal{D}}$ sont les paramètres ajustables qui doivent être optimisés lors de l'ajustement du modèle. Le tenseur $\underline{\underline{\mathcal{M}}}$ étant symétrique, le tenseur $\underline{\underline{\mathcal{W}}}$ est aussi symétrique. Le tenseur $\underline{\underline{\mathcal{W}}}_{\mathcal{D}}$ permet de coupler les différentes composantes de descripteurs ce qui n'est pas le cas du modèle linéaire simple proposé dans le chapitre précédent 3. Le modèle de régression pour la correction $\Delta S(\boldsymbol{\epsilon})$ contiendra "plus d'informations" au sens des descripteurs et permettra une prédiction plus précise qu'un modèle linéaire dans le cas des petites déformations. Cependant, il sera potentiellement moins transférable à cause de sa non-linéarité. De plus, le tenseur $\underline{\underline{\mathcal{W}}}_{\mathcal{D}}$, peut être interprété de façon plus profonde. En effet, ce tenseur permet d'effectuer la transformation permettant de passer de l'espace des descripteurs à l'espace des *modes normaux*, et fait donc le lien entre l'espace "physique" des fonctions propres des *modes normaux* et l'espace de représentation.

4.2.2 Modèle de régression du terme de correction : application à la base de données *déformée*

Nous allons maintenant construire un modèle de régression du terme de correction ΔS. Pour cela, nous utilisons la base de données *déformée* présentée dans le chapitre précédent Sec. 3.3.3. Nous rappelons que cette base de données est constituée de systèmes de fer cubique centré de volume $(8a_0)^3$ contenant des défauts ponctuels. Ces systèmes ont ensuite subi une déformation homogène et isotrope par application d'un tenseur de déformation $\boldsymbol{\epsilon}$. L'entropie vibrationnelle (harmonique) de formation de défauts a été calculée en utilisant le package `PHONDY` [170, 182-184] couplé avec `LAMMPS` [185] en utilisant le potentiel MEAM développé par Alireza et Asadi [186] (les détails concernant les simulations numériques sont donnés dans la sous-section 3.3.2).

On veut contruire un modèle de la correction d'entropie entre une configuration non déformée $\epsilon = 0$ et une configuration déformée $\epsilon \neq 0$ en se basant sur les descripteurs de ces deux configurations. En nous appuyant sur l'équation (4.20), nous proposons un modèle de la forme :

$$S_f(\boldsymbol{\epsilon}) = w_0 S_f(\mathbf{0}) + \mathrm{Tr}\left\{\underline{\underline{\mathcal{W}}}_{\mathcal{D}} \cdot \underline{\underline{\mathcal{M}}}\right\}. \tag{4.21}$$

Ici, $S_f(\boldsymbol{\epsilon})$ et $S_f(\mathbf{0})$ représentent respectivement les entropies de formation des configurations dans l'état déformé et dans l'état non-déformé. Le tenseur $\underline{\underline{\mathcal{M}}}$ est défini par l'équation (4.19). Nous considèrons que la distribution des paramètres ajustables de $\underline{\underline{\mathcal{W}}}_{\mathcal{D}}$ doit être indépendante des deux configurations d'entrée. Nous introduisons aussi un paramètre de contrôle w_0. Ce paramètre est exactement égal à 1 dans le cas théorique. Ce degré de liberté supplémentaire permet de vérifier que le modèle n'entre pas dans un régime de *sur-ajustement*. Si le modèle entre en régime de *sur-ajustement*, ce paramètre sera significativement différent de 1. Nous veillerons donc à vérifier que $w_0 \to 1$.

Afin d'éprouver ce nouveau modèle, nous utilisons la base de données *ARTn* déformée décrite dans la section 3.3.3. Nous allons comparer deux modèles de régression pour estimer la quantité $S_f(\boldsymbol{\epsilon})$: (i) le modèle linéaire décrit par l'équation (3.32) et (ii) le modèle corrigé décrit par l'équation (4.21). Le modèle linéaire permet de faire une estimation directe de $S_f(\boldsymbol{\epsilon})$ en utilisant les descripteurs de la configuration $\mathcal{C}^{(\epsilon)}$. Le modèle corrigé nécessite la valeur de $S_f(\mathbf{0})$ associée à la configuration $\mathcal{C}^{(0)}$ ainsi que le calcul de la matrice $\underline{\underline{\mathcal{M}}}$ à partir des descripteurs des configurations $\mathcal{C}^{(\epsilon)}$ et $\mathcal{C}^{(0)}$. Dans le cadre du modèle corrigé, on peut exprimer le problème de régression de l'équation (4.21) sous forme d'une régression linéaire grâce aux propriétés de la trace. Nous utilisons même le descripteur $bSO(4)_4$ avec un rayon de coupure $r_{cut} = 5$ Å pour (i) le modèle linéaire et (ii) le modèle corrigé. Les régressions linéaires ont été ajustées par la méthode Bayésienne régularisée grâce au package `scikit-learn` [165]. Les indicateurs statistiques ainsi que la procédure de *training/testing* sont détaillés dans la sous-section 3.4.2. La figure 4.4(a) présente les résultats du modèle linéaire direct entre l'espace des descripteurs et l'entropie de formation introduite dans le chapitre (3). Les résultats de la procédure de *training/testing* associés sont présentés en encart de la figure 4.4.(a). De même, la figure 4.4.(b) présente les résultats obtenus pour le modèle de régression décrit par l'équation (4.21), et les résultats de la procédure de *training/testing* sont représentés en encart.

On constate que les deux modèles, que ce soit le modèle direct ou le modèle corrigé présentent de bons indicateurs statistiques en termes de MAE et de RMSE. Le modèle corrigé possède une meilleure précision avec un RMSE de 0.31 k_B contre 0.82 k_B pour le modèle direct. En termes de transférabilité, la procédure de *training/testing* montre que les indicateurs statistiques restent stables, même pour des grandes proportions de l'ensemble de vérification (cf. encart de la figure 4.4). Les indicateurs du modèle corrigé

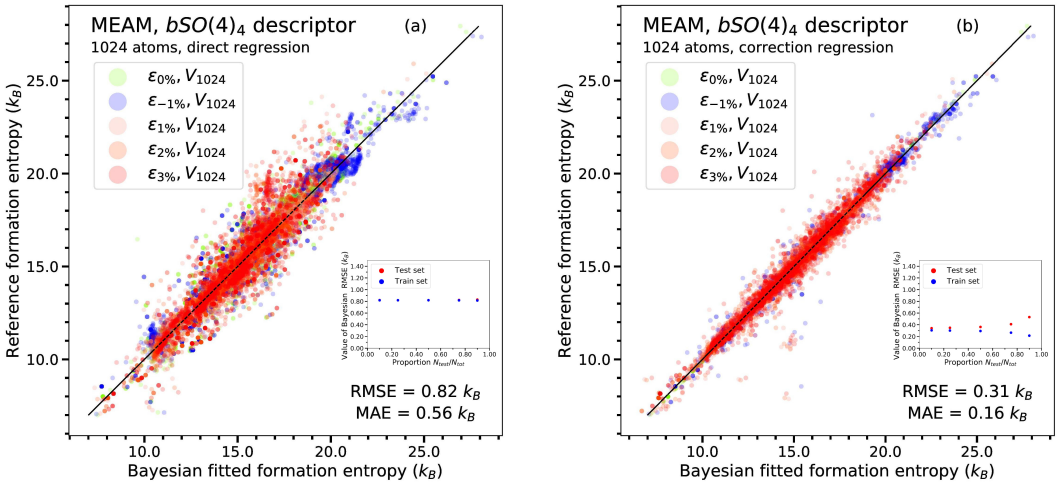

Figure 4.4: Comparaison des résultats du modèle linéaire et du modèle corrigé de la quantité $S_f(\boldsymbol{\epsilon})$ pour la base de données *déformée* pour les I_{2-4}/V_4 amas de défauts (pour le potentiel MEAM) en utilisant le descripteur $bSO(4)_4$. Les systèmes initiaux possèdaient un volume de $(8a_0)^3$ et ont été déformés par application d'une déformation homogène et isotrope. Le taux de déformation varie de -1% à 3%. La figure (a) illustre les résultats du modèle linéaire de régression décrit par l'équation (3.32) en fonction du type de défauts ponctuels ; la figure (b) illustre les résultats obtenus avec le modèle corrigé décrit par l'équation (4.21). Les résultats de la procédure de *training/testing* des deux modèles sont présentés en encart. On constate que les deux modèles présentent une bonne transférabilité et que le modèle corrigé possède une meilleure précision.

sont un peu moins stables (au sens des courbes de *train/test*) que ceux du modèle direct à cause de sa non-linéarité. Néanmoins, les indicateurs statistiques du modèle corrigé sont meilleurs que ceux du modèle linéaire direct quelque soit la proportion de l'ensemble de vérification choisi. Concernant la valeur du paramètre de contrôle w_0 discuté plus haut, on obtient $w_0 = 0.984$ pour le modèle présenté dans la figure 4.4.(b) ce qui montre que le modèle n'est pas en régime de *sur-ajustement*. La comparaison des résultats entre le modèle direct Fig 4.4.(a) et le modèle corrigé Fig 4.4.(b) montre que l'utilisation du modèle corrigé permet d'obtenir de meilleurs modèles de régression. L'approche pertubative mise en oeuvre illustre la nécessité de l'introduction d'un couplage entre les différentes composantes des modes de vibration. Ce couplage dans l'espace des phases nécessite l'introduction d'un couplage des différentes composantes des descripteurs afin de rendre compte de la physique du problème. Ce modèle, basé sur la correction du *formalisme de Green* pour la *densité d'état* de modes de vibration, permet d'obtenir un modèle précis et transférable pour le cas des petites déformations. Enfin, le degré de liberté w_0 lors de l'ajustement montre que le modèle corrigé n'est pas en régime de *sur-ajustement* et confirme que l'espace des descripteurs encode très précisément l'espace des phases.

On constate que pour les modèles EQML - Eq. (4.2) et corrigé - Eq. (4.21) -, l'erreur quadratique moyenne est divisée par deux par rapport à un modèle linéaire. De plus, les courbes de *training/testing* montrent que ces modèles restent transférables. Ainsi

l'utilisation d'un modèle d'ordre supérieur en descripteur - ici d'ordre 2 - permet d'augmenter la **capacité du modèle**, c'est-à-dire la réduction d'erreur - au sens du RMSE et du MAE - substantielle par rapport au modèle linéaire tout en conservant une bonne transférabilité. En termes de comparaison entre les deux modèles quadratiques, il est remarquable de noter que le modèle corrigé (qui est purement quadratique) présente la même RMSE que le EQML (qui implique l'utilisation d'un modèle linéaire). Ce constat met en lumière le fait que la physique des déformations peut être traduite directement par une forme quadratique dans l'espace des descripteurs.

4.3 Modèle de régression des fréquences d'attaque

Dans cette section, nous allons étendre le lien structural qui existe entre l'espace des phases et l'espace des descripteurs en travaillant sur les fréquences d'attaque. Dans un premier temps, nous définissons le cadre théorique nécessaire au calcul des propriétés cinétiques des matériaux et des fréquences d'attaque 4.3.1. Dans la sous-section 4.3.2, nous faisons le lien entre le *formalisme de Green* pour la *densité d'état* de *modes normaux* et un modèle de régression simple des fréquences d'attaque. Enfin dans la sous-section (4.3.3), nous présentons les résultats de ce modèle de régression pour un système de silicium amorphe qui présente un paysage énergétique complexe.

4.3.1 Rappels et définitions sur la Théorie Harmonique de l'État de Transition (HTST)

Les grandeurs décrites dans les chapitres précédents, notamment l'énergie, l'entropie et l'*énergie libre* sont des grandeurs thermodynamiques qui définissent l'état du système à l'équilibre. Or, l'atteinte de l'équilibre thermodynamique est entièrement conditionnée par les **propriétés cinétiques** du système. Dans certaines situations, les transformations peuvent être *cinétiquement bloquées* et le système n'atteint pas son état d'équilibre thermodynamique pendant la durée de l'expérience. C'est par exemple le cas du diamant qui n'est pas l'état le plus stable du carbone (dans des conditions normales de température et de pression) mais dont la dégradation est *cinétiquement bloquée*. La Théorie de l'État de Transition (TST) a pour but de décrire les grandeurs cinétiques des matériaux. Dans le cadre général de la TST, on s'intéresse à l'ensemble des transitions $m \to s \to m'$, c'est-à-dire les transitions qui vont du minium m vers le minimum m' en passant par le point col s. On devrait donc associer deux probabilités : (i) $p_{m\to s}$ et (ii) $p_{s\to m'}$ pour décrire la transition $m \to s \to m'$. **Dans le cadre de la TST harmonique, la probabilité associée à l'événement $s \to m'$ est toujours égale à 1 ce qui va nous permettre de définir plus simplement le concept d'événement cinétique**. On définit la notion d'événement cinétique $\mathcal{E}$. Un événement $\mathcal{E}$ est associé à deux instances : (i) une configuration d'énergie minimum notée $\mathcal{E}, m$ et une configuration de point de selle connectée à $\mathcal{E}, m$ et notée $\mathcal{E}, s$. Notons alors $X_{\mathcal{E},m\to s}$ la variable aléatoire comptant le nombre de transitions entre l'état $\mathcal{E}, m$ et

l'état $\mathcal{E}, s$ intervenant entre les instants t et $t + \delta t$. On introduit le taux de transition de $R_{\mathcal{E},m\to s}$ entre un l'état $\mathcal{E}, m$ et l'état $\mathcal{E}, s$ de la façon suivante :

$$X_{\mathcal{E},m\to s} \overset{\text{loi}}{\sim} \mathcal{P}\left(R_{\mathcal{E},m\to s}\right) \tag{4.22}$$

où $\mathcal{P}\left(R_{\mathcal{E},m\to s}\right)$ est une loi de Poisson de paramètre $R_{\mathcal{E},m\to s}$. Dans le cadre de l'approximation harmonique, il est possible de calculer analytiquement la valeur de $R_{\mathcal{E},m\to s}$ entre un état minimal d'énergie $\mathcal{E}, m$ et un point de selle du premier ordre $\mathcal{E}, s$ en faisant l'intégration thermodynamique sur l'espace des phases. Ce calcul a été effectué par Vineyard [190] et permet d'écrire le taux de transition de la façon suivante :

$$R_{\mathcal{E},m\to s} = \nu^*_{\mathcal{E},ms} e^{-\beta \Delta E_{\mathcal{E},m\to s}} \tag{4.23}$$

où $\Delta E_{\mathcal{E},m\to s}$ est la différence d'énergie entre le point de selle et le minimum et $\nu^*_{\mathcal{E},ms}$ est appelée la **fréquence d'attaque**. Dans le cadre de l'approximation harmonique, $\nu^*_{\mathcal{E},ms}$ prend l'expression suivante :

$$\nu^*_{\mathcal{E},ms} = \frac{\prod_{\nu_{m'} \in \mathcal{S}(\mathcal{E},m)} \nu_{m'}}{\prod_{\nu_{s'} \in \mathcal{S}(\mathcal{E},s)} \nu_{s'}} \tag{4.24}$$

Ici, le numérateur représente le produit des fréquences réelles des *modes normaux* (courbures positives) notées $\mathcal{S}(\mathcal{E}, m)$ à l'état de minimum $\mathcal{E}, m$. Le dénominateur est le produit des fréquences réelles (courbures positives) notées $\mathcal{S}(\mathcal{E}, s)$ des *modes normaux* au point de selle $\mathcal{E}, s$. On nomme $\nu^*_{\mathcal{E},ms}$ la fréquence d'attaque car, dans le cadre d'un point de selle du premier ordre, celle-ci est homogène à une fréquence. Le taux de transition peut être interprété comme une **fréquence moyenne au sens de la thermodynamique statistique**. $R_{\mathcal{E},m\to s}\, \delta t$ est alors le nombre de tentatives de sauts entre t et $t + \delta t$. De par leur définition, les fréquences d'attaque sont des images directes de différences de courbures du paysage énergétique entre l'état d'énergie minimale et le point de selle. En effet, cette grandeur peut être appréhendée comme le ratio entre le produit des courbures positives du bassin et le produit des courbures positives du point de selle.

La connaissance des états de transitions et des **taux de transition** permet de remonter à ses propriétés cinétiques macroscopiques. Dans le domaine des matériaux sous irradiation, les propriétés cinétiques des défauts permettent d'expliquer leur réarrangement sous forme d'objets étendus comme les boucles de dislocation [11] et leur mobilité à l'échelle macroscopique [14]. Les taux de transitions permettent de calculer les **coefficients de diffusion macroscopiques** dans un système [191-196].

4.3.2 Formalisme de *Green* local et régression des fréquences d'attaque

Composons l'expression de la fréquence d'attaque dans le cadre de l'approximation harmonique Eq. (4.24) par le logarithme. On obtient alors l'expression suivante :

$$\ln\left(\nu^*_{\mathcal{E},ms}\right) = \sum_{\nu_{m'}\in\mathcal{S}(\mathcal{E},m)} \ln\left(\nu_{m'}\right) - \sum_{\nu_{s'}\in\mathcal{S}(\mathcal{E},s)} \ln\left(\nu_{s'}\right) \tag{4.25}$$

Pour un état j donné, et en supposant que le spectre de fréquences est suffisament "dense" pour respecter l'hypothèse du continu, on peut directement relier l'expression (4.25) à la *densité d'état* de *modes normaux* Ω_j :

$$\sum_{\nu_{j'}\in\mathcal{S}(\mathcal{E},j)} \ln\left(\nu_{j'}\right) = \int_0^{+\infty} \ln\left(\frac{\omega}{2\pi}\right)\Omega_j(\omega)d\omega \tag{4.26}$$

Nous avons montré dans le chapitre 3 qu'il existe un lien entre la décomposition locale de la *densité d'état* de modes de vibration et les *fonctions de Green* solutions du problème (3.20). L'existence de cette décomposition locale dans le cadre de l'approximation harmonique nous a permis de construire un modèle de régression linéaire simple de cette grandeur dans l'espace de représentation des descripteurs 3.4.2. L'équation (4.26) suggère que le logarithme de la fréquence d'attaque peut aussi être ajusté par un modèle linéaire se basant sur les descripteurs locaux. Ce modèle s'exprime de façon analytique sous la forme suivante :

$$\ln\left(\nu^*_{\mathcal{E},ms}\right) = \underline{w}_1 \cdot \left(\underline{D}_{\mathcal{E},m} \oplus \underline{D}_{\mathcal{E},s}\right) \tag{4.27}$$

Ici, $\underline{D}_{\mathcal{E},m/s} = \sum_{d\in\mathcal{E},m/s} \underline{D}^d \in \mathbb{R}^D$ est le vecteur total de descripteurs de la configuration $\mathcal{E},m$ ou $\mathcal{E},s$. Afin d'enrichir le modèle et de capter les effets non-corrélés entre le minimum et le point-selle, nous avons choisi d'utiliser la somme directe entre les vecteurs de représentation $\underline{D}^{(\cdot)} \in \mathbb{R}^{\mathcal{D}}$ de l'état $\mathcal{E},m$ et de l'état $\mathcal{E},s$. Finalement, nous devons optimiser un vecteur de poids $\underline{w}_1 \in \mathbb{R}^{2\mathcal{D}}$. Pour cela, nous utilisons la même procédure de régression et la même procédure de *training/testing* que dans la sous-section (3.4.2). Il nous reste maintenant à construire une base de données représentative des états de transitions d'un système. Nous avons choisi de nous intéresser au cas du silicium amorphe car celui-ci possède un paysage énergétique très complexe et où il est "aisé" de trouver un grand nombre de trajectoires reliant un minimum à des points-selle du premier ordre.

4.3.3 Application au cas du silicium amorphe : base de données et modèle linéaire

Nous choisissons d'étudier des systèmes de silicium amorphe. Le silicium amorphe a fait l'objet de nombreuses études exhaustives notamment grâce à l'algorithme *ARTn* [73, 175, 176, 192, 197, 198]. L'utilisation de cet algorithme permet d'obtenir des **configurations distinctes** et la **connectivité entre le minimum et les**

différents points de selle. Une base de données a été générée à partir de 20 systèmes indépendants de silicium amorphe en utilisant le potentiel développé par Stillinger-Weber [199]. Ces systèmes ont été obtenus en effectuant des simulations dans l'ensemble canonique en utilisant le package `LAMMPS` [185] dans des super-cellules cubiques contenant 4096 atomes de silicium à une densité fixée à 2.192 g.cm^{-3}. Le pas de temps de la dynamique moléculaire a été fixé à 1 fs. Les configurations aléatoires ont d'abord été thermalisées à 2300 K pendant une durée de 20 ns et directement relaxées à 700 K pendant une durée de 100 ns. Finalement, l'énergie des systèmes a été minimisée en utilisant l'algorithme FIRE [200], jusqu'à ce que toutes les composantes de forces soit inférieures à 10^{-9} eV.Å^{-1}. Une fois ces minima atteints, l'algorithme de *ARTn* a été utilisé pour échantillonner les points de selle partant de ces minima (la convergence de *ARTn* en termes de force a été fixée à 10^{-7} eV.Å^{-1}). L'ensemble des configurations obtenues sont alors non-équivalentes grâce à la méthode developpée par Gelin *et al.* [201]. Chaque minima de la base de données est connecté en moyenne à 420 points-selle pour un total de 10502 événements distincts. Concernant *ARTn*, le spectre de la matrice dynamique a été estimé par utilisation de la méthode de Lanczos décrite par Marinica *et al.* [179]. L'erreur admissible sur les valeurs propres issues de la méthode de Lanczos a été fixée à $10^{-6}(1.018049 \times 10^{-2}\text{ps})^{-2}$.

Nous proposons de construire notre modèle de régression de la base de données de *silicium amorphe* en utilisant le descripteur $bSO(4)_4$ avec un rayon de coupure égal à celui du *potentiel semi-empirique* utilisé [199], $r_{cut} = 3.77$ Å. Le descripteur somme directe résultant possède $35 + 35$ composantes Eq. (4.27). Les figures 4.5.(a) et 4.5.(b) comparent respectivement le logarithme de la fréquence d'attaque et la fréquence d'attaque obtenue par diagonalisation directe de la matrice dynamique au minimum et au point de selle à celui et celle prédite par le modèle linéaire. Le modèle linéaire est capable de prédire précisément la valeur de $\log(\nu/\nu_0)$ où ν_0 est fixé à 1 THz. La valeur du RMSE est de 0.65 $\log(\nu/\nu_0)$ pour une plage de valeurs de 10 $\log(\nu/\nu_0)$. Les résultats de la procédure de *training/testing* sont présentés en encart de la figure 4.5.(a). Les valeurs des indicateurs statistiques restent stables même pour de larges proportions d'ensemble de vérification. Comme dans le cas de l'entropie vibrationnelle, les résultats de la procédure de *training/testing* tendent à montrer que le modèle linéaire pour ajuster le logarithme de la fréquence d'attaque sera transférable à des configurations non-présentes dans la base de données initiale. Les résultats sur les fréquences d'attaque sont présentés par la figure 4.5.(b). Si on note $\hat{\nu}$ la fréquence d'attaque prédite par le modèle de régression et ν la fréquence d'attaque issue de la base de données, on peut définir la variable $\log(\nu/\hat{\nu})$. Cette variable aléatoire suit la distribution présentée dans l'encart de la figure 4.5.(b) et possède une moyenne $\mu = 0$ et un écart-type $\sigma = 0.653$. On peut donc en déduire que 98 % des valeurs de la variable $\nu/\hat{\nu}$ vérifient l'encadrement suivant : $0.27 \leq \nu/\hat{\nu} \leq e^{2\sigma} \approx 3.7$. **Le modèle de régression permet donc d'obtenir une estimation rapide des fréquences d'attaque dont la valeur est proche (même ordre de grandeur) que les fréquences d'attaque calculée par diagonalisation directe du Hessien**. Notre approche linéaire dans l'espace des descripteurs permet d'obtenir un modèle relativement précis des fréquences

d'attaques en ne se basant que sur des informations géométriques du système et ne nécessite pas d'échantillonner et de diagonaliser la matrice dynamique du système. **La complexité numérique de notre approche Machine Learning évolue comme $\mathcal{O}(N)$ et pourrait être utilisée dans un code de Monte Carlo cinétique pour estimer "à la volée" les fréquences d'attaque**.

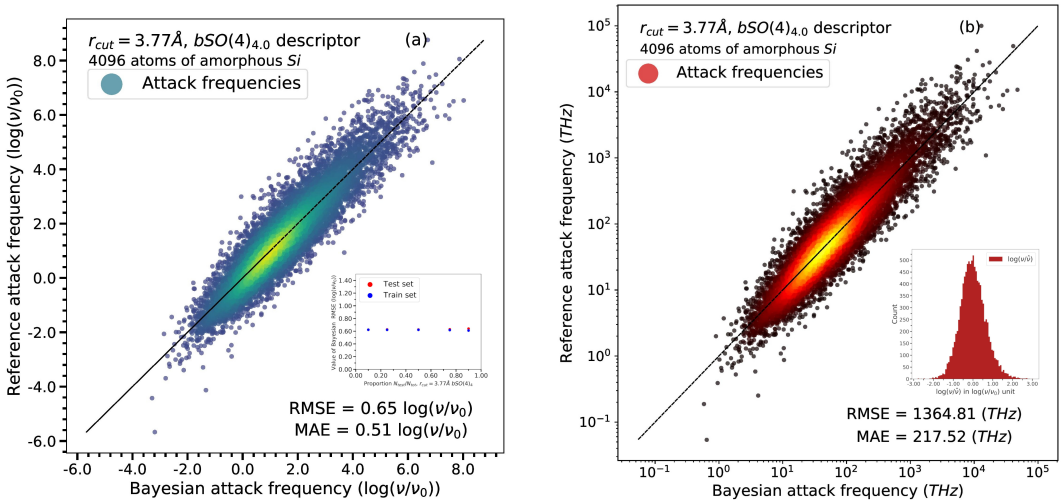

Figure 4.5: Illustration des résultats obtenus avec le modèle linéaire pour ajuster les (logarithmes des) fréquences d'attaques pour la base de données de *Si amorphe*. Le gradient de couleur représente la densité de données, le jaune correspond aux zones denses en données. Á l'opposé, le bleu (a) et le rouge (b) représentent des zones peu denses en données. Le modèle de régression sur le logarithme de la fréquence d'attaque (a) ainsi que les résultats de la procédure de *training/testing* montrent que le modèle possède une bonne transférabilité. Le modèle direct de régression des fréquences d'attaque (b) montre une bonne reconstruction de celles-ci. On note $\hat{\nu}$ l'estimation de ν par le modèle linéaire. La distribution de la variable $\log(\nu/\hat{\nu})$ est donnée dans l'encart avec $\mu = 0$ et $\sigma = 0.653$.

4.4 Barrières d'énergie associées et loi de Meyer-Neldel

La littérature [201-204] décrit l'existence d'une loi empirique nommée "loi de compensation" ou "loi de Meyer-Neldel". Cette loi générale est décrite dans un grand nombre de matériaux comme les métaux [202], les semi-conducteurs [205], minéraux ou encore dans le domaine de la biologie [204]. La loi de Meyer-Neldel est avant tout une loi expérimentale qui décrit une proportionnalité des taux de transition pour un groupe de transformations donné et leur énergies d'activation associées. Nous allons nous intéresser à une formulation microscopique de cette loi reliant le logarithme de la fréquence d'attaque et l'énergie de la barrière pour une transition donnée (et que nous appellerons, par extension, loi de Meyer-Neldel). On peut donner deux définitions quantitatives de

cette même loi : (i) une définition **"point-à-point"** et (ii) une définition **marginale**. Pour une transition donnée $\mathcal{E}, m \to s$, la relation "point-à-point" est donnée par :

$$\log(\nu^*_{\mathcal{E},ms}/\nu_0) = \gamma \Delta E_{\mathcal{E},m\to s} + \log(\nu_\gamma/\nu_0) \tag{4.28}$$

où γ a la dimension inverse d'une énergie et $\log(\nu_\gamma/\nu_0)$ est sans unité. Cette instance de la loi de Meyer-Neldel est observée dans les métaux [206] et peut être démontrée analytiquement comme étant la relation entre le gap d'énergie et la conductivité électrique pour un système modèle [207]. L'autre instance de la loi de Meyer-Neldel est sa définition marginale. Cette expression de la loi de compensation a été mise en évidence par Gelin *et al.* [201] et peut être exprimée de la façon suivante :

$$\mathbb{E}\left[\log(\nu^*_{\mathcal{E},ms}(\Delta E_{\mathcal{E},m\to s})/\nu_0)|\Delta E_{\mathcal{E},m\to s}\right] = \gamma^* \Delta E_{\mathcal{E},m\to s} + \log(\nu^*_\gamma/\nu_0), \tag{4.29}$$

où γ^* a la dimension inverse d'une énergie et $\log(\nu^*_\gamma/\nu_0)$ est sans unité. Nous notons $\mathbb{E}\left[\log(\nu^*_{\mathcal{E},ms}(\Delta E_{\mathcal{E},m\to s})/\nu_0)|\Delta E_{\mathcal{E},m\to s}\right]$ la moyenne empirique de $\log(\nu^*_{\mathcal{E},ms})$ pour chaque configuration dont l'énergie de barrière associée est comprise entre $\Delta E_{\mathcal{E},m\to s}$ et $\Delta E_{\mathcal{E},m\to s} + \delta\epsilon$, où $\delta\epsilon$ est la largeur de la fenêtre d'énergie associée à cette moyenne empirique. Cette loi marginale n'est donc plus **ponctuelle** mais porte sur une moyenne par des fenêtres énergétiques $\delta\epsilon$. Les définitions "point-à-point" et marginale sont deux façons différentes de quantifier la corrélation entre $\nu^*_{\mathcal{E},ms}$ et $\Delta E_{\mathcal{E},m\to s}$. La loi de Meyer-Neldel est généralement utilisée afin de calculer directement la valeur du préfacteur d'une transition sans avoir à diagonaliser la matrice dynamique du système. On notera que la définition "point-à-point" implique la définition marginale.

Dans cette section, nous construisons une extension de la loi de Meyer-Neldel dans l'espace des descripteurs. Pour cela, nous construisons d'abord un modèle linéaire de régression des barrières énergétiques dans le même espace de représentation que celui des fréquences d'attaque. En travaillant dans le même espace de descripteurs, nous donnons une nouvelle analyse statistique des corrélations associées à la loi de Meyer-Neldel.

4.4.1 Modèle de régression des barrières pour la base de données *Si amorphe*

Afin de mettre en évidence l'existence d'une relation de type Meyer-Neldel dans l'espace des descripteurs, nous commençons par construire un modèle de régression de l'énergie des barrières pour la base de données de *Si amorphe*. Par analogie avec le modèle de régression des fréquences d'attaque, nous choisissons la forme analytique suivante pour le modèle :

$$\Delta E_{\mathcal{E},m\to s} = \underline{w}_2 \cdot \left(\underline{D}_{\mathcal{E},m} \oplus \underline{D}_{\mathcal{E},s}\right) \tag{4.30}$$

Ici $\underline{D}_{\mathcal{E},m/s} = \sum_{d\in\mathcal{E},m/s} \underline{D}^d \in \mathbb{R}^D$ est le vecteur total de descripteurs de la configuration $\mathcal{E}, m$ ou $\mathcal{E}, s$. Comme dans la sous-section 4.3.2, nous devons ajuster le vecteur de poids $\underline{w}_2 \in \mathbb{R}^{2\mathcal{D}}$. De même que pour les fréquences d'attaque, nous construisons notre modèle

de régression de la base de données de *silicium amorphe* en utilisant le descripteur $bSO(4)_4$ avec un rayon de coupure égal à celui du *potentiel semi-empirique* utilisé c'est-à-dire $r_{cut} = 3.77$ Å. Le descripteur somme directe résultant possède $35 + 35$ composantes. Les résultats obtenus sont présentés dans la figure 4.6. On constate que le modèle linéaire de régression pour les barrières d'énergie possède une bonne précision avec une RMSE de 0.09 eV pour une gamme d'énergie de 0 à 6 eV. On peut remarquer que le modèle de régression des barrières possède une dispersion plus faible que celui des fréquences d'attaque. On peut identifier deux causes à l'origine à cette différence de variance entre les deux modèles. Le potentiel *semi-empirique* utilisé possède une régularité faible (cf. annexe C) notamment au niveau de sa dérivée seconde ; ainsi la valeur de la barrière d'énergie calculée par la méthode *ARTn* est cohérente avec la physique du système mais la valeur de la courbure (associée aux *modes normaux*) est entachée par une erreur intrinsèque liée à la régularité. Le choix du critère de la méthode *ARTn* n'est pas assez précis pour les courbures à calculer dans le paysage énergétique. La qualité des résultats obtenus pour les barrières d'énergie en utilisant la même forme analytique de régression que pour le logarithme des fréquences d'attaque nous pousse à investiguer plus profondément la loi de Meyer-Neldel.

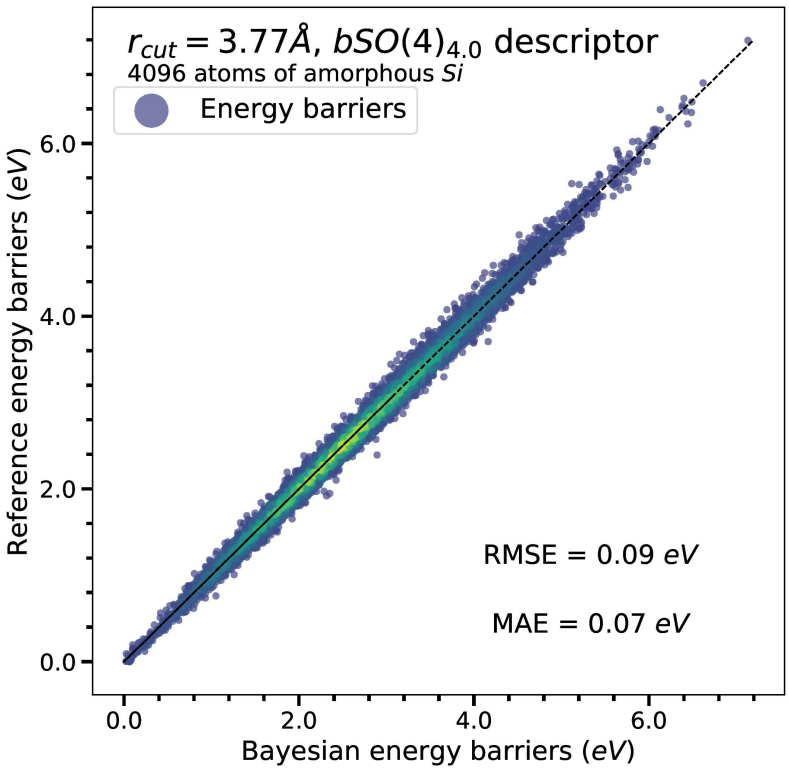

Figure 4.6: Modèle linéaire de régression pour les barrières d'énergie pour la base de données *Si amorphe*. Le descripteur et le rayon de coupure utilisés pour la régression sont les mêmes que pour les fréquences d'attaque Sec. 4.3. On note que le modèle des barrières possède une dispersion plus faible que celui des fréquences d'attaque.

4.4.2 Extension de la loi de Meyer-Neldel

Dans cette section nous construisons une extension de la loi de Meyer-Neldel - pour ses instances "point-à-point" et marginale - dans l'espace des descripteurs. Nous commençons par une propriété simple portant sur les distributions de poids entre $\underline{w}_1$ (poids pour le modèle de régression du logarithme des fréquences d'attaque) et $\underline{w}_2$ (poids du modèle de régression pour les barrières d'énergies). Pour cela, nous introduisons $\underline{\hat{w}}_1 = \frac{\underline{w}_1 - \mu_1 \underline{1}}{\sigma_1}$ et $\underline{\hat{w}}_2 = \frac{\underline{w}_2 - \mu_2 \underline{1}}{\sigma_2}$ où μ_1, μ_2 et σ_1, σ_2 sont respectivement les moyennes et les écarts types pour les vecteurs $\underline{w}_1$ and $\underline{w}_2$ et où $\underline{1} \in \mathbb{R}^{2\mathcal{D}}$ est le vecteur identité. Les vecteurs $\underline{\hat{w}}_1$ et $\underline{\hat{w}}_2$ sont les vecteurs centrés et réduits issus de $\underline{w}_1$ et $\underline{w}_2$. S'il existe une relation linéaire entre $\underline{w}_1$ et $\underline{w}_2$ alors les distributions issues de $\underline{\hat{w}}_1$ et $\underline{\hat{w}}_2$ doivent être identiques. Ces distributions sont présentées dans la figure 4.7. On constate que celles-ci sont presques identiques, ce qui suppose l'existence d'une relation linéaire entre les vecteurs $\underline{w}_1$ et $\underline{w}_2$.

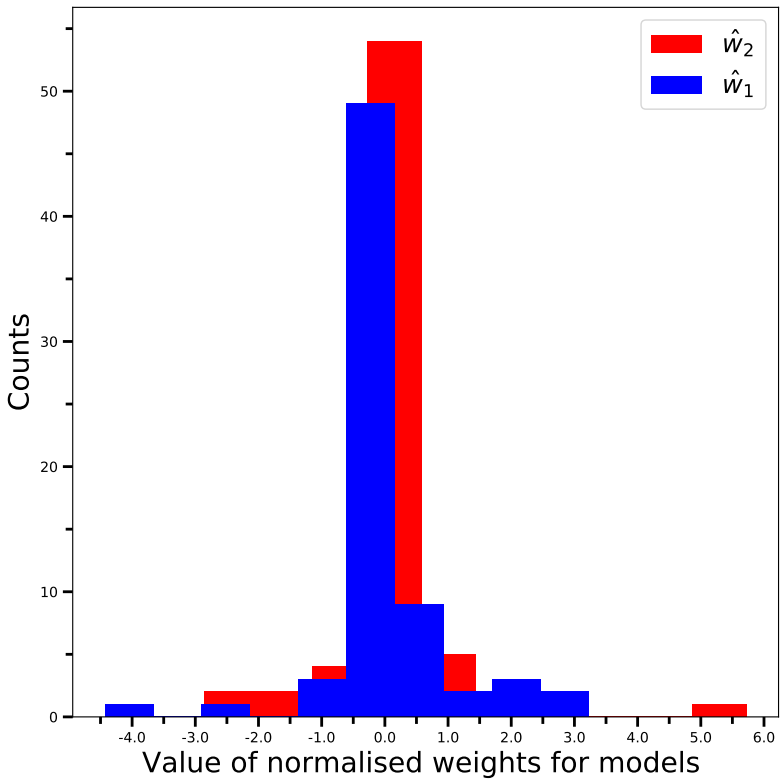

Figure 4.7: Distributions des poids centrés réduits $\underline{\hat{w}}_1$ et $\underline{\hat{w}}_2$. On constate que les distributions sont presque identiques, ce qui suppose l'existence d'une relation linéaire entre les vecteurs $\underline{w}_1$ et $\underline{w}_2$.

Nous allons donner un sens plus quantitatif à la relation "point-à-point" de Meyer-Neldel dans l'espace des descripteurs. Pour cela, nous estimons la corrélation, pour

toutes transitions de la base de données, entre les barrières d'énergie ΔE et leur fréquences d'attaque $\log(\nu/\nu_0)$. Nous avons calculé ces corrélations pour la base de données de *Si amorphe* et pour les résultats prédits par nos modèles de régression. La figure 4.8 représente la loi empirique "point-à-point" de Meyer-Neldel pour la base de données de *Si amorphe*. Les corrélations entre ΔE et $\log(\nu/\nu_0)$ issues de la base de données de *Si amorphe* et les prédictions des modèles linéaires dans l'espace des descripteurs sont quantitativement les mêmes. Nous avons calculé le coefficient de corrélation suivant :

$$r(\nu, \Delta E) = \frac{\mathbb{C}(\log(\nu/\nu_0), \Delta E)}{\sqrt{\mathbb{V}(\log(\nu/\nu_0))\mathbb{V}(\Delta E)}} \tag{4.31}$$

avec $\mathbb{C}(\log(\nu/\nu_0), \Delta E)$ la covariance entre $\log(\nu/\nu_0)$ et ΔE. Ce coefficient de corrélation est de 0.61 pour la base de données de *Si amorphe* et de 0.65 pour les données prédites par les modèles linéaires.

Il est légitime de se demander si la relation "point-à-point" de Meyer-Neldel peut se généraliser à l'espace des descripteurs. Considérons deux événements disctincts $\mathcal{E}^1$ et $\mathcal{E}^2$ et leurs barrières d'énergie associées $\Delta E_{\mathcal{E}^1,m\to s}$ et $\Delta E_{\mathcal{E}^2,m\to s}$ telles que $\Delta E_1 = \alpha \Delta E_2$. La relation (4.30) implique alors :

$$\underline{w}_2 \cdot (\underline{D}_{\mathcal{E}^1,m} \oplus \underline{D}_{\mathcal{E}^1,s} - \alpha \underline{D}_{\mathcal{E}^2,m} \oplus \underline{D}_{\mathcal{E}^2,s}) = 0 \tag{4.32}$$

Ici, $\underline{D}_{\mathcal{E},m/s} = \sum_{d\in\mathcal{E},m/s} \underline{D}^d \in \mathbb{R}^D$ est le vecteur total de descripteur de la configuration $\mathcal{E}, m$ ou $\mathcal{E}, s$. $\underline{w}_2$ est le vecteur de paramètres ajustables donné par l'équation (4.30). Si on suppose que pour deux événements i et j on a : $\Delta E_i - \Delta E_0 = \gamma \log(\nu_i/\nu_0)$ et $\Delta E_j - \Delta E_0 = \gamma^{-1} \log(\nu_j/\nu_0)$, on peut déduire de l'équation (4.28) que $\log(\nu_i/\nu_0) = \frac{\Delta E_i - \Delta E_0}{\Delta E_j - \Delta E_0} \log(\nu_j/\nu_0)$. Ainsi, si la relation de Meyer-Nedel est généralisable dans l'espace des descripteurs, on peut définir une relation d'orthogonalité analogue à l'équation (4.32) :

$$\underline{w}_1 \cdot \left(\underline{D}_{\mathcal{E}^i,m} \oplus \underline{D}_{\mathcal{E}^i,s} - \frac{\Delta E_i - \Delta E_0}{\Delta E_j - \Delta E_0} \underline{D}_{\mathcal{E}^j,m} \oplus \underline{D}_{\mathcal{E}^j,s} \right) = 0 \tag{4.33}$$

ici $\mathcal{E}^i$, $\underline{D}_{\mathcal{E}^i,m} \oplus \underline{D}_{\mathcal{E}^i,s}$ sont les événements et les descripteurs associés de la base de données de *Si amorphe*. $\underline{w}_1$ est le vecteur de paramètres ajustables défini par l'équation (4.30). Cette relation d'orthogonalité est valide si la loi "point-à-point" de Meyer-Neldel définie par l'équation (4.28) est exacte. La figure 4.8 montre que cette relation (4.28) est qualitativement vraie mais induit une corrélation imparfaite (0.61 pour la base de données au lieu de 1.0 dans le cas théorique). Pour introduire la notion de loi "point-à-point" de Meyer-Neldel généralisée dans l'espace des descripteurs avec l'équation (4.33), nous devons introduire une notion "d'orthogonalité" prenant en compte les corrélations (imparfaites) de la base de données.

Afin de simplifier les notations, nous définissons le vecteur suivant : $\underline{\mathfrak{D}}^{ij} = \underline{D}_{\mathcal{E}^i,m} \oplus$

$\underline{D}_{\mathcal{E}^i,s} - \frac{\Delta E_i - \Delta E_0}{\Delta E_j - \Delta E_0} \underline{D}_{\mathcal{E}^j,m} \oplus \underline{D}_{\mathcal{E}^j,s}$. Pour quantifier la relation "point-à-point" de Meyer-Nedel dans l'espace des descripteurs, nous calculons le produit scalaire $\underline{w}_1 \cdot \underline{\mathfrak{D}}^{ij}$ pour toutes les configurations de la base de données et nous introduisons le ratio suivant :

$$\kappa_{ij} = \frac{\left|\underline{w}_1 \cdot \underline{\mathfrak{D}}^{ij}\right|}{\left|\underline{w}_1 \cdot \left(\underline{D}_{\mathcal{E}^i,m} \oplus \underline{D}_{\mathcal{E}^i,s} + \underline{D}_{\mathcal{E}^j,m} \oplus \underline{D}_{\mathcal{E}^j,s}\right)\right|} \tag{4.34}$$

Si la loi de Meyer-Nedel est valide dans l'espace des descripteurs, nous devons avoir :

$$\langle \kappa \rangle \equiv \frac{1}{S^2} \sum_{i,j=1}^{S,S} \kappa_{ij} \ll 1 \tag{4.35}$$

où S est le nombre de configurations dans la base de données de *Si amorphe*. Nous calculons la valeur moyenne de κ pour toute la base de données et nous obtenons $\langle \kappa \rangle = 0.11$. Cette valeur faible de $\langle \kappa \rangle$ induit une généralisation de la loi "point-à-point" de Meyer-Neldel dans l'espace des descripteurs.

Dans un deuxième temps, nous nous intéressons à la définition marginale de la loi de Meyer-Neldel. Une analyse de la marginale en ΔE a été effectuée sur la base de données. Pour cela, nous avons défini 50 intervalles d'énergie de barrière uniformement distribués sur la base de données de Si amorphe. Nous avons alors calculé la marginale $\mathbb{E}\left[\log(\nu/\nu_0)|\Delta E\right]$ pour chaque valeur de ΔE. Cette analyse a aussi été menée sur la base de données prédite par notre modèle linéaire. Afin de quantifier l'erreur sur la définition marginale de la loi de Meyer-Neldel, nous avons aussi calculé la variance marginale $\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]$.
Les résultats de cette analyse sont donnés dans la figure 4.8. La figure 4.8.(a) et la figure 4.8.(b) présentent la corrélation entre $\log(\nu/\nu_0)$ et ΔE pour les deux bases de données. Les points gris présentent la marginale $\mathbb{E}\left[\log(\nu/\nu_0)|\Delta E\right]$ ainsi qu'une estimation d'erreur sur cette variable. La loi marginale de Meyer-Neldel définie par l'équation (4.29) est donnée en encart de chaque figure. Les modèles marginaux de la loi de Meyer-Neldel sont quantitativement similaires entre la base de données de *Si amorphe* et celle prédite par notre modèle. Les figures 4.8.(c) et 4.8.(d) montrent les variances marginales $\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]$ respectivement pour la base de données Si amorphe et pour celle prédite par le modèle. Une estimation d'erreur sur la variance marginale est aussi présentée. Les Fig. 4.8.(c) et Fig. 4.8.(d) montrent que la variance marginale reste stable quelque soit la valeur de la barrière d'énergie ΔE. De plus, cet indicateur $\sigma[\log(\nu/\nu_0)|\Delta E]^2 = \mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]$ est quantitativement le même pour les deux bases de données. On note seulement des variations de $\sigma[\log(\nu/\nu_0)|\Delta E]^2$ pour les valeurs de ΔE très faibles ou très élevées, ce qui correspond aux **bords de la distribution de ΔE. Le manque de statistiques consolidées pour ces valeurs de ΔE est corroborré par les grandes valeurs de** $\mathbb{V}\left[\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]\right]$.

En supposant que tous les événements de la base de données de *Si amorphe* sont

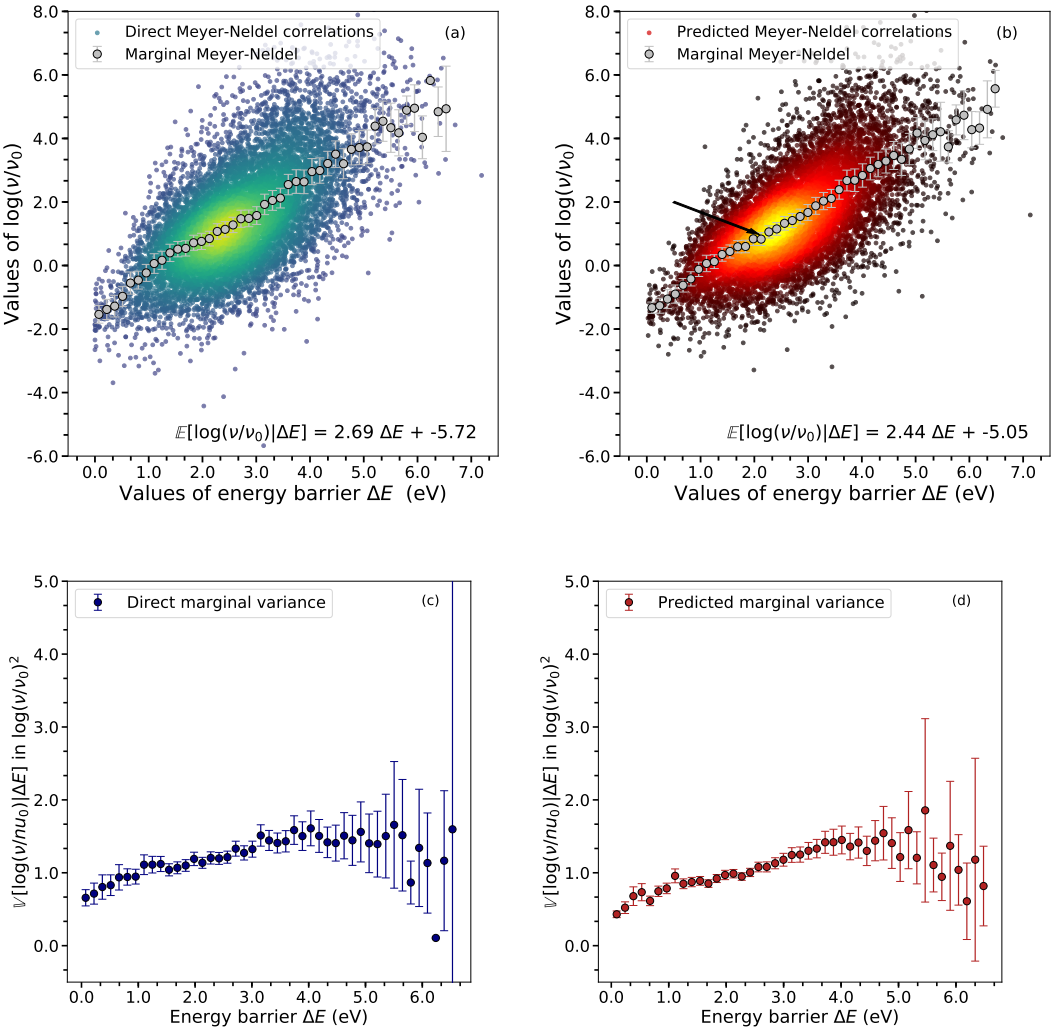

Figure 4.8: Les figures 4.8.(a) et (b) représentent les corrélations entre les barrières d'énergie ΔE et les fréquences d'attaque associées $\log(\nu/\nu_0)$. Le gradient de couleur représente la densité de données, le jaune correspond aux zones denses en données. La figure (a) correspond à la base de données de *Si amorphe* et la figure (b) correspond aux données prédites par nos modèles linéaires. Une corrélation linéaire existe entre ces deux grandeurs pour la base de données de *Si amorphe* et pour les prédictions faites par nos modèles de régression. Les figures 4.8.(c) et (d) représentent le calcul des variances marginales $\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]$ pour la base de données de *Si amorphe* resp. les prédictions du modèle linéaire. Nous donnons aussi une estimation de la quantité $\mathbb{V}\left[\mathbb{V}\left[\log(\nu/\nu_0)|\Delta E\right]\right]$ représentée par la barre d'erreur pour les figures (c) et (d).

indépendants, on peut alors appliquer le *théorème Central Limite* et donner une nouvelle formulation de la loi "point-à-point" de Meyer-Neldel :

$$\log(\nu/\nu_0) = \gamma\Delta E + \log(\nu_\gamma/\nu_0) + \mathcal{N}(0, \sigma[\log(\nu/\nu_0)|\Delta E]), \tag{4.36}$$

Ici, $\mathcal{N}(0, \sigma[\log(\nu/\nu_0)|\Delta E])$ est une variable aléatoire suivant une loi normale de moyenne 0 et d'écart-type $\sigma[\log(\nu/\nu_0)|\Delta E]$. Cette nouvelle loi revient à la loi "point-à-point" originale dans le cas où $\sigma[\log(\nu/\nu_0)|\Delta E] \to 0$. Dans le cadre de la base de données

Si amorphe, il est possible de quantifier le terme stochastique de l'équation (4.36). La quantification de ce terme n'est possible qu'en combinant l'expression de la loi "point-à-point" et marginale de la loi de Meyer-Neldel.

Ici, nous proposons d'aller plus loin dans notre démarche quantitative sur la validité de la loi de Meyer-Neldel dans l'espace des descripteurs. Nous avons introduit l'indicateur κ - Eq. (4.34) - que nous pouvons étendre au cas de l'expression marginale de la loi de Meyer-Neldel. Pour cela, nous allons introduire la notion de distribution marginale de κ. On définit la marginale de κ par rapport à ΔE, $\mathbb{E}[\kappa|\Delta E]$, de la façon suivante :

$$\begin{aligned} \mathcal{B}_{\Delta E} &= \{\mathcal{E}_k \,|\Delta E \leq \Delta E_k < \Delta E + \delta E\} \\ \forall i,j \in \mathcal{B}_{\Delta E},\, \mathbb{E}[\kappa|\Delta E] &= \frac{1}{\mathrm{card}(\mathcal{B}_{\Delta E})^2} \sum_{i,j=1}^{\mathrm{card}(\mathcal{B}_{\Delta E})} \kappa_{ij} \end{aligned} \tag{4.37}$$

Ici, $\mathrm{card}(\cdot)$ est le cardinal de l'ensemble. De même, on peut définir la marginale de κ selon $\log(\nu/\nu_0)$ par l'expression :

$$\begin{aligned} \mathcal{L}_{\log(\nu/\nu_0)} &= \{\mathcal{E}_k \,|\, \log(\nu/\nu_0) \leq \log(\nu_k/\nu_0) < \log(\nu/\nu_0) + \delta \log(\nu/\nu_0)\} \\ \forall i,j \in \mathcal{L}_{\log(\nu/\nu_0)},\, \mathbb{E}[\kappa|\log(\nu/\nu_0)] &= \frac{1}{\mathrm{card}(\mathcal{L}_{\log(\nu/\nu_0)})^2} \sum_{i,j=1}^{\mathrm{card}(\mathcal{L}_{\log(\nu/\nu_0)})} \kappa_{ij} \end{aligned} \tag{4.38}$$

On peut alors donner un critère quantitatif de la notion d'orthogonalité dans l'espace des descripteurs décrite par l'équation (4.33) en termes de loi marginale. Nous avons procédé aux calculs de ces deux indicateurs pour la base de données de Si amorphe. Les résultats de cette analyse sont donnés dans les Fig. 4.9.(a) et Fig. 4.9.(b) présentant respectivement la marginale de l'indicateur κ sur la variable ΔE et sur la variable $\log(\nu/\nu_0)$. Le gradient de couleur représente la densité de données comme dans la figure 4.8. Le trait bleu correspond à la valeur "seuil" fixée à 10^{-1} grâce au calcul de l'indicateur $\langle\kappa\rangle$ défini par l'équation (4.35). On considére que la relation d'orthogonalité Eq. (4.32) est vérifiée si la marginale de l'indicateur se trouve sous ce seuil. On constate que pour un certain nombre de domaines de ΔE et de $\log(\nu/\nu_0)$, la relation d'orthogonalité n'est pas respectée. Dans le cas de la marginale sur ΔE, on peut faire un lien avec la figure 4.8.(b). Pour une valeur de $\Delta E \simeq 2.1$ eV la valeur de la marginale de κ est très élevée (50.0), sur la figure 4.8.(b). Cette valeur de barrière correspond à une **rupture de pente de la loi marginale de Meyer-Neldel** (cf. flèche figure 4.8.(b)). [1] Les indicateurs marginaux de κ permettent donc de définir quantitativement les déviations des données par rapport à la loi marginale de Meyer-Neldel.

1. Dans le cas de cette base de données, cette déviation est expliquée par le faible nombre d'événements. Ce nombre d'événements réduit a pour conséquence d'introduire une variance plus grande sur la loi marginale et donc une déviation par rapport aux résultats de Gelin *et al.* [201] utilisant une version non-tronquée de cette base de données.

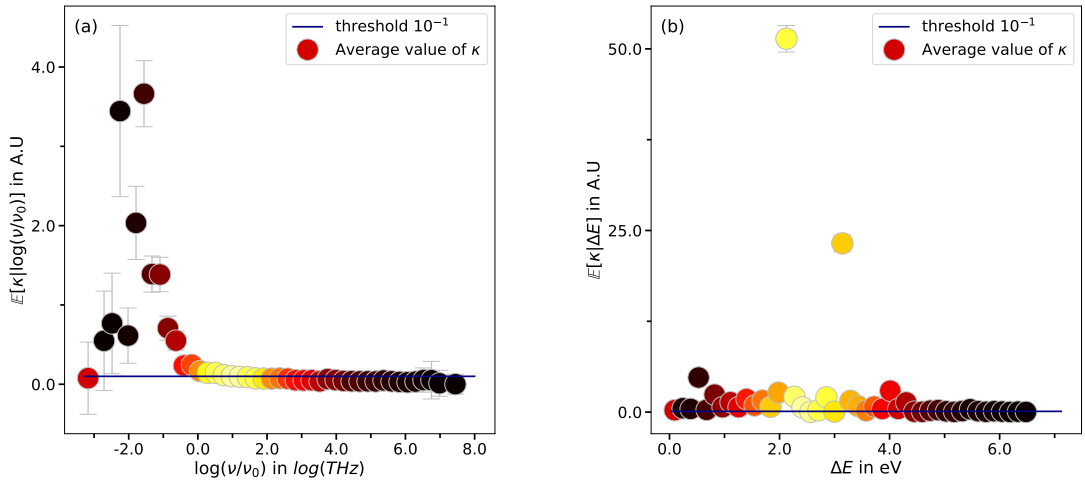

Figure 4.9: Distributions marginales de l'indicateur κ pour la base de données de Si amorphes. Le gradient de couleur représente la densité de données, le jaune correspond aux zones denses en données. La trait bleu symbolise le critère d'orthogonalité défini par $\langle\kappa\rangle$. On constate que les deux marginales possèdent des domaines dans lesquels cette relation d'orthogonalité - Eq. (4.32) - n'est pas valide.

Pour conclure cette section, la loi de Meyer-Neldel est une conjecture générale et non-triviale observée dans de nombreux matériaux [202, 204, 205]. Cette loi exprime, pour une transition donnée, le lien entre l'amplitude d'un bassin du paysage énergétique (la valeur de ΔE) et sa courbure (la valeur de $\log(\nu/\nu_0)$). Cette loi peut être apprenhendée sous deux points de vue différents : (i) une vision "point-à-point" et (ii) une vision marginale. Notre extension de la loi "point-à-point" de Meyer-Neldel dans l'espace des descripteurs met en évidence la capacité des modèles linéaires à capter des informations complexes du paysage énergétique. En effet, les données prédites par ces modèles présentent les mêmes corrélations que celles de la base de données que nous avons étudiée et sont seulement basées sur des considérations géométriques du système. De plus, l'analyse de l'instance marginale de la loi de Meyer-Neldel permet de donner une définition étendue - en introduisant un terme stochastique - de la loi "point-à-point" au travers de l'équation (4.36). Enfin, les indicateurs statistiques que nous avons développés pour quantifier les déviations de la loi "point-à-point" peuvent être étendus à la marginale de Meyer-Neldel et permettent de quantifier la déviation des données par rapport à la loi marginale. Il est important de noter que l'ensemble de ces indicateurs statistiques sont formulés seulement à l'aide d'informations géométriques du système.

4.5 Conclusions de chapitre

Dans ce chapitre, nous avons étudié plus en profondeur le lien existant entre la structure de l'espace des phases et l'espace de représentation des descripteurs. Les grandeurs que nous avons choisi d'étudier - l'entropie vibrationnelle et les fréquences

d'attaques - sont directement reliées à la structure de l'espace des phases. Elles sont diffuses et directement reliées à la notion de courbure des bassins d'énergie et des points-selle. Nous avons montré que l'utilisation d'un modèle linéaire dans l'espace des descripteurs - basé sur la décomposition locale de la *densité d'état* de modes de vibration grâce au formalisme de *Green* - permet de construire des modèles de régression précis et transférables. De plus, dans le cas des déformations, nous avons réussi à construire un lien direct entre la structure de l'espace des phases et de l'espace de représentation. La projection des coordonnées dans un espace de plus grande dimension a permis "d'aplanir" la topologie de l'espace des phases et de passer d'un problème à $3N$ dimensions à un problème à $\mathcal{D}$ dimensions. Le developpement analytique décrit dans la sous-section 4.2.2 permet d'esquisser la structure de l'espace des phases en se basant sur une transformation de la matrice de covariance des descripteurs pour une configuration donnée. Ce type d'approche permet de lever un peu le voile sur la construction et l'utilisation "en aveugle" de l'espace de représentation. Il est aussi intéressant de noter que l'approche de correction purement quadratique - basée sur l'équation (4.21) - décrit avec la même précision (au sens d'erreur quadratique moyenne) le problème de régression des déformations que le modèle EQML (Eq. 4.2). La physique de l'entropie vibrationnelle pour les petites déformations est bien décrite par le couplage des composantes des descripteurs locaux. **Cette approche pourrait donc être employée pour quantifier les corrections d'énergies élastiques induites par des déformations et dont la formulation analytique est analogue à celle développée pour l'équation (4.21)** [15].

Nous avons aussi fourni une nouvelle analyse géométrique de la loi empirique de Meyer-Neldel dans l'espace des descripteurs. Les indicateurs statistiques que nous avons définis permettent de mesurer quantitativement la déviation des données par rapport à une loi empirique donnée (ici la loi de Meyer-Neldel). **Ce type d'approche basé sur les modèles linéaire pourra être généralisé pour d'autres lois impliquant d'autres types de corrélation. Cette "correspondance" entre l'espace des descripteurs et l'espace des phases n'est possible que grâce à la structuration particulière de ces deux espaces.**

Il est important de noter que les modèles présentés dans les chapitres 3 et 4 sont valables dans le cadre de l'approximation harmonique grâce à la décomposition locale de la *densité d'état* de *modes normaux*. Cette décomposition n'est a priori pas évidente dans le cadre anharmonique en raison du couplage des différents *modes de vibration* du système.

Like a bird in a cage, tryin' to fly away
Is this the price that we have to pay ?
Overflowing with rage, yet we still obey
'Cause we're asleep in a hurricane.

— Royal Beggars, Architects

5

Au-delà de l'approximation harmonique : méthodologie

Sommaire

5.1 Nécessité de la prise en compte de l'anharmonicité

Les modèles de régression présentés dans le chapitre 3 et le chapitre 4 se basent sur l'hypothèse essentielle que la *densité d'état* de vibration possède une **décomposition locale exacte** sur la base centrée sur les atomes du système. Dans le cadre de la théorie anharmonique, les modes de phonons sont couplés. Il est donc difficile de revenir à une formulation locale des *modes de vibrations* du système. Les modèles developpés dans le chapitre 3 et le chapitre 4 ne peuvent donc pas être utilisés pour décrire les effets anharmoniques.

En effet, l'approximation harmonique représente un cadre relativement restreint permettant l'étude des **propriétés vibrationnelles** des systèmes. L'approximation des petits déplacements autour de la position d'équilibre des atomes est généralement respectée mais de nombreux effets ne peuvent être compris sans prendre en compte des termes d'ordre supérieur à 2 dans le développement de Taylor proposé dans l'équation (3.7). On peut d'ores et déjà citer deux effets physiques importants ne pouvant être expliqués dans le cadre de l'approximation harmonique : (i) la dépendance en température du **coefficient d'expansion thermique des matériaux cristallins**, qui est indépendant de la température dans le cadre de l'approximation harmonique [168] ; l'**élargissement des pics** observés en *diffusion des neutrons aux petits angles* indiquant la nécessité de prendre en compte des corrections du *Hamiltonien* du système à des ordres plus élévés (> 2) pour décrire l'interaction phonon-phonon [168]. Dans le cadre des températures finies, la prise en compte des effets anharmoniques se fait de façon implicite par calcul direct - pour une température T du système - de l'*énergie libre* $F = U - TS$, dont on rappelle la définition :

$$F = -\beta^{-1} \ln \left(\frac{1}{h^{3N}} \int_{\mathcal{Q}\times\mathcal{P}} e^{-\beta \mathcal{H}(\boldsymbol{q},\boldsymbol{p})} d\boldsymbol{q} d\boldsymbol{p} \right) \tag{5.1}$$

Dans la suite, on prendra par convention $h^{3N} = 1$. La démonstration de ce résultat important est donnée dans l'annexe A. Dans ce calcul, l'*Hamiltonien* complet du système apparait $\mathcal{H}(\boldsymbol{q}, \boldsymbol{p})$ sans aucune approximation sur la forme du potentiel du système $V(\boldsymbol{q})$ autour d'une position d'équilibre. Le calcul direct de l'*énergie libre* d'un système permet donc de prendre implicitement en compte les **effets anharmoniques**. Il est important de noter que ce calcul direct reste aujourd'hui difficile d'un point de vue numérique malgré l'apparition de méthodes efficaces et élégantes que nous allons présenter dans la section suivante 5.2.

Nous proposons dans la section 5.2 de rappeler les différentes méthodes développées dans la littérature afin de calculer l'*énergie libre* d'un système. Nous insistons sur la difficulté majeure de ces méthodes : l'**échantillonnage de la mesure canonique** dans la sous-section 5.2.1. Nous donnons ensuite un exemple concret de la mise en place et de la convergence de ces méthodes dans le cadre d'un potentiel EAM Sec. 5.3.5.

Dans ce chapitre, nous ne traiterons que le cas anharmonique complet et des méthodes de calcul d'*énergie libre*. Il existe en effet des modèles permettant de prendre en compte la dépendance des modes de phonons en fonction du volume du système. Cet ensemble de méthodes est appelé **quasi-harmonique**. Ces méthodes permettent de dépasser le cadre de l'approximation harmonique mais nous faisons le choix de nous intéresser au cadre le plus général possible.

5.2 Méthodes numériques de calcul de l'*énergie libre*

Nous présentons dans cette section les bases théoriques des méthodes d'évaluation de l'*énergie libre* dont la clef de voûte est l'échantillonnage de la mesure canonique Sec. (5.2.1). Nous présentons également ces différentes méthodes d'échantillonnage et les difficultés en résultant. Dans la sous-section (5.2.3), nous décrivons un sous-ensemble de méthodes numériques dites à **biais adaptatifs** permettant de réaliser l'échantillonnage. Enfin, dans la sous-section (5.3.5) nous mettons en place un cas pratique d'utilisation de ce type de méthodes à travers l'exemple concret d'un potentiel EAM (en calculant son paramètre de maille et son module isostatique d'équilibre à des températures finies).

5.2.1 Méthodes et difficultés d'échantillonnage de la mesure canonique

En premier lieu, nous faisons une hypothèse sur la forme de l'*énergie libre* calculée. Nous choisissons de travailler uniquement avec la partie potentielle de l'*énergie libre* F_p et nous supposons que l'*Hamiltonien* du système est séparable c'est-à-dire $\mathcal{H}(\boldsymbol{q},\boldsymbol{p}) = E_c(\boldsymbol{p}) + V(\boldsymbol{q})$. L'équation (5.1) peut alors être ré-écrite sous la forme suivante :

$$F \equiv F_p + F_c = -\beta^{-1}\left[\ln\left(\int_{\mathcal{Q}} e^{-\beta V(\boldsymbol{q})} d\boldsymbol{q}\right) + \ln\left(\int_{\mathcal{P}} e^{-\beta E_c(\boldsymbol{p})} d\boldsymbol{p}\right)\right] \tag{5.2}$$

Cette approximation est justifiable dans les systèmes cristallins où le potentiel périodique "vu" par les particules du système dépend peu de leur vitesse. La contribution cinétique de l'*énergie libre* F_c est relativement aisée à calculer par une intégration directe sur l'espace des vitesses mais l'évaluation de la contribution F_p reste difficile à l'heure actuelle à cause de la complexité des paysages énergétiques des systèmes étudiés. Toute la difficulté de l'évaluation de la contribution F_p réside dans **l'échantillonnage de l'ensemble de l'espace des coordonnées du système**. Dans la suite, nous noterons souvent $F = F_p$ par abus de langage. Nous allons décrire ici deux méthodes stochastiques permettant de réaliser cet échantillonnage.

Nous introduisons ici la notion de "dynamique" dans le cadre de la théorie stochastique. Considérons un point initial de coordonnées $\boldsymbol{q_0}$ dans l'espace des configurations. On appelle "dynamique" ou noyau de transition $\mathcal{P}_{0,t}$ l'application qui vérifie pour toute observable $\mathcal{O}$ de l'espace des configurations :

$$\mathcal{O}(\boldsymbol{q_t}) = \mathcal{P}_{0,t}(\mathcal{O})(\boldsymbol{q}) \equiv \mathbb{E}(\mathcal{O}(\boldsymbol{q_t})|\boldsymbol{q} = \boldsymbol{q_0}) \tag{5.3}$$

Dans le cas de l'observable de coordonnées, on obtient simplement la trajectoire induite par la dynamique de coordonnées initiales $\boldsymbol{q_0}$ par application de l'intégrale de *Itô* [66] :

$$\boldsymbol{q_t} \equiv \mathcal{I}_{0,t}(\boldsymbol{q_0}) = \boldsymbol{q_0} + \int_0^t \boldsymbol{a}(\boldsymbol{q_s})ds + \int_0^t \boldsymbol{b}(\boldsymbol{q_s})d\boldsymbol{W}_s \tag{5.4}$$

Ici $\boldsymbol{a}(\boldsymbol{q_s}) \in \mathbb{R}^{3N}$ et $\boldsymbol{b}(\boldsymbol{q_s}) \in \mathbb{R}^{3N\times 3N}$ sont les paramètres de la dynamique stochastique. $\boldsymbol{W}_s \in \mathbb{R}^{3N}$ est un processus de Wiener tel que $\forall t_n, t_{n+1}$ indépendants, $\int_{t_n}^{t_{n+1}} d\boldsymbol{W}_s \overset{\text{loi}}{\sim} \boldsymbol{\mathcal{N}}(0, \sqrt{t_{n+1}-t_n})$ où $\boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{\Sigma}_{t_n}^{t_{n+1}})$ est une loi normale multi-dimensionnelle de moyenne nulle et dont la matrice de covariance vérifie $\boldsymbol{\Sigma}_{t_n}^{t_{n+1}} = (t_{n+1}-t_n)\mathbf{1}_{3N}$. L'intégral de *Itô* permet de générer des trajectoires dans l'espace des phases qui vont permettre l'échantillonnage [1]. L'équivalence entre l'échantillonnage par le noyau de transition $\mathcal{P}_{0,t}$ et la mesure canonique sur l'espace des phases est due au théorème ergodique dont nous allons présenter les principales hypothèses. Nous commençons par décrire les notions de *stationnarité* et d'*irréductibilité* d'une mesure par rapport à une dynamique stochastique qui vérifie l'équation (5.3).

Définition 5.1. *Soit χ un espace mesurable, μ une mesure sur χ, $\mathcal{P}_{0,t} : \chi \to \chi$ et L_1 est l'ensemble des fonctions μ-intégrables sur χ. On dit que $\mathcal{P}_{0,t}$ préserve la mesure μ si :*

$$\forall f \in L_1, \forall \boldsymbol{x} \in \chi \quad \int_\chi \mathcal{P}_{0,t}(f)(\boldsymbol{x})\, d\mu = \int_\chi f d\mu \tag{5.5}$$

Le concept de *stationnarité* traduit le fait que le noyau $\mathcal{P}_{0,t}$ conserve le volume d'un élément $d\mu(\boldsymbol{x})$ de l'espace χ, quelque soit $\boldsymbol{x} \in \chi$. Par exemple dans le cas d'une dynamique *Hamiltonnienne*, la mesure Boltzmannienne (μ) est conservée dans l'espace des phases (χ) en vertu du théorème de Liouville. **Dans le cas de mesure Boltzmannienne, la dynamique ϕ peut être : (i) une *dynamique Hamiltonienne*, (ii) une *dynamique de Langevin* ou (ii) un algorithme Metropolis-Hasting.**

Définition 5.2. *Soit χ un espace mesurable, μ une mesure sur χ et $\mathcal{P}_{0,t} : \chi \to \chi$ et A un ensemble de Borel de mesure de Lebesgue postive. $P_{0,t}$ est dit irréductible si :*

$$\forall A, \forall \boldsymbol{x'} \in \chi \quad \mathcal{P}_{0,t}(\mathbf{1}_A(\boldsymbol{x'})) > 0 \tag{5.6}$$

La notion d'*irréductibilité* traduit le fait que la dynamique $\mathcal{P}_{0,t}$ permet de générer une trajectoire dans l'espace χ qui pourra explorer n'importe quelle boule - n'importe quel volume de l'espace χ - de χ. L'échantillonnage direct de l'espace des configurations **n'est pas possible dans le cas d'un grand nombre de degrés de liberté**. Supposons que l'on veuille échantillonner l'hypercube de l'espace des coordonnées $\in \mathbb{R}^{3N}$ avec une précision ϵ dans toutes les directions. On définit alors une grille uniforme de paramètre ϵ. Le nombre d'évaluations d'énergies nécessaires à cet échantillonnage évolue alors comme $\mathcal{O}(\epsilon^{-3N})$. Ce calcul est bien évidemment impossible dans le cas de

1. Une autre méthode d'intégration équivalente à celle d'*Itô* a été développée par *Stratonovich* [66]. Nous nous limiterons à la méthode d'intégration d'*Itô* sans perdre de généralités.

systèmes de grandes tailles ($> 10^3$ atomes).

L'ensemble des dynamiques $\mathcal{P}_{0,t}$ vérifie les conditions décrites par Lelièvre *et al.* [66] dans la section 2.2.1.2, ainsi la *stationnarité* et l'*irréductibilté* des noyaux de transitions $\mathcal{P}_{0,t}$ impliquent leur *ergodicité*. Nous pouvons alors donner l'énoncé du *théorème ergodique* pour ces "dynamiques". En 1931, G. Birkhoff démontre ce résultat essentiel basé sur les propriétés de *stationnarité* et d'*ergodicité* d'une "dynamique" préservant une mesure que l'on veut échantillonner. Ce théorème porte le nom de *théorème ergodique* [65] et peut être énoncé de la façon suivante (dans le cadre stochastique) :

Théorème 5.1. *Soit χ un espace mesurable, $\mathcal{P}_{0,t} : \chi \to \chi$ une dynamique préservant la mesure μ et vérifiant les critères définis par Lelièvre et al. [66] dans la section 2.2.1.2. Alors si existe μ une mesure sur χ de mesure de Lebesgue positive alors on a l'égalité suivante quleque soit le point de départ $\boldsymbol{x'}$ de la dynamique :*

$$\forall f \in L_1 \, , \forall \boldsymbol{x'} \in \chi \quad \lim_{T \to +\infty} \frac{1}{T} \int_0^T \mathcal{P}_{0,t}(f)(\boldsymbol{x'}) dt = \int_\chi f d\mu \tag{5.7}$$

Le *théorème ergodique* exprime la possibilité d'échantillonner la **valeur moyenne d'une observable** par rapport à une mesure de probabilité sur un espace χ à l'aide d'une **moyenne empirique de cette observable** à différents points de l'espace χ grâce à une dynamique ϕ. Quand le nombre de points de la trajectoire générée par ϕ tend vers l'infini, on converge vers la valeur théorique obtenue par intégration directe sur $d\mu$. Le *théorème ergodique* est le point de départ de deux méthodes stochastiques importantes permettant de calculer les observables d'un système. Nous allons rapidement décrire ces méthodes faisant intervenir des *processus Markoviens* sans entrer dans les détails techniques. Les processus utilisés vérifient les conditions du *théorème ergodique* : ils préservent la mesure canonique Def. 5.1 et sont irréductibles Def. 5.2. Plus concrétement, le théorème ergodique se traduit de la façon suivante pour le cas de la mesure canonique et par exemple pour une observable du système $\mathcal{O}(\boldsymbol{q})$ (en commençant l'échantillonnage aux coordonnées $\boldsymbol{q_0}$) :

$$\langle \mathcal{O} \rangle_\pi \equiv \int_\mathcal{Q} \mathcal{O}(\boldsymbol{q}) \frac{e^{-\beta V(\boldsymbol{q})}}{\int_\mathcal{Q} e^{-\beta V(\boldsymbol{q'})} d\boldsymbol{q'}} d\boldsymbol{q} = \lim_{n \to +\infty} \frac{1}{\tau_n} \sum_{k=0}^{n} \int_{\tau_k}^{\tau_{k+1}} \mathcal{P}_{\tau_k,t}(\mathcal{O})(\mathcal{I}_{0,\tau_k}(\boldsymbol{q_0})) dt \tag{5.8}$$

où $\mathcal{I}_{0,\tau_k}(\boldsymbol{q_0}) \equiv \boldsymbol{q_{\tau_k}}$ est la coordonnée générée par le schéma d'intégration d'*Itô* Eq. (5.4) de la coordonnée $\boldsymbol{q_0}$. Dans la pratique nous retiendrons l'expression plus synthétique suivante basée sur les coordonnées $\boldsymbol{q}_\tau$ générées par l'intégration d'*Itô* Eq. (5.4) pendant la procédure.

$$\langle \mathcal{O} \rangle_\pi \equiv \int_\mathcal{Q} \mathcal{O}(\boldsymbol{q}) \frac{e^{-\beta V(\boldsymbol{q})}}{\int_\mathcal{Q} e^{-\beta V(\boldsymbol{q'})} d\boldsymbol{q'}} d\boldsymbol{q} = \lim_{\tau \to +\infty} \frac{1}{\tau} \sum_{0 \leq \tau' \leq \tau} \mathcal{O}(\boldsymbol{q_{\tau'}}) \tag{5.9}$$

Concentrons-nous maintenant à décrire les deux méthodes principales permettant de générer des dynamiques stochastiques vérifiant les conditions du *théorème ergodique*. La

première - sûrement la plus connue d'entre-elle - est la méthode dite de **Metropolis-Hasting** [68, 69]. Dans cet algorithme, la trajectoire générée utilise des processus stochastiques discrets : les *chaînes de Markov discrètes*. Nous allons décrire le principe général de cette méthode pour le cas de mesure canonique. On part d'un point $\boldsymbol{q}^\tau$ de l'espace des coordonnées, auquel on applique un noyau de transition $T(\boldsymbol{q}^\tau, \cdot)$ vérifiant $p(\tilde{\boldsymbol{q}}^{\tau+1}|\boldsymbol{q}^\tau) = T(\boldsymbol{q}^\tau, d\tilde{\boldsymbol{q}}^{\tau+1})$ et permettant ainsi de proposer une nouvelle coordonnée $\tilde{\boldsymbol{q}}^{\tau+1}$. On considère ensuite les deux mesures suivantes : (i) $T(\boldsymbol{q}^\tau, d\tilde{\boldsymbol{q}}^{\tau+1})\pi(\boldsymbol{q}^\tau)$ et (ii) $T(\tilde{\boldsymbol{q}}^{\tau+1}, d\boldsymbol{q}^\tau)\pi(\tilde{\boldsymbol{q}}^{\tau+1})$ où $\pi(\boldsymbol{q}^\tau) = \mathfrak{N}^{-1}e^{-\beta V(\boldsymbol{q}^\tau)}$ est la mesure canonique ($\mathfrak{N}$ est une constante de normalisation). On calcule ensuite le coefficient $r(\boldsymbol{q}^\tau, \tilde{\boldsymbol{q}}^{\tau+1})$ défini de la façon suivante :

$$r(\boldsymbol{q}^\tau, \tilde{\boldsymbol{q}}^{\tau+1}) = \min\left(1, \frac{T(\tilde{\boldsymbol{q}}^{\tau+1}, d\boldsymbol{q}^\tau)\pi(\tilde{\boldsymbol{q}}^{\tau+1})}{T(\boldsymbol{q}^\tau, d\boldsymbol{q}^{\tau+1})\pi(\boldsymbol{q}^\tau)}\right) \tag{5.10}$$

Dans le cas d'un processus réversible c'est-à-dire : $T(\boldsymbol{q}, d\boldsymbol{q'})d\boldsymbol{q} = T(\boldsymbol{q'}, d\boldsymbol{q})d\boldsymbol{q'}$, alors l'équation (5.10) se réduit à l'expression suivante :

$$r(\boldsymbol{q}^\tau, \tilde{\boldsymbol{q}}^{\tau+1}) = \min\left(1, \pi(\tilde{\boldsymbol{q}}^{\tau+1})/\pi(\boldsymbol{q}^\tau)\right) \tag{5.11}$$

On reconnait ici le ratio des poids de Boltzmann associé au deux points de l'espace de coordonnées qui peut être interprété comme une probabilité de passage de $\boldsymbol{q}^\tau$ à $\tilde{\boldsymbol{q}}^{\tau+1}$. On effectue un tirage aléatoire d'une variable $U^\tau \overset{\text{loi}}{\sim} \mathcal{U}(0,1)$ de loi uniforme sur $[0,1]$. Si $U^\tau \leq r(\boldsymbol{q}^\tau, \tilde{\boldsymbol{q}}^{\tau+1})$, on accepte le point $\tilde{\boldsymbol{q}}^{\tau+1}$ sinon on garde le point $\boldsymbol{q}^\tau$. On peut remarquer que les points $\boldsymbol{q}^{\tau+1}$ ayant une valeur d'énergie inférieure à $\boldsymbol{q}^\tau$ sont systématiquement acceptés (dans le cas où le noyau de transition T est symétrique). On ré-itère ensuite la procédure et, dans la limite d'un grand nombre de tirages, la valeur de la moyenne empirique des valeurs d'énergie explorées du système sera la valeur réelle de cette observable vis-à-vis de la mesure canonique. Par exemple, le critère de Metropolis-Hasting est très utilisé dans la simulation des sytèmes sur réseaux rigides car il est alors simple de construire la *chaîne de Markov discrète* [208, 209] entre les différents points du réseau. La méthode des *chaînes de Markov discrètes* est néanmoins peu utilisée dans le cas des méthodes d'*énergie libre* où l'on préfère les processus continus [66, 103].

On peut aussi exprimer le *théorème ergodique* à l'aide de *processus de Markov continus*. Dans le cadre des processus continus, la dynamique engendrée peut prendre n'importe quelle valeur de coordonnées et peut donc échantillonner l'espace des configurations. La *dynamique de Langevin* vérifie les hypothèses du *théorème ergodique* et est décrite par les équations d'évolution suivantes pour un système de N atomes :

$$\left\{\begin{array}{rcl} d\boldsymbol{q} & = & \boldsymbol{M}^{-1}\cdot\boldsymbol{p}\,dt \\ d\boldsymbol{p} & = & -\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q})\,dt - \gamma\boldsymbol{M}^{-1}\cdot\boldsymbol{p}\,dt + \sqrt{2\gamma\beta^{-1}}\,d\boldsymbol{W} \end{array}\right. \tag{5.12}$$

Ici, $\boldsymbol{M} \in \mathbb{R}^{3N\times 3N}$ est la matrice de masse du système, $\boldsymbol{\nabla}_{\boldsymbol{q}}$ est l'opérateur gradient par rapport aux coordonnées, γ est le coefficient de viscosité. $\boldsymbol{W} \in \mathbb{R}^{3N}$ est un processus de Wiener tel que $\forall t_n, t_{n+1}$ indépendants, $\int_{t_n}^{t_{n+1}} d\boldsymbol{W} \overset{\text{loi}}{\sim} \boldsymbol{\mathcal{N}}(0, \sqrt{t_{n+1}-t_n})$ où $\boldsymbol{\mathcal{N}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{t_n}^{t_{n+1}})$

est une loi normale multi-dimensionnelle de moyenne nulle et dont la matrice de covariance vérifie $\mathbf{\Sigma}_{t_n}^{t_{n+1}} = (t_{n+1} - t_n)\mathbf{1}_{3N}$. On peut démontrer que cette dynamique vérifie les définitions (5.1) et (5.2) [66]. La *dynamique de Langevin* permet de générer une trajectoire continue dans l'espace des configurations et est largement utilisée dans les méthodes d'évaluation de l'*énergie libre*. Ces méthodes d'échantillonnage efficace de l'espace des phases sont utilisées dans le cadre de méthodes méscopiques décrites dans le chapitre 1. Nous allons appliquer une approche de type *Langevin* pour la suite de nos calculs.

Les deux méthodes d'échantillonnage de la mesure canonique impliquent la variation d'énergie du système entre deux points de coordonnées généralisées ($\boldsymbol{q}$). Dans le cas d'un paysage énergétique simple (un seul bassin par exemple) l'échantillonnage va être relativement aisé mais dans le cadre d'un paysage énergétique complexe, la distribution Boltzmannienne va **présenter un caractère multi-modal** ce qui va rendre difficile l'échantillonnage. **Plus la distribution Boltzmanienne s'éloigne d'une distribution simple -par exemple Gaussienne - plus l'échantillonnage sera long**. Même dans le cadre de systèmes *modèles* tels que des doubles puits de potentiels en deux dimensions [66, 210], un échantillonnage direct par la dynamique de *Langevin* ou par l'algorithme de Metropolis-Hasting présente des convergences très lentes - dans le cas optimal, la complexité numérique de ces méthodes est de l'ordre de $\mathcal{O}(N)$ avec N le nombre d'atomes du système - et peut aboutir à un mauvais échantillonnage. Afin d'effectuer un échantillonnage optimum il nous faut donc modifier - biaiser - la distribution Boltzmanienne afin de la rendre plus simple à échantillonner. Ces méthodes *biaisantes adaptatives* sont aujourd'hui grandement utilisées dans le domaine de l'*énergie libre*. Dans ces méthodes, on ajoute un **potentiel fictif** $\mathfrak{V}(\boldsymbol{q})$ (resp. une **force fictive**) au potentiel réel $V(\boldsymbol{q})$ (resp. aux forces réelles). Le potentiel biaisant resp. la force biaisante a pour but d'"aplanir" le paysage énergétique afin de pouvoir passer d'un bassin à un autre durant l'échantillonnage. Il faut ensuite dé-biaiser le calcul. Nous allons décrire les deux grandes méthodes *biaisantes adaptatives* : (i) les méthodes dites ABP (Adaptative Biasing Potential) et (ii) les méthodes dites ABF (Adaptative Biasing Forces).

5.2.2 Notion de *coordonnée de réaction*

Comme il a été décrit dans la sous-section précédente 5.2.1 l'échantillonnage de la mesure canonique dans un espace de grande dimension $\mathcal{Q} \in \mathbb{R}^{3N}$, avec N le nombre d'atomes dans le système, représente la plus grande difficulté d'évaluation de l'*énergie libre*. L'échantillonnage uniforme direct est irréaliste et des méthodes employant des *dynamiques stochastiques* sont nécessaires. Néanmoins, la vitesse de convergence de ces méthodes est aussi grandement liée à la dimension de l'espace à échantillonner. Les méthodes d'*énergie libre* utilisent la notion de *coordonnée de réaction* afin de **réduire la dimensionalité** du problème. On s'intéresse à calculer la différence d'*énergie libre* entre un état initial $\mathcal{S}_i$ et un état final $\mathcal{S}_f$. On définit un état (minimal de coordonnées

$\boldsymbol{q}_{\mathcal{S}}$) $\mathcal{S}$ de la façon suivante. On introduit la *dynamique dissipative* suivante qui va minimiser l'énergie potentielle du système $V(\boldsymbol{q})$ [57] et vérifiant $\forall t \geq 0$:

$$\begin{aligned} \phi^t_{\mathcal{D}} : \mathbb{R}^{3N} \times \mathbb{R}^{3N} &\to \mathbb{R}^{3N} \times \mathbb{R}^{3N} \\ (\boldsymbol{q}_0, \boldsymbol{p}_0) &\to (\boldsymbol{q}(t) + d\boldsymbol{q}(t), \boldsymbol{p}(t) + d\boldsymbol{p}(t)) \end{aligned}$$

$$\begin{cases} d\boldsymbol{q}(t) = -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}(t))\, dt \\ d\boldsymbol{p}(t) = \mathbf{0} \end{cases} \tag{5.13}$$

On peut alors décrire un état $\mathcal{S}$ d'énergie minimale et de coordonnées $\boldsymbol{q}_{\mathcal{S}}$ de la façon suivante :

$$\mathcal{S} = \left\{ \boldsymbol{q_i} \in \mathbb{R}^{3N}, \boldsymbol{p_i} = \mathbf{0} \in \mathbb{R}^{3N} | \lim_{t \to +\infty} \phi^t_{\mathcal{D}}(\boldsymbol{q_i}, \boldsymbol{p_i}) = (\boldsymbol{q}_{\mathcal{S}}, \mathbf{0}) \right\} \tag{5.14}$$

La notion d'état permet de definir des partitions de l'espace des phases et de calculer de façon formelle des fonctions de partition. Ainsi, dans le cas simple de la différence d'*énergie libre* entre $\mathcal{S}_i$ et $\mathcal{S}_f$ on a :

$$\Delta F(\mathcal{S}_i \to \mathcal{S}_f) = -\beta^{-1} \ln \left(\frac{Z(\mathcal{S}_f)}{Z(\mathcal{S}_i)} \right) \tag{5.15}$$

Ici $Z(\mathcal{S}_i)$ et $Z(\mathcal{S}_f)$ sont respectivement les fonctions de partitions à l'état $\mathcal{S}_i$ et à l'état $\mathcal{S}_f$ (cf. Annexe (A)). On appelle *coordonnée de réaction* l'application $\xi : \mathbb{R}^{3N} \to \mathbb{R}^l$ permettant de construire la transformation entre l'état $\mathcal{S}_i$ et l'état $\mathcal{S}_f$. Le cas le plus simple de *coordonnée de réaction* est la coordonnée dite alchimique, qui correspond au cas $l = 1$. Les méthodes alchimiques peuvent être utilisées pour calculer des grandeurs thermodynamiques telles que des *énergies libres* ou des transitions de phases [103]. Un exemple simple d'utilisation d'une coordonnée alchimique est la différence d'*énergie libre* entre un système modèle $\mathcal{M}$ de potentiel $V_{\mathcal{M}}$ et le système réel $\mathcal{R}$ de potentiel $V_{\mathcal{R}}$. On peut alors poser $\xi : \mathbb{R}^{3N} \to [0, 1]$ tel que $\xi(\mathcal{M}) = 0$ et $\xi(\mathcal{R}) = 1$. Le choix d'un potentiel mixé $V(z) = (1 - z)V_{\mathcal{M}} + zV_{\mathcal{R}}$ permettra d'échantillonner la différence d'*énergie libre* entre le système modèle et le système réel. On peut par exemple calculer l'*énergie libre* associée à un champ de force en utilisant ce potentiel comme $V_{\mathcal{R}}$ et en utilisant le *Hamiltonien* d'un modèle de type Einstein pour les phonons comme $V_{\mathcal{M}}$. Dans ce cas, on peut calculer analytiquement la valeur d'*énergie libre* du modèle Einstein.

Dans le cas général, la *coordonnée de réaction* définit le chemin de transition entre l'état initial et l'état final du système. Ainsi, pour $l > 0$, on définit une nouvelle observable d'*énergie libre* pour la valeur z de la *coordonnée de réaction* ξ :

$$F(z) = -\beta^{-1} \ln \left(\int_{\Sigma(z)} e^{-\beta V(\boldsymbol{q})} \delta_{\xi(\boldsymbol{q})-z}(d\boldsymbol{q}) \right) \tag{5.16}$$

Ici, $\Sigma(z) = \{\boldsymbol{q} \,|\, \xi(\boldsymbol{q}) = z\}$ est la variété de dimension l définie par l'ensemble des coordonnées du système pour la valeur z de $\xi(\boldsymbol{q})$. La mesure $\delta_{\xi(\boldsymbol{q})-z}(d\boldsymbol{q})$ peut être exprimée à l'aide de la formule de la co-aire [66] :

$$\delta_{\xi(\boldsymbol{q})-z}(d\boldsymbol{q}) = \det \left([\boldsymbol{\nabla}_{\boldsymbol{q}} \xi]^T \cdot \boldsymbol{\nabla}_{\boldsymbol{q}} \xi \right)^{-\frac{1}{2}} \sigma_{\Sigma(z)}(d\boldsymbol{q}) \tag{5.17}$$

où, $\sigma_{\Sigma(z)}(d\boldsymbol{q})$ est la mesure de comptage sur le sous-espace $\Sigma(z)$ telle que $\sigma_{\Sigma(z)}(d\boldsymbol{q}') = 1$ si $\boldsymbol{q}'$ appartient à $\Sigma(z)$ et 0 sinon et $\boldsymbol{\nabla}_{\boldsymbol{q}}\xi \in \mathbb{R}^{3N\times l}$ est la matrice de Jacobienne de la *coordonnée de réaction* par rapport aux coordonnées généralisées $\boldsymbol{q}$.

La construction d'une *coordonné de réaction* "optimale" au sens où elle permet d'approximer avec une précision correcte la différence d'*énergie libre* dans un temps de calcul raisonnable est aujourd'hui encore un grand défi des méthodes d'*énergie libre*. La méthode des *strings* [107-110] permet de construire la *coordonnée de réaction* grâce aux chemins d'énergies minimales. Dans le domaine de construction "à la volée" de *coordonnées de réaction* en température, nous citons le travail de Swinburne et Marinica [111] permettant le calcul de différences d'*énergie libre* pour des systèmes de grandes tailles telles que des lignes de dislocation. Les avancées récentes dans le domaine du *Machine Learning* permettent la construction directe de la *coordonnée de réaction* dans l'espace de représentation des données [113] ou utilisent directement des réseaux de neurones artificiels [112, 114, 211, 212]. On peut alors distinguer deux grandes approches concernant les méthodes d'*énergie libre* utilisant des réseaux de neurones. La première approche est de construire un espace de représentation du paysage énergétique où la densité de probabilité de l'*énergie libre* du système est une distribution simple (ici une gaussienne multi-dimensionnelle) [114, 212]. Dans ce cas, l'échantillonnage se fait dans l'espace de représentation et les moyennes sont calculées dans l'espace des configurations grâce à des réseaux de neurones inversibles. L'autre grande approche est la construction automatique de *coordonnées de réaction* [112, 211] à l'aide d'auto-encodeur. Le "bottleneck" de l'auto-encodeur est alors la *coordonnée de réaction*. L'efficacité de ces méthodes reste tout de même cantonnée à des systèmes "simples" et la **construction systématique** de *coordonnées de réaction* pour une transformation dans un système quelconque reste au coeur de la recherche sur les méthodes d'*énergie libre*.

5.2.3 Méthodes à biais adaptatifs présentes dans la littérature

Dans cette sous-section, nous allons présenter les deux principales méthodes à biais adaptatifs utilisées dans la littérature afin d'estimer la variation d'*énergie libre* entre deux états d'un système. Nous commençons par décrire les méthodes de type **potentiels biaisants adaptatifs** Sec. 5.2.3 (Adaptative Biasing Potential) : ces méthodes agissent directement sur le paysage d'énergie potentielle du système. Dans un deuxième temps, nous allons décrire les méthodes à **forces biaisantes adaptatives** Sec. 5.2.3 (Adaptative Biasing Forces) : ce type de méthodes modifie la dynamique stochastique et permet un échantillonnage plus rapide de l'énergie libre. La différence d'*énergie libre* est ensuite reconstruite par intégration de la force biaisante par rapport à la *coordonnée de réaction*.

Méthodes à potentiels biaisants adaptatifs (ABP)

Nous commençons par décrire la méthode ABP (Adaptative Biasing Potential). Comme il a été décrit dans la sous-section précédente (5.2.1), l'échantillonnage de la mesure canonique d'un système est conditionné par le potentiel vu par ce même système. Si des bassins sont trop attractifs, ils deviennent de véritables pièges pour la dynamique d'exploration. Si la trajectoire générée par la dynamique reste bloquée dans un bassin, la valeur de l'observable calculée sera fausse car la dynamique ne sera plus *irréductible* au sens de la définition 5.2. Les méthodes ABP consistent à ajouter un potentiel biaisant $\mathfrak{V}(\boldsymbol{q})$ pour un point de l'espace $\boldsymbol{q}$. Ce potentiel est construit de façon itérative au fur et à mesure de la dynamique générée. Cette méthode permet de construire un potentiel "miroir" à celui du système, qui dans l'idéal serait $\forall \boldsymbol{q},\ \mathfrak{V}(\boldsymbol{q}) = -V(\boldsymbol{q})$. Dans cette situation la mesure canonique sera simple à échantillonner car le paysage énergétique du système sera "plat". La méthode ABP est illustrée par la figure 5.1 issue de Bussi *et al.* [102] : on voit la construction itérative du potentiel biaisant $\mathfrak{V}(\boldsymbol{q})$ en miroir du potentiel réel.

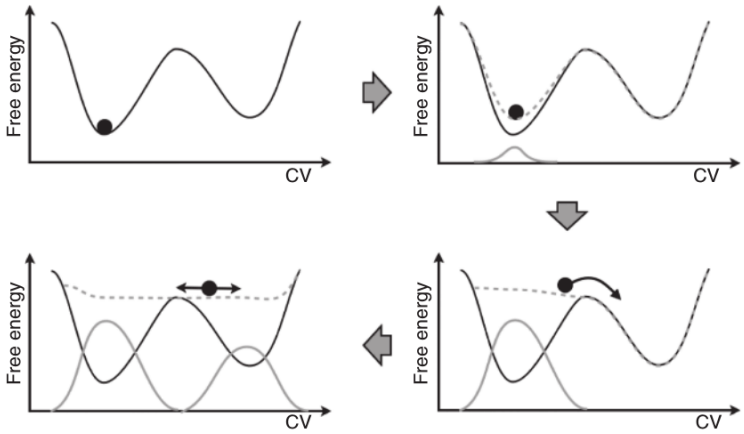

Figure 5.1: Construction itérative du potentiel biaisant $\mathfrak{V}(\boldsymbol{q})$ afin d'échantillonner le profil d'*énergie libre* le long d'une *coordonnée de réaction* d'après Bussi *et al.* [102]. Le potentiel réel est représenté en gris foncé et le potentiel biaisant en gris clair. Le potentiel biaisant tend à devenir le "miroir" du potentiel réel.

Considérons $\xi : \mathbb{R}^{3N} \to \mathbb{R}^l$ la *coordonnée de réaction* associée la transformation du système entre son état initial et son état final et introduisons le **temps numérique** t associé à la procédure d'estimation de l'*énergie libre*. Dans le cas le plus simple de la méthode ABP, on pose $\mathfrak{V}_t(\boldsymbol{q}, z) = \tilde{F}_t(\xi(\boldsymbol{q}))$ et dépendant du temps numérique t. Posons la mesure suivante, $\psi_t(\boldsymbol{q}) = Z_{\psi_t}^{-1} e^{-\beta\left[V(\boldsymbol{q}) - \tilde{F}_t(\xi(\boldsymbol{q}))\right]}$ avec Z_{ψ_t} la fonction de partition associée à ψ. Nous introduisons alors la mesure $\psi_t^{\xi}(\boldsymbol{q}, z)$ construite à l'aide de la formule de la co-aire Eq. (5.17) et vérifiant $\psi_t^{\xi}(\boldsymbol{q}, z)d\boldsymbol{q} = \psi_t(\boldsymbol{q})\delta_{\xi(\boldsymbol{q})-z}(d\boldsymbol{q})$. On

veut alors construire une estimation de l'observable $F(z)$ - Eq. (5.16) - en fonction du temps numérique t et que l'on note $F_t(z)$. On pose alors :

$$\frac{dF_t(z)}{dt} = -\beta^{-1} \ln \left(\int_{\Sigma(z)} \psi_t^{\xi}(\boldsymbol{q}, z) d\boldsymbol{q} \right) \tag{5.18}$$

Dans le cas d'un équilibre instantané, l'équation (5.18) assure que $F_t(z)$ converge exponentiellement vers $F(z)$ l'*énergie libre* - pour la valeur de *coordonnée de réaction* $\xi(\boldsymbol{q}) = z$ - du système à une constante additive près [66]. La méthode ABP converge et permet donc de calculer la différence d'*énergie libre* entre deux états du système dans ce cas simple. Intuitivement, on comprend que l'utilisation d'un potentiel biaisant ne dépendant pas des coordonnées $\boldsymbol{q}$ n'est pas optimal pour la convergence des méthodes d'*énergie libre* dans le cas d'un paysage énergétique complexe. En effet, le paysage énergétique sera "aplani" en moyenne pour un potentiel biaisant constant et l'échantillonnage de la mesure canonique pourra encore poser des problèmes. Les méthodes ABP présentes dans la littérature introduisent une dépendance en $\boldsymbol{q}$ permettant de créer un potentiel en "miroir" comme présenté en Figure 5.1. La procédure décrite dans l'équation (5.18) présente le grand intérêt de donner une esquisse de la convergence de la méthode dans ce cas de figure simple. On retiendra le schéma de mise à jour de l'estimation de l'*énergie libre* $F_t(z)$ issu de Lelièvre *et al.* [66] (bien qu'il est en existe d'autres dans la littérature [102]) :

$$\begin{cases} d\boldsymbol{q}_t = -\boldsymbol{\nabla}_{\boldsymbol{q}} \left(V(\boldsymbol{q}) - F_t(\xi(\boldsymbol{q})) \right) dt + \sqrt{2\beta^{-1}} d\boldsymbol{W}_t \\ \frac{dF_t(z)}{dt} = -\beta^{-1} \ln \left(\int_{\Sigma(z)} \psi_t^{\xi}(\boldsymbol{q}, z) d\boldsymbol{q} \right) \end{cases} \tag{5.19}$$

La méthode ABP la plus utilisée dans la littérature est la métadynamique développée par Laio *et al.* [100] et la démonstration de la convergence de cette méthode est donnée par Bussi *et al.* [101]. Dans cette méthode, le potentiel biaisant consiste en une somme de distributions Gaussiennes multi-dimensionnelles. La construction itérative du potentiel biaisant et sa convergence vers l'*énergie libre* du système sont illustrées par la figure 5.1. D'autres méthodes utilisant des potentiels biaisants adaptatifs sont présentées dans la littérature, on citera notamment la méthode *Wang-Landau* [99].

Méthodes à forces biaisantes adaptatifs (ABF)

De même que les méthodes ABP, les méthodes ABF tendent à pénaliser - en augmentant leur énergie - les portions du paysage énergétique du système ayant **déjà été explorées** par la dynamique stochastique. Dans ce cas, le biais est directement ajouté sur les forces appliquées au système. Si on note $\Gamma(z)$ la force biaisante appliquée au système dans l'espace des phases pour la valeur de *coordonnées de réaction* $\xi(\boldsymbol{q}) = z$. On peut alors écrire l'expression de $\Gamma(z)$ optimale pour le problème d'échantillonnage de l'*énergie libre* :

$$\Gamma_t(z) = \frac{\int_{\Sigma(z)} f(\boldsymbol{q}) \psi_t^{\xi}(\boldsymbol{q}, z) d\boldsymbol{q}}{\int_{\Sigma(z)} \psi_t^{\xi}(\boldsymbol{q}, z) d\boldsymbol{q}} \tag{5.20}$$

où $f(\boldsymbol{q})$ est la force par rapport à la *coordonnée de réaction* dont les composantes $f_i(\boldsymbol{q})$ vérifient :

$$f^i_{\xi(\boldsymbol{q})=z}(\boldsymbol{q}) = \sum_{j=1}^{l} \left(\nabla_q \xi_i \nabla_q \xi_j\right)^{-1} \boldsymbol{\nabla}_{\boldsymbol{q}} \xi_j \cdot \boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}) + \beta^{-1} \mathrm{div} \left(\sum_{j=1}^{l} \left(\nabla_q \xi_i \nabla_q \xi_j\right)^{-1} \boldsymbol{\nabla}_{\boldsymbol{q}} \xi_j \right) \tag{5.21}$$

Ici, div$(\cdot)$ est l'opérateur divergence. La valeur d'*énergie libre* nécessaire dans l'expression de $\psi_t(\boldsymbol{q})$ est alors obtenue par intégration de la force biaisante Γ_t le long de la *coordonnée de réaction* - entre la valeur $\xi(\boldsymbol{q}) = 0$ et $\xi(\boldsymbol{q}) = z$ -, c'est-à-dire :

$$\tilde{F}_t(z) = \tilde{F}_t(0) + \int_0^1 \Gamma_t(s(\lambda))\dot{s}(\lambda) d\lambda \tag{5.22}$$

Où, $s : [0,1] \to \mathbb{R}^l$ est une fonction lisse décrivant la *coordonnée de réaction* et vérifiant $s(0) = 0$ et $s(1) = z$. $\tilde{F}(z)$ et $\tilde{F}(0)$ sont respectivement les valeurs d'*énergie libre* du système pour $\xi(\boldsymbol{q}) = 0$ et $\xi(\boldsymbol{q}) = z$. Il est important de noter que pour l'équation (5.22) soit valide la force moyenne Γ_t dérive d'un gradient c'est-à-dire : $\boldsymbol{\nabla}_\xi \times \Gamma_t = \mathbf{0}$ où $\boldsymbol{\nabla}_\xi \in \mathbb{R}^{l \times l}$ est l'opérateur rotationnel par rapport à la *coordonnées de réaction*. Si la force moyenne n'est pas conservative *i.e* ne dérive pas d'un gradient il est nécessaire soit : (i) de projeter la force moyenne sur un gradient [213] ou (ii) de résoudre itérativement une équation de type Poisson vérifiée par la force moyenne [214]. Nous retiendrons alors le schéma de mise à jour suivant de la force biaisante $\Gamma(z)$ issu de Lelièvre *et al.* [66] :

$$\begin{cases} d\boldsymbol{q}_t = \left(-\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}) + \sum_{j=1}^{l} \Gamma_t \left[\xi(\boldsymbol{q}_t)\right]_j \boldsymbol{\nabla}_{\boldsymbol{q}} \xi_j(\boldsymbol{q}_t)\right) dt + \sqrt{2\beta^{-1}} d\boldsymbol{W}_t \\ \Gamma_t(z) = \frac{\int_{\Sigma(z)} f(\boldsymbol{q}) \psi_t^\xi(\boldsymbol{q}, z) d\boldsymbol{q}}{\int_{\Sigma(z)} \psi_t^\xi(\boldsymbol{q}, z) d\boldsymbol{q}} \end{cases} \tag{5.23}$$

Une illustration de l'échantillonnage de la mesure $\psi_t^\xi(\boldsymbol{q}, z)$ est donnée dans la figure 5.2. Dans ce cas particulier le **sous-espace** à échantillonner est l'hyperplan de normale $\partial_z \boldsymbol{q}(z)$. Dans le cadre d'une *coordonnée de réaction* de type *Hamiltonien* mixé (**alchimique**), comme présentée dans la section précedente, l'expression (5.20) se réduit à la formulation suivante :

$$\Gamma_t(z) = \frac{\int_{\mathcal{Q}} \left[V_{\mathcal{R}}(\boldsymbol{q}) - V_{\mathcal{M}}(\boldsymbol{q})\right] \psi_t(\boldsymbol{q}, z) d\boldsymbol{q}}{\int_{\mathcal{Q}} \psi_t(\boldsymbol{q}, z) d\boldsymbol{q}} \tag{5.24}$$

Ici, $V_{\mathcal{R}}(\boldsymbol{q})$ et $V_{\mathcal{M}}(\boldsymbol{q})$ sont respectivement les potentiels du système réel $\mathcal{R}$ et au potentiel du modèle $\mathcal{M}$. $\psi_t(\boldsymbol{q}, z) = Z_\psi^{-1} e^{-\beta\left[\mathcal{H}(\boldsymbol{q},z) - \tilde{F}_t(z)\right]}$ est la mesure biaisée associée à l'Hamiltonien alchimique $\mathcal{H}(\boldsymbol{q}, z)$ décrit dans les sections précédentes. Nous utiliserons la méthode ABF alchimique afin de calculer des *énergies libres* de formation dans la suite de ce manuscrit.

Les méthodes de type ABF sont en plein développement depuis les années 2010. Nous avons décrit plus haut le principe général de la méthode ABF mais plusieurs variantes ont été developpées dans la littérature. La méthode ABF étendue (eABF), proposé

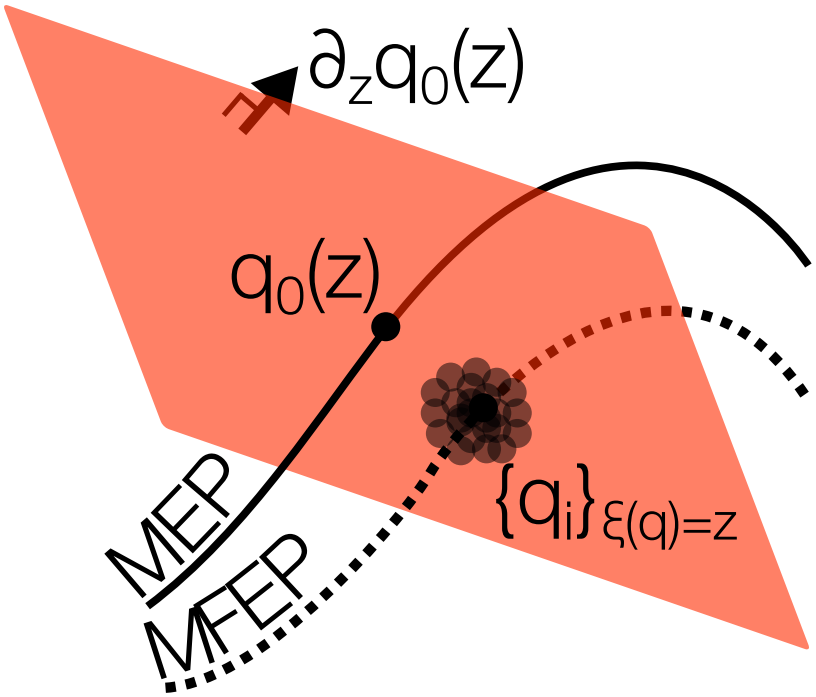

Figure 5.2: Illustration du sous-espace sur lequel est définie la mesure $\psi^{\xi}(\boldsymbol{q}, z)d\boldsymbol{q}$. Dans ce cas particulier, ce sous-espace est réduit à un hyperplan de normale $\partial_z \boldsymbol{q}(z)$. L'illustration issue de *Swinburne* [215] montre que le chemin d'énergie minimale (MEP) peut différer du chemin d'*énergie libre* minimale (MFEP) pour une transformation donnée.

par Lelièvre *et al.* [66, 216], introduit un degré de liberté supplémentaire au système et permet un échantillonnage direct des valeurs de la coordonnées de réaction. La méthode ABF généralisée (gABF), aussi introduite par Lelièvre *et al.* [66], décompose l'expression de la force biasante sur chaque dimension de la *coordonnée de réaction*. Ces deux méthodes peuvent être couplées dans l'ABF généralisée et étendue (egABF) afin de tirer avantage de la vitesse de convergence de la méthode gABF et de la flexibilité d'échantillonnage de la *coordonnée de réaction* de la méthode eABF [217]. La méthode ABF projetée (pABF), traite le problème de la recontruction de l'estimation de l'*énergie libre* par l'intégration de la force biaisante comme un problème de Poisson. La méthode pABF présente une réduction de variance par rapport à la méthode ABP classique, ce qui augmente sa vitesse de convergence [213, 214]. Des méthodes de type Machine Learning ont été utilisées afin de reconstruire l'*énergie libre* de différents systèmes. Cette reconstruction porte directement sur l'observable d'*énergie libre* en utilisant : (i) des réseaux de neurones artificiels [218] ou (ii) des processus Gaussiens [219, 220].

5.3 Méthode à force moyenne : méthode, convergence, difficultés et aspects pratiques

Nous allons nous intéresser, dans cette section, à la convergence de la méthode ABF Bayésienne - développée par Cao *et al.* [103] - dans le cadre du *mélange d'Hamiltonien*

alchimique décrit plus haut (Sec. 5.2.2). En effet, nous allons utiliser cette méthode afin de calculer des propriétés thermodynamiques dans le chapitre suivant. Nous devons donc déterminer avec précision la vitesse de convergence d'une telle méthode afin de réaliser nos calculs. Cette section est organisée de la façon suivante : (i) nous décrivons la méthode Bayésienne dans le cadre alchimique utilisé dans `FEAR` (Sec. 5.3.1) pour le calcul d'*énergie libre de formation* [103] ; (ii) nous décrivons la méthode d'intégration thermodynamique développée dans `PAFI` [111] pour le calcul d'*énergie libre de migration* ; (iii) nous présentons le modèle parallèle de la méthode ABF Bayésienne qui permet de réduire le temps "humain" [2] de calcul (Sec. 5.3.3) ; (iv) nous décrivons les points critiques de la méthode ABF Bayésienne dans le cadre d'un *Hamiltonien alchimique* et nous formulons quelques recommandations pour arriver à la convergence "rapide" de la méthode (Sec. 5.3.4) ; (v) nous exposons un cas pratique d'évaluation de la vitesse de convergence de la méthode pour des grandeurs telles que le paramètre de maille en température et le module élastique en température pour le cas d'un potentiel EAM (Sec. 5.3.5).

Dans le cas d'un échantillonnage optimal, les méthodes de calcul d'*énergie libre de migration* sont plus rapides à converger à cause de la distribution de la variance en énergie. **En effet, dans le cadre d'une migration, seule une partie des atomes contribue grandement au profil d'*énergie libre* ce qui n'est pas le cas pour l'*énergie libre de formation* où tous les atomes du système apportent une contribution importante**. Nous ne débattrons donc que des cas de convergence pour la méthode Bayésienne [103] appliquée au calcul de l'*énergie libre de formation*.

5.3.1 Principe de la méthode : cas de la formation (alchimique)

Les méthodes ABF standards nécessitent l'utilisation d'une *dynamique* pour la coordonnée de réaction comme nous l'avons décrit dans la sous-section 5.2.2. L'*ergodicité* pour la *coordonnée de réaction* est ainsi assurée par une *dynamique jointe*. Pour des bassins peu profonds d'un paysage énergétique, les méthodes standards ne permettent pas l'échantillonnage de ces dits bassins. En effet, les méthodes standards vont avoir tendance à aplanir trop vite certaines portions du paysage énergétique et provoquer le déplacement du système dans un autre bassin. Nous proposons ici de décrire la méthode développée par Cao *et al.* [103] permettant un échantillonnage plus sélectif de la *coordonnée de réaction* tout en conservant l'*ergodicité* de la *dynamique stochastique* pour les coordonnées du système.

Introduisons la notion de probabilité jointe $P_t(\boldsymbol{q}, z)$ pour la coordonnée $\boldsymbol{q}$ du système et la *coordonnée alchimique* z. On peut alors introduire la notion de probabilité marginale par rapport à $\boldsymbol{q}$ notée $P_t(\boldsymbol{q})$ telle que $P_t(\boldsymbol{q}) = \int_{\mathcal{Z}} P_t(\boldsymbol{q}, z)dz$ et par rapport à z, $P_t(z) = \int_{\mathcal{Q}} P_t(\boldsymbol{q}, z)d\boldsymbol{q}$. On peut alors naturellement introduire la notion de **probabilité conditionnelle** reliant la probabilité jointe et les probabilités marginales. On définit

2. Nous définissons le temps "humain" (*wallclock*) comme étant le temps d'éxecution du calcul

ainsi $p_t(\boldsymbol{q}|z)$ la probabilité conditionnelle de la coordonnée $\boldsymbol{q}$ sachant la *coordonnée de réaction* z par l'équation $p_t(\boldsymbol{q}|z) = P_t(\boldsymbol{q}, z)/P_t(z)$. Réciproquement, on peut définir la probabilité conditionnelle de z sachant $\boldsymbol{q}$: $p_t(z|\boldsymbol{q}) = P_t(\boldsymbol{q}, z)/P_t(\boldsymbol{q})$. Dans le cadre de la méthode ABF, la probabilité jointe que nous allons utiliser découle de la mesure $\psi_t(\boldsymbol{q}, z)$ que nous avons définie plus haut Sec. 5.2.2 :

$$P_t(\boldsymbol{q}, z) = \frac{\psi_t(\boldsymbol{q}, z)}{\int_{\mathcal{Q}\times\mathcal{Z}} \psi_t(\boldsymbol{q}, z) d\boldsymbol{q} dz} \tag{5.25}$$

Nous pouvons alors reformuler l'expression de la force biaisante donnée par l'équation (5.20) :

$$\Gamma_t(z) = \int_{\mathcal{Q}} \left[V_{\mathcal{R}}(\boldsymbol{q}) - V_{\mathcal{M}}(\boldsymbol{q})\right] p_t(\boldsymbol{q}|z) d\boldsymbol{q} \tag{5.26}$$

En remarquant que $P_t(z) = \int_{\mathcal{Q}} p_t(z|\boldsymbol{q}) P_t(\boldsymbol{q}) d\boldsymbol{q}$ et en utilisant le **théorème de Bayes** sous la forme suivante : $p_t(z|\boldsymbol{q})P_t(\boldsymbol{q}) = p_t(\boldsymbol{q}|z)P_t(z)$ nous pouvons ré-écrire l'expression donnée par l'équation (5.26) en faisant intervenir $P_t(\boldsymbol{q})$ que nous pouvons échantillonner grâce à une *dynamique stochastique* sur les coordonnées :

$$\Gamma_t(z) = \int_{\mathcal{Q}} \frac{\left[V_{\mathcal{R}}(\boldsymbol{q}) - V_{\mathcal{M}}(\boldsymbol{q})\right] p_t(z|\boldsymbol{q}) P_t(\boldsymbol{q})}{\int_{\mathcal{Q}} p_t(z|\boldsymbol{q'}) P_t(\boldsymbol{q'}) d\boldsymbol{q'}} d\boldsymbol{q} \tag{5.27}$$

L'estimation de l'*énergie libre* $\tilde{F}_t(z)$ est obtenue par une simple intégration $\tilde{F}_t(z) = F(0) + \int_0^z \Gamma_t(z')dz'$ avec $F(0)$ correspondant à l'*énergie libre* du modèle de référence. L'expression de la *dynamique stochastique* associée découle de la distribution $P_t(\boldsymbol{q})$ et se réduit à une *dynamique de Langevin* [103] :

$$d\boldsymbol{q} = \boldsymbol{\nabla}_{\boldsymbol{q}} \left\{\beta^{-1} \ln\left[P_t(\boldsymbol{q})\right]\right\} dt + \sqrt{2\gamma\beta^{-1}}\, d\boldsymbol{W} \tag{5.28}$$

On peut alors montrer que cette dynamique ne dépend que de la probabilité conditionnelle $p_t(z|\boldsymbol{q})$ de telle sorte que :

$$d\boldsymbol{q} = \left[\int_{\mathcal{Z}} \left[-z\boldsymbol{\nabla}_{\boldsymbol{q}} V_{\mathcal{R}}(\boldsymbol{q}) - (1-z)\boldsymbol{\nabla}_{\boldsymbol{q}} V_{\mathcal{M}}(\boldsymbol{q})\right] p_t(z|\boldsymbol{q}) dz\right] dt + \sqrt{2\gamma\beta^{-1}}\, d\boldsymbol{W} \tag{5.29}$$

Ici, $-\boldsymbol{\nabla}_{\boldsymbol{q}} V_{\mathcal{R}}(\boldsymbol{q})$ et $-\boldsymbol{\nabla}_{\boldsymbol{q}} V_{\mathcal{M}}(\boldsymbol{q})$ sont respectivement les forces au point de coordonnée $\boldsymbol{q}$ de l'espace des phases dues au potentiel du système réel $\mathcal{R}$ et les forces dues au potentiel du modèle $\mathcal{M}$. Cette méthode nécessite seulement la détermination - à chaque pas de la *dynamique stochastique* - de la valeur de la probabilité conditionnelle $p_t(z|\boldsymbol{q})$ dont l'expression est donnée par :

$$p_t(z|\boldsymbol{q}) = \frac{\psi_t(\boldsymbol{q}, z)}{\int_{\mathcal{Z}} \psi_t(\boldsymbol{q}, z) dz} \tag{5.30}$$

Cette méthode, basée sur le théorème de Bayes, permet de se passer de la notion de *dynamique jointe* et permet d'obtenir un meilleur échantillonnage de l'espace des phases. En effet, la force utilisée dans la *dynamique de Langevin* est construite par un raisonnement Bayésien et va naturellement se rapprocher de la force du potentiel réel ou bien de la force du potentiel modèle dépendamment de l'échantillonnage déjà effectué et contenu dans l'*énergie libre*. Cette méthode permet de stabiliser certains échantillonnages complexes et sera donc utilisée dans la suite de ce chapitre. La méthode Bayésienne est implémentée dans le code `FEAR` [103].

5.3.2 Principe de la méthode : cas de la migration

Dans les chapitres suivants, nous avons effectué des calculs d'*énergie libre de migration* en utilisant le package `PAFI` développé par Swinburne *et al.* [111]. Nous allons décrire brièvement la méthode utilisée dans cette sous-section bien que plus de détails techniques puissent être trouvés dans l'article précédemment cité.

Dans le cadre de l'évaluation de l'*énergie libre de migration* d'un état initial à un état final donné, l'équation (5.20) n'est pas aussi simple que dans le cas alchimique et peut être exprimée de la façon suivante (dans le cas d'une coordonnée de réaction à une dimension) :

$$\Gamma(z) = \left\langle \frac{\boldsymbol{v}\cdot\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q})}{\boldsymbol{v}\cdot\boldsymbol{\nabla}_{\boldsymbol{q}}\xi} + \beta^{-1}\boldsymbol{\nabla}\cdot\frac{\boldsymbol{v}}{\boldsymbol{v}\cdot\boldsymbol{\nabla}_{\boldsymbol{q}}\xi}\right\rangle_{\psi^{\xi}(\boldsymbol{q},z)} \tag{5.31}$$

Cette formulation est exacte si le vecteur arbitraire $\boldsymbol{v}$ vérifie $\boldsymbol{v}\cdot\boldsymbol{\nabla}_{\boldsymbol{q}}\xi > 0$ [105]. Ce vecteur est colinéaire à la *coordonnée de réaction* à 0 K, $\boldsymbol{\nabla}$ est l'opérateur gradient et $\langle\cdot\rangle_{\psi^{\xi}(\boldsymbol{q},z)}$ représente la moyenne d'ensemble pour la mesure $\psi^{\xi}(\boldsymbol{q},z)$ décrite dans la sous-section (5.2.3). Cette équation fait intervenir l'inverse de la matrice de Gram de la *coordonnée de réaction* par rapport aux coordonnées du système $[\boldsymbol{\nabla}_{\boldsymbol{q}}\xi]^T\cdot\boldsymbol{\nabla}_{\boldsymbol{q}}\xi \in \mathbb{R}^{3N\times 3N}$ dont le coût numérique de la pseudo-inversion évolue comme $\mathcal{O}(Nl^2)$. Cette inversion nécessite de connaître la *coordonnée de réaction* de la transformation. Le package `PAFI` se base sur une *coordonnée de réaction* calculée à $T = 0$ K par la méthode NEB [221-223]. L'algorithme a pour but de construire une *coordonnée de réaction* $\tilde{\xi}(\boldsymbol{q}) = z$ dépendante de la température en partant de la *coordonnée* à 0 K, $\xi(\boldsymbol{q}) = z$. Dans l'espace des phases, on comprend que la *coordonnée de réaction* va évoluer de proche en proche et de façon "continue" avec la température. La construction itérative de l'algorithme de `PAFI` se base sur ces conditions locales d'évolution. Nous allons décrire quantitativement cette construction. On note $\boldsymbol{q_0}(z)$ les coordonnées du système correspondant au *chemin d'énergie minimal* (construit par NEB) pour la valeur z de ξ^0. Cette indexation possible par un paramètre z à une dimension des coordonnées du chemin d'énergie minimale du système $\boldsymbol{q_0}(z)$ est essentielle pour l'algorithme de `PAFI`. On introduit ensuite, $\boldsymbol{q}(z)$ les coordonnées du système pour la valeur z de ξ. Si on note $\boldsymbol{q_T}$ les coordonnées du système à la température T, la *coordonnée de réaction* $\tilde{\xi}(\boldsymbol{q_T}) = z$ est solution du problème de minimisation suivant dans l'espace des phases :

$$\tilde{\xi}(\boldsymbol{q_T}) = \arg\min_{z} \|\boldsymbol{q_T} - \boldsymbol{q_0}(z)\|^2 \tag{5.32}$$

Ainsi, les coordonnées $\boldsymbol{q_T}$ correspondent aux coordonnées les plus proches - au sens de la distance Euclidienne - de celles du *chemin d'énergie minimale* à $0\,K$ pour une valeur z donnée de la *coordonnée de réaction* $\tilde{\xi}$. Cette solution du problème de minimisation Eq. (5.32) peut être exprimée par la relation d'orthogonalité suivante :

$$\partial_z\boldsymbol{q_0}(z)\cdot[\boldsymbol{q_T} - \boldsymbol{q_0}(z)] = 0 \tag{5.33}$$

En utilisant cette relation d'orthogonalité, on peut calculer $\boldsymbol{\nabla}_{\boldsymbol{q}}\xi$ en supposant que celui-ci est toujours colinéaire à $\partial_z \boldsymbol{q_0}(z)$ [3]. Par conservation de la relation (5.33) quelque soit $\delta\boldsymbol{q}$ et quelque soit $\delta\xi$ et par la relation de colinéarité, on peut calculer analytiquement $\boldsymbol{\nabla}_{\boldsymbol{q}}\xi$:

$$\begin{cases} \boldsymbol{\nabla}_{\boldsymbol{q}}\xi|_{\xi=z} = & \frac{\partial_z \boldsymbol{q_0}(z)}{\eta(\boldsymbol{q},T,z)\|\partial_z \boldsymbol{q_0}(z)\|^2} \\ \eta(\boldsymbol{q},T,z) = & 1 - \frac{\partial_z^2 \boldsymbol{q_0}(z)}{\|\partial_z \boldsymbol{q_0}(z)\|^2} \cdot [\boldsymbol{q_T} - \boldsymbol{q_0}(z)] \end{cases} \tag{5.34}$$

La condition de positivité $\partial_z \boldsymbol{q_0}(z) \cdot \boldsymbol{\nabla}\xi|_{\xi=z} > 0$ se traduit maintenant par $\eta(\boldsymbol{q},T,z) > 0$. Cette condition traduit le fait que la *coordonnée de réaction* évolue de proche en proche. Si cette condition est violée, on ne peut pas garantir que le système soit resté dans le bassin métastable de la *coordonnée de réaction* à 0 K et/ou que l'échantillonnage reste efficace. La vérification de ce critère de positivité est une condition nécessaire pour la convergence de la méthode. Finalement, en injectant l'équation (5.34) dans l'équation (5.31), on obtient la formulation suivante :

$$\Gamma(z) = \left\langle \eta(\boldsymbol{q},T,z)\, \partial_z \boldsymbol{q_0}(z) \cdot \boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}) + \beta^{-1}\partial_z \ln \frac{|\eta(\boldsymbol{q},T,z)|}{|\partial_z \boldsymbol{q_0}(z)|} \right\rangle_{\psi^{\xi}(\boldsymbol{q},z)} \tag{5.35}$$

L'équation (5.35) ne fait plus intervenir la matrice de Gram de $\xi(\boldsymbol{q})$. On peut maintenant construire, de façon itérative, la *coordonnée de réaction* pour une température T donnée et évaluer l'*énergie libre de migration* par l'intégration de la force moyenne. Pour chaque mise à jour de la *coordonnée de réaction*, on vérifie la **condition de positivité** et si elle n'est pas respectée, on recommence l'échantillonnage. Des détails techniques supplémentaires sont directement donnés dans [111, 224] et nous ne décrirons pas de façon plus exhaustive cette méthode afin de nous consacrer plus en profondeur aux méthodes alchimiques.

Notons que la méthode développée par Swinburne et Marinica se base sur la construction numérique de la *coordonnée de réaction* - à une dimension - correspondant au chemin d'énergie minimale. Cette construction est effectuée par une interpolation de fonction *splines* d'ordre 3 pour l'ensemble des atomes du système. La *coordonnée de réaction* ainsi construite permet de calculer numériquement la quantité $\partial_z \boldsymbol{q_0}(z)$. L'efficacité de cette méthode a été montrée pour de nombreux systèmes [111]. Néanmoins, dans le cas de systèmes et de transitions impliquant une évolution rapide de $\partial_z \boldsymbol{q}(z)$, l'interpolation par les fonctions *splines* ne permettra pas d'assurer les conditions de réalisation de l'équation (5.35) et ne pourra donc pas être utilisée.

3. Dans ce cas, tous les déplacements perpendiculaires $\delta\boldsymbol{q}^{\perp}$ vérifient $\delta\xi = \boldsymbol{\nabla}\xi \cdot \delta\boldsymbol{q}^{\perp} = 0$ grâce à l'équation (5.33) et laissent donc inchangé ξ

5.3.3 Parallélisation de la méthode ABF Bayésienne alchimique

Comme nous l'avons décrit plus haut, les méthodes d'*énergie libre* basées sur les biais adaptatifs sont **coûteuses** en temps numérique. En effet, le *théorème ergodique* donne une convergence de ces méthodes dans la limite où le nombre de pas de dynamique stochastique tend vers $+\infty$. Sans surprise, il faudra donner un nombre de pas finis à nos simulations et nous n'aurons donc qu'une estimation $\tilde{F}$ de l'*énergie libre* du système. La question est la suivante : notre estimation $\tilde{F}$ est-elle proche de la valeur réelle de l'*énergie libre* F ?
Si on considère que l'on est au pas J de la dynamique stochastique et que les événements sont indépendants ; le théorème Central Limite nous donnent une estimation de l'erreur commise entre l'estimation $\tilde{F}^J$ de l'*énergie libre* au pas J et de la valeur réelle de l'*énergie libre* F :

$$|F - \tilde{F}^N| = \mathcal{O}\left(\frac{1}{\sqrt{J}}\right) \tag{5.36}$$

La **vitesse de convergence** de ces méthodes est donc **faible** ce qui nécessite d'effectuer un grand nombre d'itérations. Nous proposons d'utiliser une méthode déja développée par Lelièvre *et al.* [225] et Raiteri *et al.* [226] et appelée "méthode des répliques parallèles". L'idée est simple : on crée L répliques du même système auxquelles on applique la dynamique stochastique. Ces L répliques vont alors évoluer et explorer le paysage énergétique du système de façon indépendante. Après τ' pas de dynamique stochastique pour chaque système, on calcule la nouvelle force biaisante de la façon suivante :

$$\Gamma_{\tau'}(z) = \frac{\sum_{l\leq L}\sum_{\tau\leq\tau'}\left[\mathcal{H}_{\mathcal{R}}(\boldsymbol{q}) - \mathcal{H}_{\mathcal{M}}(\boldsymbol{q})\right] p(z|\boldsymbol{q}_{\tau}^{l})}{\sum_{l\leq L}\sum_{\tau\leq\tau'} p(z|\boldsymbol{q}_{\tau}^{l})} \tag{5.37}$$

où $\boldsymbol{q}_{\tau}^{l}$ sont les coordonnées de la réplique l pour le pas de temps τ et où les autres notations ont été décrites dans l'équation (5.27). Cette méthode permet - dans le cas idéal où tous les événements peuvent être considérés comme indépendants - d'obtenir la nouvelle estimation d'erreur suivante entre l'estimateur de l'*énergie libre* après J pas de temps pour L répliques du sytème $\tilde{F}^{J,L}$ et l'*énergie libre* du système F :

$$|F - \tilde{F}^{J,L}| = \mathcal{O}\left(\frac{1}{\sqrt{JL}}\right) \tag{5.38}$$

L'erreur sur l'estimation de l'*énergie libre* se comporte donc - dans le cas idéal - comme si on avait effectué JL pas de dynamique stochastique sur une réplique du système. **Dans les faits, les trajectoires des différentes répliques ne sont pas indépendantes car elles partagent toutes l'estimation d'*énergie libre* comme biais.** Cette méthode ne permet pas de réduire le temps numérique de calcul mais permet un **véritable gain de *scalabilité***. Cette méthode de parallélisation a été implémentée dans le code `FEAR`. Il nous reste maintenant à déterminer un nombre de pas de dynamique stochastique permettant un bon compromis entre : précision, temps machine et temps "humain".

5.3.4 ABF Bayésienne Alchimique : points critiques de la méthode et recommandations

Le premier choix "arbitraire" - et non des moindres - est le **nombre de pas de dynamique stochastique** que l'on doit fixer pour une simulation. Comme il a été décrit plus haut, celui-ci est un compromis entre **précision, temps machine et temps "humain"**. L'équation (5.38) permet d'estimer un ordre de grandeur du nombre de pas pour une précision donnée. Néanmoins, cette estimation est très approximative d'un point de vue quantitatif et varie grandement d'un système à l'autre. Il est donc nécessaire d'adopter une démarche heuristique [4] pour déterminer le meilleur couple J et L pour calculer une bonne estimation de l'*énergie libre*.

Une deuxième contrainte s'impose à nous si nous voulons calculer des propriétés thermodynamiques à l'équilibre. Pour une valeur de température donnée T, un système va avoir un volume d'équilibre $V^{eq}(T)$. De façon générale, on sait que ce volume va avoir une dépendance par rapport à la température mais on ne connait pas cette dépendance *a priori*. Afin de calculer des propriétés thermodynamiques, il faut déterminer l'expression de $\mathcal{V}^{eq}(T)$ de façon itérative (une méthode sera décrite dans la sous-section suivante (5.3.5)) **en partant d'un choix *a priori* de la dépendance du volume d'équilibre en fonction de la température**. Ce choix pourra se faire de façon arbitraire ou en fonction des données expérimentales. Plus le choix *a priori* de $\mathcal{V}^{eq}(T)$ sera proche de la réalité, plus la procédure itérative convergera rapidement. En termes de conditionnement, plus un champ de force présentera un volume d'équilibre $\mathcal{V}^{eq}(T)$ proche des données expérimentales, plus la procédure itérative sera rapide. Nous reviendrons sur ce point dans le chapitre suivant.

La méthode de *mixage d'Hamiltonien alchimique* nécessite un autre choix arbitraire : celui du **modèle de référence**. On choisit en général un potentiel inter-atomique dont l'expression est simple et souvent analytique. On pourra citer l'utilisation de cristaux harmoniques de type Einstein par Frenkel *et al.* [227] et l'utilisation de potentiels de type Morse par Glensk *et al.* [228]. Ces potentiels permettent de déterminer une expression analytique des forces appliquées à chaque atome et la valeur de leur énergie libre en fonction de la température. Dans le code `FEAR`, le choix du potentiel de référence est un potentiel de type Einstein. L'*Hamiltonien* de ce système de type Einstein $\mathcal{H}_{ein}$ pour un système de N atomes s'exprime de la façon suivante :

$$\mathcal{H}_{ein}(\boldsymbol{p},\boldsymbol{q}) = \sum_{i=1}^{N} \frac{1}{2m_i}\|\boldsymbol{p}_i\|^2 + \sum_{i=1}^{N} \frac{1}{2} m_i\omega^2 \|\boldsymbol{q}_i - \boldsymbol{q}_i^{eq}\|^2 \tag{5.39}$$

où $\boldsymbol{q}_i^{eq}$ sont les coordonnées d'équilibre de l'atome i et ω est la fréquence d'Einstein du modèle. Le modèle d'Einstein pour les phonons est un cas très simple de modèle harmonique. Tous les phonons locaux sont considérés comme étant sans interaction les uns avec les autres et possédant la même pulsation ω. La fréquence du modèle d'Einstein peut être déterminée directement à l'aide d'un calcul d'*énergie libre*. Ce

4. c'est-à-dire en essayant un grand nombre de couple J, L

calcul est détaillé dans l'annexe A et permet de déduire la fréquence d'Einstein par la relation suivante :

$$F_{ein,pot} = 3(N - N_{corr})\beta^{-1} \ln\left(\frac{2\pi}{\beta\hbar\omega}\right) \tag{5.40}$$

Ici N_{corr} est un terme de correction permettant de prendre en compte des contraintes sur la fonction de partition du système. Par exemple, on peut imposer la position du centre masse du système et/ou empêcher sa rotation. Dans le cas des conditions périodiques, on a $N_{corr} = 3$. Le système doit avoir une taille suffisante pour estimer correctement cette *énergie libre* : la taille minimum est de l'ordre de 100 atomes. Dans certains cas, le calcul direct de cette *énergie libre* sera trop long et on pose $\ln(\omega) = \frac{1}{3N}\sum_\nu \ln(\omega_\nu)$ la valeur de moyenne des log pulsations $\ln(\omega_\nu)$ du système après un calcul harmonique. Le choix de ω doit être précis. En effet, un choix de ω trop petit ou trop grand va **conditionner la "vitesse" et la "qualité" de l'ergodicité** pour la *coordonnée de réaction alchimique* c'est-à-dire la description de la distribution $P(z)$. Dans la limite d'un nombre de pas de *dynamique stochastique* infini, $P(z)$ devrait tendre vers une distribution uniforme $\mathcal{U}([0,1])$. Si ω est mal pré-conditionné, il faudra un échantillonnage plus long pour obtenir une valeur précise de l'estimation de l'*énergie libre*. Nous allons détailler l'ensemble de ces problématiques en nous basant sur un exemple concret dans la sous-section suivante.

Nous illustrons le problème du conditionnement du modèle de type Einstein par un exemple très simple en **une dimension**. Considérons le potentiel suivant centré sur la position d'équilibre q_0 (cette forme de potentiel est grandement inspirée de Hellman *et al.* [229]) :

$$V_{1D}(q, \delta\alpha, \sigma) = \alpha(q - q_0)^2 + \delta\alpha e^{-\frac{(q-q_0)^2}{\sigma^2}} \tag{5.41}$$

Ce potentiel devient purement harmonique dans le cas où $\delta\alpha = 0$. Sinon, il est bruité par une fonction Gaussienne de largeur σ^2. Le terme Gaussien représente donc la partie **anharmonique** de ce potentiel. Que se passe t-il maintenant si nous voulons ajuster un modèle de type Einstein (autour de la postion q_0) avec un tel potentiel ? On cherche une solution simple de la forme $\tilde{V}_{1D}(q) = a(q - q_0)^2 + b$ et vérifiant les conditions suivantes : (i) $\tilde{V}_{1D}(q_0) = V_{1D}(q_0, \delta\alpha, \sigma)$ et (ii) $\tilde{V}''_{1D}(q_0) = V''_{1D}(q_0, , \delta\alpha, \sigma)$. On peut calculer analytiquement (a, b) en fonction de $\delta\alpha, \alpha$ et σ^2. Le potentiel $\tilde{V}_{1D}$ est purement harmonique et est **l'image exacte d'un modèle Einstein ajusté** à l'aide d'un calcul de phonons dans un système à une dimension. Nous choisissons de présenter les résultats obtenus pour $\alpha = 1.0$, $\delta\alpha = 0.1$ et pour un ensemble de valeurs de $\sigma^2 = \{1.0, 0.5, 0.1\}$ dans la figure 5.3.

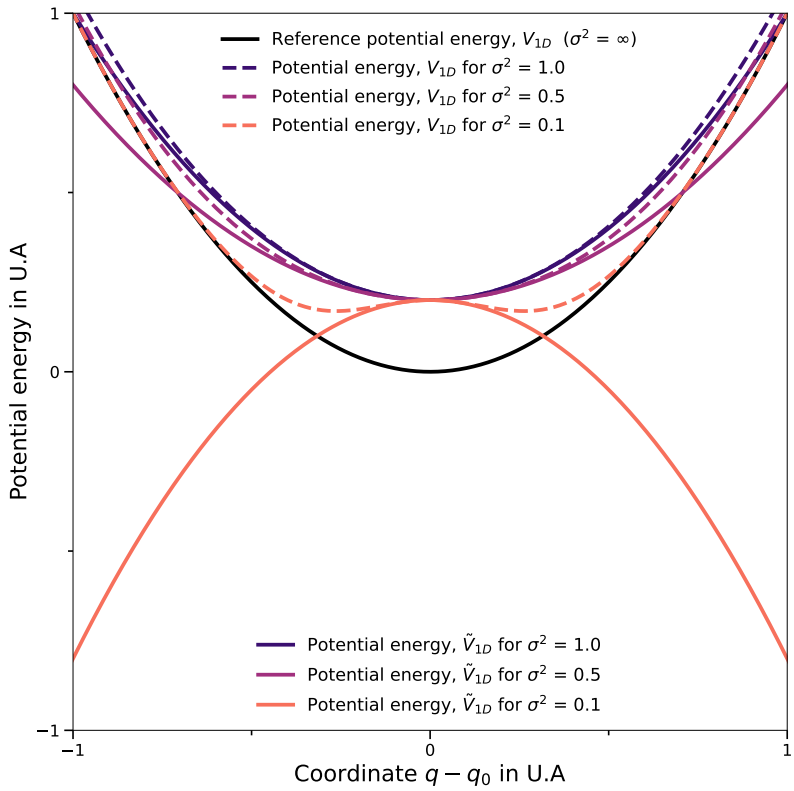

Figure 5.3: Illustration de l'ajustement du potentiel $\tilde{V}_{1D}$ (en traits pleins) en fonction des différentes valeurs de σ^2. Les potentiels V_{1D} sont présentés en pointillés. Le potentiel V_{1D} purement harmonique est présenté en trait plein noir. On constate que pour les faibles valeurs de σ^2, la courbure du potentiel peut être grandement impactée et même changer de signe.

On constate que la forme du potentiel $\tilde{V}_{1D}$ ajusté dépend grandement de la valeur de σ^2. Ainsi, pour des valeurs faibles de σ^2, la courbure du potentiel peut même subir un changement de signe. Il est évident que dans ce genre de **situations pathologiques** l'échantillonnage par une méthode de type *alchimique* sera beaucoup plus difficile à mettre en oeuvre en raison de l'erreur sur la valeur de ω. Il est donc important de bien préconditionner la valeur de ω afin d'obtenir un échantillonnage correct et efficace. Lelièvre *et al.* [66] donne une estimation de la variance dans le cas d'un *Hamiltonien alchimique* linéaire $\mathcal{H}(z) = (1-z)\mathcal{H}_{\mathcal{M}} + z\mathcal{H}_{\mathcal{R}}$. La variance pour l'estimation de l'*énergie libre* $\tilde{F}(z)$ à la valeur de *mixage* z est $\sigma^2(z) = -\beta^{-1}\frac{d^2\tilde{F}}{dz^2}(z)$. On comprend que la valeur de $\frac{d^2\tilde{F}}{dz^2}(z)$ sera plus élevée dans les zones où le potentiel réel et le potentiel modèle sont très différents. Ces zones devront donc être échantillonnées plus longtemps afin de réduire la variance sur l'observable d'*énergie libre*.

Des **situations pathologiques** de préconditionnement du modèle de réference peuvent être évitées en appliquant les recommandations suivantes : (i) éviter l'utilisation de potentiels directement tabulés ; (ii) éviter les potentiels de type EAM ajustés à l'aide de fonctions spline cubiques et (iii) utiliser des potentiels dont les fonctions de bases sont lisses tels que des potentiels MEAM ou Machine Learning. D'un point de vue

plus général, les potentiels utilisés **ne doivent pas être rugueux**. Les potentiels tabulés et ajustés avec des splines cubiques fournissent de **bonnes estimations de l'énergie et des forces d'un système pour un coût numérique faible**. Néanmoins, leur utilisation pour estimer des propriétés à la courbure du paysage énergétique est plus délicate (cf. annexe C).

5.3.5 Cas pratique : calcul du paramètre de maille d'équilibre et du module élastique isostatique à une température finie d'un potentiel EAM

Nous allons détailler un cas concret de l'utilisation de la méthode ABF dans le cadre d'un *Hamiltonien alchimique* pour le calcul de deux propriétés importantes à une température donnée T : (i) la valeur du **paramètre de maille d'équilibre** $a_0(T)$; (ii) la valeur du **module élastique isostatique** $B(T)$. Nous allons donner quelques définitions quantitatives pour ces paramètres en nous plaçant dans le cas d'un système cubique.

Introduisons $\mathcal{V}\{a(T), K\}$, la fonctionnelle du volume d'un système en fonction du paramètre de maille $a(T)$ et du nombre de répliques de la maille élémentaire du système K. Dans le cas des métaux cubiques, on a simplement $\mathcal{V}_{bcc}(a_0(T), K) = Ka(T)^3$. Considérons le tenseur $\boldsymbol{C} = a\mathbf{1} \in \mathbb{R}^{3\times3}$. Ce tenseur correspond à cellule unité d'un réseau cubique de paramètre de maille a. On définit alors l'estimation de l'*énergie libre* pour un système répliqué K fois $\tilde{F}(\boldsymbol{C}, T, K)$ et le paramètre de maille d'équilibre à la température T comme solution du problème de minimisation suivant :

$$a_0(T) = \lim_{K\to+\infty} \arg\min_{\boldsymbol{C}\in\mathbb{R}^{3\times3}} \tilde{F}(\boldsymbol{C}, T, K) \tag{5.42}$$

On peut alors définir l'estimation du module élastique $B(T)$ isostatique d'équilibre à la température T par la relation :

$$B(T) = \lim_{K\to+\infty} \frac{1}{9} \frac{\partial^2 F(\boldsymbol{C}, T, K)}{\partial a(T)^2}\bigg|_{a(T)=a_0(T)} K a_0^{-1}(T) \tag{5.43}$$

Afin d'évaluer de façon quatitative ces deux grandeurs, nous avons utilisé la méthode des *courbes énergie libre/volume*. On se place à un paramètre maille $a_i(T)$ pour une température donnée T. On crée alors L répliques du système, auxquelles on applique un tenseur de déformation homogène isotrope $\boldsymbol{\epsilon}_l = \epsilon_l\,\mathbf{1}$ avec $\epsilon_{min} \leq \epsilon_1 \leq \cdots \leq \epsilon_l \leq \cdots \leq \epsilon_L \leq \epsilon_{max}$. On obtient un ensemble de systèmes dont le volume du système l est $\mathcal{V}\{(1+\epsilon_l)a_i(T)\}$ (pour un système cubique). On peut alors calculer l'*énergie libre* $F(\mathcal{V}\{(1+\epsilon_l)a_i(T)\})$ pour chacun de ces volumes. On construit alors ce qu'on appelle une *courbe énergie libre/volume*. Cette courbe permet de déduire le paramètre de maille d'équilibre et le module élastique isostatique d'équilibre. Le paramètre de maille d'équilibre correspond simplement au minimum de cette courbe qui est lui même solution du problème de minimisation donné par l'équation (5.42). Le module élastique isostatique d'équilibre peut alors être calculé numériquement si la discrétisation en L répliques est suffisament fine. Une illustration de la méthode des *courbes énergie libre/volume* est donnée en Figure 5.4.

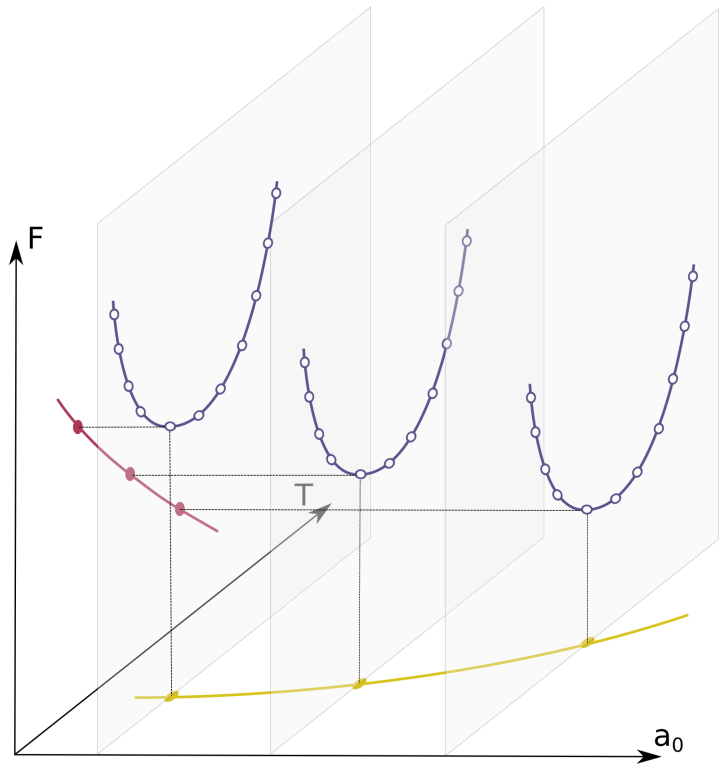

Figure 5.4: Illustration de la méthode des *courbes énergie libre/volume* pour le calcul du paramètre de maille d'équilibre en température $a_0(T)$. Le paramètre de maille d'équilibre en température correspond à la valeur de a_0 tel que l'énergie libre F est minimale pour une température donnée T. De plus, pour la valeur $a_0(T)$ on assure que $P = 0$ et donc que l'*énergie libre* coïncide avec l'*enthalpie libre* du système.

Nous allons appliquer cette méthodologie au cas d'un simple potentiel EAM pour le tungstène, afin de calculer le paramètre de maille d'équilibre et le module élastique isostatique de 0 K à la température de fusion. Cet exemple rapide va permettre de fixer le nombre de pas de dynamique stochastique que nous allons utiliser pour d'autres potentiels afin d'obtenir une bonne convergence de l'*énergie libre* et du paramètre de maille d'équilibre. **En effet, nous nous cantonnons à des métaux de transitions, cubiques centrés et non-magnétiques qui sont relativement simples et similaires**. Nos calculs portent sur une gamme de température allant de 0 K à la température de fusion du matériau T_f. Le cas du tungstène est particulièrement intéressant car sa température de fusion (expérimentale) est de 3695 K. Cette température particulièrement élevée est donc un cas complexe d'application de la *dynamique de Langevin*, où les déplacements seront d'amplitudes très importantes. Pour de tels déplacements, le système peut quitter son bassin d'origine pendant la dynamique ce qui est catastrophique pour l'échantillonnage. Dans le cadre de cette "expérience numérique" nous avons choisi de travailler avec le potentiel développé par Marinica *et al.* [230] (potentiel EAM 4). Nous travaillons avec la méthode ABP

avec un *Hamiltonien mixé alchimique* à l'aide du package `FEAR` [103]. Le potentiel de référence est un potentiel de type Einstein dont la fréquence ω a été qualibrée sur un calcul harmonique complet. Nous utilisons la méthode Bayésienne décrite précedemment en Sec. 5.3.1 et une *dynamique de Langevin* dite *suramortie* ; plus de détails techniques peuvent être trouvés dans Lelièvre *et al.* [66]. Le choix *a priori* des $a(T)$ a été les paramètres de maille d'équilibre expérimentaux du tungstène. Les résultats obtenus pour la convergence de $a_0(T)$ et de $B(T)$ sont présentés dans la figure 5.5 pour un calcul de référence avec 2×10^6 pas de *dynamique de Langevin*. La figure 5.5 (gauche) représente le paramètre de maille d'équilibre en fonction de la température. La figure 5.5 (droite) représente le module élastique isostatique d'équilibre en fonction de la température. On constate qu'après seulement quelques itérations de la méthode des *courbes énergie libre/volume*, ces deux grandeurs convergent. On peut aussi noter que le paramètre de maille d'équilibre de ce potentiel est très différent du paramètre de maille d'équilibre expérimental du tungstène et ne présente pas, par exemple, la dépendance quadratique en température de l'expansion thermique. Nous faisons ici deux remarques : (i) une première d'ordre physique, ce potentiel EAM [230] n'est pas représentatif du comportement du paramètre de maille d'équilibre du tungstène ; (ii) une deuxième d'ordre numérique, la méthode ABF Bayésienne permet d'échantillonner très précisément deux grandeurs nécessitant une très **grande précision numérique** ($a_0(T)$ et $B(T)$).

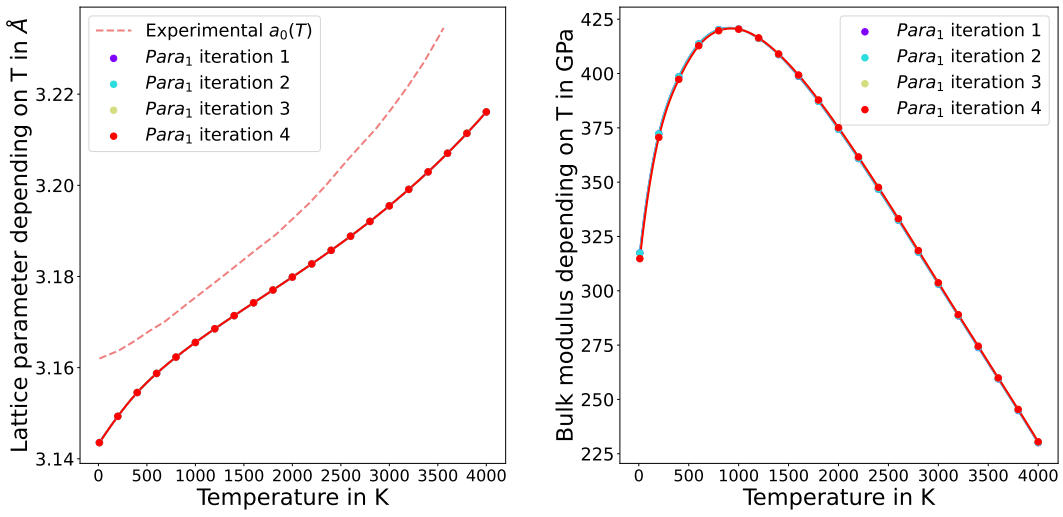

Figure 5.5: Illustration de la convergence du paramètre de maille d'équilibre (à gauche) et du module élastique isostatique d'équilibre (à droite) pour le calcul de référence à ($Para_1$ de la table 5.1) pas de dynamique stochastique. Nous présentons les itérations successives de la méthode des *courbes énergie libre volume* jusqu'à la convergence de $a_0(T)$ et de $B(T)$. La ligne pointillée orange présente l'expansion thermique expérimentale.

La méthode ABF Bayésienne pour un *Hamiltonien alchimique* permet donc déterminer deux propriétés thermodynamiques d'importance que sont $a_0(T)$ et $B(T)$. Ces deux grandeurs sont respectivement des images de la dérivée première $a_0(T)$ et de la dérivée seconde $B(T)$ de l'*énergie libre*, ce qui nécessite une excellente précision de la méthode. Cet exemple concret montre l'efficacité de la méthode ABF. Nous allons maintenant nous intéresser à l'estimation de la vitesse de convergence de cette méthode et à "construire" un jeu de paramètres - nombre de processeurs en parallèle, nombre de pas *Langevin* par processeur - permettant d'obtenir une convergence de ces deux grandeurs pour un "temps humain" et un temps machine raisonnable.

Table 5.1: Liste des différentes parallélisations $(L_{proc}, J_{Langevin})$ utilisées pour effectuer les tests de convergence pour le paramètre de maille d'équilibre et le module élastique isostatique d'équilibre pour le potentiel EAM [230]. L_{proc} représente le nombre de répliques parallèles utilisées pour la simulation et $J_{Langevin}$ le nombre de pas de *dynamique de Langevin* effectués pour chaque réplique.

Parallélisation	Couples $(L_{proc}, J_{Langevin})$		Total $(L_{proc}J_{Langevin})$
	L_{proc}	$J_{Langevin}$	
$Para_1$	4.0×10^1	5.0×10^4	2.0×10^6
$Para_2$	3.0×10^1	3.2×10^4	9.6×10^5
$Para_3$	2.0×10^1	3.2×10^4	6.4×10^5
$Para_4$	1.0×10^1	3.2×10^4	3.2×10^5
$Para_5$	1.0×10^1	1.6×10^4	1.6×10^5
$Para_6$	5.0×10^0	1.6×10^4	8.0×10^4
$Para_7$	1.0×10^0	4.0×10^4	4.0×10^4

Afin de calculer un couple (nombre de pas *Langevin*, nombre de répliques parallèles) nous définissons une valeur de référence pour $a_0(T)$ et $B(T)$. Cette valeur de référence est donnée par le résultat de la $4^{ième}$ itération de la méthode des *courbes énergie libre/volume* pour un total de 2×10^6 pas *Langevin*. Nous présentons les couples (nombre de pas *Langevin*, nombre de répliques parallèles) pour les différents tests que nous avons effectués dans le tableau 5.1. Nous considérons que ces simulations sont "convergées" à partir de la $3^{ième}$ itération de la méthode des *courbes énergie libre/volume*. Les résultats de ces simulations sont donnés par la figure 5.6. La figure 5.6 (gauche) représente l'erreur relative $\Delta a_0/a_0 = |a_0(T) - a_0^{ref}(T)|/a_0^{ref}(T)$ sur le paramètre de maille d'équilibre. La figure 5.6 (droite) représente l'erreur relative $\Delta B/B = |B(T) - B^{ref}(T)|/B^{ref}(T)$ sur le module élastique isostatique d'équilibre. On constate que l'erreur sur $a_0(T)$ et sur $B(T)$ diminue quand le nombre d'itérations *Langevin* effectives augmente. Néanmoins, même pour un faible nombre d'itérations *Langevin* - de l'ordre de 10^5 - l'erreur sur $a_0(T)$ et sur $B(T)$ est très faible. Nous allons donc considérer que l'ordre de grandeur de 10^5 pas *Langevin* sera suffisant pour obtenir la convergence dans la suite de nos calculs. Nous soulignons qu'un nombre de pas *Langevin* minimum par processeur est

nécessaire afin d'obtenir une trajectoire aboutissant à la thermalisation du système [5]. Nous estimons que dans le cadre d'un potentiel de type EAM, ce nombre d'itérations par processeur est de l'ordre de 10^3.

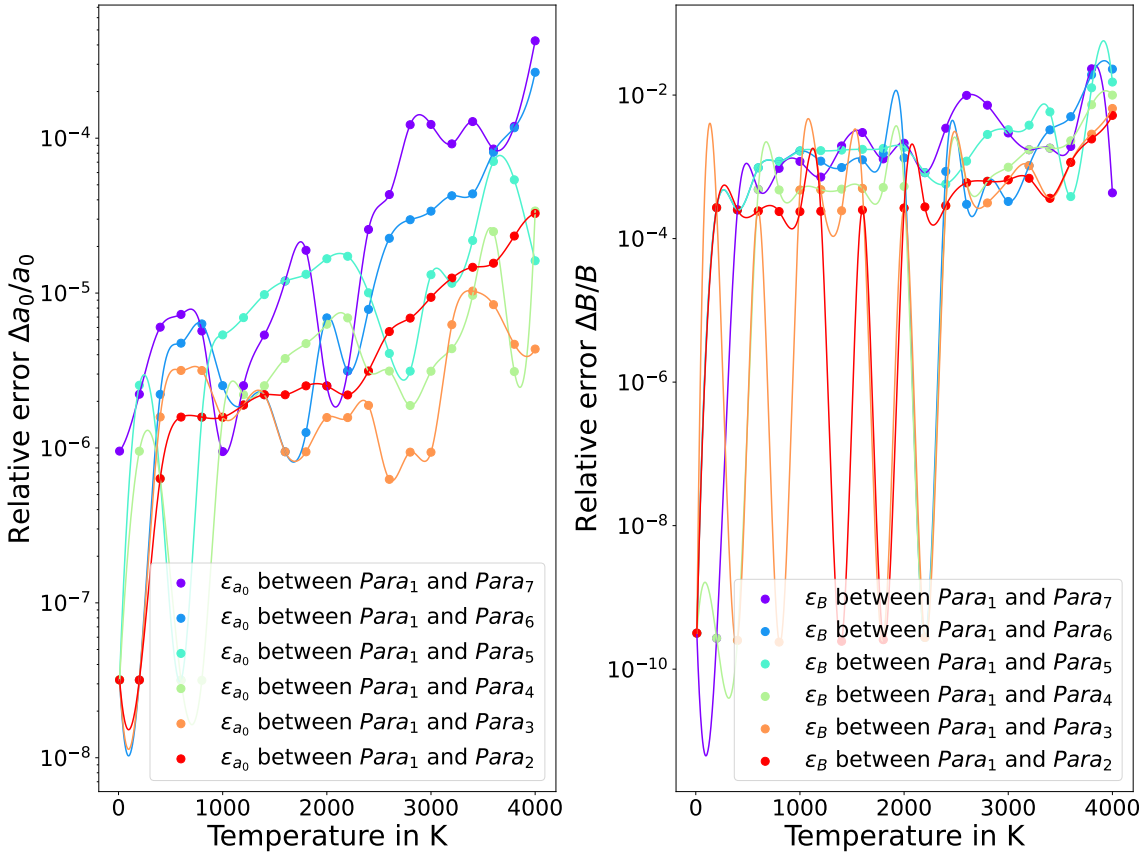

Figure 5.6: Convergence de $a_0(T)$ et de $B(T)$ en fonction du nombre de pas de dynamique stochastique. La figure de gauche représente l'erreur sur $a_0(T)$ (ϵ_{a_0}) et la figure de droite celle de l'erreur sur $B(T)$ (ϵ_B). La valeur de référence pour $a_0(T)$ et de $B(T)$ est donnée par la valeur calculée à $4^{ième}$ itération du calcul de référence. (2×10^6 pas de *dynamique de Langevin*). Nous considérons les valeurs de $a_0(T)$ et de $B(T)$ pour la $3^{ième}$ itération pour un nombre de pas *Langevin* donné. Les différentes parallélisations comparées sont décrites dans la table 5.1

5.4 Conclusions de chapitre

Dans ce chapitre, nous avons proposé une revue des méthodes de calculs d'*énergie libre* basées sur les dynamiques stochastiques présentes dans la littérature. Nous avons mis l'accent sur la méthode Bayésienne dans le cadre d'un *Hamiltonien alchimique*.

5. On considère qu'un système est thermalisé quand sa température cinétique atteint la valeur imposée par le thermostat de Langevin

Nous avons montré les capacités de la méthode Bayésienne pour le calcul de propriétés thermodynamiques d'équilibre tels que le paramètre de maille d'équilibre et le module élastique isostatique d'équilibre. Le calcul de ces propriétés thermodynamiques a été mis en oeuvre dans le cadre d'un potentiel de type EAM pour le tungstène. Ce "cas d'école" relativement simple nous a permis de montrer le cadre et les limites de convergence de la méthode. Cette étude préliminaire va nous permettre de nous fixer un couple : nombre de processeurs, nombre de pas *Langevin* permettant d'obtenir le meilleur compromis entre temps numérique et "temps humain". L'étude du potentiel EAM nous permet de mettre en évidence qu'un nombre total de 10^5 pas de *dynamique de Langevin* permet d'obtenir la convergence du paramètre de maille d'équilibre et du module élastique isostatique d'équilibre. Nous retiendrons ce nombre d'itérations pour la suite.

Il est remarquable de noter que la méthode ABF Bayésienne, dans le cadre d'un *Hamiltonien mixé*, permette de calculer avec une grande précision des quantités telles que le paramètre de maille et le module élastique isostatique d'équilibre pour une température donnée T. Ces résultats montrent la grande précision et la robustesse de cette méthode lorsqu'elle est bien pré-conditionnée. Dans les chapitres suivants, nous allons utiliser cette méthode implémentée dans le code `FEAR` [103] afin de calculer d'autres propriétés d'équilibre thermodynamiques telles que des *énergies libres de formation* de défauts. Nous utiliserons aussi le package `PAFI` [111] afin de calculer des propriétés cinétiques telles que des *énergies libres de migration*. Nous allons plus précisément nous intéresser au cas des lacunes dans les métaux cubiques centrés.

I am the lion and I want to be free
Do you see a lion when you look inside of me ?
Outside the window, just to watch you as you sleep
'Cause I am a lion born from things you cannot be

— Lion, Hollywood Undead

6

Au-delà de l'approximation harmonique : auto-diffusion, métaux cubiques centrés et potentiels *Machine Learning*

Sommaire

Dans ce chapitre, nous nous intéressons au cas spécifique du coefficient d'auto-diffusion dans les métaux cubiques centrés. Nous commençons par introduire la problématique du comportement non-Arrehnius haute température du coefficient d'auto-diffusion dans les métaux cubiques centrés Sec. 6.1. L'anomalie de comportement du coefficient d'auto-diffusion dans les métaux cubiques centrés est un sujet d'études depuis près de 50 ans [231]. Aujourd'hui encore, il est difficile de mettre en place un calcul direct d'*énergie libre d'activation* afin d'estimer numériquement la valeur du coefficient d'auto-diffusion en fonction de la température. Le groupe de Grabowski et Neugebauer a pu réaliser des calculs complets, et prenant en compte les **effets anharmoniques**, pour l'*énergie libre de formation* de la lacune dans plusieurs métaux en utilisant la *théorie de la fonctionnelle de la densité* [228, 232, 233]. Néanmoins, il n'existe pas à notre connaissance une méthodologie **complète et universelle** - dans la littérature - permettant d'effectuer un calcul direct de l'*énergie libre d'activation* de la mono-lacune pour différents métaux-cubiques centrés et conservant une précision *ab initio*. Dans la suite de ce chapitre, nous construisons une telle méthodologie.

Nous proposons de construire un ensemble de potentiels *Machine Learning* adaptés aux bassins énergétiques des lacunes dans les métaux cubiques centrés et d'effectuer un calcul complet du coefficient d'auto-diffusion en estimant l'*énergie libre de formation* et l'*énergie libre de migration* de la mono-lacune. Le but est d'ajuster, le plus fidèlement possible, le paysage énergétique des lacunes décrit par la *théorie de la fonctionnelle de la densité* et de comparer directement les résultats obtenus par les potentiels - issus de chaque fonctionnelle d'échange-corrélation [1] - et les données expérimentales. Dans la section 6.2 nous décrivons la méthodologie que nous avons utilisée pour construire les bases de données d'entraînement et nous donnons les principales caractéristiques des potentiels que nous avons ajustés. Dans la section 6.3, nous comparons directement le calcul du coefficient d'auto-diffusion avec les données expérimentales.

6.1 Coefficients d'auto-diffusion dans les métaux cubiques centrés : une histoire bien incurvée !

Nous commençons par rappeler la définition du coefficient d'auto-diffusion dans un matériau. Le coefficient d'auto-diffusion est avant tout une grandeur **mesurable** expérimentalement et se traduit de la façon suivante :

$$D_\infty = \lim_{t \to +\infty} \frac{\langle \|\Delta \boldsymbol{q}(t)\|^2 \rangle}{6t} \tag{6.1}$$

Ici, $\langle \|\Delta \boldsymbol{q}(t)\|^2 \rangle$ représente le **déplacement quadratique moyen** de l'ensemble des atomes dont on veut suivre la diffusion. Dans le cas du coefficient d'auto-diffusion, on suit l'ensemble des atomes du matériau pur.

1. Nous assimilons parfois, par abus de langage, la fonctionnelle d'échange-corrélation et le potentiel ajusté pour cette fonctionnelle

Cette grandeur a été mesurée expérimentalement, et avec grande précision, pour l'ensemble des métaux de transitions cubiques centrés et pour des gammes de températures allant de l'ordre de 1/10 de la température de fusion à la température de fusion des dits métaux. **En effet, pour des températures inférieures, les durées d'expériences nécessaires pour mesurer - avec une incertitude statistique - raisonnable sont trop grands**. Les résultats expérimentaux obtenus pour ces différents métaux cubiques centrés sont donnés par la figure 6.1 issue de Neumann *et al.* [234]. Dans la limite des basses températures, le comportement du coefficient d'auto-diffusion suit les prévisions de la loi d'**Arrhenius**. Mais, pour les hautes températures, on constate que le comportement du coefficient d'auto-diffusion dévie de la loi d'Arrhenius et que cette déviation augmente avec la température. Plusieurs explications à ce phénomène ont été proposées dans la littérature. Dans le cas du β-Titane et du β-Zirconium, Sanchez *et al.* [235] proposent d'intégrer la contribution de la phase métastable ω et obtiennent un très bon accord avec les résultats expérimentaux dans le cas du zirconium. Cette hypothèse a ensuite été critiquée par Petry et Volg [236, 237] suite à leurs mesures expérimentales ne mettant pas en évidence la phase ω dans le cas du Ti et Zr. Dans les autres métaux de transition - hors du groupe *IV B* - l'explication ne fait pas consensus. Ainsi, deux mécanismes principaux sont mis en avant par la littérature : (i) l'implication de deux groupes de défauts, la mono et la di-lacune [238-242] ; (ii) l'implication du mode de phonons $\frac{2}{3}\langle 111\rangle$ pour la mono-lacune [236, 237]. On pourra notamment citer les travaux de Smirnova *et al.* [243] mettant en avant les effets anharmoniques de la mono-lacune dans le molybdène. **Dans les faits, il n'existe pas de consensus pour rendre compte du comportement hautes températures du coefficient d'auto-diffusion pour tous les métaux cubiques centrés.**

Plusieurs explications possibles ont été proposées dans la littérature mais nous allons nous concentrer sur la suivante : les propriétés à température finie de la mono-lacune. On peut exprimer le coefficient d'auto-diffusion à l'aide de grandeurs thermodynamiques et cinétiques de la mono-lacune, si on considère que c'est ce défaut qui est dominant dans le processus d'auto-diffusion. On a alors d'après la théorie de l'État de transition [244] :

$$D(T) = \mathfrak{Z}\frac{k_B T}{h} f d^2(T) e^{-\frac{1}{k_B T}\left[F_f(T)+F_{mig}(T)\right]} \tag{6.2}$$

Ici, $\mathfrak{Z}$ est le nombre de chemins de diffusion possibles de la mono-lacune, f est le facteur de corrélation et $d(T)$ est la distance de saut pour le chemin de diffusion. Dans le cas du coefficient d'auto-diffusion de la mono-lacune, on ne considère que les sauts en premiers voisins comme chemins de diffusion possibles. Le facteur pré-exponentiel correspond au coefficient d'auto-diffusion calculé à l'aide de la TST [14] anharmonique. Le facteur exponentiel est lié à la distribution Boltzmanienne de la concentration de lacune à une température fixée T. C'est dans ce terme exponentiel que sont quantifiés les **effets anharmoniques** par l'intermédiaire de l'*énergie libre d'activation.*

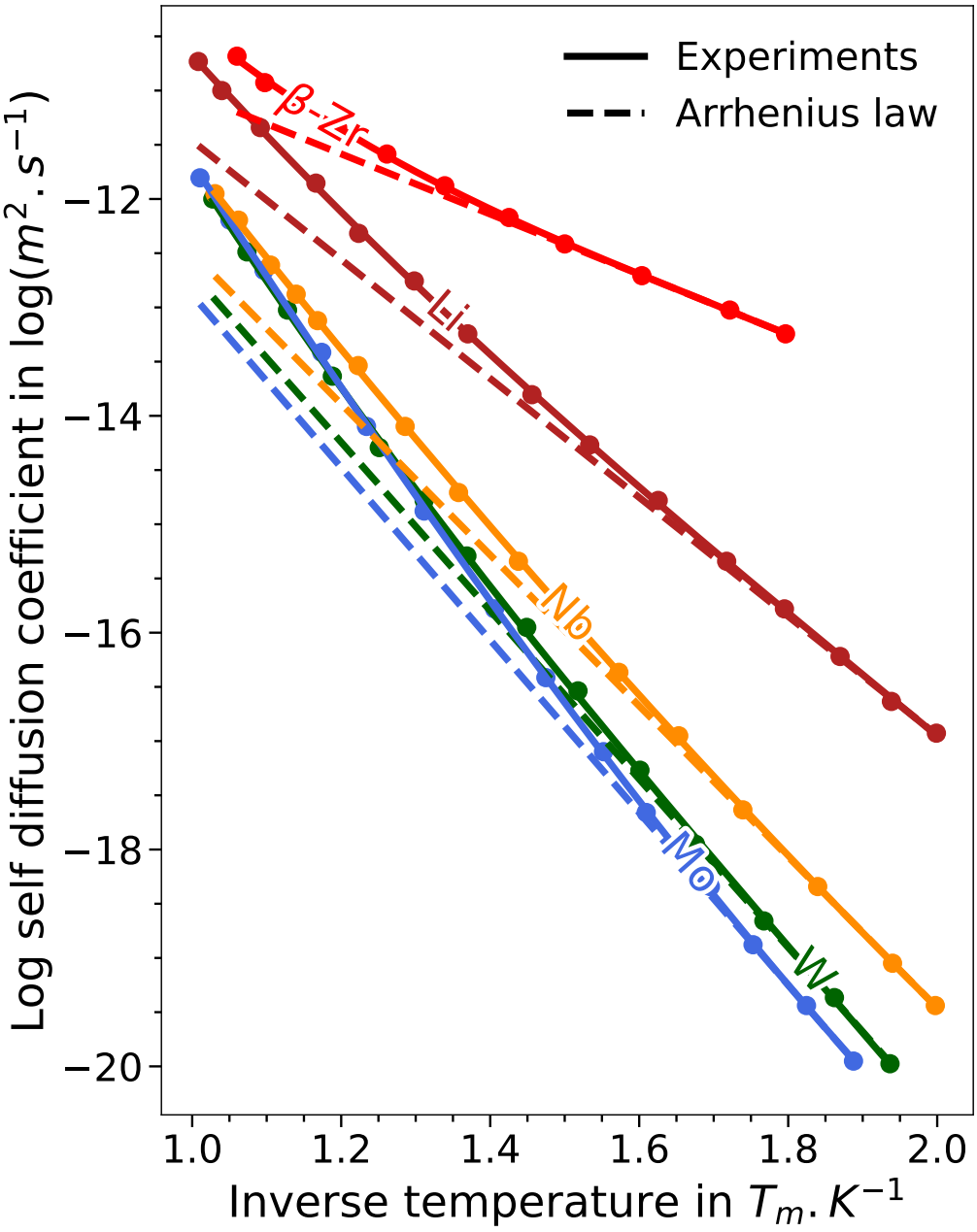

Figure 6.1: Évolution du coefficient d'auto-diffusion des métaux de transitions cubiques centrés en fonction de T_m/T [240]. On constate que le comportement du coefficient de diffusion dévie de la loi d'Arrhenuis pour les hautes températures, et ce pour tous les métaux cubiques centrés étudiés. Ce graphique est issu de Neumann *et al.* [240]. Nous avons figuré en trait bleu, le comportement Arrhenuis attendu pour le molybdène, le tantale, le lithium et le β-zirconium.

Afin de rendre compte du comportement à hautes températures du coefficient d'auto-diffusion, nous allons mettre en place un schéma de calcul en deux étapes. Dans un premier temps, nous allons construire des potentiels Machine Learning (i) représentatifs des propriétés des lacunes en température dans les métaux cubiques centrés. Pour

cela nous allons constituer des bases de données DFT. Dans un deuxième temps nous allons effectuer des calculs d'*énergies libres de formation et de migration* (ii) pour la mono-lacune. L'ensemble de cette procédure permettra de calculer de façon numérique le coefficient d'auto-diffusion grâce à l'équation (6.2). Dans la section 6.2, nous décrivons la procédure de génération des bases de données que nous avons mises en place (Sec. 6.2.1) et nous comparons directement des grandeurs clefs des potentiels ajustés avec les valeurs issues de la DFT (Sec. 6.2.2) en nous limitant pour le moment au cas du tungstène et du molybdène. Enfin, nous nous intéressons à la comparaison directe du coefficient d'auto-diffusion numérique obtenu pour le tungstène avec les données expérimentales Sec. 6.3. Nous comparons les résultats obtenus pour des potentiels ajustés pour différentes fonctionnelles d'échange-corrélation.

6.2 Ajustement de potentiels représentatifs des propriétés des lacunes dans les métaux cubiques centrés

Le point essentiel de l'ajustement des potentiels de type Machine Learning est la génération de la base de données d'entraînement. Celle-ci doit être cohérente et représentative des propriétés à étudier. Dans cette section, nous allons d'abord décrire la méthode de génération que nous avons utilisée pour créer des bases de données représentatives du comportement des lacunes en température pour différents métaux cubiques centrés Sec. 6.2.1. Puis, nous comparons les données *ab initio* et les données prédites par les potentiels concernant les propriétés thermodynamiques et cinétiques des lacunes à $T = 0\ K$ Sec. 6.2.2. Pour notre première application, nous nous intéressons au cas du W mais cette procédure est génerale et peut être appliquée pour les autres métaux cubiques centrés.

6.2.1 Méthodologie : génération de base de données et ajustement

Afin de mettre en place les méthodes de calcul d'*énergie libre* présentées dans le chapitre précédent (5), nous devons nous assurer que le potentiel ajusté est représentatif des propriétés thermodynamiques telles que l'énergie de formation et l'énergie de migration des lacunes à $T = 0$ K. De plus, afin d'être le plus représentatif possible, nous voulons que l'expansion thermique associée à notre potentiel soit au plus proche possible des données expérimentales afin de prendre en compte de façon la plus précise possible le caractère extensif de l'*énergie libre*. Pour cela, nous allons mettre en place une méthodologie structurée applicable pour l'ensemble des métaux cubiques centrés. Dans la suite de cette section, nous allons nous atteler à décrire la construction d'une base de données "élémentaire" représentative des propriétés des lacunes dans les métaux cubiques centrés pour une gamme de températures allant de $T = 0$ K jusqu'à $T = T_m$ où T_m est la température de fusion du métal. Pour cela, nous allons nous baser sur les méthodes *ab initio*.

Nous voulons construire une base de données respectant les conditions suivantes : (i) les données doivent permettre d'ajuster correctement les constantes élastiques du métal à 0 K ; (ii) les données doivent être représentatives de l'expansion thermique quadratique constatée expérimentalement de 0 K à T_m ; (iii) les données doivent permettre de prendre en compte le comportement des petits amas de lacunes de 0 K à T_m.

Nous allons décrire comment générer une base de données vérifiant ces conditions dans les paragraphes suivants. Afin de générer nos bases de données, nous avons utilisé le code `VASP` [245]. L'ensemble de nos calculs a été effectué en utilisant des pseudo-potentiels de type PAW [246] et incluant les électrons de semi-coeur. Nous avons effectué des calculs pour deux tailles de boîtes différentes : $a_0 \times a_0 \times a_0$ et $4a_0 \times 4a_0 \times 4a_0$. Afin de conserver une cohérence de calcul entre ces deux tailles de boîtes, nous utilisons deux grilles de *points-k* de type Monkhorst-Pack [247] pour conserver le même échantillonnage dans l'*espace réciproque* : $24 \times 24 \times 24$ et $6 \times 6 \times 6$. L'utilisation d'une grille très dense, relativement atypique pour des boîtes de cette taille, est justifiée par la convergence de grandeurs thermodynamiques d'intérêts pour le comportement des lacunes. Nous avons fixé l'énergie de cut-off à 500 eV et un calcul est considéré comme convergé si : (i) d'un point de vue **électronique** la différence d'énergie entre deux itérations est inférieure à 10^{-8} eV, (ii) d'un point de vue **ionique** la différence d'énergie entre deux itérations est inférieure à 2×10^{-2} eV et (iii) notre paramètre de smearing est fixé à 0.3 en utilisant la méthode de Methfessel et Paxton [248]. Nous justifions le choix d'une grille de point-k très dense ($6 \times 6 \times 6$ pour des systèmes de volume $4a_0 \times 4a_0 \times 4a_0$) et de l'utilisation d'un paramètre de smearing élevé (0.3) dans l'annexe (D) afin d'assurer la convergence de grandeurs d'importance : (i) **les énergies de formation de la mono-lacune et des di-lacunes premier et deuxième voisins** ; (ii) **les énergies de liaison entre la mono-lacune et les di-lacunes premier et deuxième voisins**.

Constantes élastiques et déformations

Afin de rendre compte des constantes élastiques du matériau, nous allons appliquer la méthode des *courbes énergie/volume* décrite dans le chapitre précédent (5). Le module élastique isostatique B s'obtient en appliquant un tenseur de déformation homogène et isotrope et à l'aide de l'expression donnée par l'équation (5.43). Dans notre cas, nous voulons aussi être représentatif des constantes élastiques anisotropes C_{11}, C_{12} et C_{44}. Pour cela, nous allons utiliser des tenseurs de déformations différents.

Pour le calcul de C_{11} et C_{12}, nous utilisons le module élastique isostatique calculé par la méthode des courbes *énergie/volume*, l'équation (5.43) ainsi que le tenseur de déformation suivant :

$$\boldsymbol{\epsilon}^{\mathbf{11}}(\delta) = \begin{pmatrix} \delta/2 & 0 & 0 \\ 0 & -\delta/2 & 0 \\ 0 & 0 & \delta^2/(1-\delta^2) \end{pmatrix} \tag{6.3}$$

On peut noter que $\boldsymbol{\epsilon}^{\mathbf{11}}(\delta)$ conserve le volume du système. Dans le cadre de la théorie élastique, on peut facilement relier la différence d'énergie entre une configuration non déformée et l'énergie d'une configuration sur laquelle on applique la déformation $\boldsymbol{\epsilon}^{\mathbf{11}}(\delta)$. Cette différence d'énergie $\Delta E^{11}(\delta)$ fait directement intervenir les constantes élastiques C_{11} et C_{12} dans le cas d'une structure cubique et on a alors :

$$\Delta E^{11}(\delta) = V(C_{11} - C_{12})\delta^2 + \mathcal{O}(\delta^4) \tag{6.4}$$

où V est le volume du système considéré. De plus, en appliquant la relation valable pour les structures cubiques $B = (C_{11} + 2C_{12})/3$, on peut alors calculer C_{11} et C_{12}. La détermination de $(C_{11} - C_{12})$ dans l'équation (6.4) se fait par ajustement d'une fonction quadratique entre $\Delta E^{11}(\delta)$ et δ pour une grille de δ donnée. Dans notre cas, nous avons utilisé 13 valeurs de δ distribuées uniformément entre -1.0% et 1.0%. De même, pour le calcul de C_{44}, on utilise le tenseur de déformation suivant :

$$\boldsymbol{\epsilon}^{\mathbf{44}}(\delta) = \begin{pmatrix} 0 & \delta/2 & 0 \\ \delta/2 & 0 & 0 \\ 0 & 0 & \delta^2/(4-\delta^2) \end{pmatrix} \tag{6.5}$$

Cette déformation conserve elle aussi le volume du système. Grâce à la théorie élastique, on peut calculer la différence d'énergie entre une configuration non déformée et l'énergie d'une configuration sur laquelle on applique la déformation $\boldsymbol{\epsilon}^{\mathbf{44}}(\delta)$. Cette expression implique simplement C_{44} dans le cas d'une structure cubique :

$$\Delta E^{44}(\delta) = \frac{1}{2} V C_{44} \delta^2 + \mathcal{O}(\delta^4) \tag{6.6}$$

De même que pour l'équation (6.4), nous utilisons un ajustement d'une fonction quadratique entre $\Delta E^{44}(\delta)$ et δ pour une grille de δ donnée. Dans notre cas, nous avons utilisé 13 valeurs de δ distribuées uniformement entre -1.0% et 1.0%. Pour rendre compte des constantes élastiques et du module élastique isostatique, nous allons utiliser 39 configurations de volume $a_0 \times a_0 \times a_0$ avec a_0 le paramètre d'équilibre obtenu grâce à la méthode des *courbes énergie/volume*. Soit 13 configurations pour évaluer le module élastique isostatique, 13 configurations pour évaluer les constantes C_{11} et C_{12} et 13 configurations pour C_{44}. L'ensemble des configurations nécessaires pour le calcul des constantes élastiques est contenu dans les boîtes de simulation de volume $4a_0 \times 4a_0 \times 4a_0$.

Nous ajoutons aussi des configurations déformées simulées dans des boîtes de simulation de volume $a_0 \times a_0 \times a_0$. Pour cela, on fixe un tenseur de déformation isotrope $\boldsymbol{\epsilon}^i$ de taux de déformation ϵ auquel on ajoute un tenseur de déformation aléatoire dont les composantes vérifient l'équation $\varepsilon_{ij} \overset{\text{loi}}{\sim} \alpha\, \mathcal{N}(0,1)$ où α donne l'amplitude du tenseur aléatoire et $\mathcal{N}(0,1)$ est une variable aléatoire de loi normale centrée et réduite. Finalement, le tenseur de déformation appliqué à la configuration est donné par :

$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^i + \frac{1}{2}\left(\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\right) \tag{6.7}$$

Ici, $\cdot^T$ est l'opérateur de transposition. Cette opération sur le tenseur de déformation aléatoire $\boldsymbol{\varepsilon}$ permet d'assurer la symétrie du tenseur des déformations totales $\boldsymbol{\epsilon}$. Nous avons appliqué des déformations allant de -5% à 5% avec un paramètre $\alpha = 0.01$ pour des boîtes de simulation de volume $a_0 \times a_0 \times a_0$ avec a_0 le paramètre de maille d'équilibre obtenu par la méthode des *courbes énergie/volume*. Au total, nous avons généré 1000 configurations déformées.

Expansion thermique et configurations "en température"

Nous voulons rendre compte de l'expansion thermique expérimentale tout en restant en adéquation avec le paramètre de maille d'équilibre calculé à 0 K par DFT. Pour cela, nous interpolons l'expansion thermique du paramètre de maille expérimental du matériau que l'on note $f(T) = a_0^{exp}(T)/a_0^{exp}(0)$ (en reprenant les notations du chapitre précédent 5). L'expansion thermique expérimentale est donnée par Touloukian *et al.* [249]. On peut alors construire "l'expansion thermique *ab inito*" par la formulation suivante :

$$a_0^{DFT}(T) = \frac{a_0^{DFT}(0)}{a_0^{exp}(0)} f(T) \tag{6.8}$$

Dans le cas d'une configuration dite "en température" en *ab initio*, on partira du paramètre de maille donné par l'équation (6.8) pour une température donnée T. Afin de générer des configurations "en température", nous devrions partir de cette formulation du paramètre de maille en température et effectuer une *dynamique moléculaire ab initio*. Cette méthode est prohibitive d'un point de vue numérique et nécessite une très longue trajectoire afin d'assurer la décorrélation des configurations. Nous allons donc utiliser une méthode basée sur les *dynamiques stochastiques* pour générer ces configurations.

En effet, un pas de *dynamique moléculaire ab initio* autour de la position d'équilibre ($\boldsymbol{q}_{min}$) peut être traduit par l'équation suivante :

$$\boldsymbol{q}(\delta t) \overset{\text{loi}}{\sim} \mathcal{N}\left(\boldsymbol{q}_{min}, 2\beta^{-1}\,\delta t \boldsymbol{M}^{-1} \cdot \mathbf{1}_{3N}\right) \tag{6.9}$$

où $\mathcal{N}\left(\boldsymbol{q}_{min}, 2\beta^{-1}\,\delta t \boldsymbol{M}^{-1} \cdot \mathbf{1}_{3N}\right)$ une loi normale multi-dimensionnelle centrée en $\boldsymbol{q}_{min}$ et dont la matrice de covariance est $2\beta^{-1}\,\delta t \boldsymbol{M}^{-1} \cdot \mathbf{1}_{3N}$. L'équation (6.9) permet de générer des configurations de *dynamique moléculaire* pour une température donnée T autour de la position d'équilibre du système. Ces configurations sont plus décorrélées qu'une *dynamique moléculaire classique* et permettent une meilleure représentativité. Pour générer ces configurations, nous fixons une amplitude maximale de déplacement sur le système $\|\boldsymbol{\delta q}\|^2$ et nous calculons le pas de temps $\delta\tilde{t}$ tel que :

$$\|\boldsymbol{\delta q}\|^2 = \mathrm{Tr}\left\{2k_B T_m \cdot \boldsymbol{M}^{-1}\right\}\,\delta\tilde{t} \tag{6.10}$$

En ayant fixé $\delta\tilde{t}$, on peut alors générer des configurations "en température" entre $T = 0$ K et $T = T_m$. D'un point de vue pratique, on part d'une configuration $\boldsymbol{q}$: (i) on minimise l'énergie de la configuration avec le paramètre de maille à la température T et on obtient la configuration $\boldsymbol{q}_{min}$; (ii) on applique les déplacements $\boldsymbol{\delta q}$ correspondant à la méthode d'échantillonnage décrite par l'équation (6.9 pour la température T et on effectue une évaluation d'énergie et des forces pour la configuration $\boldsymbol{q}_{min} + \boldsymbol{\delta q}$.

Configurations de lacunes

Afin de rendre compte des propriétés des lacunes dans le cadre des températures finies, nous adoptons la démarche suivante dans les boîtes de simulation de volume $(4a_0)^3$ pour une grille de points-k $6 \times 6 \times 6$. Pour une température donnée - donc pour un paramètre de maille donné - nous générons des configurations de mono-lacunes et de di-lacunes. Les di-lacunes considérées sont les di-lacunes, 1^{er}, $2^{ième}$ et $3^{ième}$ voisins. Ces configurations sont minimisées et on applique ensuite le bruit thermique correspondant à la température grâce à la démarche décrite dans le paragraphe précédent. Nous considérons 5 températures entre 0 K et T_m et nous générons 10 configurations de mono et di-lacunes par température (nous ne considérons qu'une configuration à 0 K). Soit un total de 164 configurations.

Les configurations précédentes sont représentatives des propriétés thermodynamiques des lacunes. Afin de rendre compte de leurs propriétés cinétiques, nous ajoutons 7 configurations issues de calcul NEB [221-223] pour la migration de la mono-lacune, les di-lacunes 1^{er} voisins et les tri-lacunes 1^{er} voisins, ce qui donne un total de 21 configurations.

Bases de données et ajustement

Nous nous intéressons d'abord à produire des potentiels Machine Learning pour le tungstène (W) en utilisant trois fonctionnelles d'échange-corrélation différentes : (i) une fonctionnelle LDA [38, 39], (ii) une fonctionnelle PBE [41] et (iii) une fonctionnelle AM [44-46]. L'interpolation des expansions thermiques expérimentales est issue de Touloukian *et al.* [249] ainsi que les constitutions des bases de données *ab initio* sont données en annexe D. Afin de réaliser les ajustements, nous utilisons le package `MILADY` [20, 138] et nous privilégions le bi-spectrum SO(4) [121] comme descripteur avec $j_{max} = 4.0$. Le package `MILADY` permet de calculer analytiquement les forces et les contraintes du potentiel grâce aux dérivées analytiques par rapport aux positions des atomes et peut donc être directement utilisé pour faire de la *dynamique moléculaire*. **Pour les trois potentiels, nous donnons un fort poids aux configurations représentatives des constantes élastiques afin de "fixer physiquement" les potentiels ajustés**. En effet, les potentiels générés sans contraintes fortes sur les constantes élastiques se révèlent souvent inutilisables car ils présentent un fort risque de divergence après un nombre restreint de pas de *dynamique moléculaire*.

6.2.2 Comparaison des grandeurs clefs avec l'expérience et la *théorie de la fonctionnelle de la densité*

Dans cette section, nous comparons les grandeurs clefs calculées par les potentiels Machine Learning développés, les données issues de la *Théorie de la fonctionnelle de la densité électronique* et les données expérimentales. Ces comparaisons vont être effectuées pour les trois fonctionnelles utilisées pour les potentiels du tungstène (W) : (i) LDA, (ii) PBE et (iii) AM.

Comparaison des données du tungstène (W)

Nous commençons par comparer les valeurs de grandeurs "clefs" - thermodynamiques et cinétiques - issues des potentiels Machine Learning ajustés et la DFT. Ces grandeurs concernent les propriétés des petits amas de lacunes, notamment la mono et la di-lacune, ainsi que les constantes élastiques et le paramètre de maille à 0 K. L'ensemble des comparaisons pour les trois fonctionnelles étudiées (AM, PBE et LDA) est récapitulé dans la table 6.1. On constate que les grandeurs prédites par le potentiels Machine Learning sont très proches des données issues de la DFT. Pour la suite de notre étude, la propriété la plus importante est l'énergie d'activation $E_a = E_f^{1\,vac} + E_{mig}^{1\,vac}$. L'erreur entre la valeur prédite par les potentiels Machine Learning et la DFT, pour l'énergie d'activation, est inférieure ou égale à 0.1 eV pour les trois fonctionnelles. **Les trois potentiels ajustés reproduisent parfaitement l'énergie d'activation DFT.**

Nous présentons l'expansion thermique calculée pour le potentiel Machine Learning pour le tungstène obtenue avec les trois fonctionnelles étudiées : (i) LDA [38, 39], (ii) PBE [41] et (iii) AM [44-46]. L'expansion thermique du potentiel a été calculée grâce à des calculs d'*énergie libre* par utilisation du package `FEAR` [103] et par la méthode des courbes *énergie libre/volume* présentée dans le chapitre précédent 5. La comparaison entre la fonction décrivant l'expansion thermique expérimentale renormalisée extraite de Touloukian *et al.* [249] décrite par l'équation (6.8) et le potentiel est donnée par la figure 6.2. La comparaison est effectuée entre les données du potentiel et l'expansion thermique renormalisée à 300 K. Le tout est exprimé en unité de a_0 *ab initio*. Comme on peut le constater, l'accord entre les données expérimentales et les données calculées pour le potentiel est très bon, même pour les hautes température. **Les données issues du potentiel sont représentatives de l'expansion thermique quadratique en température caractéristique des données expérimentales. On note ici que c'est la fonctionnelle LDA qui est la plus représentative de la propriété d'expansion thermique.**

On peut noter que l'ensemble des potentiels ajustés pour le tungstène présente une expansion thermique proche de l'expansion thermique expérimentale et des propriétés thermodynamiques des petits amas de lacunes au plus proche de la *théorie de la fonctionnelle de la densité*. Il nous reste maintenant à utiliser ces potentiels afin de calculer l'*énergie libre de formation et de migration* de la mono-lacune grâce aux méthodes décrites dans le chapitre précédent (5). Nous pourrons alors calculer le coefficient d'auto-diffusion grâce à l'équation (6.2) et comparer nos résultats numériques directement avec les données expérimentales.

Table 6.1: Comparaison des grandeurs clefs entre les calculs *ab initio* et les calcul réalisés grâce au potentiel de type Machine Learning du W pour les trois fonctionnelles : (i) LDA [38, 39], PBE [41] et AM [44-46]. Les grandeurs clefs du potentiel sont très proches des résultats obtenus par la *théorie de la fonctionnelle de la densité électronique*.

LDA	Valeur calculée		Unité
	ab initio	potentiel ML	
a_0	3.14167	3.14159	Å
B	336.0	335.8	GPa
C_{11}	568.3	563.1	GPa
C_{12}	219.9	222.1	GPa
C_{44}	150.8	146.5	GPa
$E_f^{1\,vac}$	3.28	3.24	eV
$E_f^{2\,vac\,1nn}$	6.75	6.69	eV
$E_f^{2\,vac\,2nn}$	7.04	6.99	eV
$E_{mig}^{1\,vac}$	1.76	1.76	eV
$E_{mig}^{2\,vac}$	1.39	1.27	eV

PBE	Valeur calculée		Unité
	ab initio	potentiel ML	
a_0	3.185685	3.18568	Å
B	304.7	304.6	GPa
C_{11}	515.4	513.5	GPa
C_{12}	199.3	200.2	GPa
C_{44}	139.9	139.8	GPa
$E_f^{1\,vac}$	3.20	3.17	eV
$E_f^{2\,vac\,1nn}$	6.56	6.50	eV
$E_f^{2\,vac\,2nn}$	6.89	6.83	eV
$E_{mig}^{1\,vac}$	1.72	1.67	eV
$E_{mig}^{2\,vac}$	1.30	1.20	eV

AM	Valeur calculée		Unité
	ab initio	potentiel ML	
a_0	3.15073	3.15070	Å
B	327.9	327.6	GPa
C_{11}	559.6	554.4	GPa
C_{12}	212.0	214.2	GPa
C_{44}	115.4	151.5	GPa
$E_f^{1\,vac}$	3.53	3.48	eV
$E_f^{2\,vac\,1nn}$	7.20	7.13	eV
$E_f^{2\,vac\,2nn}$	7.50	7.45	eV
$E_{mig}^{1\,vac}$	1.78	1.74	eV
$E_{mig}^{2\,vac}$	1.37	1.36	eV

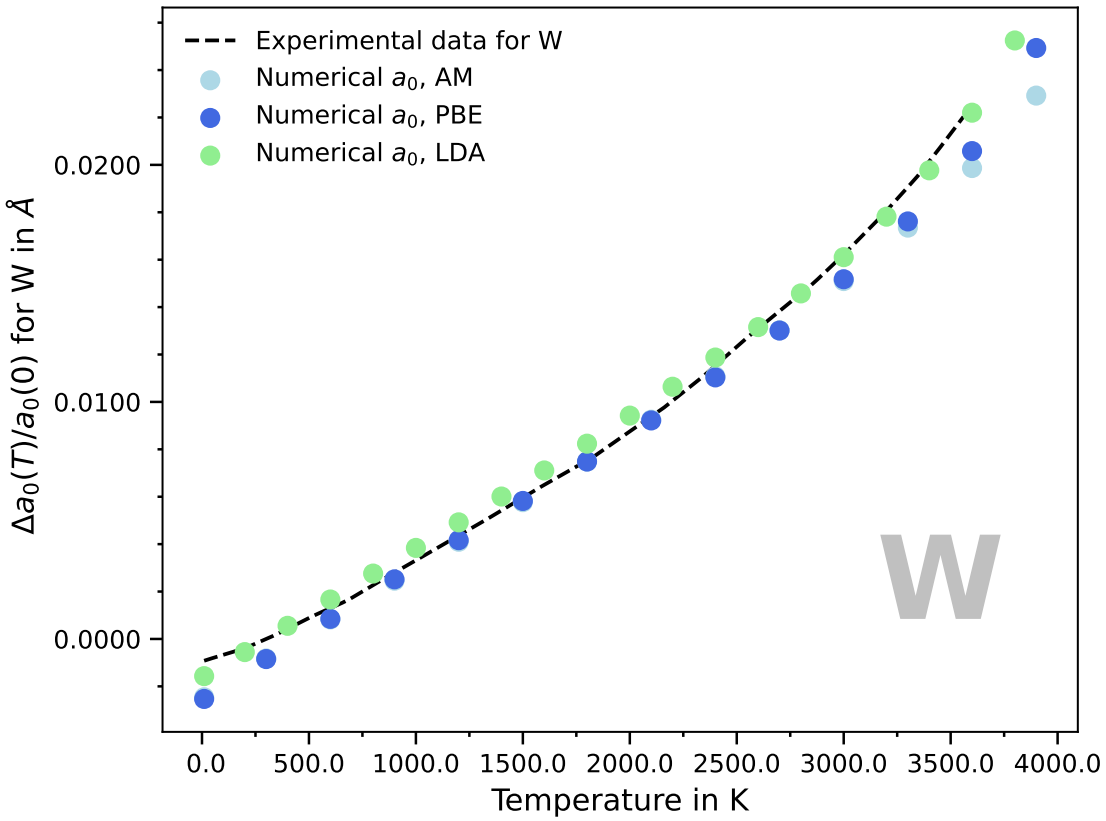

Figure 6.2: Comparaison de l'expansion thermique expérimentale renormalisée et de l'expansion thermique du potentiel Machine Learning pour le tungstène en utilisant les fonctionnelles (i) AM [44-46], (ii) PBE et (iii) LDA. Nous suivons les grandeurs : $\Delta a_0^{exp}(T)/a_0^{exp}(0) = \left(a_0^{exp}(T) - a_0^{exp}(0)\right)/a_0^{exp}(0)$ présentée en pointillées et $\Delta a_0^{ML}(T)/a_0^{ML}(0) = \left(a_0^{ML}(T) - a_0^{ML}(0)\right)/a_0^{ML}(0)$ présentée par les cercles de couleurs. Les données expérimentales sont issues de Touloukian *et al.* [249]

6.3 Application au cas de la mono-lacune dans le tungstène : *énergie libre de formation* et *énergie libre de migration*

Dans cette section, nous calculons l'*énergie libre de formation et de migration* de la mono-lacune dans le tungstène à l'aide des trois potentiels présentés précédemment. Le calcul de l'*énergie libre de formation* a été effectué grâce au package `FEAR` [103]. Nous avons utilisé une super-cellule de volume $4a_0(T) \times 4a_0(T) \times 4a_0(T)$ et nous avons effectué deux calculs différents. Un calcul dans cette super-cellule contenant un cristal parfait de 128 atomes et un calcul d'une super-cellule cubique centrée contenant une mono-lacune de 127 atomes. Afin d'utiliser l'*énergie libre* comme potentiel thermodynamique, nous devons nous assurer que nous travaillons à pression nulle $P = 0$. Si $P \neq 0$, alors le potentiel thermodynamique du système est l'*enthalpie libre* $G = F - PV$. Afin d'assurer $P = 0$, nous avons utilisé la méthode *courbes*

énergie libre/volume décrite dans le chapitre (5) et nous considérons que les calculs sont convergés quand les **quatre premières décimales du paramètre de maille d'équilibre sont inchangées entre deux itérations**. On calcule ensuite l'*énergie libre de formation* de la mono-lacune de la façon suivante :

$$F_f(T) = F_{vac}(T) - \frac{127}{128} F_{bulk}(T) \tag{6.11}$$

Ici, $F_{vac}(T)$ et $F_{bulk}(T)$ sont respectivement l'énergie libre du système contenant la lacune et l'énergie libre du cristal parfait en fonction de la température.

Une fois ce calcul effectué, nous utilisons le package `PAFI` [111] (voir Sec. 5.3.2) pour calculer l'*énergie libre de migration* de la mono-lacune. Cette méthode nécessite la connaissance d'une *coordonnée de réaction* à 0 K. Nous avons utilisé le package `MILADY-LAMMPS` [20, 138, 185] afin d'effectuer une NEB dont le paramètre de maille à 0 K est donné par la méthode *courbes énergie libre/volume* pour le système contenant la mono-lacune. De plus, afin de prendre en compte l'expansion thermique du matériau, une fonction quadratique du paramètre de maille du système contenant la mono-lacune est ajustée en fonction de la température. Nous présentons, dans la suite de cette section, les résultats obtenus pour les trois potentiels du tungstène décrits précédemment.

6.3.1 Mono-lacune dans le tungstène (W) : grandeurs thermodynamiques et cinétiques à températures finies

Nous présentons les résultats obtenus pour les trois potentiels machine learning ("LDA", "PBE" et "AM") dans la figure 6.3. Dans la figure 6.3 de gauche resp. droite nous présentons l'évolution de l'*énergie libre de formation* resp. l'*énergie libre de migration* en fonction de la température pour les trois potentiels ajustés. Nous avons tracé, en pointillés, l'évolution de ces deux grandeurs en fonction de la température dans le cadre de l'approximation harmonique. De plus, nous donnons en encart les valeurs de ΔU_0 et ΔS_0 correspondant aux modèles harmoniques et obtenus par régression linéaire grâce aux données basse températures.

$$F_{f,ha}^{1\ vac}(T) = \Delta U_0 - T\Delta S_0 \tag{6.12}$$

On constate que le comportement de l'*énergie libre de formation* resp. *énergie libre de migration* dévie grandement du modèle harmonique pour des températures $T > T_m/2$. L'ensemble des modèles harmoniques - dont les paramètres sont présentés en encart de la figure 6.3 - sont du même ordre de grandeur pour tous les potentiels. **Seul le modèle harmonique de l'*énergie libre de migration* du potentiel AM présente une faible dépendance en température (ΔS_0 et pente faibles)**.

Le potentiel ajusté pour la fonctionnelle AM présente **l'anharmonicité la plus élévée pour l'*énergie libre de formation***. En effet le potentiel "AM" dévie de la loi d'Arrhenius à partir de $T > T_m/3$, alors que les deux autres potentiels commencent à dévier de celle-ci à partir de $T > 2T_m/3$. Le potentiel ajusté pour la fonctionnelle

LDA présente un comportement peu physique avec une remontée de l'*énergie libre* à partir de 2000 K, ce qui correspond à une ***entropie anharmonique de formation* négative de ce potentiel pour les hautes températures**. Le potentiel ajusté pour la fonctionnelle LDA présente **l'anharmonicité la plus élevée pour l'*énergie libre de migration*. Nous retiendrons que le potentiel issu de la fonctionnelle AM présente l'énergie d'activation la plus élevée à basse température et le comportement des potentiels issus des fonctionnelles PBE et LDA seront similaires en température** (par un effet de compensation de l'*energie libre de migration et de formation* pour la fonctionnelle LDA).

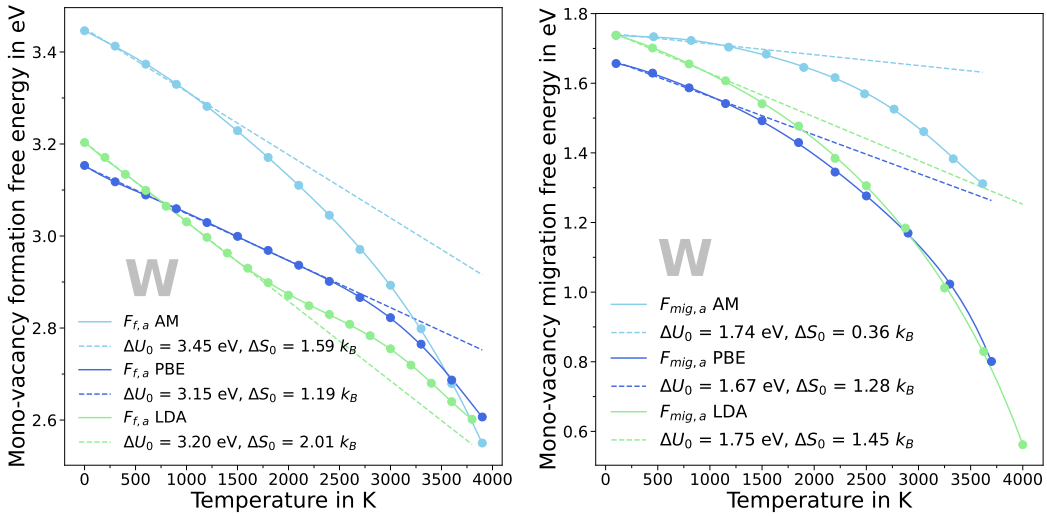

Figure 6.3: Calcul de l'*énergie libre de formation et de migration* ($F_{f,a}$ et $F_{mig,a}$) de la mono-lacune dans le tungstène en fonction de la température pour les fonctionnelles (i) LDA, (ii) PBE et (iii) AM. La droite en pointillés représente le comportement harmonique attendu de ces deux grandeurs en fonction de la température

Le "simple" calcul de l'*énergie libre de formation et de migration* de la mono-lacune ne permet que de tirer des conclusions qualitatives. On peut cependant d'ores et déjà dire avec certitude que **les effets anharmoniques sont importants - quelque soit la fonctionnelle utilisée - pour une température** $T > T_m/2$. Néanmoins, ces deux grandeurs ne sont pas accessibles ni mesurables directement d'un point de vue expérimental. Ainsi, nous devons maintenant calculer le coefficient d'auto-diffusion pour ces trois potentiels afin d'effectuer une comparaison directe avec les données expérimentales.

6.3.2 Mono-lacune dans le tungstène (W) : coefficients d'auto-diffusion et comparaison avec l'expérience

Nous allons calculer directement le coefficient d'auto-diffusion grâce à l'équation (6.2). Nous avons choisi d'ajuster un polynôme d'ordre 6 afin d'interpoler l'*énergie libre de formation et de migration* en fonction de la température et d'ajuster un polynôme d'ordre 2 afin d'interpoler l'évolution du paramètre de maille d'équilibre en

fonction de la température. Les données expérimentales concernant l'auto-diffusion dans le tungstène sont issues de Neumann *et al.* [234]. Les résultats obtenus pour les trois potentiels et leur comparaison directe avec les données expérimentales sont présentés dans la figure 6.4 (gauche). Les ajustements des coefficients d'auto-diffusion pour les différents potentiels sont représentés en traits pleins et les données expérimentales sont représentées par les cercles noirs. Nous présentons aussi le comportement harmonique ajusté sur les données numériques à basses températures pour les trois potentiels. On constate que, quelque soit la fonctionnelle utilisée, le coefficient d'auto-diffusion prenant en compte les effets anharmoniques de la mono-lacune présente **l'incurvation caractéristique des données expérimentales à hautes températures**. Ce type de résultat n'a jamais pu être obtenu en utilisant la loi d'Arrhenius qui permet d'ajuster seulement les données basses températures **ou** les données hautes températures **mais jamais les deux à la fois**.

Néanmoins, les données issues des simulations numériques ne concordent pas parfaitement avec les données expérimentales. Les résultats issus de nos potentiels (qui reproduisent les propriétés de la *théorie de la fonctionnelle de la densité* - **surestiment** toujours la valeur du coefficient d'auto-diffusion par rapport aux données expérimentales. Pour les trois potentiels utilisés, **le potentiel issu de la base de données AM est le plus proche des mesures expérimentales**. Afin d'expliquer le décalage entre les valeurs obtenues par nos potentiels et les données expérimentales, nous introduisons une correction énergétique indépendante de la température ΔE dans l'équation (6.2). Le nouveau coefficient d'auto-diffusion $D^s(T)$ est alors donné par l'équation suivante :

$$D^s(T) = 3\frac{k_B T}{h} f d^2(T) e^{-\frac{1}{k_B T}\left[F_f(T)+F_{mig}(T)+\Delta E\right]} \tag{6.13}$$

Ce nouveau coefficient d'auto-diffusion $D_s(T)$ correspond à une correction de la valeur directe obtenue par les simulations numériques en faisant **l'hypothèse qu'il existe une erreur intrinséque - ΔE - entre l'énergie d'activation expérimentale et l'énergie d'activation issue de la DFT**. Les résultats obtenus pour $D_s(T)$ sont présentés dans la figure 6.4 (droite). On constate que les trois potentiels permettent d'ajuster parfaitement les données expérimentales une fois corrigés. Le potentiel issu de la fonctionnelle AM est le plus proche de l'expérience (avec $\Delta E = 0.18$ eV) et les potentiels issus des fonctionnelles LDA et PBE présentent des comportements très similaires ($\Delta E = 0.62$ eV). **Dans le cas du tungstène, c'est le potentiel issu de la fonctionelle AM qui est le plus représentatif de la propriété d'auto-diffusion. En effet, c'est ce potentiel qui nécessite la valeur de ΔE la plus faible afin d'obtenir l'accord avec les données expérimentales**.

Ce type de schéma de calcul "complet" permettant de calculer des grandeurs thermodynamiques et cinétiques avec une erreur inférieure à 0.1 eV par rapport aux données *ab initio* à 0 K ouvre de nouveaux horizons. Grâce à cette "expérience" numérique, nous pouvons proposer un **choix éclairé d'une fonctionnelle d'échange-corrélation électronique pour un problème donné en faisant une comparaison directe avec l'expérience**. Ici, nous allons tenter de voir si la fonctionnelle AM est aussi la plus représentative pour d'autres métaux cubiques centrés.

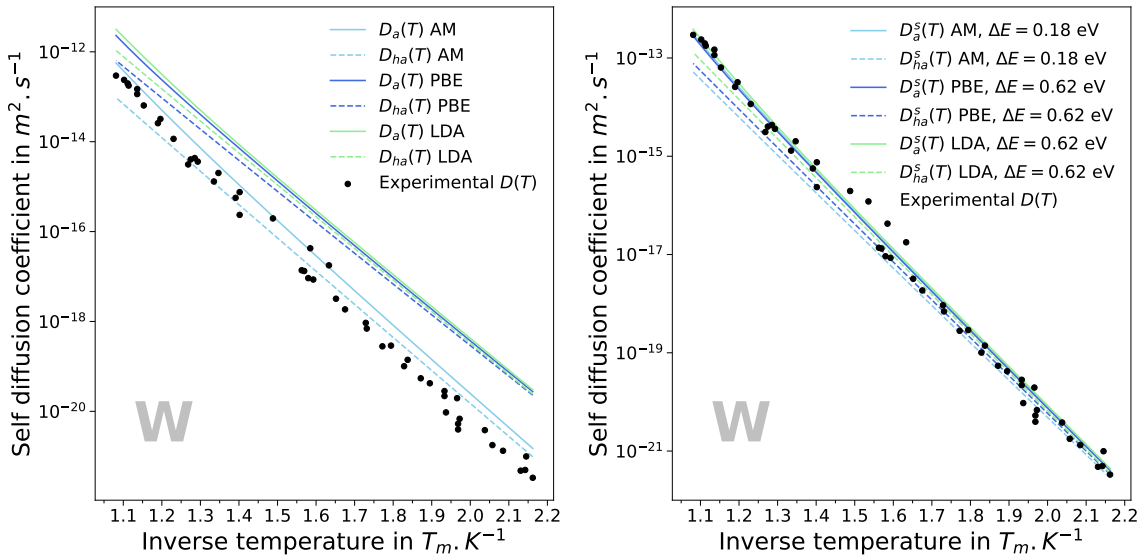

Figure 6.4: Évolution du coefficient d'auto-diffusion $D_a(T)$ du tungstène en fonction de T_m/T. Comparaison directe entre les données issues des simulations numériques calculées à l'aide des trois potentiels Machine Learning et l'expérience. Nous figurons aussi en pointillés le coefficient d'auto-diffusion harmonique $D_{ha}(T)$ attendu à partir des données à 0 K pour les trois fonctionnelles. On constate que les trois potentiels présentent des coefficients d'auto-diffusion très proches de l'expérience mais que c'est la fonctionnelle AM [44-46] qui l'approche le plus. La figure de droite présente les coefficients d'auto-diffusion $D_a^s(T)$ donnés par l'équation (6.13) pour les trois fonctionnnelles ainsi que leur comportements harmoniques $D_{ha}^s(T)$ attendus à partir des données à 0 K.

6.4 Coefficients d'auto-diffusion dans le molybdène : numérique vs. expérience

Dans le cas du molybdène, nous avons généré deux potentiels en utilisant la procédure décrite dans la section 6.2. Nous avons généré deux bases de données en utilisant la fonctionnelle (i) AM [44-46] et la fonctionnelle (ii) PBE [41]. De même que dans la section précédente, nous allons d'abord calculer l'*énergie libre de formation et de migration* de la mono-lacune pour obtenir une expression du coefficient d'auto-diffusion en fonction de la témpérature.

6.4.1 Mono-lacune dans le molybdène (Mo) : grandeurs thermodynamiques et cinétiques à températures finies

De même que pour le tungstène, nous présentons l'expansion thermique des deux potentiels ajustés en fonction de la température. Nous avons utilisé la même procédure de calcul que dans la section 6.3. Les résultats sont présentés dans la figure 6.5. Les potentiels pour le molybdène présentent le même comportement que ceux du tungstène.

Pour ces deux potentiels issus de fonctionnelle de type GGA, l'expansion thermique présente une forme quadratique spécifique des données expérimentales. **Cependant, la valeur du paramètre de maille haute température est quelque peu sous-estimée pour les hautes températures** ($T > 2T_m/3$).

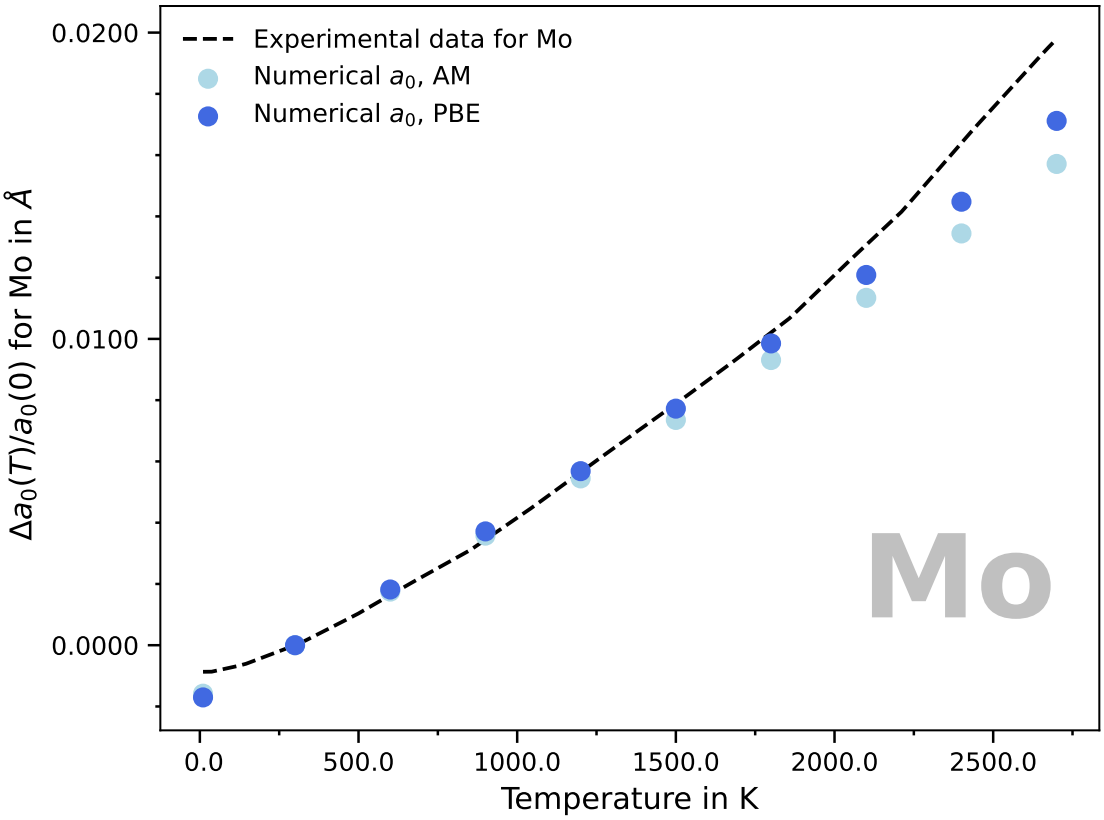

Figure 6.5: Comparaison de l'expansion thermique expérimentale renormalisée et de l'expansion thermique du potentiel Machine Learning pour le molybdène en utilisant les fonctionnelles (i) AM [44-46], (ii) PBE. Nous suivons les grandeurs : $\Delta a_0^{exp}(T)/a_0^{exp}(0) = \left(a_0^{exp}(T) - a_0^{exp}(0)\right)/a_0^{exp}(0)$ présentée en pointillés et $\Delta a_0^{ML}(T)/a_0^{ML}(0) = \left(a_0^{ML}(T) - a_0^{ML}(0)\right)/a_0^{ML}(0)$ présentée par les cercles de couleurs. Les données expérimentales sont issues de Touloukian *et al.* [249]

La figure 6.6 présente l'évolution de l'*énergie libre de formation* (gauche) resp. l'*énergie libre de migration* (droite) de la mono-lacune dans le molybène en fonction de la température. Nous donnons en pointillés, l'ajustement d'un modèle harmonique pour ces deux grandeurs. Les paramètres des modèles harmoniques ΔU_0 et ΔS_0 Eq. (6.12) sont estimés en encart. Contrairement au tungstène, le comportement de l'*énergie libre de formation* du molybdène reste **harmonique** jusqu'à environ $T = 2T_m/3$. Le comportement anharmonique - au delà de $T > 2T_m/3$ - est alors plus prononcé notamment pour le potentiel issu de la **fonctionnelle PBE**. Cette tendance se retrouve

aussi pour l'*énergie de migration* où le potentiel issu de la fonctionnelle PBE devient très anharmonique au dessus de 2000 K.

De même que pour le tungstène, la fonctionnelle AM possède l'énergie d'activation la plus grande. **On retiendra que dans le cas du molybdène c'est la "fonctionnelle PBE" qui présente le comportement le plus anharmonique**.

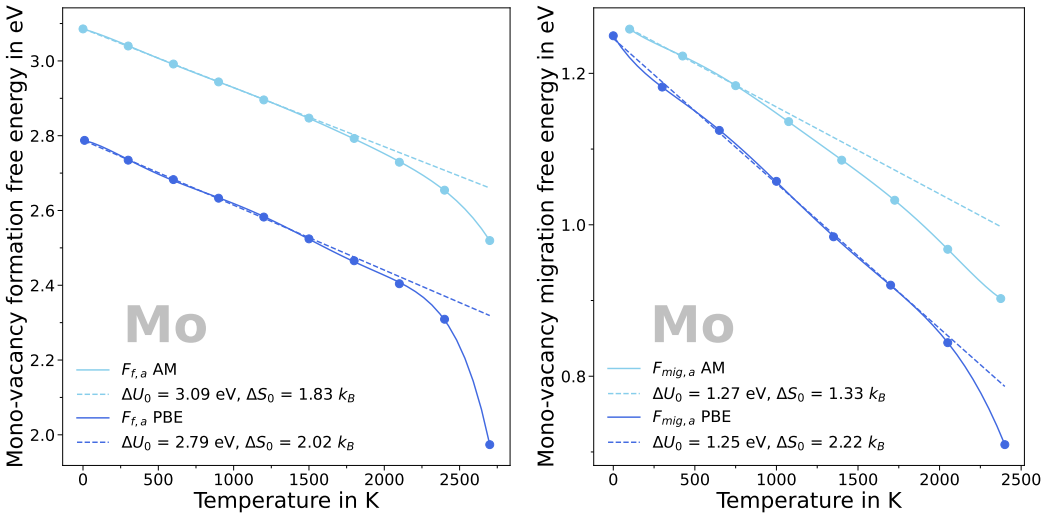

Figure 6.6: Calcul de l'*énergie libre de formation et de migration* ($F_{f,a}$ et $F_{mig,a}$) de la mono-lacune dans le molybdène en fonction de la température pour des fonctionnelles AM et PBE. La droite en pointillés représente le comportement harmonique attendu de ces deux grandeurs en fonction de la température

6.4.2 Mono-lacune dans le molybdène (Mo) : coefficients d'auto-diffusion et comparaison avec l'expérience

Nous comparons directement les données expérimentales issues de Maier *et al.* [250] avec les résultats numériques obtenus via nos simulation et l'équation (6.2). De même que dans la section 6.3.2, nous avons choisi d'ajuster un polynôme d'ordre 6 afin d'interpoler l'*énergie libre de formation et de migration* en fonction de la température et d'ajuster un polynôme d'ordre 2 afin d'interpoler l'évolution du paramètre de maille d'équilibre en fonction de la température. La comparaison directe entre les données expérimentales et les données de simulation est présentée dans la figure 6.7 (gauche). Comme dans le cas du tungstène, **l'ensemble des données issu de nos potentiels surestime le coefficient d'auto-diffusion et les données du potentiel issues de la fonctionnelle AM présentent le meilleur accord avec l'expérience**.

En utilisant l'équation (6.13), nous avons determiné la valeur de ΔE permettant d'ajuster les données expérimentales. Les valeurs de ΔE obtenues pour le molybdène sont du même ordre de grandeur que pour le tungstène (0.12 eV pour AM et 0.56 eV). Il est intéressant de noter que le **potentiel "PBE" interpole parfaitement les**

deux derniers points à hautes températures des données expérimentales grâce à sa plus grande anharmonicité. De même que pour le tungstène, c'est le potentiel "AM" qui a nécessité la valeur de ΔE la plus faible afin d'obtenir l'accord avec les données expérimentales.

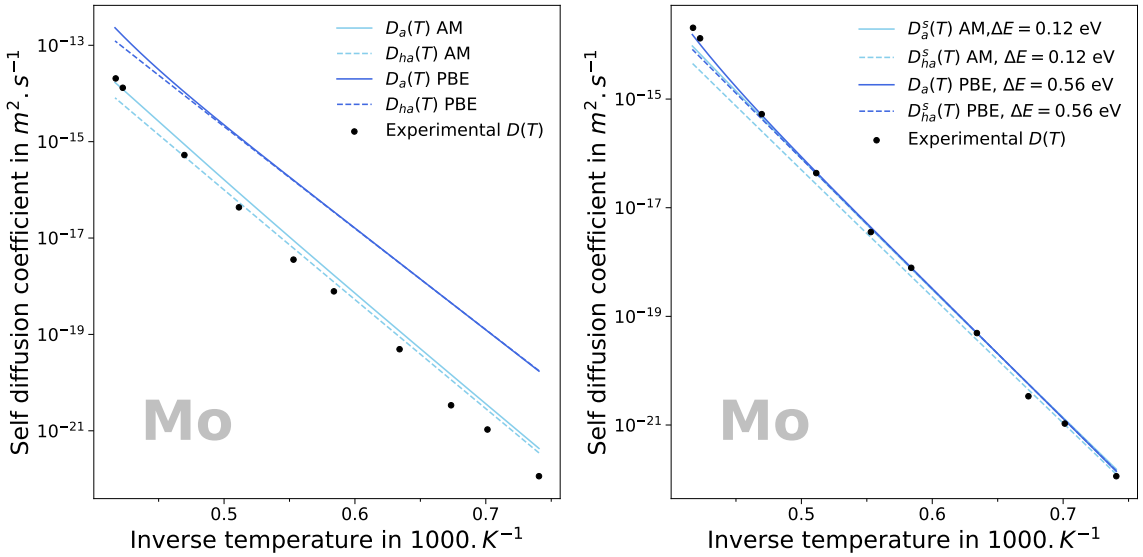

Figure 6.7: Évolution du coefficient d'auto-diffusion du molybdène $D_a(T)$ en fonction de $1000.K^{-1}$. Comparaison directe entre les données issues des simulations numériques calculées à l'aide des deux potentiels Machine Learning et l'expérience. Nous figurons aussi en pointillés le coefficient d'auto-diffusion harmonique $D_{ha}(T)$ attendu à partir des données à 0 K pour les deux fonctionnelles. On constate que les trois potentiels présentent des coefficients d'auto-diffusion très proches de l'expérience mais que c'est la fonctionnelle AM [44-46] qui approche le plus l'expérience. La figure de droite présente les coefficients d'auto-diffusion $D^s_a(T)$ donnés par l'équation (6.13) pour les deux focntionnnelles ainsi que leur comportements harmoniques $D^s_{ha}(T)$ attendus à partir des données à 0 K.

6.5 Conclusions de chapitre

Dans ce chapitre, nous avons montré que les méthodes d'*énergie libre* - de type forces biaisantes adaptatives - couplées avec des potentiels de type Machine Learning permettent de calculer des grandeurs directement quantifiables d'un point de vue expérimental. Nous nous sommes intéressés aux coefficients d'auto-diffusion dans les métaux cubiques centrés. **Nous avons estimé que le calcul direct en utilisant de la dynamique moléculaire *ab initio* afin d'échantillonner l'*énergie libre* d'activation de la mono-lacune nécessiterait environ 10^9 heures CPU par fonctionnelle d'échange-corrélation**. En utilisant notre schéma de calcul, nous pouvons effectuer ce même calcul - pour une fonctionnelle d'échange-corrélation donnée - en 5×10^6 heures CPU tout en conservant une précision énergétique "*ab initio*". Ce

calcul complet n'avait jamais été effectué jusqu'ici et nous avons montré que le **comportement non-Arrhenius haute température ($T > T_m/2$) du coefficient d'auto-diffusion est dû aux effets anharmoniques**. Les données issues de nos potentiels **surestiment** systématiquement les coefficients d'auto-diffusion par rapport aux valeurs expérimentales. Enfin, nous pouvons affirmer que c'est la fonctionnelle AM04 [44-46] qui reproduit le mieux les propriétés expérimentales de la mono-lacune dans le cas du **tungstène et du molybdène**. **Ce schéma complet de calcul peut donc aussi être utilisé de façon réciproque, c'est-à-dire en se basant sur les propriétés expérimentales d'un phénomène afin de faire un choix de fonctionnelle d'échange-corrélation électronique**.

World was on fire, and no one could save me, but you
Strange what desire will make foolish people do
I never dreamed that I'd meet somebody like you
And I never dreamed that I'd lose somebody like you.

— Wicked Game, Chris Isaak

7

Autres études utilisant les méthodes de régressions en hautes dimensions

Sommaire

Dans ce chapitre, je décris brièvement trois collaborations que j'ai effectuées au cours de mon doctorat et qui concernent trois thématiques différentes liées aux méthodes d'apprentissage automatiques appliquées à la science des matériaux :

- le développement d'une nouvelle approche statistique permettant de donner une définition **quantitative** d'un défaut dans une structure cristalline. Cette approche originale décrit la différence entre un défaut et une structure de référence par une métrique appelée *distortion score* ;
- le développement d'un nouveau type de descripteurs issu des observables GW telles que : l'énergie cinétique et les projections sur un ensemble d'orbitales atomiques. Ces descripteurs ont ensuite été utilisés afin de prédire précisément l'énergie d'ionisation de molécules en se basant sur un calcul GW sur un faible nombre de **fonctions de base** ;
- le calcul de propriétés de température finies pour des potentiels de type Machine Learning pour le fer et le tungstène (expansion thermique et l'*énergie libre* de formation de la mono-lacune) ;

Ces trois études ont mené à trois publications : (i) Goryaeva *et al.* [159], (ii) Bruneval *et al.* [21] et (iii) Goryaeva *et al.* [23].

7.1 Une nouvelle définition des défauts cristallins

Ce travail a été mené en collaboration avec Alexandra Goryaeva, Chendi Dai, Julien Dérès, Jean-Bernard Maillet et Cosmin Marinica. Un défaut cristallin est avant tout une anomalie dans une structure ordonnée. Aujourd'hui encore, il est difficile de donner un caractère quantitatif à cette notion d'anomalie. Il existe dans la littérature un certain nombre de stratégies existantes afin de détecter des défauts [251-253]. Néanmoins, chaque stratégie est adaptée à un type de défaut et il n'existe pas d'approche systématique et donc pas de définition quantitative systématique des défauts.

7.1.1 Introduction d'une distance statistique robuste

L'outil introduit ici est une distance statistique dans l'espace des descripteurs par rapport à une référence. Nous introduisons d'abord la distance de Mahalonobis, $d_{\mathcal{M}}(\underline{x})$, associée à une base de données (de matrice convariance $\underline{\underline{\Sigma}} \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ et de moyenne $\underline{\mu} \in \mathbb{R}^{\mathcal{D}}$) et à un vecteur $\underline{x} \in \mathbb{R}^{\mathcal{D}}$:

$$d_{\mathcal{M}}(\underline{x}) = \sqrt{(\underline{x}-\underline{\mu})^T \cdot \underline{\underline{\Sigma}}^{-1} \cdot (\underline{x}-\underline{\mu})} \tag{7.1}$$

La distance de Mahalonobis mesure la distance d'un vecteur $\underline{x}$ à l'enveloppe hyper-elliptique engendrée par $\underline{\underline{\Sigma}}$ et centrée sur $\underline{\mu}$. Dans le cas de la distance de Mahalonibis, l'enveloppe elliptique est peu robuste (au sens où elle intègre des données anormales à la distribution dans l'enveloppe elliptique). Ce manque de robustesse est illustré par la figure 7.1 issue de Hubert *et al.* [254] (en rouge). On constate que les anomalies dans la distribution de données peuvent tout de même être intégrées

à l'enveloppe elliptique dans le cas de la distance de Mahalonobis. Nous allons donc utiliser une autre distance statistique basée sur la notion de "minimum covariant determinant" (MCD) afin d'obtenir un indicateur statistique plus stable.

On définit la distance dite MCD, d_{MCD}, associée à une base de données (de matrice de convariance $\underline{\underline{\Sigma}} \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ et de moyenne $\underline{\mu} \in \mathbb{R}^{\mathcal{D}\times 1}$) par la formulation suivante :

$$d_{\text{MCD}}(\underline{x}) = \sqrt{(\underline{x}-\underline{\mu}_0)^T \cdot \underline{\underline{\Sigma}}_{m_0}^{-1} \cdot (\underline{x}-\underline{\mu}_0)} \tag{7.2}$$

Ici, $\underline{\underline{\Sigma}}_{m_0}$ et $\underline{\mu}_0$ sont respectivement les estimations de la matrice de covariance et du barycentre sur le sous-ensemble H_{m_0} vérifiant la condition suivante :

$$H_{m_0} = \arg \min_{H=\{\underline{x}_1,\cdots,\underline{x}_m\} \mid \text{card}(H)=m_0} \left\{ \det |\underline{\underline{\Sigma}}_H| \right\} \tag{7.3}$$

La méthode MCD permet de construire une enveloppe elliptique robuste pour une **distribution unimodale** de données. Dans le cas de la distance MCD, les anomalies de la distribution sont quantitativement mieux représentées car l'algorithme choisit le sous ensemble de données H_{m_0}, de cardinal m_0, qui **minimise la surface de l'hyper-ellipse engendrée par ce sous-ensemble**. L'application de la méthode MCD est illustrée dans la figure 7.1 issue de Hubert *et al.* [254] (en rouge). Cette méthode est beaucoup plus robuste que la distance de Mahalonobis pour identifier les anomalies.

7.1.2 Application au cas des défauts cristallins : stratification

La métrique MCD peut être directement appliquée dans l'espace des descripteurs afin de rendre compte d'anomalies de distribution. Les descripteurs étant des objets permettant la description systématique des environnements locaux, il est raisonnable de penser qu'une anomalie - au sens de la distance MCD - pour un vecteur de descripteurs correspondant à un atome i sera l'image directe d'une anomalie au sens configurationnelle - plus communément appelée **défaut** - de l'atome i. La distance MCD permet de rendre compte de façon quantitative de la différence entre une structure de référence et une anomalie. Nous appelons cette distance spécifique le *distorsion score*. Un exemple d'utilisation du *distorsion score* est donné par la figure 7.2. Dans cette figure, on calcule le *distorsion score* pour un di-amas de type $C15$ dans le fer cubique centré [7]. Les couleurs correspondent à la valeur du *distorsion score*. On constate que le *distorsion score* permet de construire une "stratification" du niveau d'anomalie de la structure. Ainsi, dans le cas du di-amas $C15$, les atomes ayant le *distorsion score* le plus élévé sont les centres des polyhèdres $Z16$. Le *distorsion score* diminue alors par "échelon" au fur et à mesure que l'on se rapproche de la structure du cristal parfait.

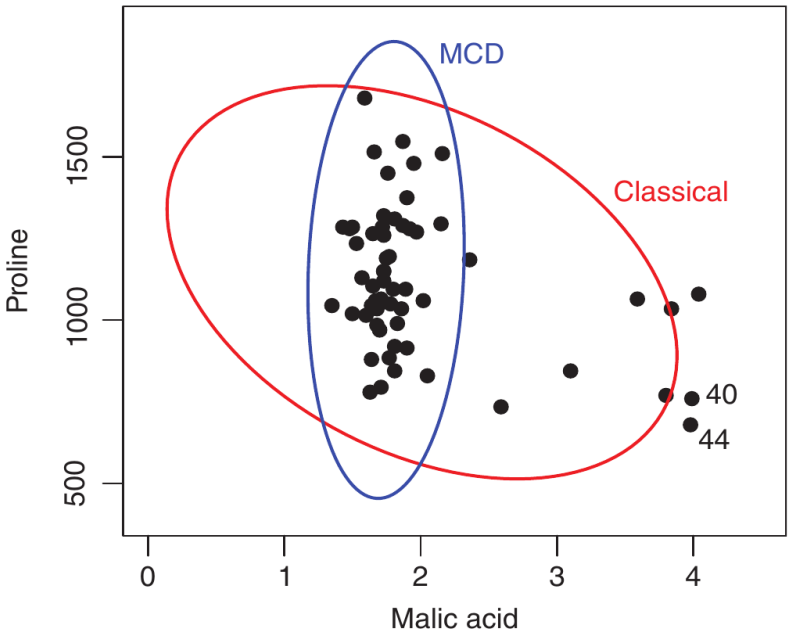

Figure 7.1: Illustration des enveloppes elliptiques engendrées par la méthode de Mahalonobis (en rouge) et par la méthode de Minimum Covariant Determinant (MCD) (en bleu). Les données présentées correspondent aux teneur des vins en proline et acide malique. La méthode dite MCD est beaucoup plus robuste pour rendre compte des anomalies dans la distribution de données. En effet, l'enveloppe définie par MCD correspond à une distribution unimodale de données. Le modèle classique va inclure des données que l'on qualifierait d'*outliers*.

La métrique induite par le *distorsion score* peut être utilisée pour identifier d'autres types de défauts, de façon quantitative. La figure 7.3 illustre la détection de (i) défauts ponctuels (lacunes, auto-interstitiels) ; (ii) défauts en deux dimensions (dislocation vis, faute d'empilement) ou (iii) défauts en trois dimensions (amas de type $C15$). Ces types de défaut ont pu être identifiés grâce à une méthode unique, ce qui était impossible jusqu'à présent avec les méthodes présentées dans la littérature [251-253].

Une telle méthode - générale - n'est possible que grâce à l'utilisation de l'**espace des descripteurs**. Celui-ci encode de façon systématique les informations géométriques d'une configuration. Cet encodage permet de traiter les données dans un même espace ce qui rend possible une **généralisation des méthodes de détections**.

7.1.3 Invariances et structure : lien entre matrice de convariance et *Hamiltonien*

L'espace des descripteurs est construit à l'aide de fonctions vérifiant les propriétés de symétrie et d'invariance du réseau cristallin. Ceci implique que l'*Hamiltonien* du système doit commuter avec l'ensemble des opérations de symétries, le laissant invariant - ce qui est naturel et connu depuis les travaux de Bloch [255]. Par définition des

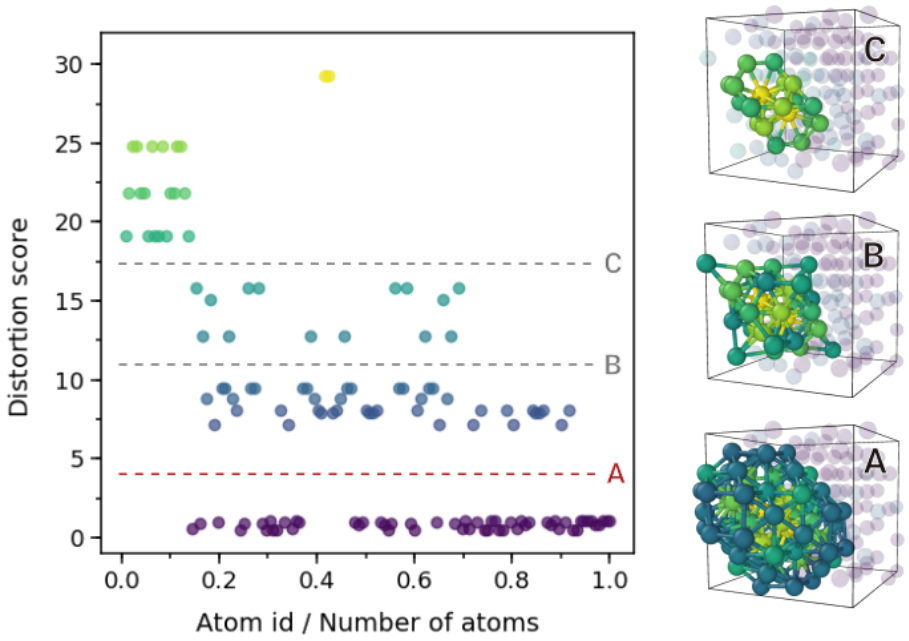

Figure 7.2: Stratification du niveau de "défaut" par rapport à une structure de réference (ici le fer cubique centré) grâce au *distortion score*. La structure représentée est un di-amas $C15$. Ce schéma est issu de Goryaeva *et al.* [159]

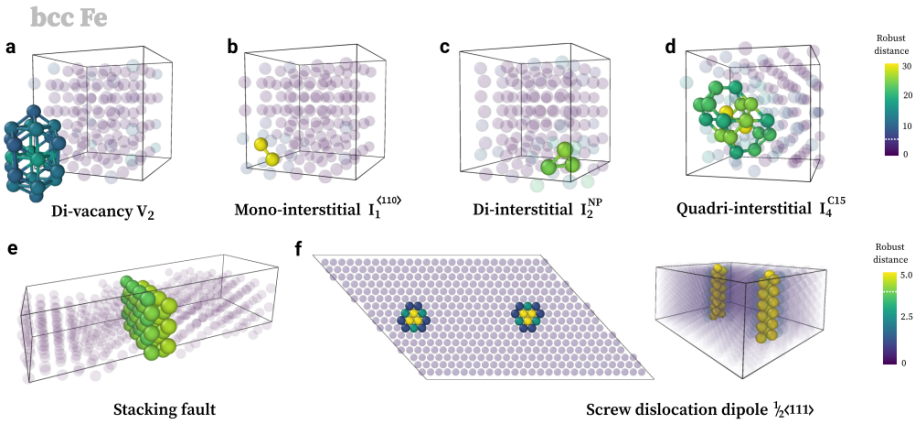

Figure 7.3: Identification de différents types de défauts cristallin grâce au *distortion score* pour une référence de fer cubique centré. L'axe des abscisses représente les identifiants des atomes normalisés de 0 à 1. Ce schéma est issu de Goryaeva *et al.* [159]

descripteurs, **la matrice de convariance des descripteurs doit aussi commuter avec lesdites opérations de symétries**. Comme deux opérateurs qui commutent peuvent être **co-diagonalisés**, on peut se demander s'il existe un lien entre la structure du *Hamiltonien* et la matrice de covariance des descripteurs. On peut alors faire un lien quantitatif direct entre des observables telles que l'énergie E d'un système et

la trace de la matrice de covariance associée $\mathrm{Tr}\left(\underline{\underline{\Sigma}}\right)$:

$$\boldsymbol{\mathcal{H}} = \sum_{\lambda=1}^{\infty} \epsilon_\lambda \boldsymbol{e}_\lambda \otimes \boldsymbol{e}_\lambda \rightleftharpoons \underline{\underline{\Sigma}} = \sum_{\lambda=1}^{\mathcal{D}} \eta_\lambda \underline{v}_\lambda \otimes \underline{v}_\lambda \tag{7.4}$$

$$E = \sum_{\lambda=0}^{\infty} \int \epsilon n(\epsilon) \delta(\epsilon - \epsilon_\lambda) d\epsilon \rightleftharpoons \mathrm{Tr}\left(\underline{\underline{\Sigma}}\right) = \sum_{\lambda=1}^{\mathcal{D}} \int \eta \delta(\eta - \eta_\lambda) d\eta \tag{7.5}$$

Nous pouvons utiliser les relations décrites dans les équations précédentes (7.5) afin de relier l'énergie locale (E_i) d'un atome i de coordonnées $\boldsymbol{q}_i$ et de vecteur de descripteur $\underline{D}_i$ à la distance statistique d_i^2 associée à cet atome :

$$\rho_i(\epsilon) = \sum_{\lambda=1}^{\infty} |\boldsymbol{e}_i \cdot \boldsymbol{e}_\lambda|^2 \delta(\epsilon - \epsilon_\lambda) \rightleftharpoons \rho_i(\eta) = \sum_{\lambda=1}^{\mathcal{D}} |\underline{D}_i \cdot \underline{v}_\lambda|^2 \delta(\eta - \eta_\lambda) \tag{7.6}$$

$$E_i = \int \epsilon \rho_i(\epsilon) n(\epsilon) d\epsilon \rightleftharpoons d_i^2 = \int \frac{1}{\eta} \rho_i(\eta) d\eta \tag{7.7}$$

La façon d'aboutir aux observables E_i et d_i^2 présente des analogies évidentes et il est intéressant de comparer directement la corrélation entre E_i et d_i^2 pour un atome i donné. La corrélation entre l'énergie locale et la distance statistique sont présentés dans la figure 7.4. Il est inconstestable que ces deux grandeurs sont fortement corrélées - et même proportionnelles -, ce qui démontre le lien structural entre l'espace de *Hilbert* associé à l'*Hamiltonien* du système et l'espace des descripteurs. Il est même possible de construire des distances statistiques *ad-hoc* à la mécanique quantique et donnant de meilleures corrélations [159].

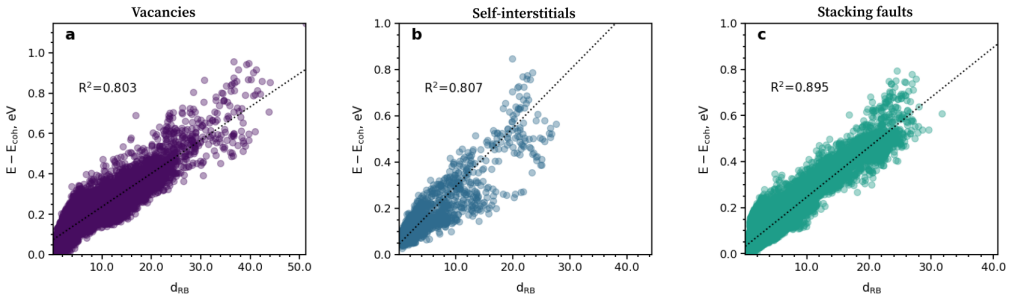

Figure 7.4: Corrélations directes entre l'énergie locale d'un atome et la distance statistique associée. Il est indéniable que ces deux grandeurs sont proportionnelles, ce qui implique un lien fort entre l'espace de *Hilbert* associé à l'*Hamiltonien* du système et l'espace des descripteurs.

En conclusion, l'introduction de distances statistiques dans l'espace des descripteurs permet de construire une méthode systématique de détection des défauts dans les matériaux cristallins grâce à l'introduction d'une métrique quantitative, les *distorsion score*, permettant de qualifier le **niveau d'anomalie d'une structure par rapport à une structure de référence**. La corrélation forte entre les distances statistiques

et l'observable d'énergie locale nous pousse à penser que **les distances statistiques peuvent être utilisées comme descripteur d'entrée afin d'engendrer des métamodèles pour certaines observables thermodynamiques** (telle que l'énergie ou l'entropie vibrationnelle harmonique). **Enfin, cette corrélation entre distance statistique et énergie locale montre une nouvelle fois le lien très étroit qui existe entre l'espace des phases et l'espace des descripteurs.**

J'ai pour ma part effectué une partie des calculs *ab initio* et de dynamique moléculaire présentés dans la publication [159]. J'ai aussi pris part à des discussions sur les distances statistiques.

7.2 Observables *GW*, Machine Learning et énergies d'ionisation

Ce travail a été mené en collaboration avec Fabien Bruneval, Ivan Maliyov et Cosmin Marinica. Cette étude porte sur le développement de nouveaux types de descripteurs issus directement de quantités provenant d'un calcul GW grâce au package `MOLGW` [256]. Ces descripteurs ont été ensuite utilisés afin de construire un métamodèle permettant de calculer l'énergie d'ionisation de molécules organiques en se basant sur l'estimation d'un calcul GW tronqué.

7.2.1 Observables *GW* et descripteurs

La méthode GW est une technique *ab initio* permettant de calculer la structure électronique d'un système. La package `MOLGW` permet d'effectuer des calculs *ab initio* dans l'approximation GW et donne accès à un certain nombre de quantités dérivées de ce calcul. Nous allons nous intéresser plus précisément à deux quantités : (i) l'énergie cinétique d'une orbitale donnée et (ii) la projection d'une fonction de base sur une orbitale donnée.

Commençons par l'énergie cinétique, on définit l'énergie cinétique de l'orbitale associée à l'état $|i\rangle$ de la façon suivante :

$$T_i = -\frac{1}{2}\langle i|\boldsymbol{\nabla}^2|i\rangle \tag{7.8}$$

Ici, $\boldsymbol{\nabla}$ est l'opération gradient. On définit la projection de Mulliken d'un état $|i\rangle$ sur une fonction de base $|\mu\rangle$ de la façon suivante :

$$p_i^\mu = \sum_\alpha \mathcal{C}_{\mu i}\mathcal{S}_{\mu\alpha}\mathcal{C}_{\alpha i} \tag{7.9}$$

Les matrices $\boldsymbol{\mathcal{C}}$ et $\boldsymbol{\mathcal{S}}$ sont respectivement la matrice de coefficients de la fonction d'onde et la matrice de recouvrement. On peut alors définir la projection de Mulliken pour un état $|i\rangle$ sur un élément donné e et pour un moment orbital donné l dans une structure $p_i^{e,l} = \sum_{\mu\in e,l} p_i^\mu$.

Les deux quantités décrites plus haut - l'énergie cinétique T_i et la projection de Mulliken $p_i^{e,l}$ - respectent les propriétés d'invariance et de symétrie de la structure par construction. Ces quantités sont dérivées de la fonction d'onde électronique solution de l'équation de *Schrödinger* dans l'approximation GW, elles contiennent donc "l'information physique" de la structure électronique. Elles présentent toutes les caractéristiques pour être d'excellents descripteurs atomiques. Pour des raisons physiques, nous retiendrons $\ln(T_i)$ et non pas T_i (cf. Bruneval *et al.* [21] pour plus de détails). Nous allons former un descripteur atomique en concaténant $\ln(T_i)$ ainsi que le projection de Mulliken de $p_i^{e,l}$. Nous nous limitons à 8 éléments de la table périodique $\mathfrak{E} = \{H, C, N, O, F, P, S, Cl\}$ et aux orbitales $\mathfrak{L} = \{s, p\}$. La dimension totale de notre nouveau descripteur est alors de 17. Le descripteur $\underline{D}_i^\dagger$ pour l'état $|i\rangle$ basé sur les quantités GW peut alors être écrit de la manière suivante avec $\bigoplus$ l'opérateur de concaténation :

$$\underline{D}_i^\dagger = \ln(T_i) \bigoplus_{e,l \in \mathfrak{E},\mathfrak{L}} p_i^{e,l} \tag{7.10}$$

7.2.2 Régression de l'énergie d'ionisation à partir d'un calcul *GW* non-convergé

La quantité d'intérêt que nous étudions est l'énergie d'ionisation [1] ϵ_i d'une molécule. Dans le calcul GW, cette quantité nécessite un grand nombre de fonctions de base initiales afin d'être convergée, ce qui est très coûteux en temps CPU.

Dans cette étude, nous cherchons à prédire des énergies d'ionisation convergées à partir de calculs GW peu coûteux. Pour cela, nous avons utilisé un nombre restreint de fonctions de base afin d'obtenir une estimation de l'énergie de ionisation $\tilde{\epsilon}_i$ et constuire un modèle de régression prédictif de la quantité $\Delta\epsilon_i = \epsilon_i - \tilde{\epsilon}_i$.

Pour cela, nous allons utiliser le descripteur $\underline{D}_i^\dagger$ (décrit plus haut) et nous construisons un modèle linéaire de régression entre $\Delta\epsilon_i$ et $\underline{D}_i^\dagger$. De même que dans le chapitre (3), nous utilisons la régression linéaire Bayésienne et nous utilisons la base de données d'entraînement utilisée par Bruneval *et al.* [21]. Les résultats de la régression pour des molécules de la base de données sont présentés dans la figure 7.5. On constate que le modèle linéaire dans l'espace des descripteurs est très précis quelque soit les éléments chimiques présents dans la base de données. Les segments gris foncés représentent l'incertitude minimale atteignable pour un calcul GW et les segments gris clair représente l'incertitude minimale expérimentale. **L'erreur commise par notre modèle linéaire est du même ordre de grandeur que la précision *GW*. De plus, l'erreur de notre modèle est - presque - toujours plus faible que l'erreur expérimentale.**

1. L'énergie d'ionisation est l'énergie nécessaire pour arracher un électron et former un cation.

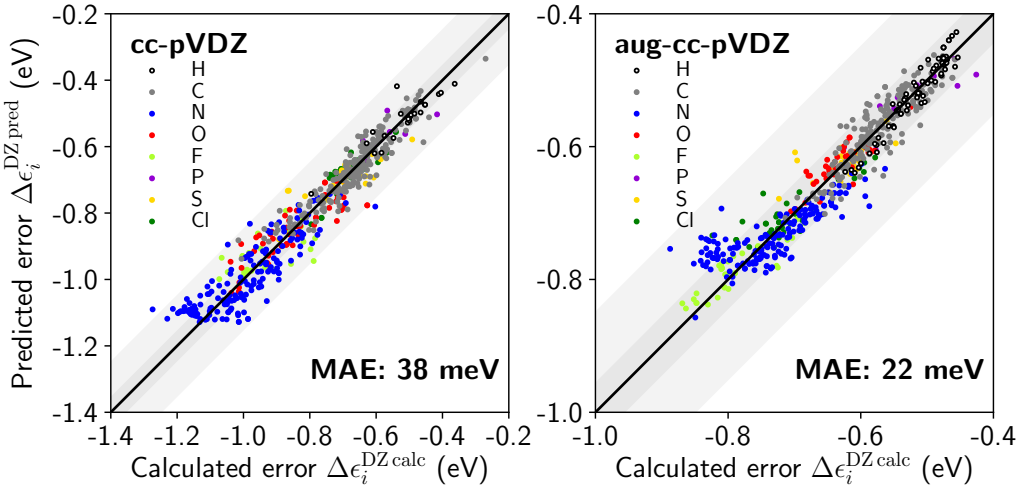

Figure 7.5: Résultats du modèle de régression pour $\Delta\epsilon_i$ pour la base de données utilisée par Bruneval *et al.* [21]. Nous utilisons le modèle linéaire dans l'espace des descripteurs présenté dans le paragraphe précedent. Ce modèle est précis pour tous les éléments présents dans la base de données. Cette figure est issue de Bruneval *et al.* [21]

Enfin, nous présentons directement les résultats obtenus pour les nucléotides formant l'ADN et/ou l'ARN (non présents dans la base de données d'entraînement). Les résultats de la convergence de la méthode sont donnés dans la figure 7.6. Nous comparons la convergence de l'énergie HOMO (Highest Occupied Molecular Orbital) en fonction de la taille de la base pour les cinq nucléotides de l'ADN et/ou ARN. Le trait noir correspond au calcul asymptotique de l'énergie HOMO, les cercles correspondent à la valeur calculée pour une taille de base de données et le carré correspond à la méthode corrigée avec notre métamodèle. On constate que le métamodèle est toujours très proche de la valeur asymptotique. Un calcul GW "classique" nécessite environ 1185 fonctions de bases, tandis que le modèle corrigé offre des résultats plus convergés avec seulement 165 fonctions de base. **Le métamodèle est donc très précis et réduit de façon importante la complexité numérique du problème.**

Pour conclure, nous avons développé un métamodèle de correction de l'énergie d'ionisation pour un ensemble de molécules organiques. Ce métamodèle s'est révélé être très précis sur la base de données d'entraînement et semble posséder une bonne capacité d'extrapolation (cf. Bruneval *et al.* [21]). **Nous nous sommes basés sur un modèle de régression simple - un modèle linéaire - et sur des descripteurs nouveaux issus des quantités calculées dans l'approximation GW. Ces descripteurs respectent les propriétés de symétries et d'invariances des systèmes étudiés. Ainsi, de par leur nature même, il contiennent "l'information" physique (ils sont des projections de la fonction d'onde solution du problème)**. Ce type de descripteurs pourrait être généralisé, par exemple en incorporant des projections magnétiques. Ils pourraient aussi être utilisés afin de construire des métamodèles pour

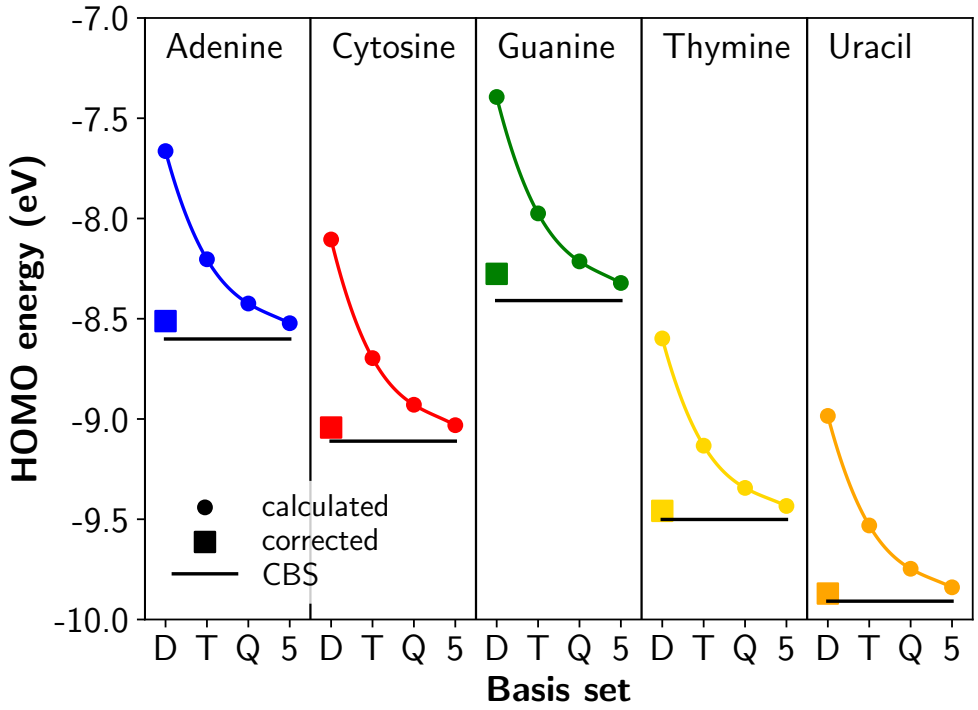

Figure 7.6: Comparaison entre convergence de l'énergie HOMO en fonction de la taille de la base pour les cinq nucléotides de l'ADN et/ou ARN. Le trait noir correspond au calcul asymptotique de l'énergie HOMO, les cercles correspondent à la valeur calculée pour une taille de base de données et le carré correspond à la méthode corrigée avec le métamodèle. On constate que le métamodèle est toujours très proche de la valeur asymptotique. Cette figure est issue de Bruneval *et al.* [21]

d'autres grandeurs liées à l'*Hamiltonien* du système.

Au cours de cette étude, j'ai participé à des discussions informelles sur les corrélations statistiques, les descripteurs atomiques et les modèles linéaires avec Fabien Bruneval. Je le remercie chaleureusement de m'avoir impliqué dans cette étude.

7.3 Potentiels Machine Learning et *énergie libre*

Ce travail a été mené en collaboration avec Alexandra Goryaeva, Julien Dérès, Petr Grigorev, Thomas D. Swinburne, James R. Kermode, Lisa Ventelon, Jacopo Baima et Cosmin Marinica. Cette étude porte sur le développement de potentiels de type Machine Learning pour le fer et le tungstène. Comme nous l'avons décrit dans le chapitre (1), les potentiels de type Machine Learning sont plus lents que les potentiels *semi-empiriques* classiques mais présentent - s'ils sont bien construits - une précision et une transférabilité plus grandes. Les potentiels construits dans cette étude sont à la fois représentatifs [23] : (i) des propriétés des défauts ponctuels ; (ii) des propriétés des structures de grandes tailles telles que les dislocations vis ou la stabilité

relative des amas $C15$ et des boucles de dislocation dans le fer ; (iii) des propriétés à températures finies telle que l'expansion thermique.Dans cette section, nous allons nous concentrer sur les propriétés de températures finies.

7.3.1 Calcul de l'expansion thermique des potentiels Machine Learning de l'étude

En utilisant la méthode des courbes *énergie libre*/volume décrite dans le chapitre 5, nous avons calculé l'expansion thermique pour quatre potentiels différents. Nous avons généré deux potentiels différents par élément (fer et tungstène) : (i) un potentiel linéaire en descripteurs et appelé LML (Linear Machine Learning) et (ii) un potentiel qui utilise le formalisme EQML présenté dans le chaptire 4 et appelé ici QNML (Quadratic Noise Machine Learning). L'ensemble des calculs d'*énergie libre* ont été effectués à l'aide du package `FEAR` en utilisant la méthode ABF Bayésienne décrite dans le chapitre 5 pour une intégration alchimique. Les résultats obtenus pour l'expansion thermique sont présentés dans la figure 7.7. Pour chaque potentiel, nous présentons la quantité $\Delta a_0(T) = a_0^{ML}(T) - a_0^{exp}(300)$. $a_0^{ML}(T)$ et $a_0^{exp}(300)$ sont respectivement les valeurs du paramètre de maille prédit par le potentiel Machine Learning et l'interpolation du paramètre de maille expérimental issu de Touloukian [249] en fonction de la température. On constate que l'expansion thermique de tous les potentiels Machine Learning est très proche de l'expansion thermique expérimentale. Contrairement à certains potentiels EAM, l'évolution quadratique de l'expansion thermique expérimentale est parfaitement reproduite par les potentiels Machine Learning (cf. chapitre 5). Cette dépendance en température est directement liée aux effets anharmoniques du champ de forces. Ainsi les potentiels Machine Learning sont capables de reproduire avec précision les propriétés d'expansion thermique des métaux, ce qui n'est pas le cas d'autres potentiels *semi-empiriques* (cf. Chap. 5).

7.3.2 Calcul de l'*énergie libre* de formation de la mono-lacune pour les potentiels Machine Learning de l'étude

Pour les quatre potentiels présentés, nous avons aussi calculés l'*énergie libre de formation* de la mono-lacune. Nous rappelons que l'*énergie libre de formation* n'est pas une grandeur directement quantifiable d'un point de vue expérimental. Néanmoins, nous souhaitons quantifier les effets anharmoniques des potentiels Machine Learning pour cette grandeur. Les résultats des calculs effectués avec le package `MAB` sont présentés dans la figure 7.8. Les traits pointillés représentent l'*énergie libre* vibrationnelle de formation harmonique calculée en se basant sur les données d'auto-diffusion à $T = 0$ K [257]. Le comportement à basse température tend à respecter la loi d'Arrhenius en bon accord avec les données expérimentales. Néanmoins, pour des températures $T > \frac{1}{3}T_f$ (avec T_f la température de fusion) on constate une déviation croissante liée aux effets anharmoniques.

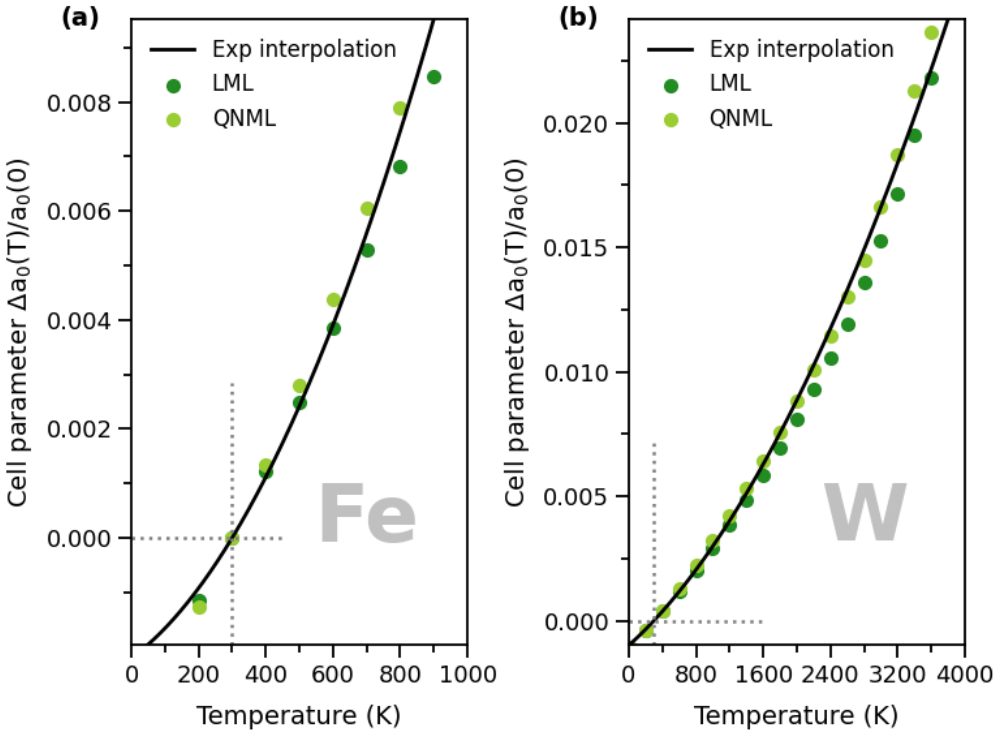

Figure 7.7: Variation des paramètres de mailles cubiques centrés $\Delta a_0(T)$ pour les potentiels Machine Learning pour le fer (à gauche) et le tunsgtène (à droite). Ceux-ci sont comparés aux données expérimentales de Touloukian [249]. On note $\Delta a_0(T) = a_0^{ML}(T) - a_0^{exp}(300)$. On constate que les données issues des potentiels Machine Learning (cercles verts) sont en très bon accord avec l'interpolation des valeurs expérimentales issues de Touloukian [249]. Cette figure est issue de Goryaeva *et al.* [23]

.

Pour conclure, nous avons réalisé des calculs de températures finies pour les quatre potentiels présentés. Nous avons pu calculer leur expansion thermique ainsi que l'*énergie libre de formation* de la mono-lacune pour un total d'environ 5×10^6 heures CPU par potentiel. **Tous les potentiels reproduisent précisément l'expansion thermique expérimentale et la valeur d'*énergie libre* vibrationnelle de formation harmonique calculées en se basant sur les données d'auto-diffusion à $T = 0$ K [257]. Ces potentiels permettent donc de calculer des grandeurs complexes et importantes dans un temps machine raisonnable.** De plus, ces potentiels mettent en évidence l'importance des effets anharmoniques et la nécessité de leur prise en compte dans les simulations à température finie. **Les potentiels Machine Learning sont donc des outils de choix pour les calculs de températures finies, grâce à leur précision et leur régularité (cf. annexe C).**

Pour cette étude [23], j'ai effectué l'ensemble des calculs d'*énergie libre* pour les quatre potentiels présentés et j'ai contribué au développement des scripts permettant de tester les potentiels ajustés.

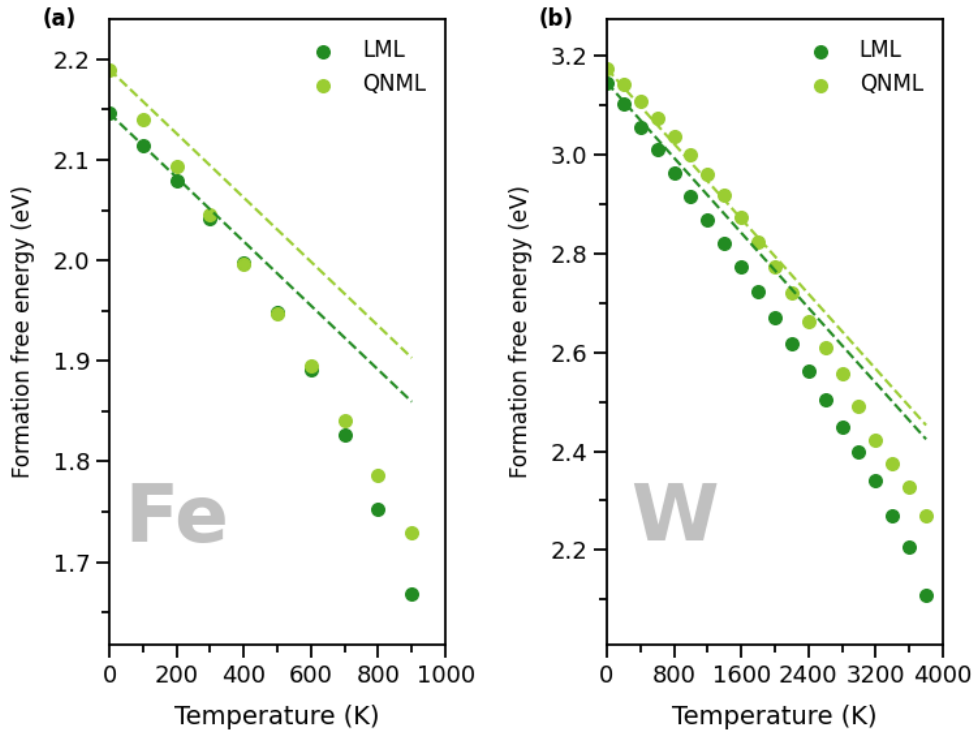

Figure 7.8: *Énergie libre* de formation anharmonique pour la mono-lacune à pression nulle pour les potentiels de fer (à gauche) et du tungstène (à droite). Ce calcul a été effectué pour les quatre potentiels présentés. Les lignes pointillées représentent l'*énergie libre* vibrationnelle de formation harmonique calculée en se basant sur les données d'auto-diffusion à $T = 0\,K$ [257]. Pour une température $T > \frac{1}{3}T_f$ (avec T_f la température de fusion), on constate la forte contribution anharmonique pour tous les potentiels Machine Learning de l'étude. Cette figure est issue de Goryaeva *et al.* [23]

7.4 Conclusion de chapitre

Dans ce chapitre, je présente l'ensemble des collaborations de ma thèse en rapport avec les méthodes d'apprentissage automatique et de régression en haute dimension. Ces collaborations sont directement reliées aux travaux décrits dans ce manuscrit. Le *distorsion score* introduit dans Goryaeva *et al.* [159] peut être utilisé comme outil quantitatif de classification d'anomalies structurelles dans les matériaux cristallins. **Le *distorsion score* - et les métriques statistiques plus généralement - peuvent être utilisées directement comme descripteurs grâce à leur analogie avec l'*Hamiltonien du système***. Dans la collaboration Bruneval *et al.* [21], nous nous sommes intéressés à construire une estimation rapide de l'énergie d'ionisation d'une molécule à partir de calcul GW peu coûteux. Nous avons montré qu'il était possible de créer des descripteurs atomiques en utilisant directement des quantités issues de calcul GW. **Ces nouveaux descripteurs atomiques peuvent être directement utilisés pour créer des métamodèles précis et transférables pour l'énergie d'ionisation**. Enfin, il est possible d'utiliser directement des potentiels Machine

Learning afin de prédire des observables thermodynamiques difficiles à calculer [23]. C'est notammment le cas du paramètre de maille d'équilibre ou de l'*énergie libre de formation* en température. **Les grandeurs calculées sont quantitativement plus proches des données expérimentales que la plupart des calculs effectués avec les potentiels *semi-empiriques classiques*.**

On a fait ce qu'on a fait comme on l'a fait
Mais on l'a fait
Tout se tranforme, rien ne se perd
Ombre et lumière.

— Civilisation, Orelsan

Conclusions et perspectives

Dans le chapitre 3, nous avons montré qu'il était possible de construire un **modèle de régression de l'entropie vibrationnelle harmonique**, **rapide et transférable**, en nous basant sur la décomposition locale de la *densité d'état des modes normaux*. Ce modèle linéaire a été testé pour le cas d'une base de données de défauts ponctuels dans le fer cubique centré. Nous avons ensuite testé la transférabilité du modèle pour des configurations contenant un très grand nombre de défauts ponctuels. Notre modèle est resté très stable - au sens de l'erreur quadratique moyenne - et très transférable. La complexité numérique de cette approche évolue comme $\mathcal{O}(N)$ - avec N le nombre d'atomes dans le système - alors que les approches "traditionnelles" évoluent comme $\mathcal{O}(N^3)$. Nous avons effectué un premier calcul d'un défaut étendu pour le cas d'une boucle de dislocation interstitielle dont l'estimation - par diagonalisation directe du Hessiene - de l'entropie vibrationnelle de formation était de 10 heures sur 3000 CPU. Notre modèle Machine Learning permet une estimation en environ 10 minutes sur un ordinateur de bureau pour une erreur d'environ 5% par rapport à la méthode de diagonalisation. **Cet exemple a aussi montré l'importance de la base de donnée d'entraînement pour l'estimation. En effet, sans prise en compte des effets de taille finies l'erreur était de** 200% **pour cette même boucle**.

Notre approche semble donc être prometteuse - au vu de sa transférabilité - pour calculer des entropies vibrationnelles harmoniques pour des défauts étendus telles que d'autres boucles de dislocations dont le calcul traditionnelle est coûteux, difficile voire impossible.

Dans le chapitre 4, nous décrivons une extension de notre modèle de régression de l'entropie vibrationnelle harmonique dans le cas de systèmes faiblement déformés. Nous montrons qu'une approche quadratique bien choisie permet d'obtenir des modèles d'une grande précision et transférables. De plus, en nous basant sur une approche analytique, nous montrons que la variation d'entropie vibrationnelle due à une petite déformation peut être intégralement décrite par **un modèle quadratique en descripteurs**. Ces modèles quadratiques ont été testés sur la base de données décrite dans le chapitre 3.

Dans un deuxième temps, nous avons montré qu'il était possible de construire un modèle de régression du logarithme des fréquences d'attaques et des barrières d'énergie dans le cadre de la *théorie de l'état de transition harmonique*. L'évaluation de ces deux quantités évolue comme $\mathcal{O}(N)$ dans le cadre de nos modèles, **ce qui rend possible leur utilisation à la volée, dans un code de Monte Carlo cinétique hors réseau, afin d'évaluer les taux de transitions**. Enfin, nous donnons une nouvelle analyse de la loi phénoménologique de Meyer-Neldel en proposant une reformulation dans l'espace des descripteurs.

Les approches purement quadratiques en descripteurs pourraient être étendues à la variation d'énergie élastique dont le formalisme est analogue [15]. La reformulation de la loi de Meyer-Neldel dans l'espace des descripteurs nous permet de donner un critère quantitatif de la validité de cette loi pour un ensemble de données. De plus, la structure de l'espace des descripteurs et l'utilisation de modèle linéaire permet de proposer de nouvelles lois de corrélations entre des quantités physiques.

Dans le chapitre 5, nous décrivons un "cas d'école" permettant d'évaluer quantitativement la vitesse de convergence et de définir un cadre "pratique de l'utilisation" des méthodes *énergie libre* à force moyenne. La vitesse de convergence a été estimée pour le cas du paramètre de maille et du module isostatique d'équilibre d'un potentiel EAM du tungstène. Nous montrons qu'il est possible de calculer ces grandeurs avec une grande précision jusqu'à la température de fusion expérimentale du tungstène. Nous proposons aussi une réformulation de la méthode ABP (Adaptative Biaising Potential) incluant une information globale sur la distribution de probabibilité associée à l'espace des phases (l'ensemble des détails techniques de cette méthode sont donnés en annexe **??**). Cette méthode est déjà implémentée en langage `python` mais n'a pas encore été mise en pratique.

La méthodologie des courbes *énergie libre/volume* pourra être utilisée pour d'autres systèmes pour calculer avec grande précision le paramètre de maille et le module isostatique d'équilibre à températures finies.

Dans le chapitre 6, nous mettons en pratique les méthodes d'*énergie libre* à biais adaptatifs couplées à l'utilisation de potentiels de type *Machine Learning* afin de calculer la dépendance en température du coefficient d'auto-diffusion de plusieurs métaux cubiques centrés. Les potentiels *Machine Learning* ont été ajustés sur des calculs *ab initio* - en utilisant plusieurs fonctionnelles d'échange-corrélation électroniques - afin de reproduire le comportement des **petits amas de lacunes**. Nous comparons directement les résultats numériques obtenus avec les mesures expérimentales. Nous montrons que la **déviation de la loi d'Arrhenius du coefficient d'auto-diffusion à hautes températures** peut être expliquée par une seule population de défauts - **la mono-lacune** - en prenant en compte les **effets anharmoniques de l'*énergie libre* d'activation**. Cette étude permet de mettre également en avant les différences entre les plusieurs fonctionnelles d'échange-corrélation testées pour ajuster les potentiels.

Ces approches couplantes non-supervisées, par calcul direct via les potentiels *Machine Learning*, pourrait permettre de sélectionner une fonc-

tionnelle d'échange-corrélation donnée pour un problème donné en se basant directement sur l'expérience.

Dans le chapitre 7, nous avons introduit une nouvelle définition quantitative et universelle des défauts critallins à l'aide des distances statistiques. Nous avons pu construire de nouveaux descripteurs basés sur des quantités GW et adaptés à des problèmes de mécanique quantique. Enfin, nous mis en oeuve la démarche du chapitre 5 afin de calculer le paramètre de maille d'équilibre et l'*énergie libre de formation* de la mono-lacune pour des potentiels de type Machine Learning.

Les distances statistiques basées sur la matrice de descripteurs pourront servir à des analyses statistiques poussées et/ou servir de nouveau descripteur atomique via le *distorsion score*.

Réflexions plus personnelles sur les méthodes d'apprenstissage automatique...

Les méthodes *Machine Learning* appliquées à la science des matériaux sont en plein essor ces dernières années. Le grand nombre de publications relatif à ces méthodes dégagent des messages parfois contradictoires et peu éclairant pour la communauté. Ainsi, nous voudrions dégager les points clefs suivants :

— La notion de descripteurs atomiques locaux ouvre la possibilité d'une démarche statistique généralisée pour les configurations atomiques. Cet espace décrit de façon systématique les environnements atomiques locaux et permet des analyses fines *de distributions et de corrélations*.

— Les potentiels *Machine Learning* sont une nouvelle classe de potentiels interatomique prometteurs. Néanmoins, ils constituent un compromis entre **précision, transférabilité et temps de calcul**. Les potentiels très précis nécessitent des descripteurs de grandes tailles et des méthodes de régression non-linéaires qui augmentent leur temps de calcul et réduisent leur transférablilité. Nous mettons en avant les modèles simples - linéaires ou quadratiques - afin de garantir la stabilité et la transférablité des potentiels ajustés. **Dans les faits, afin d'être précis et rapide, il faudrait ajuster UN potentiel *Machine Learning* pour un problème donné comme dans le cas du chapitre 6.**

— *Machine Learning* et *Fin de la physique* sont deux concepts totalement antithétiques. Les méthodes *Machine Learning*, à travers les bases de données, nécessitent plus que jamais le regard du physicien afin d'assurer la qualité de l'ajustement du potentiel ou de l'analyse statistique. **Aujourd'hui, le *Machine Learning* pour la science des matériaux a beaucoup plus besoin de la physique et du physicien que de l'inverse**. On pourra se poser la question de la validité de cette assertion dans un futur - plus ou moins - proche mais à titre personnel je suis convaincu que la *physique* a encore de beaux jours devant elle.

Annexes

Will you tell me - See you soon in a while
When my eyes fade please give me your smile
And even dark nights are ending in dawn
You'll have time to cry when I'm gone

— See You Soon, Lord of The Lost

A

Rappels et définitions de thermodynamique statistique

Sommaire

A.1 Mesure canonique

Dans cet annexe, nous définissons la convention suivante : $h^{3N} = 1$. Considérons un système $\mathcal{S}$ de température T évoluant dans l'espace des phases $\mathcal{Q} \times \mathcal{P}$ et l'*Hamiltonien* associé à $\mathcal{S}$ et noté $\mathcal{H}(\boldsymbol{q}, \boldsymbol{p})$. On définit alors la mesure canonique $\pi(\boldsymbol{q}, \boldsymbol{p})$ de la façon suivante :

$$\pi(\boldsymbol{q}, \boldsymbol{p}) = \frac{e^{-\beta \mathcal{H}(\boldsymbol{q}, \boldsymbol{p})}}{\int_{\mathcal{Q} \times \mathcal{P}} e^{-\beta \mathcal{H}(\boldsymbol{q}', \boldsymbol{p}')} d\boldsymbol{q}' d\boldsymbol{p}'} \tag{A.1}$$

Ici, nous avons $\beta = (k_B T)^{-1}$ avec k_B la constante de Boltzmann. On peut aisément vérifier que $\pi(\boldsymbol{q}, \boldsymbol{p})$ est une mesure de probabilité sur l'espace des phases. Cette mesure a été initialement introduite par Boltzmann et découle de la maximisation de l'entropie de Shannon.

Considérons le système $\mathcal{S}$ à la température T et cherchons la mesure de probabilité qui va maximiser l'entropie $S = -\int_{\mathcal{Q} \times \mathcal{P}} \rho(\boldsymbol{q}, \boldsymbol{p}) \ln \rho(\boldsymbol{q}, \boldsymbol{p}) d\boldsymbol{q} d\boldsymbol{p}$ sous la contrainte que l'énergie moyenne du système suivant la mesure $\rho(\boldsymbol{q}, \boldsymbol{p})$ doit être égale à E. La mesure $\rho^*(\boldsymbol{q}, \boldsymbol{p})$ qui vérifie ces conditions est solution du problème d'optimisation suivant :

$$\rho^*(\boldsymbol{q}, \boldsymbol{p}) = \arg \max_{\rho(\boldsymbol{q}, \boldsymbol{p}), \lambda} \left\{ -\int_{\mathcal{Q} \times \mathcal{P}} \rho(\boldsymbol{q}, \boldsymbol{p}) \ln \rho(\boldsymbol{q}, \boldsymbol{p}) d\boldsymbol{q} d\boldsymbol{p} + \lambda \left[\int_{\mathcal{Q} \times \mathcal{P}} \mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) \rho(\boldsymbol{q}, \boldsymbol{p}) d\boldsymbol{q} d\boldsymbol{p} - E \right] \right\} \tag{A.2}$$

Ici λ est un *multiplicateur de Lagrange* dont la dimension est l'inverse d'une énergie. Une solution de ce problème est donnée par la stationnarité de la fonctionnelle présentée dans l'équation (A.2) et on obtient :

$$\rho^*(\boldsymbol{q}, \boldsymbol{p}) = e^{-1 + \lambda \mathcal{H}(\boldsymbol{q}, \boldsymbol{p})} \equiv \frac{e^{-\beta \mathcal{H}(\boldsymbol{q}, \boldsymbol{p})}}{Z} \tag{A.3}$$

Où Z est une constante de normalisation. La mesure canonique $\pi(\boldsymbol{q}, \boldsymbol{p})$ est donc une solution du problème variationnel donnée par l'équation (A.2) applicable pour un système dont l'énergie moyenne est E.

A.2 Fonction de partition canonique

La fonction de partition canonique est une grandeur qui découle naturellement de la mesure canonique $\mathcal{H}(\boldsymbol{q}, \boldsymbol{p})$. En effet, elle correspond à la constante de normalisation :

$$Z = \int_{\mathcal{Q} \times \mathcal{P}} e^{-\beta \mathcal{H}(\boldsymbol{q}, \boldsymbol{p})} d\boldsymbol{q} d\boldsymbol{p} \tag{A.4}$$

Cet "artifice" de calcul est en fait lié à un grand nombre d'observables thermodynamiques macroscopiques. Nous allons en détailler quelques-unes dans les sous-sections suivantes notamment l'énergie libre Sec. A.2.1, l'énergie interne et l'entropie Sec. A.2.2.

A.2.1 Lien avec l'énergie libre

Considérons le système $\mathcal{S}$ à la température T évoluant dans l'espace des phases $\mathcal{Q}\times\mathcal{P}$ et l'*Hamiltonien* associé à $\mathcal{S}$ et noté $\mathcal{H}(\boldsymbol{q},\boldsymbol{p})$. On cherche la probabilité que le système $\mathcal{S}$ possède une énergie interne $U\pm\frac{1}{2}\delta U$ à l'équilibre thermodynamique $p_{\mathcal{S}}(U)$. Cette probabilité est donnée par l'intégrale suivante :

$$p_{\mathcal{S}}(U)=\int_{\mathcal{Q}\times\mathcal{P}}\mu_{\delta U}(|\mathcal{H}(\boldsymbol{q},\boldsymbol{p})-U|)\pi(\boldsymbol{q},\boldsymbol{p})d\boldsymbol{q}d\boldsymbol{p}=\frac{\Omega_{\mathcal{S}}(U)e^{-\beta U}}{Z} \tag{A.5}$$

Ici $\mu_{\delta U}(|\mathcal{H}(\boldsymbol{q},\boldsymbol{p})-U|)$ est la mesure de comptage définie par l'équation (3.2) et $\Omega_{\mathcal{S}}(U)$ le nombre de micro-états du système $\mathcal{S}$ possédant une énergie $U\pm\frac{1}{2}\delta U$. Si l'état d'énergie U est un état d'équilibre du système $\mathcal{S}$ alors l'événement associé à la probabilité $p_{\mathcal{S}}(U)$ est presque certain, c'est-à-dire $p_{\mathcal{S}}(U)=1$. En passant au logarithme dans l'équation (A.5) et en utilisant la définition de l'entropie de Boltzmann, on obtient la relation suivante :

$$-\beta^{-1}\ln Z\overset{\text{p.s}}{=}U-TS \tag{A.6}$$

On constate que l'énergie libre du système $F=U-TS$ est directement reliée à la fonction de partition canonique par la relation Eq. (A.6).

A.2.2 Lien avec l'énergie interne et l'entropie

On peut aussi déduire une relation simple entre l'énergie interne du système et la fonction de partition. Ainsi, utilisons la définition de la fonction partition donnée par l'équation (A.4), composons par le logarithme et dérivons par rapport à β :

$$-\partial_\beta\ln Z=\int_{\mathcal{Q}\times\mathcal{P}}\mathcal{H}(\boldsymbol{q},\boldsymbol{p})\pi(\boldsymbol{q},\boldsymbol{p})d\boldsymbol{q}d\boldsymbol{p}\equiv U \tag{A.7}$$

L'énergie interne E du système peut donc être calculée avec une simple dérivation par rapport à β du logarithme de la fonction de partition.

En utilisant la relation thermodynamique $F=U-TS$ et l'expression de l'énergie libre de la fonction de partition issue de l'équation (A.6), on peut en déduire une expression simple de S :

$$S=\partial_T\left\{\beta^{-1}\ln Z\right\} \tag{A.8}$$

La fonction de partition canonique permet donc d'accéder à de nombreuses observables thermodynamiques macroscopiques. Nous allons nous servir de ces relations pour déduire l'expression de l'entropie vibrationnelle harmonique donnée dans le chapitre (3).

A.3 Cas des phonons dans le cadre de l'approximation harmonique

Nous allons déterminer l'expression de l'entropie vibrationnelle dans le cadre de l'approximation harmonique donnée dans l'équation (3.19). Pour cela, commençons par donner l'expression de la fonction de partition d'un phonon de pulsation ω_ν à la température T :

$$z_\nu(T) = \sum_{n=0}^{\infty} e^{-\beta\hbar\omega_\nu\left(n+\frac{1}{2}\right)} = \frac{e^{-\frac{1}{2}\beta\hbar\omega_\nu}}{1-e^{-\beta\hbar\omega_\nu}} = \frac{1}{2\sinh\left(\frac{1}{2}\beta\hbar\omega_\nu\right)} \tag{A.9}$$

Dans le cadre de l'approximation harmonique, les modes de phonons sont indépendants. On peut donc écrire la fonction de partition totale du système $Z_{vib}(N,T)$ comme étant le produit des fonctions de partitions. Pour chaque phonon du système, on a alors :

$$F_{vib}(N,T) = \beta^{-1}\sum_{\nu=1}^{3N}\ln\left(2\sinh\left(\frac{1}{2}\beta\hbar\omega_\nu\right)\right) \tag{A.10}$$

En utilisant ensuite l'expression de l'entropie vibrationnelle en fonction de l'énergie libre donnée par l'équation (A.8), on obtient la formulation suivante :

$$S_{vib}(N,T) = -k_B\beta^2\sum_{\nu=1}^{3N}\left[-\beta^{-2}\ln\left(2\sinh\left(\frac{1}{2}\beta\hbar\omega_\nu\right)\right) + \frac{1}{2}\beta^{-1}\hbar\omega_\nu\coth\left(\frac{1}{2}\beta\hbar\omega_\nu\right)\right] \tag{A.11}$$

Ici, $\sinh(\cdot)$ et $\coth(\cdot)$ sont respectivement les fonctions sinus et cotangente hyperboliques. Si on se place à une température très supérieure à la température de Debye, on peut assurer que $\forall\omega_\nu,\ \beta\hbar\omega_\nu \ll 1$. On peut alors effectuer un développement limité au premier ordre de l'équation (A.11) et on obtient l'expression de l'équation (3.19) :

$$S_{vib}(T,N) = k_B\sum_{\nu=1}^{3N}\left[\ln\left(\frac{k_BT}{\hbar\omega_\nu}\right)+1\right] \tag{A.12}$$

A.4 Calcul d'*énergie libre* sous contraintes

Nous donnons l'expression de l'*énergie libre* d'un système possédant un *Hamiltonien* de type Einstein en imposant des contraintes sur la fonction de partition. En nous basant sur les travaux de Ryckaert *et al.* [258], on peut écrire la fonction de partition sous la contrainte σ :

$$Z_{Ref,c} = \frac{1}{h^{3N}}\int_{\mathcal{Q}\times\mathcal{P}}\exp\left[\beta H(\boldsymbol{q}_i,\boldsymbol{p}_i)\right]\delta\left[\sigma(\boldsymbol{q})\right]\delta(\boldsymbol{G}^{-1}\cdot\dot{\sigma}(\boldsymbol{p}))d^{3N}\boldsymbol{p}\,d^{3N}\boldsymbol{q} \tag{A.13}$$

Ici, $\dot{\sigma}$ est la dérivée temporelle de la contrainte et le tenseur $\boldsymbol{G}$ est défini par l'équation :

$$G_{kl} = \sum_{i=1}^{N}\frac{1}{m_i}\nabla_{\boldsymbol{q}_i}\sigma_k\cdot\nabla_{\boldsymbol{q}_i}\sigma_l \tag{A.14}$$

Considérons la contrainte $\sigma(\boldsymbol{q}) = \sum_{i=1}^{N} m_i \boldsymbol{q}_i$ sur le centre de masse du système. Dans le cas d'une fonction de partition de type Einstein, l'équation (A.13) va pouvoir être séparée en deux contributions : (i) la contribution cinétique et (ii) la contribution configurationnelle. Dans le cas de la contribution cinétique, la fonction de partition s'écrit :

$$Z_{Ein,c}^{kinetic} = \frac{1}{h^{3N}} \int_{\mathcal{P}} \exp\left(-\frac{\beta}{2} \sum_{i=1}^{N} \frac{\boldsymbol{p}_i^2}{2m_i}\right) \delta\left(\sum_{i=1}^{N} \boldsymbol{p}_i\right) d\boldsymbol{p} \tag{A.15}$$

Cette équation peut alors être simplifiée sous la forme :

$$Z_{Ein,c}^{kinetic} = \left(\frac{\beta h^2}{2\pi \left(\sum_{i=1}^{N} m_i\right)}\right)^{3/2} Z_{Ein}^{kinetic}, \tag{A.16}$$

L'*énergie libre cinétique* sous la contrainte du centre de masse s'écrit alors :

$$F_{Ein,c}^{kinetic} = -\beta^{-1} \ln \left[\frac{\beta h^2}{2\pi \left(\sum_{i=1}^{N} m_i\right)}\right]^{3/2} - \beta^{-1} \ln Z_{Ein}^{kinetic}. \tag{A.17}$$

Par analogie, on peut traiter la partie configurationnelle de la fonction de partition sous la contrainte σ :

$$Z_{Ein,c}^{config} = \int_{\mathcal{Q}} \exp\left(-\frac{\beta}{2} \sum_{i=1}^{N} m_i \omega_i^2 (\boldsymbol{q}_i - \boldsymbol{q}_{i,0})^2\right) \delta\left(\frac{\sum_{i=1}^{N} m_i r_i}{\sum_{i=1}^{N} m_i}\right) d\boldsymbol{q}, \tag{A.18}$$

Après calcul, on obtient :

$$Z_{Ein,c}^{config} = \frac{\beta^{\frac{3}{2}}}{2\pi \left(\sum_{i=1}^{N} \frac{\mu_i^2}{m_i \omega_i^2}\right)^{\frac{3}{2}}} Z_{Ein}^{config} \tag{A.19}$$

Ici, μ_i est la masse réduite $\mu_i = m_i / \sum_i m_i$. En combinant les deux contributions de la fonction de partition, on obtient une expression de l'*énergie libre* sous la contrainte σ sur le centre de masse :

$$F_{Ein,c} = -\beta^{-1} \frac{3}{2} \ln \left(\frac{\beta^2 h^2}{4\pi^2 (\sum_{i=1}^{N} m_i)(\sum_{i=1}^{N} \frac{\mu_i^2}{m_i \omega_i^2})}\right) + F_{Ein}. \tag{A.20}$$

Finalement, dans le cas d'oscillateurs identiques, l'équation (A.20) se réduit à l'expression suivante :

$$F_{Ein,c} = -\beta^{-1} (3N - 3) \ln \frac{2\pi}{\beta h \omega}. \tag{A.21}$$

Oh, how I wish for soothing rain
All I wish is to dream again
My loving heart was deep in thought
For hope I'd give my everything

— Nemo, Nightwish

B

Développement analytique de la correction d'entropie basée sur le formalisme de *Green* pour les déformations

Sommaire

B.1 Quelques lemmes et définitions...

Lemme B.1. *Soient* $\boldsymbol{T}^1$ *et* $\boldsymbol{T}^2 \in \mathbb{C}^{3N\times 3N}$. *On définit l'opération de contraction simple de la façon suivante :*

$$\boldsymbol{T}^1 \cdot \boldsymbol{T}^2 : \mathbb{C}^{3N\times 3N} \times \mathbb{C}^{3N\times 3N} \to \mathbb{C}^{3N\times 3N}$$
$$\left(T^1 \cdot T^2\right)_{ij} = \sum_{k=1}^{3N} T^1_{ik} T^2_{kj} \tag{B.1}$$

Lemme B.2. *Soit* $\boldsymbol{T} \in \mathbb{C}^{3N\times 3N}$. *On définit la norme associée à* $\boldsymbol{T}$ *de la façon suivante :*

$$\|\boldsymbol{T}\| \equiv \|\boldsymbol{T}\|_\infty = \max_{i,j}\left(|T_{ij}|\right) \tag{B.2}$$

Ici i, j *sont les indices du tenseur* $\boldsymbol{T}$. *La dimension finie de* $\boldsymbol{T}$ *implique que toutes les normes sont équivalentes.*

Lemme B.3. *Soit* $\boldsymbol{\lambda} \in \mathbb{C}^{3N\times 3N}$ *tel que* $\|\mathbf{1}\|_\infty \gg \|\boldsymbol{\lambda}\|_\infty$, *on peut alors approximer* $\det(\mathbf{1}+\boldsymbol{\lambda})$ *par la formule suivante :*

$$\det(\mathbf{1}+\boldsymbol{\lambda}) = 1 + \mathrm{Tr}(\boldsymbol{\lambda}) + \mathcal{O}(|\lambda|^2) \tag{B.3}$$

Ici, $\det(\cdot)$ *est l'opérateur déterminant,* $\mathrm{Tr}(\cdot)$ *est l'opérateur trace,* $\mathbf{1}$ *est le tenseur identité et* $\|\cdot\|_\infty$ *est la norme définie par le lemme (B.2).*

Lemme B.4. *Soit* $\boldsymbol{T} \in \mathbb{C}^{3N\times 3N}$ *tel que* $\mathrm{Arg}\left[\mathrm{Sp}(\boldsymbol{T})\right] \in [0, \frac{\pi}{2}]$, *nous avons l'égalité suivante :*

$$\mathrm{Tr}(\ln(\boldsymbol{T})) = \ln(\det|\boldsymbol{T}|) \tag{B.4}$$

Ici, $\mathrm{Arg}(\cdot)$ *est l'argument du nombre complexe associé,* $|\cdot|$ *est le module du nombre complexe et* $\mathrm{Sp}(\boldsymbol{T})$ *est le spectre du tenseur* $\boldsymbol{T}$ *c'est-à-dire l'ensemble des* λ *vérifiant :*

$$\mathrm{Sp}(\boldsymbol{T}) = \{\lambda \in \mathbb{C} \mid \det(\boldsymbol{T} - \lambda\mathbf{1}) = 0\} \tag{B.5}$$

Lemme B.5. *Soient* $\boldsymbol{T_1} \in \mathbb{C}^{3N\times 3N}$ *et* $\boldsymbol{T_2} \in \mathbb{C}^{3N\times 3N}$, *on dit que* $\boldsymbol{T_1} \simeq \boldsymbol{T_2}$ *si il existe* $\boldsymbol{\lambda} \in \mathbb{C}^{3N\times 3N}$ *tel que :*

$$\boldsymbol{T_1} = \boldsymbol{T_2} + \boldsymbol{\lambda} \tag{B.6}$$
$$\|\boldsymbol{\lambda}\cdot\boldsymbol{T_2}^{-1}\|_\infty \ll 1 \tag{B.7}$$

Ici, $\boldsymbol{T_2}^{-1}$ *est l'inverse du tenseur* $\boldsymbol{T_2}$ *et* $\|\cdot\|_\infty$ *est la norme définie par le lemme (B.2).*

B.2 Principales hypothèses de l'approche perturbative

La construction théorique proposée succinctement dans la sous-section (4.2.2) est basée sur les hypothèses suivantes :

— il existe une relation linéaire entre la fonction de *Green* non-perturbée $\boldsymbol{\mathcal{G}}^{(0)}(\omega)$ du système non déformé et la correction du premier ordre $\boldsymbol{\mathcal{G}}^{(1)}(\omega)$ pour le système perturbé telle qu'on a :

$$\boldsymbol{\mathcal{G}}^{(1)}(\omega) = \boldsymbol{\mathcal{G}}^{(0)}(\omega) \cdot \boldsymbol{\delta} \cdot \boldsymbol{\mathcal{G}}^{(0)}(\omega) \tag{B.8}$$

— l'opérateur linéaire $\boldsymbol{\delta} \in \mathbb{C}^{3N\times 3N}$ est une petite déformation induite dans l'espace des phases agissant sur l'équation (3.20) et vérifiant les propriétés suivantes :

$$\|\boldsymbol{\delta} \cdot \boldsymbol{\mathcal{G}}^{(0)}(\omega)\|_\infty \ll 1 \tag{B.9}$$

$$\|\boldsymbol{\delta} \cdot \boldsymbol{\mathcal{G}}^{(0)}(\omega)\|_\infty = \max_{i,j} \left(\left| \sum_{k=1}^{3N} \delta_{ik} \mathcal{G}_{kj} \right| \right) \tag{B.10}$$

Ici, i, j sont les indices du tenseur $\boldsymbol{\delta} \cdot \boldsymbol{\mathcal{G}}^{(0)}(\omega) \in \mathbb{C}^{3N\times 3N}$. Le tenseur $\boldsymbol{\delta}$ est directement relié à la déformation du système. En effet, si on note respectivement $\boldsymbol{\mathcal{G}}^{\epsilon}(\omega)$ et $\boldsymbol{\mathcal{G}}^{(0)}(\omega)$ les fonctions de *Green* solutions de l'équation (3.20) pour le système déformé et le système non déformé, on a :

$$\boldsymbol{\delta} = \left[\boldsymbol{\mathcal{G}}^{(0)}(\omega)\right]^{-1} \cdot \left[\boldsymbol{\mathcal{G}}^{\epsilon}(\omega) - \boldsymbol{\mathcal{G}}^{(0)}(\omega)\right] \cdot \left[\boldsymbol{\mathcal{G}}^{(0)}(\omega)\right]^{-1} \tag{B.11}$$

B.3 Application de l'approche pertubative pour la *densité d'état* de *modes normaux* et la variation d'entropie vibrationnelle

La variation de la *densité d'état* de *modes normaux* $\Delta\Omega(\omega)$ est directement reliée à la partie imaginaire de la perturbation au premier ordre de la fonction de *Green* solution de l'équation (3.20) que l'on note $\boldsymbol{\mathcal{G}}^{(1)}(\omega)$, ce qui se traduit par la relation suivante :

$$\Delta\Omega(\omega, \boldsymbol{\epsilon}) = -\frac{2\omega}{\pi} \operatorname{Tr}\left\{ \Im\Big(\underbrace{\boldsymbol{\mathcal{G}}^{\epsilon}(\omega) - \boldsymbol{\mathcal{G}}^{(0)}(\omega)}_{\boldsymbol{\mathcal{G}}^{(1)}(\omega)} \Big) \right\} \tag{B.12}$$

$$= -\frac{2\omega}{\pi} \operatorname{Tr}\left\{ \Im\Big(\boldsymbol{\mathcal{G}}^{(0)}(\omega) \cdot \boldsymbol{\delta} \cdot \boldsymbol{\mathcal{G}}^{(0)}(\omega) \Big) \right\} \tag{B.13}$$

En utilisant les propriétés de $\boldsymbol{\mathcal{G}}^{(0)}(\omega)$, le lemme (B.4) et la propriété de linéarité de l'opérateur trace et l'opérateur de dérivation nous obtenous un nouvelle expression de $\Delta\Omega(\omega, \boldsymbol{\epsilon})$:

$$\Delta\Omega(\omega, \boldsymbol{\epsilon}) = -\frac{2\omega}{\pi} \operatorname{Tr}\left\{ \Im\Big(\boldsymbol{\mathcal{G}}^{(0)}(\omega) \cdot \boldsymbol{\delta} \cdot \boldsymbol{\mathcal{G}}^{(0)}(\omega) \Big) \right\} \tag{B.14}$$

$$= -\frac{1}{\pi} \Im\Big(\operatorname{Tr}\left\{ \partial_\omega \ln\left[\mathbf{1} + \boldsymbol{\mathcal{G}}^{(0)}(\omega) \cdot \Delta\tilde{\boldsymbol{\mathfrak{D}}}(\omega, \boldsymbol{\epsilon}) \right] \right\} \Big) \tag{B.15}$$

$$= -\frac{1}{\pi} \Im\Big(\partial_\omega \Big[\ln\Big(\det\left\{ \mathbf{1} + \boldsymbol{\mathcal{G}}^{(0)}(\omega) \cdot \Delta\tilde{\boldsymbol{\mathfrak{D}}}(\omega, \boldsymbol{\epsilon}) \right\} \Big) \Big] \Big) \tag{B.16}$$

Ici, le tenseur $\Delta\tilde{\mathfrak{D}}(\omega,\boldsymbol{\epsilon})$ est la perturbation induite par la déformation sur la matrice dynamique du système. En utilisant la fonction de *Green* solution de l'équation (3.20), nous pouvons dériver une expression analytique pour $\tilde{\mathfrak{D}}$ (ici nous utilisons la notation abusive pour désigner les tenseurs inverses) :

$$\Delta\tilde{\mathfrak{D}}(\omega,\boldsymbol{\epsilon}) = -\frac{\Delta\boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon})}{\boldsymbol{\mathcal{G}}(\omega)\cdot\boldsymbol{\mathcal{G}}(\omega)}, \tag{B.17}$$

Avec $\Delta\boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon}) = \boldsymbol{\mathcal{G}}^{(\epsilon)}(\omega) - \boldsymbol{\mathcal{G}}^{(0)}(\omega)$ et $\boldsymbol{\mathcal{G}}(\omega) = \boldsymbol{\mathcal{G}}^{(0)}(\omega) \simeq \boldsymbol{\mathcal{G}}^{(\epsilon)}(\omega)$. Dans le régime des perturbations, nous avons $\|\boldsymbol{\delta}\cdot\boldsymbol{\mathcal{G}}^{(0)}(\omega)\|_\infty \ll 1$, en utilisant l'équation (B.17), les propriétés de $\boldsymbol{\mathcal{G}}^{(0)}(\omega)$ et le lemme (B.3) pour le tenseur $\Delta\boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon})\cdot[\boldsymbol{\mathcal{G}}(\omega)]^{-1}$ nous pouvons déduire l'expression analytique suivante pour $\Delta\Omega(\omega)$:

$$\Delta\Omega(\omega,\boldsymbol{\epsilon}) = \frac{1}{\pi}\Im\left(\partial_\omega\left\{\mathrm{Tr}\left(\frac{\Delta\boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon})}{\boldsymbol{\mathcal{G}}(\omega)}\right)\right\}\right)$$

Grâce à l'expression analytique donnée par l'équation (B.3), nous pouvons calculer la variation d'entropie ΔS en utilisant l'expression (4.10). En effectuant une intégration par partie et sous l'hypothèse que $\Delta S_0^\infty < \infty$, on obtient :

$$\Delta S(\boldsymbol{\epsilon}) = \frac{k_B}{\pi}\int_0^{+\infty}\left[\ln\left(\frac{\hbar\omega}{k_BT}\right)-1\right]\Im\left(\partial_\omega\left\{\mathrm{Tr}\left(\frac{\Delta\boldsymbol{\mathcal{G}}(\omega)}{\boldsymbol{\mathcal{G}}(\omega)}\right)\right\}\right)d\omega \tag{B.18}$$

$$= \Delta S_0^{+\infty} - \frac{k_B}{\pi}\int_0^{+\infty}\omega^{-1}\Im\left\{\mathrm{Tr}\left(\frac{\Delta\boldsymbol{\mathcal{G}}(\omega)}{\boldsymbol{\mathcal{G}}(\omega)}\right)\right\}d\omega \tag{B.19}$$

En utilisant l'égalité tensorielle $\frac{\Delta\boldsymbol{\mathcal{G}}(\omega)}{\boldsymbol{\mathcal{G}}(\omega)} = \Delta\ln\{\boldsymbol{\mathcal{G}}(\omega)\}$ nous avons :

$$\Delta S(\boldsymbol{\epsilon}) = \Delta S_0^{+\infty} - \frac{k_B}{\pi}\int_0^{+\infty}\omega^{-1}\,\mathrm{Tr}\left\{\Im\Big(\Delta\ln\{\boldsymbol{\mathcal{G}}(\omega)\}\Big)\right\}d\omega \tag{B.20}$$

$$= \Delta S_0^{+\infty} - \frac{k_B}{\pi}\int_0^{+\infty}\omega^{-1}\,\mathrm{Tr}\left\{\Delta\,\mathrm{Arg}\left(\boldsymbol{\mathcal{G}}(\omega)\right)\right\}d\omega \tag{B.21}$$

Ici $\mathrm{Arg}(\cdot)$ est l'argument de chaque composante du tenseur $\boldsymbol{\mathcal{G}}(\omega)$. Sous l'hypothèse que $\|\boldsymbol{\delta}\|_\infty \ll 1$, nous pouvons déduire les propositions suivantes :

— $\left\{\hat{\mathbf{e}}_\nu\right\}_{\hat{\mathbf{e}}_\nu\in\mathcal{V}^{(0)}} \simeq \left\{\hat{\mathbf{e}}_\nu + \Delta\hat{\mathbf{e}}_\nu\right\}_{\hat{\mathbf{e}}_\nu+\Delta\hat{\mathbf{e}}_\nu\in\mathcal{V}^{(\epsilon)}}$ où $\mathcal{V}^{(0)}$ et $\mathcal{V}^{(\epsilon)}$ sont respectivement les espaces vectoriels induits par les *modes normaux* de la configuration initiale et de la configuration déformée.

— $\|\Delta\Im(\boldsymbol{\mathcal{G}}(\omega))\| \gg \|\Delta\Re(\boldsymbol{\mathcal{G}}(\omega))\|$.

Nous obtenons une nouvelle expression de l'équation (B.21) ne dépendant que de la partie imaginaire des fonctions de *Green* et nous pouvons montrer que $\Delta S_0^{+\infty} = 0$ (voir l'annexe (B.4)), où $\xi_0^{i\alpha}$ et $\xi_\epsilon^{i\alpha}$ sont respectivement les facteurs d'occupations associés aux modes de vibration de fréquence ν^0 et ν^ϵ pour l'atome i selon la direction α pour

la configuration initiale $\mathcal{E}^{(0)}$ et la configuration déformée $\mathcal{E}^{(\epsilon)}$:

$$\Delta S(\boldsymbol{\epsilon}) = -\frac{k_B}{\pi}\int_0^{+\infty}\omega^{-1}\,\mathrm{Tr}\left\{\frac{\Delta\Im\{\mathcal{G}_{i\alpha}(\omega)\}}{|\mathcal{G}_{i\alpha}|}\,\hat{\mathbf{e}}_{i\alpha}^T\otimes\hat{\mathbf{e}}_{i\alpha}\right\}d\omega \tag{B.22}$$

$$= -k_B\left\{\sum_{i,\alpha=1}^{N,3}\left(\sum_{\nu^\epsilon=1}^{3N}\frac{\omega_{\nu^\epsilon}^{-2}|\xi_\epsilon^{i\alpha}|^2}{\sqrt{\sum_{\nu^0=1}^{3N}\frac{|\xi_0^{i\alpha}|^4}{|\omega_{\nu^0}^2-\omega_{\nu^\epsilon}^2|^2}}}-\sum_{\nu^0=1}^{3N}\frac{\omega_{\nu^0}^{-2}|\xi_0^{i\alpha}|^2}{\sqrt{\sum_{\nu^\epsilon=1}^{3N}\frac{|\xi_\epsilon^{i\alpha}|^4}{|\omega_{\nu^\epsilon}^2-\omega_{\nu^0}^2|^2}}}\right)\right\} \tag{B.23}$$

En faisant l'hypothèse que $|\sum_{\nu^\epsilon\neq\nu^0}| \ll |\sum_{\nu^\epsilon}|$, nous pouvons trouver une fonction $f : \mathbb{C}^2 \to \mathbb{R}$ vérifiant :

$$f(\omega_{\nu^0},\omega_{\nu^\epsilon}) \simeq \frac{1}{|\omega_{\nu^0}^2-\omega_{\nu^\epsilon}^2|^2} \tag{B.24}$$

$$\frac{\partial}{\partial\nu^\epsilon}\left(\sqrt{\sum_{\nu^0=1}^{3N}f(\omega_{\nu^0},\omega_{\nu^\epsilon})|\xi_0^{i\alpha}|^4}\right)\simeq 0 \tag{B.25}$$

Finalement nous obtenons l'expression de ΔS en fonction du spectre de la matrice dynamique $\tilde{\mathfrak{D}}$ du système initial et du système déformé :

$$\Delta S(\boldsymbol{\epsilon}) = -k_B\left\{\sum_{i,\alpha=1}^{N,3}\left(\sum_{1\leq\nu^\epsilon\neq\nu^0\leq 3N}\frac{\omega_{\nu^\epsilon}^{-2}|\xi_\epsilon^{i\alpha}|^2-\omega_{\nu^0}^{-2}|\xi_0^{i\alpha}|^2}{\sqrt{\sum_{\nu^0=1}^{3N}f(\omega_{\nu^0},\omega_{\nu^\epsilon})|\xi_0^{i\alpha}|^4}}\right)\right\} \tag{B.26}$$

B.4 Preuve de la convergence de $\Delta S_0^{+\infty}$

B.4.1 Existence de la limite $\omega \to +\infty$

Définissons la fonction suivante :

$$g(\omega) = \ln(\omega)\Im\left\{\mathrm{Tr}\left(\frac{\Delta\boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon})}{\boldsymbol{\mathcal{G}}(\omega)}\right)\right\}. \tag{B.27}$$

Le comportement asymptotique de la fonction de *Green* est donné par (nous utilisons la normalisation des *modes normaux* donnée dans la Sec. 3.2.2) :

$$\begin{cases}\|\boldsymbol{\mathcal{G}}(\omega)\|_\infty \underset{\omega\to+\infty}{\sim} \omega^{-2}\\ \|\Delta\boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon})\|_\infty \underset{\omega\to+\infty}{\sim} 2\Delta\omega(\boldsymbol{\epsilon})\omega^{-3}\end{cases} \tag{B.28}$$

Par utilisation de l'opérateur module sur g, nous obtenons la majoration suivante :

$$|g(\omega)| \leq |\ln(\omega)|\left|\left\{\mathrm{Tr}\left(\frac{\Delta\boldsymbol{\mathcal{G}}(\omega,\boldsymbol{\epsilon})}{\boldsymbol{\mathcal{G}}(\omega)}\right)\right\}\right| \underset{\omega\to+\infty}{\sim} 2|\Delta\omega(\boldsymbol{\epsilon})||\ln(\omega)||\omega^{-1}| \tag{B.29}$$

Finalement sous réserve que $\Delta\omega(\boldsymbol{\epsilon})$ est borné, $\omega \to +\infty$ converge vers une limite finie :

$$\lim_{\omega\to+\infty}|g(\omega)| = 0. \tag{B.30}$$

B.4.2 Existence de la limite $\omega \to 0^+$

Il est possible de montrer que $\Delta\boldsymbol{\mathcal{G}}(\omega)$ prend l'expression suivante dans la limite $\omega \to 0^+$, ici sym$(\cdot)$ est la partie symétrique :

$$\lim_{\omega\to 0^+} \Delta\boldsymbol{\mathcal{G}}(\omega) = 2\,\mathrm{sym}\left(\sum_\nu \frac{\left(\Delta\hat{\mathbf{e}}_\nu - \frac{\Delta\omega_\nu}{\omega_\nu}\hat{\mathbf{e}}_\nu\right)\otimes\hat{\mathbf{e}}_\nu}{\omega_\nu^2}\right). \tag{B.31}$$

Définissons la fonction suivante afin de simplifier les expressions :

$$h\left(\frac{\Delta\omega_\nu}{\omega_\nu},\nu\right) = |\Delta\nu|^2 + \frac{(\Delta\omega_\nu)^2}{\omega_\nu^2}|\nu|^2 + 2\frac{\Delta\omega_\nu}{\omega_\nu}|\Delta\nu||\nu|. \tag{B.32}$$

Par analogie avec la limite $\omega \to +\infty$, passons au module pour obtenir une majoration de g dans la limite $\omega \to 0^+$:

$$|g(\omega)| \leq |\ln(\omega)|\frac{2h\left(\frac{\Delta\omega_\nu}{\omega_\nu},\nu\right)}{\min_\nu|\omega_\nu^2|}\max_\nu|\omega_\nu^2|. \tag{B.33}$$

Sous l'hypothèse que le coefficient de Grüneisen du potentiel γ_p est strictement positif quelque soit le mode de vibration du système, il existe $\eta_{1,\nu}, \eta_{2,\nu}$ and $\eta_{3,\nu}$ strictement positifs tels que :

$$\begin{cases} \Delta\omega_\nu \propto \omega^{\eta_1^\nu} \ /\eta_1^\nu > 0. \\ |\Delta\nu| \propto \omega^{\eta_2^\nu} \ /\eta_1^\nu > 0. \\ |\nu| \propto \omega^{\eta_3^\nu} \ /\eta_1^\nu > 0. \end{cases} \tag{B.34}$$

Nous pouvons alors ré-écrire la majoration valable dans la limite $\omega \to 0^+$:

$$\begin{aligned} |g(\omega)| \leq & |\ln(\omega)|\frac{2h\left(\frac{\Delta\omega_\nu}{\omega_\nu},\nu\right)}{\min_\nu|\omega_\nu^2|}\max_\nu|\omega_\nu^2| \\ & \underset{\omega\to 0^+}{\sim} 2|\ln(\omega)|\frac{\max_\nu|\omega_\nu^2|}{\min_\nu|\omega_\nu^2|}\min_{\eta_1^\nu,\eta_2^\nu,\eta_3^\nu}\left(\omega^{2\eta_2^\nu}, \frac{\omega^{2(\eta_2^\nu+\eta_3^\nu)}}{\min_\nu|\omega_\nu^2|}, \frac{2\omega^{\eta_1^\nu+\eta_2^\nu+\eta_3^\nu}}{\min_\nu|\omega_\nu|}\right) \end{aligned} \tag{B.35}$$

Nous pouvons alors déduire la limite du module de g pour $\omega \to 0^+$:

$$\lim_{\omega\to 0^+}|g(\omega)| = 0. \tag{B.36}$$

Finalement, les convergences pour les limites $\omega \to 0^+$ et $\omega \to +\infty$ permettent de déduire que $\Delta S_0^{+\infty}$ converge :

$$\Delta S_0^{+\infty} = 0. \tag{B.37}$$

B.5 Manipulations matricielles

Sous l'hypothèse d'existence d'une relation linéaire (donnée par l'équation (3.31)) entre la *densité d'état* de *modes normaux* et les descripteurs atomiques locaux, nous avons :

$$\sum_{\nu}\sum_{\alpha}|\xi^{i\alpha}(\nu)|^2 = \underline{w}\cdot\underline{D}^i \tag{B.38}$$

Cette relation linéaire peut être ré-écrite sous la forme suivante :

$$\|\underline{w}\cdot\underline{D}^i\|^2 = \mathrm{Tr}\left\{\mathrm{diag}(\underline{w})^2\cdot\underline{D}^i\cdot\left[\underline{D}^i\right]^T\right\} \tag{B.39}$$

Il existe alors un tenseur $\underline{\underline{W}}^{\frac{1}{2}}$ tel que $\underline{\underline{W}}^{\frac{1}{2}}\cdot\left[\underline{\underline{W}}^{\frac{1}{2}}\right]^T = \mathrm{diag}(\underline{w})^2$, ce qui aboutit à la formulation suivante :

$$\|\underline{w}\cdot\underline{D}^i\|^2 = \mathrm{Tr}\left\{\underline{\underline{W}}^{\frac{1}{2}}\cdot\underline{D}^i\cdot\left[\underline{D}^i\right]^T\cdot\left[\underline{\underline{W}}^{\frac{1}{2}}\right]^T\right\} \tag{B.40}$$

I won't crawl on my knees for you
I won't believe the lies that hide the truth
I won't sweat one more drop for you

—Re-Education (Through Labor), Rise Against

C

Importance des propriétés de régularité des potentiels *semi-empiriques* pour la quantification des effets vibrationnels

Sommaire

Dans cette annexe, nous mettons en évidence de façon quantitative l'une des problématiques évoquée dans le chapitre 3 concernant le lien entre la qualité du modèle linéaire de régression d'entropie vibrationnelle et la **régularité du potentiel** *semi-empirique* utilisé pour générer la base de données d'entraînement. Avant toute chose, nous commençons par donner une définition plus quantitative de ce que nous appelons la régularité d'un champ de force.

Définition C.1. *Soit un potentiel* $V : \mathbb{R}^{3N} \to \mathbb{R}$ *définit sur le support* $\mathcal{Q}$*. On dit que* V *est de classe* $\mathcal{C}^k$ *sur* $\mathcal{Q}$ *si* $\forall \boldsymbol{q} \in \mathcal{Q}$ *et* $\forall \alpha_1, \ldots, \alpha_j, \ldots \alpha_k \mid \alpha_j \in [\![1, N]\!]$ *la fonction* $\mathcal{D}^k \{V(\boldsymbol{q})\}$ *est continue.*

$$\mathcal{D}^k \{V(\boldsymbol{q})\} = \frac{\partial^k}{\partial q_{\alpha_1} \ldots \partial q_{\alpha_j} \ldots \partial q_{\alpha_k}} \{V(\boldsymbol{q})\} \tag{C.1}$$

Cette définition se comprend simplement dans le cas d'une fonction à une seule variable. Dans cette situation, la fonction est de classe $\mathcal{C}^k$ si toutes ses dérivées jusqu'à la $k^{-ième}$ dérivée sont continues. Pour notre étude, on dit que notre champ de force est régulier pour quantifier les contributions vibrationnelles si celui-ci est au moins de classe $\mathcal{C}^3$. Nous allons détailler ce choix dans la section suivante.

C.1 Contributions vibrationnelles et *matrice dynamique*

Nous rappelons le lien étroit entre l'entropie vibrationnelle harmonique et l'échantillonnage de la *matrice dynamique* du système. Nous rappelons les définitions données dans le chapitre 3. On introduit la notion de *matrice dynamique* d'un système autour d'un minimum d'énergie de coordonnée $\boldsymbol{q}_0$. La matrice $\mathfrak{D} \{\boldsymbol{q}_0\}$ est définie de la façon suivante :

$$\mathfrak{D} \{\boldsymbol{q}_0\} = \sum_{i,\alpha=1}^{N,3} \sum_{j,\beta=1}^{N,3} \frac{1}{\sqrt{m_i m_j}} \frac{\partial^2}{\partial q_{i\alpha} \partial q_{j\beta}} \{V(\boldsymbol{q}_0)\} \, \boldsymbol{q}'_{i\alpha} \otimes \boldsymbol{q}'_{j\beta} \tag{C.2}$$

La *matrice dynamique* du système fait intervenir la *matrice Héssienne* de l'énergie du système évaluée à la coordonnée $\boldsymbol{q}_0$, $\mathbb{H}_{ij} \{\boldsymbol{q}_0\} = \frac{\partial^2}{\partial q_i \partial q_j} \{V(\boldsymbol{q}_0)\}$. L'entropie vibrationnelle harmonique est alors directement reliée aux valeurs propres de la *matrice dynamique* du système (cf. chapitre 3). Afin d'obtenir une valeur fiable de l'entropie vibrationnelle du système à la coordonnée $\boldsymbol{q}_0$ il est nécessaire que la *matrice Héssienne* du système soit correctement définie à la coordonnée $\boldsymbol{q}_0$, ce qui implique que le potentiel V doit être au minimum de classe $\mathcal{C}^2$.

Dans le cadre de notre modèle linéaire de régression de l'entropie vibrationnelle harmonique dans l'espace des descripteurs, nous considérons qu'il est possible d'établir le lien entre l'espace des phases et l'espace des descripteurs en utilisant **seulement le minimum d'un bassin donné**. Traditionnellement, l'échantillonnage de la *matrice Héssienne* du système nécessite un grand nombre de perturbations de la coordonnée

$\boldsymbol{q}_0$ [14, 174]. Une hypothèse forte de notre modèle est donc que le minimum du bassin doit être suffisamment représentatif de la *matrice Héssienne* du système. Une condition nécessaire est d'imposer le potentiel V du système soit au moins de classe $\mathcal{C}^3$ à la coordonnée $\boldsymbol{q}_0$. Dans ce cas, la courbure de l'espace des phases décrite par la *matrice Héssienne* est continue et est lisse au sens où elle est dérivable et ses dérivées partielles sont continues. Cette hypothèse impose donc que les potentiels utilisés soient non-rugueux, comportement qui est souvent observé pour les potentiels EAM [25, 26] utilisant des tables numériques ou qui sont ajustés avec des fonctions *splines* cubiques. Nous allons illustrer ces problèmes de régularité en utilisant deux exemples concrets dans les sections suivantes.

C.2 Rappels des résultats du chapitre 3 : entropie vibrationnelle

Dans le chapitre 3, nous avons construit un modèle de régression de l'entropie vibrationnelle harmonique pour des défauts ponctuels dans le fer cubique centré. Pour cela, nous avons utilisé la base de données *ARTn* et deux potentiels *semi-empiriques* différents. Un potentiel EAM [25, 26] développé par Ackland *et al.* [180] et un potentiel MEAM [63] développé par Alizera *et al.* [186]. Nous avons constaté que le modèle linéaire est plus précis, au sens de l'erreur quadratique moyenne, pour le potentiel MEAM que pour le potentiel EAM. Pour opérer cette comparaison, nous avons utilisé le même descripteur, les Angular Fourrier Series [121], avec le même nombre de composantes radiales et angulaires ansi que le même rayon de coupure. Les résultats de cette comparaison sont donnés dans la figure C.1, avec à gauche les résultats donnés par le potentiel EAM [180] et à droite les résultats obtenus avec le potentiel MEAM [186]. Dans cette situation, on constate que l'erreur quadratique moyenne est presque deux fois plus grande pour le potentiel EAM que pour le potentiel MEAM.

Le potentiel AM04 développé par Ackland *et al.* [180] utilise des interpolations par *splines* cubiques. Dans ce cadre, la courbure au point $\boldsymbol{q}$ décrite par la *matrice Héssienne* $\mathbb{H}\{\boldsymbol{q}\}$ est une fonction continue par morceau. Dans le cas du potentiel MEAM développé par Alizera *et al.* [186] la courbure au point $\boldsymbol{q}$ décrite par la *matrice Héssienne* $\mathbb{H}\{\boldsymbol{q}\}$ est de classe $\mathcal{C}^\infty$ car les fonctions de bases de l'ajustement sont analytiques. Le comportement continu par morceau de la courbure du potentiel EAM rend bancale notre hypothèse de travail selon laquelle la courbure d'un bassin peut être décrite par son seul minimum. En effet, il est impossible - dans le cas du potentiel EAM [180] - d'assurer une corrélation entre les "courbures" $\mathbb{H}\{\boldsymbol{q_1}\}$ et $\mathbb{H}\{\boldsymbol{q_2}\}$ de deux points de coordonnées $\boldsymbol{q_1}$ et $\boldsymbol{q_2}$. Cette absence de corrélation peut s'exprimer quantitativement de la façon suivante :

$$\forall \alpha \in \mathbb{R}^+ \quad \exists \boldsymbol{q_1}, \boldsymbol{q_2} \in \mathcal{Q} : \quad \frac{\|\mathbb{H}\{\boldsymbol{q_2}\} - \mathbb{H}\{\boldsymbol{q_1}\}\|}{\|\boldsymbol{q_2} - \boldsymbol{q_1}\|} \geq \alpha \tag{C.3}$$

Ici $\|\cdot\|$ est l'opérateur de norme. Cette relation traduit le fait que même si la différence au sens de la norme $\|\boldsymbol{q_2} - \boldsymbol{q_1}\|$ est petite cela n'implique pas que la différence de norme $\|\mathbb{H}\{\boldsymbol{q_2}\} - \mathbb{H}\{\boldsymbol{q_1}\}\|$ est petite.

Dans le cadre de régression relative à la courbure d'un paysage énergétique, il est donc nécessaire de s'assurer que le potentiel utilisé est suffisamment lisse. Dans la section suivante, nous allons illustrer un cas pathologique de ce type de comportement, dit rugueux, de certains potentiels *semi-empiriques*.

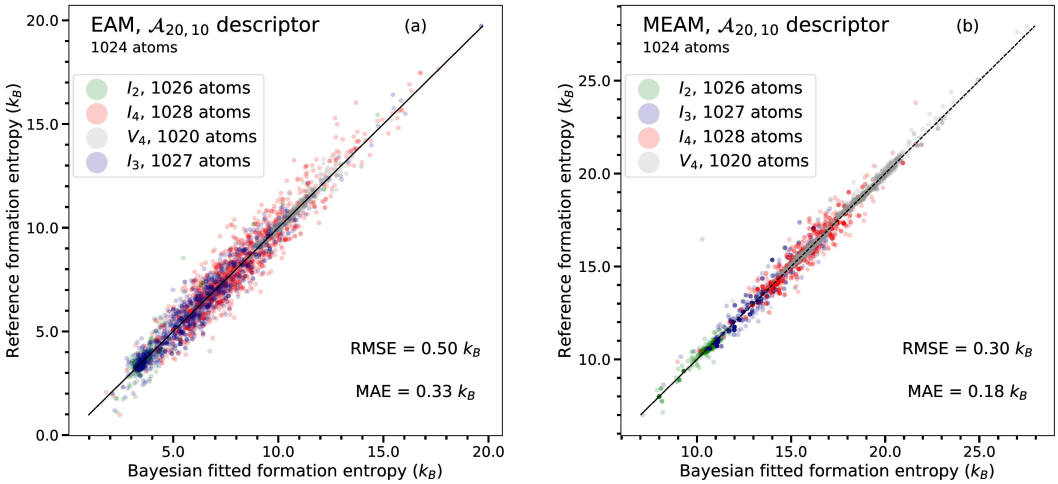

Figure C.1: Comparaison directe entre les résultats obtenus pour la base de données *ARTn* pour les deux potentiels différents testés dans le chapitre (3). Le descripteur utilisé est le même pour les deux potentiels $\mathcal{A}_{20,10}$ pour $r_{cut} = 5$Å. On constate une meilleure présision du modèle pour le potentiel MEAM [186] (à droite) que pour le potentiel EAM [180] (à gauche)

C.3 Système binaire *Cu-Zr* : mise en évidence de la nécessité de la régularité

Nous allons illustrer un exemple pathologique de régularité d'un potentiel *semi-empirique* et ses conséquences sur la régression des propriétés de courbures de ce champ de forces. Dans le chapitre 4, nous avons construit un modèle de régression dans l'espace des descripteurs des fréquences d'attaques pour la base de données de *Si amorphe*. Une autre base de données de fréquences d'attaques nous a été fournie par Normand Mousseau et son post-doctorant Simon Gelin. Cette base de données a aussi été construite par la méthode *ARTn* [175-179] et porte sur le verre métallique *Cu-Zr*. Le calcul numérique des fréquences d'attaques par l'algorithme d'*ARTn* a été effectué de la même façon que celui présenté dans le chapitre (4). Le potentiel EAM utilisé pour mener à bien cette étude a été developpé par Cheng *et al.* [259]. De même que pour la base de données de *Si amorphe*, nous avons voulu construire un modèle de régression des fréquences d'attaque de la base de données de *Cu-Zr*.

Pour le cas d'un système à un seul élément, le modèle de régression des fréquences d'attaque pour une collection d'événements cinétiques $\{\mathcal{E}^i\}$ décrit dans le chapitre 4 se traduit de la façon suivante :

$$\ln\left(\nu^*_{\mathcal{E},ms}\right) = \underline{w}_1 \cdot \left(\underline{D}_{\mathcal{E},m} \oplus \underline{D}_{\mathcal{E},s}\right) \tag{C.4}$$

Ici, $\underline{D}_{\mathcal{E},m/s} = \sum_{d\in\mathcal{E},m/s} \underline{D}^d \in \mathbb{R}^D$ est le vecteur total de descripteur de la configuration $\mathcal{E}, m$ ou $\mathcal{E}, s$. Dans le cas d'un système à plusieurs éléments, nous devons incorporer l'information chimique dans la construction des descripteurs. Pour cela, nous introduisons le descripteur suivant :

$$\underline{D}^*(\boldsymbol{w}) = \underline{D} \oplus \underline{D}_{\boldsymbol{w}} \tag{C.5}$$

$\underline{D}$ est le vecteur de descripteurs de la configuration et $\underline{D}_{\boldsymbol{w}}$ est le vecteur de descripteurs dit "enrichi" par le vecteur de poids $\boldsymbol{w}$, $\oplus$ est l'opérateur de concaténation. Cet enrichissement s'écrit de la façon suivante dans le cadre du descripteur bi-spectrum SO(4) décrit dans le chapitre (2). La densité atomique $\rho(\boldsymbol{q})$ se décompose de la façon suivante dans le cas du bi-specteurm SO(4) :

$$\rho_i(\boldsymbol{q}) = \sum_{k\in\mathcal{R}_i} w_k \delta(\boldsymbol{q} - \boldsymbol{q_k}) \tag{C.6}$$

$$= \sum_{k\in\mathcal{R}_i} \sum_{j=0}^{\infty} \sum_{m,m'=-j}^{j} \boldsymbol{c}_{i,j}^{m,m'} U_j^{m,m'} \tag{C.7}$$

Dans le cas simple comportant une seule espèce chimique $w_k = 1$ pour tous les atomes. Dans le cas du descripteur enrichi, on associe un vecteur $\boldsymbol{w} \in \mathbb{R}^{N_\chi}$ avec N_χ le nombre d'espèces chimiques de la configuration. La densité locale autour de l'atome i se traduit alors de la façon suivante :

$$\rho_i(\boldsymbol{q}) = \sum_{\chi=1}^{N_\chi} \sum_{k\in\mathcal{R}_i\cap\mathcal{S}_\chi} w_\chi \delta(\boldsymbol{q} - \boldsymbol{q_k}) \tag{C.8}$$

Ici $\mathcal{S}_\chi$ est l'ensemble des atomes appartenant à l'espèce chimique χ. Le calcul des descripteurs s'effectue ensuite de la même manière que pour le descripteur $\underline{D}$. Cette méthode de concaténation entre la description géométrique ($\underline{D}$) du système et sa description chimique $\underline{D}_{\boldsymbol{w}}$ permet de construire des modèles de régressions riches et dont le nombre de composantes reste relativement faible. Pour un descripteur de dimension $\mathcal{D}$, le nombre de composantes évolue comme $\mathcal{O}(\mathcal{D})$ alors que l'approche traditionnelle par "paires" d'intéractions évolue comme $\mathcal{O}(N_\chi^2\mathcal{D})$. Ce type d'approche se montre d'ailleurs aussi efficace qu'une approche par "paires" car la concaténation de l'information chimique et géométrique permet de décrire la majorité de l'information du système [260]. Finalement, le modèle de régression pour un système à plusieurs éléments se formule de la façon suivante :

$$\ln\left(\nu^*_{\mathcal{E},ms}\right) = \underline{w}_3 \cdot \left(\underline{D}^*_{\mathcal{E},m}(\boldsymbol{w}) \oplus \underline{D}^*_{\mathcal{E},s}(\boldsymbol{w})\right) \tag{C.9}$$

Afin de procéder à la régression des fréquences d'attaque dans le système *Cu-Zr*, nous avons choisi le descripteur bi-spectrum SO(4). Nous avons choisi de ne présenter qu'une partie des résultats obtenus pour ce système. Les résultats présentés sont obtenus en utilisant le bSO(4) non-diagonal avec $j_{max} = 5.0$ et pour $r_{cut} = 4.04$Å. Ce qui représente un vecteur de descripteur de dimension $\mathcal{D} = 364$. Pour ce cas d'école, nous avons délibérément augmenté la dimension du descripteur afin d'obtenir le pouvoir interpolant maximum du modèle. Des résultats avec des j_{max} plus faibles ne sont pas présentés ici mais font apparaître des problèmes d'ajustement des données. Afin de nous placer dans un cas typique de *sur-ajustement*, nous choisissons aussi une approche de type quadratique avec ce même descripteur ce qui nous amène au nombre (absurde) de $\mathcal{D} = 16744$. La base de données *Cu-Zr* comporte envrion 9000 configurations, ce qui implique que le modèle quadratique est largement en régime de *sur-ajustement*. Les résultats obtenus pour ces deux modèles sont présentés dans la figure C.2, à gauche les résultats du modèle linéaire et à droite ceux du modèle quadratique. On constate que les deux modèles sont incapables d'ajuster correctement les données avec notamment un écart net entre le grand axe de l'ellipse des données et la droite $y = x$. On note même que l'erreur quadratique moyenne est presque la même pour les deux modèles alors que le modèle quadratique présente plus de paramètres ajustables qu'il n'existe de configurations dans la base de données ! L'interrogation se porte maintenant sur la base de données car ce type de modèle a fait ses preuves pour la base de données de *Si amorphe*.

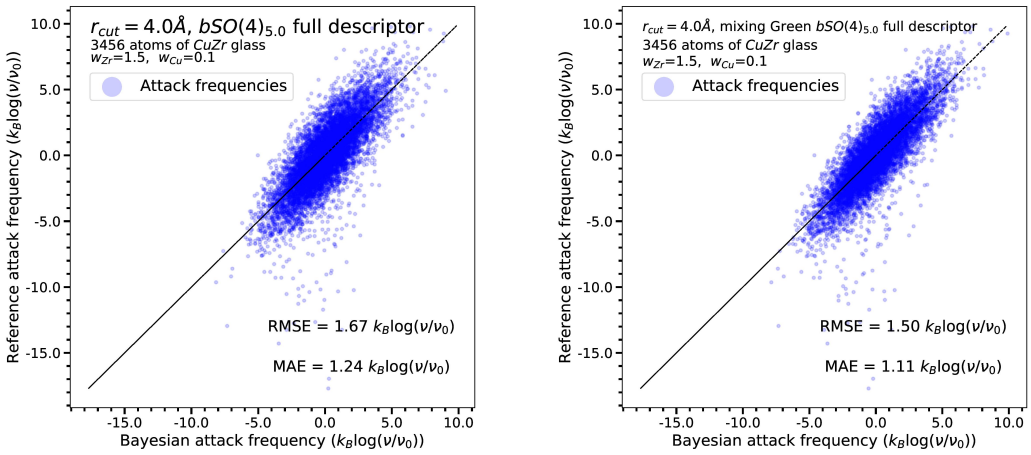

Figure C.2: Régressions des fréquences d'attaque dans le verre métallique *Cu-Zr*, à gauche le modèle linéaire et à droite le modèle quadratique. On constate que les deux modèles, malgrè leur grand nombre de composantes, sont incapables d'ajuster correctement les données.

La question de la régularité du potentiel EAM se pose aussi dans le cas de la base de données *Cu-Zr*. Afin de se donner une idée de la régularité du potentiel développé par

Cheng *et al.* [259], nous avons calculé par différences finies la valeur de la dérivée seconde du potentiel V par rapport à la distance r pour les paires *Cu-Cu* et *Zr-Zr*. Le résultat de cette procédure est présenté dans la figure C.3. Le résultat est édifiant : les dérivées secondes fluctent sur plusieurs ordres de grandeurs en fonction de la distance r entre paires. On constate que celles-ci vérifient clairement le critère de "non-corrélation" fourni par l'équation (C.3). Cette absence de corrélation entre l'environnement atomique et la valeur de la fréquence d'attaque calculée par la méthode *ARTn* permet de comprendre l'impossibilité d'ajustement du modèle de régression des fréquences d'attaque dans l'espace des descripteurs. **Le paysage énergétique de ce potentiel est tellement rugueux, qu'une erreur faible sur la coordonnée q du système implique une grande erreur sur la valeur du *Hessien*** $\mathbb{H}\{q_1\}$ conformement au critère de l'équation (C.3). Dans le cadre de ce type de potentiels peu-lisses, une méthode dite à un **seul point**, basée sur les descripteurs atomiques locaux, est incapable de fournir un modèle de régression de qualité suffisante pour permettre une extrapolation.

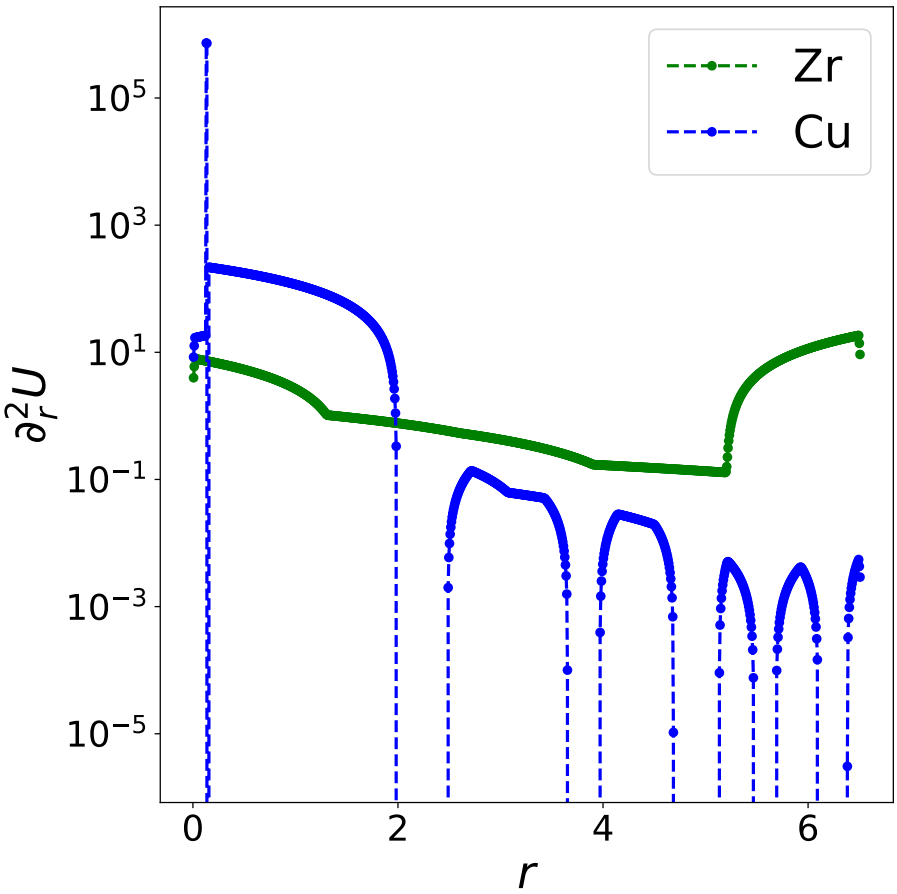

Figure C.3: Illustration des dérivées secondes pour les paires *Cu-Cu* et *Zr-Zr* en fonction de la distance entre paires en Å. On constate que la valeur des dérivées évolue rapidement sur plusieurs ordres de grandeurs.

C.4 Potentiels *semi-empiriques* et températures finies ?

À la lueur des constatations présentées dans les sections précédentes, il est légitime de se poser un certain nombre de questions sur la validité des potentiels *semi-empiriques* pour la quantification d'effets vibrationnels. Les sections précédentes montrent que le modèle de régression basé sur les descripteurs atomiques locaux aux points de coordonnées du minimum ou du point col du système ne fonctionne que si le potentiel *semi-empirique* utilisé pour générer la base de données est suffisamment lisse. Néanmoins, cette constatation met en lumière un autre point de réserve sur les potentiels *semi-empiriques* rugueux. En effet, les méthodes de températures finies nécéssitent d'évaluer la contribution entropique d'un système (qu'elle soit harmonique ou anharmonique). Les expériences numériques présentées mettent en évidence que les propriétés de courbure des potentiels rugueux sont peu fiables et présentent une forte erreur intrinsèque. Il est donc raisonnable de questionner l'utilisation des potentiels *semi-empiriques* pour quantifier les effets de températures finies. Cette question est d'autant plus légitime que ce type d'approche au faible coût numérique a pour but de rendre possible cette quantification nécessitant un grand nombre d'évaluations de forces. **Une étude plus systématique sur les propriétés de régularités des potentiels devrait être effectuée avant leur utilisation**. Une autre alternative est de développer des potentiels qui sont intrinséquement plus lisses et qui ne présentent donc pas ce type de rugosité. C'est la cas des potentiels de type *Machine Learning*.

When the dreamers dies
So dies the dream
Turn me inside out
Soak me in bleach.

— Soak Me in Bleach, The Amity Affliction

D

Construction des bases de données du Chap. 6 : convergence, constitution

Sommaire

D.1 Vérification de la convergence des grandeurs thermodynamiques dans l'espace réciproque

Dans la section suivante, nous nous intéressons à vérifier la convergence de grandeurs thermodynamiques d'interêt lors des calculs *ab initio*. En effet, la surface de Fermi du tungstène présente une structure très complexe qui nécessite un échantillonnage dense dans l'espace réciproque. Nous présentons l'influence de deux paramètres : (i) la **densité de la grille de points-k**, dans la majorité de la bibliographie incluant des calculs dans des super-cellules de $4a_0 \times 4a_0 \times 4a_0$, la grille choisie est une $4 \times 4 \times 4$ ou $5 \times 5 \times 5$; (ii) la valeur du **paramètre de smearing** qui est peu discuté dans les études précédentes. Nos calculs ont été effectués dans des super-cellules de volume $4a_0 \times 4a_0 \times 4a_0$ et dont le paramètre maille a_0 a été calculé par la méthode *courbes énergie/volume*. Le pseudo-potentiel utilisé est de type PAW [246] incluant les électrons de semi-coeur, l'énergie de cut-off est fixée à 500 eV et le critère de convergence à 1×10^{-6} eV pour la convergence électronique et 1×10^{-2} eV pour la convergence ionique. Les grilles de points-k sont générées par la méthode de Monkhorst-Pack [247]. Nous utilisons la méthode de smearing de Methfessel et Paxton [248] pour approximer la statistique de Fermi-Dirac. Les calculs ont été réalisés à l'aide du package `VASP` [245].

D.1.1 Convergence de l'énergie de formation et de liaison : espace réciproque et paramètre de smearing

Nous nous intéressons à la convergence de deux grandeurs d'importances (pour les résultats du Chap. 6) en fonction de la grille de points-k utilisée et de la valeur du paramètre de smearing. Les grandeurs étudiés sont : (i) **les énergies de formation de la mono-lacune et des di-lacunes premier et deuxième voisins** ; (ii) **les énergies de liaison entre la mono-lacune et les di-lacunes premier et deuxième voisins**. Nous nous baserons sur la convergence du tungstène pour la paramétrisation DFT des autres métaux cubiques centrés. Les résultats de la procédure de convergence pour la fonctionnelle AM [45, 46] pour les valeurs de paramètres smearing 0.1 et 0.3 sont présentés respectivement dans les figures D.1 et D.2. On constate que les grandeurs convergent difficilement en utilisant le paramètre de smearing égal à 0.1. Dans le cas du paramètre de smearing 0.3, les grandeurs convergent avec l'utilisation d'une grille de point-k de type Monkhorst-Pack [247] $6 \times 6 \times 6$.

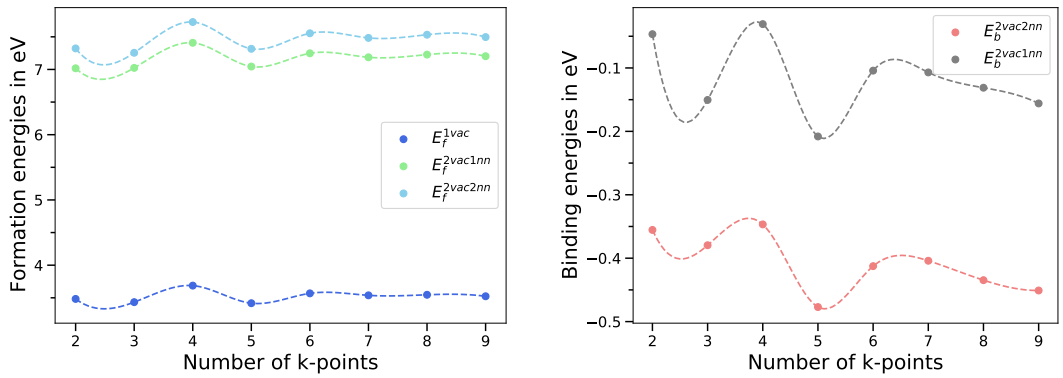

Figure D.1: Convergence des énergies de formation et de liaison pour la mono-lacune et les di-lacunes premier et deuxième voisins pour un paramètre de smearing de 0.1 pour la fonctionnelle AM.

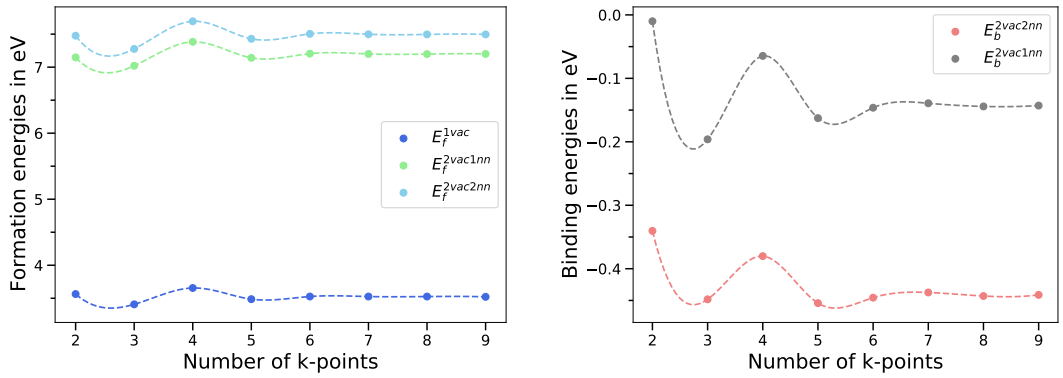

Figure D.2: Convergence des énergies de formation et de liaison pour la mono-lacune et les di-lacunes premier et deuxième voisins pour un paramètre de smearing de 0.3 pour la fonctionnelle AM.

Afin de s'assurer que cette paramétrisation (smearing à 0.3 et grille de points-k $6 \times 6 \times 6$) sera valide pour d'autres types de fonctionnelles, nous avons effectué le même test de convergence pour la fonctionnelle LDA. Les résultats de la procédure sont présentés dans la figure D.3. De même que pour la fonctionnelle AM, les énergies de formation et les énergies de liaison sont convergées à partir d'une grille de points-k $6 \times 6 \times 6$ et pour un paramètre de smearing égal à 0.3. **Nous allons donc considérer cette paramétrisation DFT pour toutes les fonctionnelles du tungstène**.

D.2 Constitution des bases de données DFT pour les métaux cubiques centrés

Nous décrivons la composition des bases de données DFT utilisées dans le chapitre (6) pour ajuster nos potentiels de type Machine Learning. Pour chaque fonc-

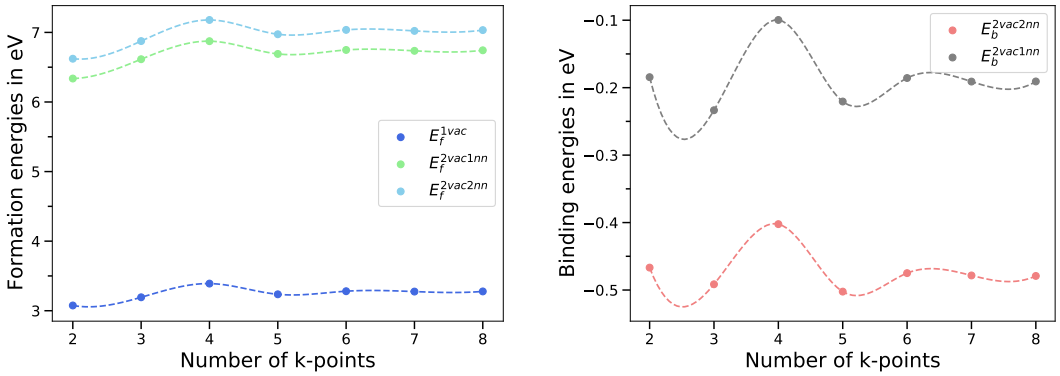

Figure D.3: Convergence des énergies de formation et de liaison pour la mono-lacune et les di-lacunes premier et deuxième voisin pour un paramètre de smearing de 0.3 pour la fonctionnelle LDA.

tionnelle - et élément - nous avons déterminé le paramètre de maille d'équilibre DFT avec la paramétrisation décrite dans chaque section suivante. Le calcul du paramètre de maille d'équilibre a été réalisé en utilisant la méthode de courbes *énergie/volume* décrite dans le chapitre (5). Nous décrivons ensuite la structure de toutes les bases de données DFT que nous avons générées.

D.2.1 Base de données pour le tungstène (W)

Conditions de simulation :

— Super-cellules de volume $(4a_0)^3$, avec une grille de points-k $6 \times 6 \times 6$.

— Pseudo-potentiel : semi-coeur incluant 14 électrons.

— Critère de convergence : auto-cohérence à 1×10^{-8} eV et smearing=0.3

— Super-cellules de volume $(a_0)^3$ pour les déformations.

Table D.1: W : paramètres de maille et constantes élastiques en fonction de la fonctionnelle d'échange-correlation. Critère de convergence : 1×10^{-7} eV, grille de points-k MP translatée $6 \times 6 \times 6$ pour des super-cellules $(4a_0)^3$

W (14 électrons)	Échange-correlation		
	PBE	AM05	LDA
a_0	3.1859	3.1507	3.1417
C_{11}	515.4	559.6	563.1
C_{12}	199.3	212.0	222.1
C_{44}	139.9	115.4	146.5

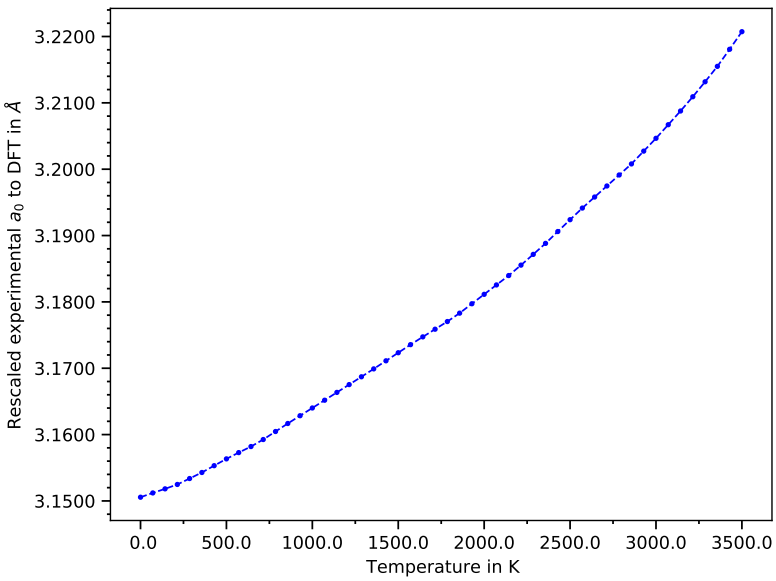

Figure D.4: Extension thermique du paramètre de maille pour le W basée sur les données expérimentales renormalisées pour être en accord avec les données *ab initio*. La valeur expérimentale permettant la renormalisation est $a_0 = 3.1648$Å à 293 K [249].

Table D.2: Liste des configurations pour la base de données minimale pour le W

Système W (N_d,ϵ)	Température en 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 lacune, $\epsilon = 0\%$	10	10	10	10	10	50
2 lacunes, 1^{st} NN	10	10	10	10	10	50
2 lacunes, 2^{nd} NN	10	10	10	10	10	50
2 lacunes, 3^{rd} NN	10	10	10	10	10	50
NEB 1 lacunes, 1^{st} NN	7	0	0	0	0	7
NEB 2 lacunes, 1^{st} NN	7	0	0	0	0	7
NEB 3 lacunes, 1^{st} NN	7	0	0	0	0	7
Total	1062	50	50	50	50	1262

D.2.2 Base de données pour le molybdène (Mo)

Conditions de simulation :

— Super-cellules de volume $(4a_0)^3$, avec une grille de points-k $6 \times 6 \times 6$.

— Pseudo-potentiel : semi-coeur incluant 14 électrons.

— Critère de convergence : auto-cohérence à 1×10^{-8} eV et smearing=0.3

— Super-cellules de volume $(a_0)^3$ pour les déformations.

Table D.3: Mo : paramètres de maille et constantes élastiques en fonction de la fonctionnelle d'échange-correlation. Critère de convergence : 1×10^{-7} eV, grille de points-k MP translatée $6 \times 6 \times 6$ pour des super-cellules $(4a_0)^3$

Mo (14 électrons)	Échange-correlation	
	PBE	AM05
a_0	3.16324	3.12709
C_{11}	461.2	499.9
C_{12}	158.8	172.1
C_{44}	99.2	107.8

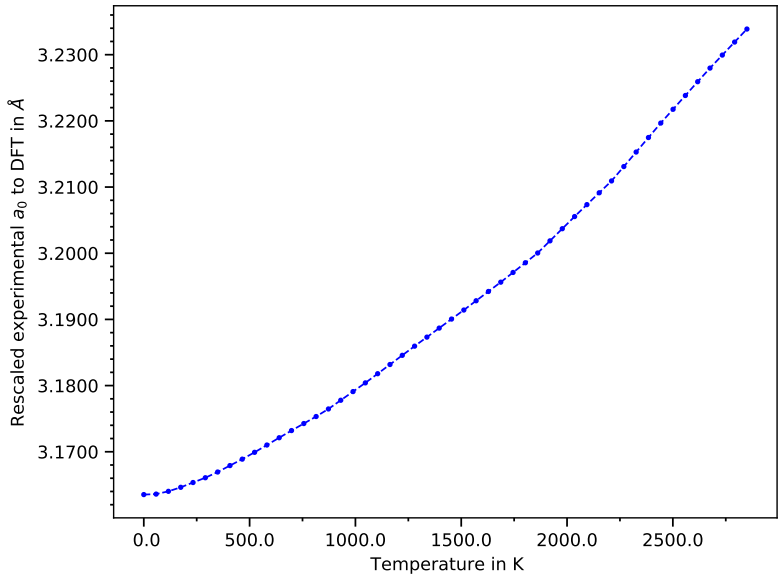

Figure D.5: Extension thermique du paramètre de maille pour le Mo basée sur les données expérimentales renormalisées pour être en accord avec les données *ab initio*. La valeur expérimentale permettant la renormalisation est $a_0 = 3.14700$Å à 298 K [249].

Table D.4: Liste des configurations pour la base de données minimale pour le Mo

Système Mo (N_d,ϵ)	Température en K					Total
	0	712.5	1425	2135.5	2850	
bulk, $\epsilon = +0\%$	1	10	10	10	10	41
bulk, $-5\% \leq \epsilon \leq 5\%$	1000	0	0	0	0	1000
1 lacune, $\epsilon = 0\%$	10	10	10	10	10	50
2 lacunes, 1^{st} NN	10	10	10	10	10	50
2 lacunes, 2^{nd} NN	10	10	10	10	10	50
2 lacunes, 3^{rd} NN	10	10	10	10	10	50
NEB 1 lacunes, 1^{st} NN	7	0	0	0	0	7
NEB 2 lacunes, 1^{st} NN	7	0	0	0	0	7
NEB 3 lacunes, 1^{st} NN	7	0	0	0	0	7
Total	1062	50	50	50	50	1262

I tried so hard and got so far
But in the end, it doesn't even matter.

— In The End, Linkin Park

Bibliographie

[1] C.-C. Fu, J. Dalla Torre, F. Willaime, J.-L. Bocquet et A. Barbu. « Multiscale modelling of defect kinetics in irradiated iron ». In : *Nat. Mater.* 4.1 (2005), p. 68-74.

[2] C. Domain et G. Monnet. « Simulation of screw dislocation motion in iron by molecular dynamics simulations ». In : *Phys. Rev. Lett.* 95 (21 2005), p. 215506.

[3] L. Malerba, C.S. Becquart, H. Hou et C. Domain. « Comparison of algorithms for multiscale modelling of radiation damage in Fe–Cu alloys ». In : *Phil. Mag.* 85.4-7 (2005), p. 417-428.

[4] R. Alexander. « Energy landscape of defects in body-centered cubic metals. » Theses. Université Paris Saclay (COmUE), nov. 2016.

[5] L. Malerba, M.J. Caturla, Gaganidze E., C. Kaden, M.J. Konstantinović, P. Olsson, C. Robertson, D. Rodney, A.M. Ruiz-Moreno, M. Serrano, J. Aktaa, N. Anento, S. Austin, A. Bakaev, J.P. Balbuena, F. Bergner, F. Boioli, M. Boleininger, G. Bonny, N. Castin, J.B.J. Chapman, P. Chekhonin, M. Clozel, B. Devincre, L. Dupuy, G. Diego, S.L. Dudarev, C.-C. Fu, R. Gatti, L. Gélébart, B. Gómez-Ferrer, D. Gonçalves, C. Guerrero, P.M. Gueye, P. Hähner, S.P. Hannula, Q. Hayat, M. Hernández-Mayoral, J. Jagielski, N. Jennett, F. Jiménez, G. Kapoor, A. Kraych, T. Khvan, L. Kurpaska, A. Kuronen, N. Kvashin, O. Libera, P.-W. Ma, T. Manninen, M.-C. Marinica, S. Merino, E. Meslin, F. Mompiou, F. Mota, H. Namburi, C.J. Ortiz, C. Pareige, M. Prester, R.R. Rajakrishnan, M. Sauzay, A. Serra, I. Simonovski, F. Soisson, P. Spätig, D. Tanguy, D. Terentyev, M. Trebala, M. Trochet, A. Ulbricht, M.Vallet, K. Vogel, T. Yalcinkaya et J. Zhao. « Multiscale modelling for fusion and fission materials : The M4F project ». In : *Nucl. Mater. Ener.* 29 (2021), p. 101051.

[6] M.-C. Marinica, F. Willaime et N. Mousseau. « Energy landscape of small clusters of self-interstitial dumbbells in iron. » In : *Phys. Rev. B.* 83 (2011), p. 094119.

[7] M.-C. Marinica, F. Willaime et J.-P. Crocombette. « Irradiation-induced formation of nanocrystallites with C15 laves phase structure in bcc iron ». In : *Phys. Rev. Lett.* 108.2 (2012), p. 025501.

[8] T. Jourdan, G. Bencteux et G. Adjanor. « Efficient simulation of kinetics of radiation induced defects : A cluster dynamics approach ». In : *J. Nucl. Mater.* 444.1 (2014), p. 298-313.

[9] R. Alexander, M.-C. Marinica, L. Proville, F. Willaime, K. Arakawa, M.R. Gilbert et S.L. Dudarev. « Ab initio scaling laws for the formation energy of nanosized interstitial defect clusters in iron, tungsten, and vanadium ». In : *Phys. Rev. B* 94.2 (2016), p. 024103.

[10] C. Domain et C.S. Becquart. *Object Kinetic Monte Carlo (OKMC) : A Coarse-Grained Approach to Radiation Damage*. Springer International Publishing, 2019.

[11] A. Chartier et M.-C. Marinica. « Rearrangement of interstitial defects in alpha-Fe under extreme condition ». In : *Acta Mater.* 180 (2019), p. 141 -148.

[12] E. Clouet, L. Ventelon et F. Willaime. « Dislocation core energies and core fields from first principles ». In : *Phys. Rev. Lett.* 102.5 (2009), p. 55502.

[13] R.G.A. Veiga, M. Perez, C.S. Becquart, E. Clouet et C. Domain. « Comparison of atomistic and elasticity approaches for carbon diffusion near line defects in α-iron ». In : *Acta Mater.* 59.18 (2011), p. 6963-6974.

[14] L. Proville, D. Rodney et M.-C. Marinica. « Quantum effect on thermally activated glide of dislocations ». In : *Nat. Mater.* 11 (2012), 845–849.

[15] C. Varvenne, F. Bruneval, M.-C. Marinica et E. Clouet. « Point defect modeling in materials : Coupling ab initio and elasticity approaches ». In : *Phys. Rev. B* 88 (13 2013), p. 134102.

[16] A. De Backer, D.R. Mason, C. Domain, D. Nguyen-Manh, Marinica M.-C., L. Ventelon, C.S. Becquart et Dudarev S.L. « Hydrogen accumulation around dislocation loops and edge dislocations : from atomistic to mesoscopic scales in BCC tungsten ». In : *Phys. Script.* T170 (2017), p. 014073.

[17] D. Carpentier, T. Jourdan, P. Terrier, M. Athènes et Y. Le Bouar. « Effect of sink strength dispersion on cluster size distributions simulated by cluster dynamics ». In : *J. Nucl. Mater.* 533 (2020), p. 152068.

[18] L. Messina, T. Schuler, M. Nastar, M.-C. Marinica et P. Olsson. « Solute diffusion by self-interstitial defects and radiation-induced segregation in ferritic Fe–X (X=Cr, Cu, Mn, Ni, P, Si) dilute alloys ». In : *Acta Mater.* 191 (2020), p. 166-185.

[19] L. Huang, M. Nastar, T. Schuler et L. Messina. « Multiscale modeling of the effects of temperature, radiation flux, and sink strength on point-defect and solute redistribution in dilute Fe-based alloys ». In : *Phys. Rev. Materials* 5 (3 2021), p. 033605.

[20] A.G. Goryaeva, J.B. Maillet et M.-C. Marinica. « Towards better efficiency of interatomic linear machine learning potentials ». In : *Comput. Mater. Sci.* 166 (2019), p. 200 -209.

[21] F. Bruneval, I. Maliyov, C. Lapointe et M.-C. Marinica. « Extrapolating unconverged GW energies up to the complete basis set limit with linear regression ». In : *J. Chem. Theory Comput.* 16.7 (2020), p. 4399-4407.

[22] C. Lapointe, T.D. Swinburne, L. Thiry, S. Mallat, L. Proville, C.S. Becquart et M.-C. Marinica. « Machine learning surrogate models for prediction of point defect vibrational entropy ». In : *Phys. Rev. Materials* 4 (6 2020), p. 063802.

[23] A.M. Goryaeva, J. Dérès, C. Lapointe, P. Grigorev, T.D. Swinburne, J.R. Kermode, L. Ventelon, J. Baima et M.-C. Marinica. « Efficient and transferable machine learning potentials for the simulation of crystal defects in bcc Fe and W ». In : *Phys. Rev. Materials* 5 (10 2021), p. 103803.

[24] P.H. Dederichs, R. Zeller et K. Schroeder. *Point defects in metals II, Dynamical properties and diffusion controlled reactions*. Springer Tracts in Modern Physics, Berlin, 1980.

[25] M.S. Daw et M.I. Baskes. « Semiempirical, quantum mechanical calculation of hydrogen embrittlement in metals ». In : *Phys. Rev. Lett.* 50 (17 1983), p. 1285-1288.

[26] M.S. Daw et M.I. Baskes. « Embedded-atom method : Derivation and application to impurities, surfaces, and other defects in metals ». In : *Phys. Rev. B* 29 (12 1984), p. 6443-6453.

[27] R.E. Peierls. « The size of a dislocation ». In : *Proc. R. Soc. London* 52 (1940), p. 34.

[28] H. Curien. *Structure des Métaux, méthodes, principes et résultats cristallographiques, par C. S. Barrett, traduit par C. Leymonie, 1957*. 1958.

[29] V. Volterra. « Sur l'équilibre des corps élastiques multiplement connexes ». In : *Annales scientifiques de l'École Normale Supérieure* 3e série, 24 (1907), p. 401-517.

[30] J. Friedel. « General properties of dislocations ». In : *Dislocations*. International Series of Monographs on Solid State Physics. Pergamon, 1964, p. ii.

[31] O. Senninger, F. Soisson, E. Martínez, M. Nastar, C.-C. Fu et Y. Bréchet. « Modeling radiation induced segregation in iron–chromium alloys ». In : *Acta Mater.* 103 (2016), p. 1-11.

[32] M. Nastar et F. Soisson. *Comprehensive nuclear materials*. Oxford, 2012, p. 471-496.

[33] M.R. Gilbert, K. Arakawa, Z. Bergstrom, M.J. Caturla, S.L. Dudarev, F. Gao, A.M. Goryaeva, S.Y. Hu, X. Hu, R.J. Kurtz, A. Litnovsky, J. Marian, M.-C. Marinica, E. Martinez, E.A. Marquis, D.R. Mason, B.N. Nguyen, P. Olsson, Y. Osetskiy, D. Senor, W. Setyawan, M.P. Short, T. Suzudo, J.R. Trelewicz, T. Tsuru, G.S. Was, B.D. Wirth, L. Yang, Y. Zhang et S.J. Zinkle. « Perspectives on multiscale modelling and experiments to accelerate materials development for fusion ». In : *J. Nucl. Mater.* 554 (2021), p. 153113.

[34] P. Echenique et J.L. Alonso. « A mathematical and computational review of Hartree–Fock SCF methods in quantum chemistry ». In : *Mol. Phys.* 105.23-24 (2007), 3057–3098.

[35] P. Hohenberg et W. Kohn. « Inhomogeneous electron gas ». In : *Phys. Rev.* 136 (3B 1964), B864-B871.

[36] P.A.M. Dirac et R.H. Fowler. « Quantum mechanics of many-electron systems ». In : *Proc. R. Soc. London. Series A, Containing Papers of a Mathematical and Physical Character* 123.792 (1929), p. 714-733.

[37] J. C. Slater. « A simplification of the Hartree-Fock method ». In : *Phys. Rev.* 81 (3 1951), p. 385-390.

[38] J. P. Perdew et A. Zunger. « Self-interaction correction to density-functional approximations for many-electron systems ». In : *Phys. Rev. B* 23 (10 1981), p. 5048-5079.

[39] S. H. Vosko, L. Wilk et M. Nusair. « Accurate spin-dependent electron liquid correlation energies for local spin density calculations : a critical analysis ». In : *Can. J. Phys* 59 (1980), p. 1200.

[40] A.D. Becke. « Density-functional exchange-energy approximation with correct asymptotic behavior ». In : *Phys. Rev. A* 38 (6 1988), p. 3098-3100.

[41] J.P. Perdew, K. Burke et M. Ernzerhof. « Generalized gradient approximation made simple ». In : *Phys. Rev. Lett.* 77 (18 1996), p. 3865-3868.

[42] C. Adamo et V. Barone. « Exchange functionals with improved long-range behavior and adiabatic connection methods without adjustable parameters : The mPW and mPW1PW models ». In : *J. Chem. Phys.* 108.2 (1998), p. 664-675.

[43] J.P. Perdew et W. Yue. « Accurate and simple density functional for the electronic exchange energy : Generalized gradient approximation ». In : *Phys. Rev. B* 33 (12 1986), p. 8800-8802.

[44] R. Armiento et A.E. Mattsson. « Functional designed to include surface effects in self-consistent density functional theory ». In : *Phys. Rev. B* 72.8 (2005), p. 085108.

[45] A.E. Mattsson, R. Armiento, J. Paier, G. Kresse, J.M. Wills et T.R. Mattsson. « The AM05 density functional applied to solids ». In : *J. Chem. Phys.* 128.8 (2008), p. 084714.

[46] A.E. Mattsson et R. Armiento. « Implementing and testing the AM05 spin density functional ». In : *Phys. Rev. B* 79 (15 2009), p. 155101.

[47] J. Tao, J.P. Perdew, V.N. Staroverov et G.E. Scuseria. « Climbing the density functional ladder : Nonempirical meta–generalized gradient approximation designed for molecules and solids ». In : *Phys. Rev. Lett.* 91 (14 2003), p. 146401.

[48] G. Sun, J.B. Khurgin et A. Bratkovsky. « Coupled-mode theory of field enhancement in complex metal nanostructures ». In : *Phys. Rev. B* 84 (4 2011), p. 045415.

[49] B. Patra, S. Jana, L.A. Constantin et P. Samal. « Efficient band gap prediction of semiconductors and insulators from a semilocal exchange-correlation functional ». In : *Phys. Rev. B* 100 (4 2019), p. 045147.

[50] A.D. Becke. « Density-functional thermochemistry. III. The role of exact exchange ». In : *J. Chem. Phys.* 98.7 (1993), p. 5648-5652.

[51] R. Iftimie, P. Minary et M.E. Tuckerman. « Ab initio molecular dynamics : Concepts, recent developments, and future trends ». In : *Proc. Natl. Acad. Sci.* 102.19 (2005), p. 6654-6659.

[52] J. Neugebauer et T. Hickel. « Density functional theory in materials science ». In : *WIREs Comput. Mol. Sci.* 3.5 (2013), p. 438-448.

[53] F. Ducastelle et F. Cyrot-Lackmann. « Moments developments and their application to the electronic charge distribution of d bands ». In : *J. Phys. Chem. Solids* 31.6 (1970), p. 1295-1306.

[54] R.P. Gupta. « Lattice relaxation at a metal surface ». In : *Phys. Rev. B* 23 (12 1981), p. 6265-6270.

[55] J. E. Jones et S. Chapman. « On the determination of molecular fields. From the variation of the viscosity of a gas with temperature ». In : *Proc. R. Soc. London. Series A, Containing Papers of a Mathematical and Physical Character* 106.738 (1924), p. 441-462.

[56] D. J. Wales. *Energy landscapes*. Sous la dir. de Cambridge University Press. 2003.

[57] D.J. Wales. *Energy landscapes : Applications to clusters, biomolecules and glasses*. Cambridge Molecular Science. Cambridge University Press, 2004.

[58] D.J. Wales. « Some further applications of discrete path sampling to cluster isomerization ». In : *Mol. Phys.* 102.9-10 (2004), p. 891-908.

[59] D.J. Wales. *Cambridge cluster database*. URL : http://www-wales.ch.cam.ac.uk/CCD.html.

[60] J. Friedel. « Electronic structure of primary solid solutions in metals ». In : *Adv. Phys.* 3.12 (1954), p. 446-507.

[61] M. W. Finnis et J. E. Sinclair. « A simple empirical N-body potential for transition metals ». In : *Phil. Mag., Part A* 50.1 (1984), p. 45-55.

[62] A. P. Sutton et J. Chen. In : *Phil. Mag. Lett.* 61 (1984), p. 139.

[63] M. I. Baskes. « Modified embedded-atom potentials for cubic materials and impurities ». In : *Phys. Rev. B* 46.5 (1992), p. 2727.

[64] L. Dezerald, L. Ventelon, E. Clouet, C. Denoual, D. Rodney et F. Willaime. « Ab initio modeling of the two-dimensional energy landscape of screw dislocations in bcc transition metals ». In : *Phys. Rev. B* 89.2 (2014), p. 24104.

[65] G.D. Birkhoff. « Proof of the ergodic theorem ». In : *Proc. Natl. Acad. Sci.* 17.12 (1931), p. 656-660.

[66] T. Lelièvre, G. Stoltz et M. Rousset. *Free energy computations : A mathematical perspective*. Imperial College Press, London, 2010.

[67] P. Langevin. « Sur la theorie du mouvement brownien (On the theory of Brownian motion) ». In : *C. R. Acad. Hebd. Séances Acad. Sci.* 146 (1908), p. 530-533.

[68] N. Metropolis et S. Ulam. « The Monte Carlo method ». In : *J. Am. Stat. Assoc.* 44.247 (1949), p. 335-341.

[69] W.K. Hastings. « Monte Carlo sampling methods using markov chains and their applications ». In : *Biometrika* 57.1 (1970), p. 97-109.

[70] C. Sutton et S.V. Levchenko. « First-principles atomistic thermodynamics and configurational entropy ». In : *Front. Chem.* 8 (2020), p. 757.

[71] C.S. Becquart, N. Mousseau et C. Domain. *Atomistic Kinetic Monte Carlo and solute effects*. Sous la dir. de Wanda Andreoni et Sidney Yip. Cham : Springer International Publishing, 2019, p. 1-20.

[72] F. El-Mellouhi, N. Mousseau et L.J. Lewis. « Kinetic activation-relaxation technique : An off-lattice self-learning kinetic Monte Carlo algorithm ». In : *Phys. Rev. B* 78.15 (2008), p. 153202.

[73] N. Mousseau, L. Béland, P. Brommer, F. El-Mellouhi, J.F. Joly, G.K. N'Tsouaglo, O. Restrepo et M. Trochet. « Following atomistic kinetics on experimental timescales with the kinetic Activation–Relaxation Technique ». In : *Comput. Mater. Sci.* 100 (2015). Special Issue on Advanced Simulation Methods, p. 111 -123.

[74] T. Jourdan, F. Soisson, E. Clouet et A. Barbu. « Influence of cluster mobility on Cu precipitation in alpha-Fe : A cluster dynamics modeling ». In : *Acta Mater.* 58.9 (2010), p. 3400-3405.

[75] B. Gámez, L. Gámez, C.J. Ortiz, M.J. Caturla et J.M. Perlado. « Object Kinetic Monte Carlo calculations of electron and He irradiation of nickel ». In : *J. Nucl. Mater.* 386-388 (2009). Fusion Reactor Materials, p. 90-92.

[76] N. Castin, G. Bonny, A. Bakaev, C.J. Ortiz, A.E. Sand et D. Terentyev. « Object kinetic Monte Carlo model for neutron and ion irradiation in tungsten : Impact of transmutation and carbon impurities ». In : *J. Nucl. Mater.* 500 (2018), p. 15-25.

[77] V. Jansson, L. Malerba, A. De Backer, C.S. Becquart et C. Domain. « Sink strength calculations of dislocations and loops using OKMC ». In : *J. Nucl. Mater.* 442.1 (2013), p. 218-226.

[78] M. Chiapetto, C. S. Becquart, C. Domain et L. Malerba. « Kinetic Monte Carlo simulation of nanostructural evolution under post-irradiation annealing in dilute FeMnNi ». In : *Phys. Status Solidi C* 12.1-2 (2015), p. 20-24.

[79] R. Herschberg, C.-C. Fu, M. Nastar et F. Soisson. « Atomistic modelling of the diffusion of C in FeCr alloys ». In : *Acta Mater.* 165 (2019), p. 638-653.

[80] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles et G. J. Tucker. « Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials ». In : *J. Comp. Phys.* 285 (2015), p. 316.

[81] M.A. Wood et A.P. Thompson. « Quantum-accurate molecular dynamics potential for tungsten ». In : *arXiv :1702.07042v1 [physics.comp-ph]* (2017).

[82] M.A. Wood et A.P. Thompson. « Extending the accuracy of the SNAP interatomic potential form ». In : *J. Chem. Phys.* 148.24 (2018).

[83] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu et S. P. Ong. « Accurate force field for molybdenum by machine learning large materials data ». In : *Phys. Rev. Materials* 1 (4 2017), p. 043603.

[84] J. Behler et M. Parrinello. « Generalized neural-network representation of high-dimensional potential-energy surfaces ». In : *Phys. Rev. Lett.* 98 (14 2007), p. 146401.

[85] J. Behler. « Atom-centered symmetry functions for constructing high-dimensional neural network potentials ». In : *J. Chem. Phys.* 134.7 (2011), p. 074106.

[86] N. Artrith et A. Urban. « An implementation of artificial neural-network potentials for atomistic materials simulations : Performance for TiO2 ». In : *Comput. Mater. Sci.* 114 (2016), p. 135-150.

[87] J. Behler. « Perspective : Machine learning potentials for atomistic simulations ». In : *J. Chem. Phys.* 145.17 (2016), p. 170901.

[88] E. D. Cubuk, S. S. Schoenholz, J. M. Rieser, B. D. Malone, J. Rottler, D. J. Durian, E. Kaxiras et A. J. Liu. « Identifying structural flow defects in disordered solids using machine learning methods ». In : *Phys. Rev. Lett.* 114.10 (2015), p. 108001.

[89] A. P. Bartók et G. Csányi. « Gaussian approximation potentials : A brief tutorial introduction ». In : *Int. J. Quantum Chem.* 115.16 (2015), p. 1051-1057.

[90] T. Hofmann, B. Schölkopf et A.J. Smola. « Kernel methods in machine learning ». EN. In : *Ann. Statist.* 36.3 (2008), p. 1171-1220.

[91] V. Botu, R. Batra, J. Chapman et R. Ramprasad. « Machine learning force fields : Construction, validation, and outlook ». In : *J. Phys. Chem. C* 121.1 (2017), p. 511-522.

[92] V. Botu et R. Ramprasad. « Adaptive machine learning framework to accelerate ab initio molecular dynamics ». In : *Int. J. Quantum Chem.* 115.16 (2015), p. 1074-1083.

[93] V. Botu et R. Ramprasad. « Learning scheme to predict atomic forces and accelerate materials simulations ». In : *Phys. Rev. B* 92.9 (2015), p. 094306.

[94] Z. Li, J.R. Kermode et A. De Vita. « Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces ». In : *Phys. Rev. Lett.* 114.9 (2015), p. 096405.

[95] A.P. Bartók, S. De, C. Poelking, N. Bernstein, J.R. Kermode, G. Csányi et M. Ceriotti. « Machine learning unifies the modeling of materials and molecules ». In : *Sci. Adv.* 3.12 (2017), e1701816.

[96] L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl et M. Scheffler. « Big data of materials science : Critical role of the descriptor ». In : *Phys. Rev. Lett.* 114.10 (2015), p. 105503.

[97] A.P. Bartók. « Gaussian Approximation Potential : an interatomic potential derived from first principles Quantum Mechanics ». Thèse de doct. University of Cambridge, Cambridge, 2009.

[98] A.P. Bartók, M.C. Payne, R. Kondor et G. Csányi. « Gaussian approximation potentials : The accuracy of quantum mechanics, without the electrons ». In : *Phys. Rev. Lett.* 104 (13 2010), p. 136403.

[99] F. Wang et D. P. Landau. « Efficient, multiple-range random walk algorithm to calculate the density of states ». In : *Phys. Rev. Lett.* 86 (10 2001), p. 2050-2053.

[100] A. Laio et M. Parrinello. « Escaping free-energy minima ». In : *Proc. Natl. Acad. Sci.* 99.20 (2002), p. 12562-12566.

[101] G. Bussi, A. Laio et M. Parrinello. « Equilibrium free energies from nonequilibrium metadynamics ». In : *Phys. Rev. Lett.* 96 (9 2006), p. 090601.

[102] G. Bussi et D. Branduardi. « Free-energy calculations with metadynamics : theory and practice ». In : *Reviews in Computational Chemistry Volume 28*. John Wiley et Sons, Ltd, 2015. Chap. 1, p. 1-49.

[103] L. Cao, G. Stoltz, T. Lelièvre, M.-C. Marinica et M. Athènes. « Free energy calculations from adaptive molecular dynamics simulations with adiabatic reweighting ». In : *J. Chem. Phys.* 140.10 (2014), p. 104108.

[104] E. Darve et A. Pohorille. « Calculating free energies using average force ». In : *J. Chem. Phys.* 115.20 (nov. 2001), p. 9169-9183.

[105] E. Darve, D. Rodríguez-Gómez et A. Pohorille. « Adaptive biasing force method for scalar and vector free energy calculations ». In : *J. Chem. Phys.* 128.14 (2008), p. 144120.

[106] C. Chipot et J. Hénin. « Exploring the free-energy landscape of a short peptide using an average force ». In : *J. Chem. Phys.* 123.24 (2005), p. 244906.

[107] E. Weinan, W. Ren et E. Vanden-Eijnden. « String method for the study of rare events ». In : *Phys. Rev. B* 66.5 (2002), p. 052301.

[108] L. Maragliano et E. Vanden-Eijnden. « On-the-fly string method for minimum free energy paths calculation ». In : *Chem. Phys. Lett.* 446.1 (2007), p. 182-190.

[109] L. Maragliano, A. Fischer, E. Vanden-Eijnden et G. Ciccotti. « String method in collective variables : minimum free energy paths and isocommittor surfaces ». In : *J. Chem. Phys.* 125.2 (2006), p. 024106.

[110] E. Vanden-Eijnden et M. Venturoli. « Revisiting the finite temperature string method for the calculation of reaction tubes and free energies ». In : *J. Chem. Phys.* 130.19 (2009), 05B605.

[111] T. D. Swinburne et M.-C. Marinica. « Unsupervised calculation of free energy barriers in large crystalline systems ». In : *Phys. Rev. Lett.* 120 (13 2018), p. 135503.

[112] Z. Belkacemi, P. Gkeka, T. Lelièvre et G. Stoltz. *Chasing collective variables using autoencoders and biased trajectories*. 2021. arXiv : 2104.11061 [physics.bio-ph].

[113] A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis et P. G. Debenedetti. « Nonlinear dimensionality reduction in molecular simulation : The diffusion map approach ». In : *Chem. Phys. Lett.* 509.1 (2011), p. 1-11.

[114] F. Noé, S. Olsson, J. Köhler et H. Wu. « Boltzmann generators : Sampling equilibrium states of many-body systems with deep learning ». In : *Science* 365.6457 (2019).

[115] H. Chen, H. Liu, H. Feng, H. Fu, W. Cai, X. Shao et C. Chipot. « MLCV : Bridging machine-learning-based dimensionality reduction and free-energy calculation ». In : *J. Chem. Inf. and Model.* 62.1 (2022), p. 1-8.

[116] M. Rupp, A. Tkatchenko, K.-R. Müller et O. A. von Lilienfeld. « Fast and accurate modeling of molecular atomization energies with machine learning ». In : *Phys. Rev. Lett.* 108 (5 2012), p. 058301.

[117] M. Rupp. « Machine learning for quantum mechanics in a nutshell ». In : *Int. J. Quantum Chem.* 115.16 (2015), p. 1058-1073.

[118] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller et O.A. Von Lilienfeld. « Machine learning of molecular electronic properties in chemical compound space ». In : *New J. Phys.* 15.9 (2013), p. 095003.

[119] J. Behler. « Atom-centered symmetry functions for constructing high-dimensional neural network potentials ». In : *J. Chem. Phys.* 134.7 (2011), p. 074106.

[120] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler et M. Ceriotti. « Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials ». In : *J. Chem. Phys.* 148.24 (2018), p. 241730.

[121] A. P. Bartók, R. Kondor et G. Csányi. « On representing chemical environments ». In : *Phys. Rev. B* 87 (18 2013), p. 184115.

[122] H. Zong, G. Pilania, X. Ding, G. J. Ackland et T. Lookman. « Developing an interatomic potential for martensitic phase transformations in zirconium by machine learning ». en. In : *Npj Comput. Mater.* 4.1 (2018), p. 1-8.

[123] J. R. Kermode, A. Gleizer, G. Kovel, L. Pastewka, G. Csányi, D. Sherman et A. De Vita. « Low speed crack propagation via kink formation and advance on the silicon (110) cleavage plane ». In : *Phys. Rev. Lett.* 115.13 (2015), p. 135501.

[124] G. Ferré, J.-B. Maillet et G. Stoltz. « Permutation-invariant distance between atomic configurations ». In : *J. Chem. Phys.* 143.10 (2015), p. 104114.

[125] F. Noe et C. Clementi. « Kinetic distance and kinetic maps from molecular dynamics simulation ». In : *J. Chem. Theory Comput.* 11.10 (2015), p. 5002-5011.

[126] K. T. Schutt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko et K.-R. Müller. « SchNet A deep learning architecture for molecules and materials ». In : *J. Chem. Phys.* 148.24 (2018), p. 241722.

[127] W. F. Reinhart, A. W. Long, M. P. Howard, A. L. Ferguson et A. Z. Panagiotopoulos. « Machine learning for autonomous crystal structure identification ». In : *Soft Matter* 13.27 (2017), p. 4733-4745.

[128] M. Eickenberg, G. Exarchakis, M. Hirn et S. Mallat. « Solid Harmonic Wavelet Scattering : Predicting Quantum Molecular Energy from Invariant Descriptors of 3D Electronic Densities ». In : *Advances in neural information processing systems 30*. 2017, 6540–6549.

[129] M. Eickenberg, G. Exarchakis, M. Hirn, S. Mallat et L. Thiry. « Solid harmonic wavelet scattering for predictions of molecule properties ». In : *J. Chem. Phys.* 148.24 (2018), p. 241732.

[130] M. Hirn, S. Mallat et N. Poilvert. « Wavelet scattering regression of quantum chemical energies ». In : *Multiscale Model. Simul.* 15 (2016).

[131] E. Homer, D. Hensley, C. Rosenbrock, A. Nguyen et G. Hart. « Machine-learning informed representations for grain boundary structures ». In : *Front. Mater.* 6 (2019).

[132] A. Shapeev. « Moment tensor potentials : A class of systematically improvable interatomic potentials ». In : *Multiscale Model. Sim.* 14.3 (2016), p. 1153-1173.

[133] E. V. Podryabinkin et A. V. Shapeev. « Active learning of linearly parametrized interatomic potentials ». In : *Comput. Mater. Sci.* 140 (2017), p. 171-180.

[134] A. E. A. Allen, G. Dusson, C. Ortner et G. Csányi. « Atomic permutationally invariant polynomials for fitting molecular force fields ». en. In : *Mach. Learn. : Sci. Technol.* 2.2 (2021), p. 025017.

[135] C. Van Der Oord, G. Dusson, G. Csányi et C. Ortner. « Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials ». In : *Mach. Learn. : Sci. Technol.* 1.1 (2020), p. 015004.

[136] R. Drautz. « Atomic cluster expansion for accurate and transferable interatomic potentials ». In : *Phys. Rev. B* 99.1 (2019), p. 014104.

[137] R. Drautz. « Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer ». In : *Phys. Rev. B* 102.2 (2020), p. 024104.

[138] M.-C. Marinica, A.M. Goryaeva et W. Unn-Toc. *MiLaDy - Machine Lerning Dynamics*. Saclay : CEA, 2015-2022.

[139] Ramakrishna K. « The bispectrum as a source of phase-sensitive invariants for Fourier descriptors : a group-theoretic approach ». Thèse de doct. Irvine University, Irvine, 1992.

[140] J. Schmidt, M. R. G. Marques, S. Botti et M. A. L. Marques. « Recent advances and applications of machine learning in solid-state materials science ». In : *Npj Comput. Mater.* 5.83 (2019).

[141] M. H. Hassoun. *Fundamentals of artificial neural networks*. 1st. Cambridge, MA, USA : MIT Press, 1995.

[142] C. E. Rasmussen. *Gaussian processes in machine learning*. Springer, Berlin, Heidelberg, 2004.

[143] F. Rosenblatt. « The perceptron : A probabilistic model for information storage and organization in the brain. » In : *Psychol. Rev.* 65.6 (1958), p. 386-408.

[144] H. J. Kelley. « Gradient theory of optimal flight paths ». In : *ARS Journal* 30.10 (1960), p. 947-954.

[145] P. J. Werbos. « Backpropagation through time : what it does and how to do it ». In : *Proceedings of the IEEE* 78.10 (1990), p. 1550-1560.

[146] *ImageNet*. URL : www.image-net.org.

[147] Y. LeCun et C. Cortes. « MNIST handwritten digit database ». In : (2010). URL : http://yann.lecun.com/exdb/mnist/.

[148] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard et L. D. Jackel. « Backpropagation applied to handwritten zip code recognition ». In : *Neural Comput.* 1.4 (déc. 1989), 541–551.

[149] G. Cybenko. « Approximation by superpositions of a sigmoidal function ». In : *Math. Control, Signals Syst.* 2 (1989), p. 303-314.

[150] M. Stone. « Cross-validatory choice and assessment of statistical predictions ». In : *J. R. Stat. Soc. : Series B (Methodological)* 36.2 (1974), p. 111-133.

[151] M. Stone. « An asymptotic equivalence of choice of model by cross-validation and akaike's criterion ». In : *J. R. Stat. Soc. : Series B (Methodological)* 39.1 (1977), p. 44-47.

[152] B. Neyshabur, S. Bhojanapalli, D. McAllester et N. Srebro. « Exploring generalization in deep learning ». In : *NIPS*. 2017.

[153] K ; Kawaguchi, L.P. Kaelbling et Y. Bengio. « Generalization in deep learning ». In : *ArXiv* abs/1710.05468 (2018).

[154] D.C. Ciresan, U. Meier et J. Schmidhuber. « Multi-column deep neural networks for image classification ». In : *CoRR* abs/1202.2745 (2012).

[155] J. Mercer et A. R. Forsyth. « XVI. Functions of positive and negative type, and their connection the theory of integral equations ». In : *Phil. Trans. R. Soc. London. Series A, Containing Papers of a Mathematical or Physical Character* 209.441-458 (1909), p. 415-446.

[156] B. Schölkopf, R. Herbrich et A. J. Smola. « A Generalized Representer Theorem ». In : *Comput. Learn. Theory*. Berlin, Heidelberg, 2001, p. 416-426.

[157] M.A. Aizerman, E.M. Braverman et L.I. Rozonoèr. « Theoretical foundation of potential functions method in pattern recognition ». In : *Avtomat. i Telemekh.* 25.6 (1964), p. 917-936.

[158] A. P. Bartók, J. Kermode, N. Bernstein et G. Csányi. « Machine learning a general-purpose interatomic potential for silicon ». In : *Phys. Rev. X* 8 (4 2018), p. 041048.

[159] A. M. Goryaeva, C. Lapointe, C. Dai, J. Dérès, J.-B. Maillet et M.-C. Marinica. « Reinforcing materials modelling by encoding the structures of defects in crystalline solids into distortion scores ». In : *Nat. Comm.* 11 (2020), p. 4691.

[160] A. C. Aitken. « IV.—On least squares and linear combination of observations ». In : *Proc. R. Soc. Edinburgh* 55 (1936), 42–48.

[161] D. L. Phillips. « A technique for the numerical solution of certain integral equations of the first kind ». In : *J. ACM* 9.1 (1962), p. 84-97.

[162] R. Tibshirani. « Regression shrinkage and selection via the lasso ». In : *J. R. Stat. Soc. : Series B (Methodological)* 58.1 (1996), p. 267-288.

[163] H. Zou et T. Hastie. « Regularization and variable selection via the elastic net ». In : *J. R. Stat. Soc. : Series B (Statistical Methodology)* 67.2 (2005), p. 301-320.

[164] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever et R. Salakhutdinov. « Dropout : A simple way to prevent neural networks from overfitting ». In : *J. Mach. Learn. Res.* 15.56 (2014), p. 1929-1958.

[165] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot et E. Duchesnay. « Scikit-learn : Machine learning in python ». In : *J. Mach. Learn. Res.* 12 (2011), p. 2825-2830.

[166] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu et X. Zheng. *TensorFlow : Large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org. 2015. URL : http://tensorflow.org/.

[167] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai et S. Chintala. « PyTorch : An imperative style, high-performance deep learning library ». In : *Advances in Neural Information Processing Systems 32*. 2019, p. 8024-8035.

[168] N. W. Ashcroft et N. D. Mermin. *Solid state physics*. Holt-Saunders International Editions : Science : Physics. Holt, Rinehart et Winston, 1976.

[169] G. Lucas et R. Schäublin. « Vibrational contributions to the stability of point defects in bcc iron : A first-principles study ». In : *Nucl. Instrum. Methods Phys. Res. B : Beam Interact. Mater. At.* 267.18 (2009), p. 3009 -3012.

[170] M.-C. Marinica et F. Willaime. « Orientation of interstitials in clusters in α-Fe : A comparison between empirical potentials ». In : *Solid State Phenom.* 129 (2007), p. 67.

[171] S. Chiesa, P. M. Derlet et S. L. Dudarev. « Free energy of a $\langle 110 \rangle$ dumbbell interstitial defect in bcc Fe : Harmonic and anharmonic contributions ». In : *Phys. Rev. B* 79 (21 2009), p. 214109.

[172] J. Villain, M. Lavagna et P. Bruno. « Jacques Friedel and the physics of metals and alloys ». In : *C. R. Phys.* 17.3 (2016). Physique de la matière condensée au XXIe siècle : l'héritage de Jacques Friedel, p. 276-290.

[173] J. M. Ziman. *Principles of the Theory of Solids*. 2e éd. Cambridge University Press, 1972.

[174] C. Huang, A. F. Voter et D. Perez. « Scalable kernel polynomial method for calculating transition rates ». In : *Phys. Rev. B* 87.21 (2013), p. 214106.

[175] G. T. Barkema et N. Mousseau. « Event-based relaxation of continuous disordered systems ». In : *Phys. Rev. Lett.* 77 (21 1996), p. 4358-4361.

[176] R. Malek et N. Mousseau. « Dynamics of Lennard-Jones clusters : A characterization of the activation-relaxation technique ». In : *Phys. Rev. E* 62.6 (2000), 7723–7728.

[177] E. Cancès, F. Legoll, M.-C. Marinica, K. Minoukadeh et F. Willaime. « Some improvements of the activation-relaxation technique method for finding transition pathways on potential energy surfaces ». In : *J. Chem. Phys.* 130.11 (2009), p. 114711.

[178] E. Machado-Charry, L. K. Béland, D. Caliste, L. Genovese, T. Deutsch, N. Mousseau et P. Pochet. « Optimized energy landscape exploration using the ab initio based activation-relaxation technique ». In : *J. Chem. Phys.* 135.3 (2011), p. 034102.

[179] M-C Marinica, F Willaime et N Mousseau. « Energy landscape of small clusters of self-interstitial dumbbells in iron ». In : *Phys. Rev. B* 83.9 (2011), p. 094119.

[180] G. J Ackland, M. I. Mendelev, D. J. Srolovitz, S. Han et A. V. Barashev. « Development of an interatomic potential for phosphorus impurities in agr-iron ». In : *J. Phys. : Cond. Mat.* 16 (27 2004), S2629.

[181] E. Wigner et F. Seitz. « On the Constitution of Metallic Sodium ». In : *Phys. Rev.* 43 (10 1933), p. 804-810.

[182] M-C. Marinica et C. Lapointe. *Phondy - Phonons Dynamics*. Saclay : CEA, 2007-2022.

[183] A. Soulié, F. Bruneval, M.-C. Marinica, S. Murphy et J.-P. Crocombette. « Influence of vibrational entropy on the concentrations of oxygen interstitial clusters and uranium vacancies in nonstoichiometric UO_2 ». In : *Phys. Rev. Materials* 2.8 (2018), p. 083607.

[184] F. Berthier, J. Creuze, T. Gabard, B. Legrand, M.-C. Marinica et C. Mottet. « Order-disorder or phase-separation transition : Analysis of the Au-Pd system by the effective site energy model ». In : *Phys. Rev. B* 99.1 (2019), p. 014108.

[185] S. Plimpton. « Fast parallel algorithms for short-range molecular dynamics ». In : *J. Comput. Phys.* 117 (1995), p. 1-19.

[186] S. Alireza Etesami et E. Asadi. « Molecular dynamics for near melting temperatures simulations of metals using modified embedded-atom method ». In : *J. Phys. Chem. Solid.* 112 (2018), 61–72.

[187] A. Stukowski. « Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool ». In : *Model. Simul. Mater. Sci. Eng.* 18.1 (2010), p. 015012.

[188] Y. Mishin, M.R. Sørensen et A.F. Voter. « Calculation of point-defect entropy in metals ». In : *Phil. Mag. A* 81.11 (2001), 2591–2612.

[189] D. J Wales. « Discrete path sampling ». In : *Mol. phys.* 100.20 (2002), p. 3285-3305.

[190] G.H. Vineyard. « Frequency factors and isotope effects in solid state rate processes ». In : *J. Phys. Chem. Solid.* 3.1 (1957), p. 121 -127.

[191] O.A. Restrepo, C.S. Becquart, F. El-Mellouhi, O. Bouhali et N. Mousseau. « Diffusion mechanisms of C in 100, 110 and 111 Fe surfaces studied using kinetic activation-relaxation technique ». In : *Acta Mater.* 136 (2017), p. 303 -314.

[192] M. Trochet, L. K. Béland, J.-F. Joly, P. Brommer et N. Mousseau. « Diffusion of point defects in crystalline silicon using the kinetic activation-relaxation technique method ». In : *Phys. Rev. B* 91 (22 2015), p. 224106.

[193] T. D. Swinburne, D. Kannan, D. J. Sharpe et D. J. Wales. « Rare events and first passage time statistics from the energy landscape ». In : *J. Chem. Phys.* 153.13 (2020), p. 134115.

[194] T. Schuler, L. Messina et M. Nastar. « KineCluE : A kinetic cluster expansion code to compute transport coefficients beyond the dilute limit ». In : *Comput. Mater. Sci.* 172 (2020), p. 109191.

[195] T. Schuler et M. Nastar. « Transport properties of dilute $\alpha-\mathrm{Fe}(X)$ solid solutions (X = C, N, O) ». In : *Phys. Rev. B* 93 (22 2016), p. 224101.

[196] L. Huang, T. Schuler et M. Nastar. « Atomic-scale modeling of the thermodynamic and kinetic properties of dilute alloys driven by forced atomic relocations ». In : *Phys. Rev. B* 100 (22 2019), p. 224103.

[197] A. Jay, C. Huet, N. Salles, M. Gunde, L. Martin-Samos, N. Richard, G. Landa, V. Goiffon, S. De Gironcoli, A. Hémeryck et N. Mousseau. « Finding reaction pathways and transition states : r-ARTn and d-ARTn as an efficient and versatile alternative to string approaches ». In : *J. Chem. Theory Comput.* 16.10 (2020), p. 6726-6734.

[198] L. K. Béland, Y. Anahory, D. Smeets, M. Guihard, P. Brommer, J.-F. Joly, J.-C. Pothier, L. J. Lewis, N. Mousseau et F. Schiettekatte. « Replenish and relax : Explaining logarithmic annealing in ion-implanted c-Si ». In : *Phys. Rev. Lett.* 111 (10 2013), p. 105502.

[199] R.L.C. Vink, G.T. Barkema, W.F. van der Weg et Normand Mousseau. « Fitting the Stillinger–Weber potential to amorphous silicon ». In : *J. Non-Cryst. Solids* 282.2-3 (2001), p. 248-255.

[200] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler et P. Gumbsch. « Structural relaxation made simple ». In : *Phys. Rev. Lett.* 97.17 (2006), p. 170201.

[201] S. Gelin, A. Champagne-Ruel et N. Mousseau. « Enthalpy-entropy compensation of atomic diffusion originates from softening of low frequency phonons ». In : *Nat. Comm.* 11 (1 2020).

[202] J. Philibert. « Some thoughts and/or questions about activation energy and pre-exponential factor ». In : *Diffusion in Solids - Past, Present and Future*. T. 249. Defect and Diffusion Forum. Trans Tech Publications Ltd, 2006, p. 61-72.

[203] A. Yelon, B. Movaghar et R. S. Crandall. « Multi-excitation entropy : its role in thermodynamics and kinetics ». In : *Rep. Prog. Phys.* 69.4 (2006), p. 1145-1194.

[204] A. G. Jones. « Compensation of the Meyer-Neldel compensation law for H diffusion in minerals ». In : *Geochem. Geophys.* 15.6 (2014), p. 2616-2631.

[205] L. Shcherbak, O. Kopach, P. Fochuk, A. E. Bolotnikov et R. B. James. « Empirical correlations between the Arrhenius' parameters of impurities' diffusion coefficients in CdTe crystals ». In : *J. Ph. Equilibria Diffus.* 36 (4 2015), 99–109.

[206] Y. Lubianiker et I. Balberg. « Observation of a Meyer-Neldel rule for hopping conductivity ». In : *Phys. Status Solidi B* 205.1 (1998), p. 119-124.

[207] D. Emin. « Phonon-assisted jump rate in noncrystalline solids ». In : *Phys. Rev. Lett.* 32 (6 1974), p. 303-307.

[208] M. Athènes. « Computation of a chemical potential using a residence weight algorithm ». In : *Phys. Rev. E* 66 (4 2002), p. 046705.

[209] G. Adjanor et M. Athènes. « Gibbs free-energy estimates from direct path-sampling computations ». In : *J. Chem. Phys.* 123.23 (2005), p. 234104.

[210] M. Athenes. « Conditioning and enhanced sampling schemes for simulating thermodynamic and kinetic properties of condensed matter ». Thèse de doct. 2018.

[211] M. M. Sultan, H. K. Wayment-Steele et V. S. Pande. « Transferable neural networks for enhanced sampling of protein dynamics ». In : *J. Chem. Theory Comput.* 14.4 (2018), p. 1887-1894.

[212] J. M. L. Ribeiro, P. Bravo, Y. Wang et P. Tiwary. « Reweighted autoencoded variational Bayes for enhanced sampling (RAVE) ». In : *J. Chem. Phys.* 149.7 (2018), p. 072301.

[213] H. Alrachid et T. Lelièvre. « Long-time convergence of an adaptive biasing force method : Variance reduction by Helmholtz projection ». In : *J. Comput. Math.* 1 (2015), p. 55-82.

[214] J. Hénin. « Fast and accurate multidimensional free energy integration ». In : *J. Chem. Theory Comput.* 17.11 (2021), p. 6789-6798.

[215] T.D. Swinburne. « Uncertainty and anharmonicity in thermally activated dynamics ». In : *Comput. Mater. Sci.* 193 (2021), p. 110256.

[216] A. Lesage, T. Lelièvre, G. Stoltz et J. Hénin. « Smoothed biasing forces yield unbiased free energies with the extended-system adaptive biasing force method ». In : *J. Phys. Chem. B* 121.15 (2017), p. 3676-3685.

[217] T. Zhao, H. Fu, T. Lelièvre, X. Shao, C. Chipot et W. Cai. « The extended generalized adaptive biasing force algorithm for multidimensional free-energy calculations ». In : *J. Chem. Theory Comput.* 13.4 (2017), p. 1566-1576.

[218] H. Sidky et J. K. Whitmer. « Learning free energy landscapes using artificial neural networks ». In : *J. Chem. Phys.* 148.10 (2018), p. 104111.

[219] L. Mones, N. Bernstein et G. Csányi. « Exploration, sampling, and reconstruction of free energy surfaces with gaussian process regression ». In : *J. Chem. Theory Comput.* 12.10 (2016), p. 5100-5110.

[220] T. Stecher, N. Bernstein et G. Csányi. « Free energy surface reconstruction from umbrella samples using gaussian process regression ». In : *J. Chem. Theory Comput.* 10.9 (2014), p. 4079-4097.

[221] G. Henkelman, B.P. Uberuaga et H. Jonsson. « A climbing image nudged elastic band method for finding saddle points and minimum energy paths ». In : *J. Chem. Phys.* 113.22 (2000), p. 9901-9904.

[222] G. Henkelman. « Atomistic simulations of activated processes in materials ». In : *Annu. Rev. Mater. Res.* 0 (2017).

[223] G. Henkelman, G. Johannesson et H. Jonsson. *Methods for finding saddle points and mimnum energy paths.* Springer, p. 269-302.

[224] Y. Sato, T.D. Swinburne, S. Ogata et D. Rodney. « Anharmonic effect on the thermally activated migration of 1012 twin interfaces in magnesium ». In : *Mater. Res. Lett.* 9.5 (2021), p. 231-238.

[225] T. Lelièvre, M. Rousset et G. Stoltz. « Computation of free energy profiles with parallel adaptive dynamics ». In : *J. Chem. Phys.* 126.13 (2007), p. 134111.

[226] P. Raiteri, A. Laio, F. L. Gervasio, C. Micheletti et M. Parrinello. « Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics ». In : *J. Phys. Chem. B* 110.8 (2006), p. 3533-3539.

[227] D. Frenkel et 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 ». In : *J. Chem. Phys.* 81.7 (1984), p. 3188-3193.

[228] A. Glensk, B. Grabowski, T. Hickel et J. Neugebauer. « Understanding anharmonicity in fcc materials : From its origin to ab initio strategies beyond the quasiharmonic approximation ». In : *Phys. Rev. Lett.* 114.19 (2015), p. 195901.

[229] O. Hellman, P. Steneteg, I. A. Abrikosov et S. I. Simak. « Temperature dependent effective potential method for accurate free energy calculations of solids ». In : *Phys. Rev. B* 87 (10 2013), p. 104111.

[230] M.-C. Marinica, L. Ventelon, M. R. Gilbert, L. Proville, S. L. Dudarev, J. Marian, G. Bencteux et F. Willaime. « Interatomic potentials for modelling radiation defects and dislocations in tungsten ». In : *J. Phys. : Cond. Mat.* 25.39 (2013), p. 395502.

[231] B. Sharma, G. Tiwari et S. Ray. *Diffusion in bcc Transition Metals.* 1969.

[232] B. Grabowski, L. Ismer, T. Hickel et J. Neugebauer. « Ab initio up to the melting point : Anharmonicity and vacancies in aluminum ». In : *Phys. Rev. B* 79.13 (2009), p. 134106.

[233] A. Glensk, B. Grabowski, T. Hickel et J. Neugebauer. « Breakdown of the Arrhenius law in describing vacancy formation energies : The importance of local anharmonicity revealed by ab initio thermodynamics ». In : *Phys. Rev. X* 4 (1 2014), p. 011018.

[234] G. Neumann et C. Tuijn. « Self-diffusion : Self-diffusion in BCC metals ». In : *Impurity Diffusion in Metals*. T. 88. Solid State Phenomena. Trans Tech Publications Ltd, 2002, p. 40-50.

[235] J. M. Sanchez et D. de Fontaine. « Model for anomalous self-diffusion in group-IVB transition metals ». In : *Phys. Rev. Lett.* 35 (4 1975), p. 227-230.

[236] W. Petry, T. Flottmann, A. Heiming, J. Trampenau, M. Alba et G. Vogl. « Atomistic study of anomalous self-diffusion in bcc β-titanium ». In : *Phys. Rev. Lett.* 61 (6 1988), p. 722-725.

[237] G. Vogl, W. Petry, Th. Flottmann et A. Heiming. « Direct determination of the self-diffusion mechanism in bcc β-titanium ». In : *Phys. Rev. B* 39 (8 1989), p. 5025-5034.

[238] G. Smirnov. « Non-Arrhenius diffusion in bcc titanium : Vacancy-interstitialcy model ». In : *Phys. Rev. B* 102 (18 2020), p. 184110.

[239] C. Herzig et U. Köhler. « Anomalous self-diffusion in BCC IVB metals and alloys ». In : *Vacancies and Interstitials in Metals and Alloys*. T. 15. Materials Science Forum. Trans Tech Publications Ltd, 1987, p. 301-322.

[240] G. Neumann et V. Tölle. « Self-diffusion in body-centred cubic metals : Analysis of experimental data ». In : *Phil. Mag. A* 61.4 (1990), p. 563-578.

[241] J. N. Mundy, S. J. Rothman, N. Q. Lam, H. A. Hoff et L. J. Nowicki. « Self-diffusion in tungsten ». In : *Phys. Rev. B* 18 (12 1978), p. 6566-6575.

[242] R. E. Einziger, J. N. Mundy et H. A. Hoff. « Niobium self-diffusion ». In : *Phys. Rev. B* 17 (2 1978), p. 440-448.

[243] D. Smirnova, S. Starikov, G. D. Leines, Y. Liang, N. Wang, M. N. Popov, I. A. Abrikosov, D. G. Sangiovanni, R. Drautz et M. Mrovec. « Atomistic description of self-diffusion in molybdenum : A comparative theoretical study of non-Arrhenius behavior ». In : *Phys. Rev. Materials* 4 (1 2020), p. 013605.

[244] K. J. Laidler et M. C. King. « Development of transition-state theory ». In : *J. Phys. Chem.* 87.15 (1983), p. 2657-2664.

[245] G. Kresse et J. Furthmüller. « Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set ». In : *Phys. Rev. B* 54.16 (1996), p. 11169.

[246] G. Kresse et D. Joubert. « From ultrasoft pseudopotentials to the projector augmented-wave method ». In : *Phys. Rev. B* 59.3 (1999), p. 1758.

[247] H. J. Monkhorst et J. D. Pack. « Special points for Brillouin-zone integrations ». In : *Phys. Rev. B* 13 (12 1976), p. 5188-5192.

[248] M. Methfessel et A. T. Paxton. « High-precision sampling for Brillouin-zone integration in metals ». In : *Phys. Rev. B* 40 (6 1989), p. 3616-3621.

[249] Y. S. Touloukian et C. Y. Ho. *Thermal expansion. Metallic elements and alloys.* 1970.

[250] K. Maier, M. Peo, B. Saile, H.E. Schaefer et A. Seeger. « High–temperature positron annihilation and vacancy formation in refractory metals ». In : *Phil. Mag. A* 40.5 (1979), p. 701-728.

[251] E. A. Lazar, J. Han et D. J. Srolovitz. « Topological framework for local structure analysis in condensed matter ». In : *Proc. Natl. Acad. Sci.* 112.43 (2015), E5769-E5776.

[252] P.M. Larsen, S. Schmidt et J. Schiøtz. « Robust structural identification via polyhedral template matching ». In : *Model. Simul. Mater. Sci. Eng.* 24.5 (2016), p. 055007.

[253] A. Stukowski. « Computational analysis methods in atomistic modeling of crystals ». In : *JOM* 66.3 (2014), p. 399-407.

[254] M. Hubert, M. Debruyne et P. J. Rousseeuw. « Minimum covariance determinant and extensions ». In : *WIREs Comp. Stats.* 10.3 (2017).

[255] F. Bloch. « Über die quantenmechanik der elektronen in kristallgittern ». In : *Zeitschrift für Physik* 52.7 (1929), p. 555-600.

[256] F. Bruneval, T. Rangel, S.M. Hamed, M. Shao, C. Yang et J.B. Neaton. « molgw 1 : Many-body perturbation theory software for atoms, molecules, and clusters ». In : *Comput. Phys. Comm.* 208 (2016), p. 149-161.

[257] P. Ehrhart, P. Jung, H. Schultz et H. Ullmaier. *Atomic defects in metals.* Berlin : Springer-Verlag, 1991.

[258] J. P. Ryckaert et G. Ciccotti. « Introduction of Andersen's demon in the molecular dynamics of systems with constraints ». In : *J. Chem. Phys.* 78.12 (1983), p. 7368-7374.

[259] Y.Q. Cheng et E. Ma. « Atomic-level structure and structure–property relationship in metallic glasses ». In : *Prog. Mater. Sci.* 56.4 (2011), p. 379-473.

[260] N. Artrith, A. Urban et G. Ceder. « Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species ». In : *Phys. Rev. B* 96 (1 2017), p. 014112.

Résumé :

Modélisation multi-échelles des défauts d'irradiation dans les métaux cubiques centrés

La modélisation des métaux sous conditions extrêmes nécessite une approche de type multi-échelles. En effet, pour des raisons de complexité numérique, il n'est pas possible d'utiliser un formalisme unique, précis et transférable pour toutes les échelles de simulation. Il correspond, en général, une ou plusieurs méthodes utilisables pour une échelle spatiale et temporelle donnée. Ces méthodes se basent sur des approximations - physiques et/ou numériques - dont le nombre croît lorsque les échelles d'espace et de temps simulées augmentent. La modélisation multi-échelle peut donc se résumer comme un - utopique - équilibre entre échelles "spatio-temporelles" simulées et représentatitivité des phénomènes mis en jeu lors des transformations du système étudié. Les méthodes multi-échelles appliquées à la science des matériaux doivent aussi prendre en compte les effets de températures finies afin de simuler des structures dans leurs conditions nominales d'utilisation industrielles. Ces dernières années, les effets de température ont été traités dans le cadre d'approximations locales : harmonique et/ou quasiharmonique. La prise en compte des effets anharmoniques reste difficile - mais nécessaire pour rendre compte de certains phénomènes physiques - et est un sujet de recherche à part entière pour l'amélioration des modèles multi-échelles. Les objectifs de cette thèse sont de (i) développer de nouveaux outils de simulations afin d'étendre les domaines d'applicabilité des méthodes multi-échelles (ii) et d'estimer des grandeurs de températures finies. Nous nous basons sur un ensemble de méthodes en plein essor dans le domaine de la science des matériaux : le Machine Learning. Ces méthodes permettent de développer des outils statistiques systématiques et d'étudier plus facilement des corrélations. Dans un premier temps, nous développons des méthodes d'estimations rapides de quantités dérivées de propriétés vibrationnelles harmoniques : l'entropie de formation de défauts et les fréquences d'attaque. Le formalisme développé est précis, transférable et permet de réduire grandement le coût numérique (évoluant traditionnellement comme $O(N^3)$ où N est le nombre de particules dans le système) dont la complexité numérique évolue comme O(N). Dans un deuxième temps, nous nous intéressons à quantifier l'influence des effets anharmoniques pour des systèmes métalliques. Nous développons une approche de calcul direct (couplant Machine Learning et méthodes d'énergie libre) permettant de calculer le coefficient d'auto-diffusion, avec une précision "ab initio", des métaux cubiques centrés. Nous confrontons directement nos résultats avec l'expérience et nous donnons une explication, générale pour les métaux cubiques centrés, du comportement anormal du coefficient d'auto-diffusion à hautes températures.

Mots clefs :
Températures finies, modélisation, machine learning, paysage énergétique, défauts

Abstract :

Multi-scale modelling of point defects in bcc metals

The modelling of metals under extreme conditions requires a multi-scale approach. Indeed, for reasons of numerical complexity, it is not possible to use a unique formalism, accurate and transferable for all scales of simulation. In general, one or more methods have to be used for a given spatial and temporal scale. These methods are based on approximations - physical and/or numerical - which are more and more numerous when the simulated space and time scales increase. Multi-scale modelling can therefore be summarised as a - utopian - balance between simulated "space-time" scales and representativeness of phenomena involved in the transformations of the system studied. Multi-scale methods applied to materials science must also take into account the effects of finite temperatures in order to simulate structures in their nominal conditions of industrial use. In recent years, temperature effects have been treated in the context of local approximations: harmonic and/or quasi-harmonic. Taking into account anharmonic effects remains difficult - but necessary to be representative of physical phenomena - and is a main research topic for the improvement of multi-scale models. The objectives of this thesis are (i) to develop new simulation tools to extend the applicability of multi-scale methods (ii) and to estimate finite temperature quantities. We rely on a set of methods, in full expansion in the field of materials science: Machine Learning. These methods allow to develop systematic statistical tools and to study correlations more easily. In a first step, we develop methods for fast estimation of quantities derived from the harmonic vibrational properties: the defect formation entropy and the attack frequencies. The formalism developed is accurate, transferable and allows to greatly reduce the numerical cost (traditionally evolving as $O(N^3)$ where N is the number of particles in the system) whose numerical complexity evolves as O(N). In a second step, we are interested in quantifying the influence of anharmonic effects for metallic systems. We develop a direct computational approach (coupling Machine Learning and free energy methods) to calculate the self-diffusion coefficient, with an accuracy of ab initio, in body-centered cubic metals. We directly confront our results with the experiment and we give a - general - explanation, for the cubic centered metals, of the anomalous behavior of the self-diffusion coefficient at high temperatures.

Keywords :
Finite temperatures, modelling, machine learning, energy landscape, defects